# Unveiling Hidden Customer Segments in E-Commerce Using DBSCAN Clustering on Demographic and Behavioral Insights

Adimas Aglasia[1,*], Isnandar Agus[2]

[1,2]*Institute Informatics and Business Darmajaya, Indonesia, Jln ZA Pagar Alam 93 A, Bandar Lampung and 35136, Indonesia*

(Received: June 4, 2024; Revised: July 31, 2024; Accepted: August 11, 2024; Available online: September 1, 2024)

**Abstract**

Customer segmentation is a crucial process in e-commerce that allows businesses to tailor their marketing strategies to specific customer groups. This research applies the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to segment customers based on their demographic and behavioral data. The dataset used includes variables such as age, annual income, total spending, and campaign engagement, which are essential for identifying meaningful patterns within the customer base. The DBSCAN algorithm was chosen due to its ability to detect clusters of arbitrary shapes and handle noise, making it ideal for complex e-commerce datasets. The analysis identified one dominant customer segment, with a small portion of the data labeled as noise, indicating that the majority of customers exhibit similar behaviors. However, the results also highlight the challenge of parameter selection for DBSCAN, as the clustering outcome was sensitive to the chosen values of ε (epsilon) and MinPts. The segmentation revealed valuable insights, such as the fact that most customers share similar characteristics in terms of spending habits and engagement, yet a few outliers exist who do not align with these patterns. These findings provide a foundation for businesses to develop targeted marketing strategies based on customer segmentation. Despite the promising results, the study acknowledges limitations in the segmentation process, particularly with the influence of outliers and the need for further tuning of the algorithm's parameters. Future research could explore hybrid clustering models that combine DBSCAN with other techniques, as well as incorporating additional behavioral features for more refined segmentation. The insights gained from this research can guide businesses in crafting personalized marketing campaigns that cater to distinct customer segments.

*Keywords:* DBSCAN, Customer Segmentation, E-Commerce, Clustering, Personalized Marketing

## 1. Introduction

The rapid growth of e-commerce necessitates an effective approach to customer segmentation, which is integral to personalized marketing strategies. As the landscape of e-commerce evolves, businesses are recognizing that a one-size-fits-all approach is no longer viable. Instead, successful retailers are leveraging customer segmentation to tailor their marketing efforts toward specific consumer groups based on shared characteristics, ultimately enhancing customer interactions and driving sales.

Customer segmentation involves categorizing consumers into groups with similar traits, allowing businesses to deliver personalized products and marketing campaigns that resonate with distinct consumer needs. Sharma and Aggarwal emphasize the significance of understanding factors that influence online shopping behavior, including customer satisfaction and personalization, which can be essential determinants of e-commerce success [1]. By incorporating these factors into segmentation strategies, e-commerce platforms can optimize their marketing efforts.

In light of the increasing consumer sophistication and varied preferences, it is imperative for businesses to employ advanced analytical techniques to refine their marketing strategies. Research conducted by Serwah et al. illustrates that traditional marketing methods may fall short in addressing the nuanced requirements of today's consumers, presenting opportunities for more targeted strategies through data analytics and customer behavior analysis [2]. Techniques such as RFM (Recency, Frequency, Monetary) analysis facilitate the identification of key customer segments, ensuring that personalized marketing efforts are relevant and timely. Additionally, the importance of leveraging data-driven

*Corresponding author: Adimas Aglasia (dimas@darmajaya.ac.id)

DOI: https://doi.org/10.47738/ijaim.v4i3.85

approaches cannot be overstated. Wei discusses how analyzing consumers' varied preferences and behaviors allows e-commerce firms to adapt their marketing and product offerings to meet individual consumer needs effectively [3]. By employing tools such as deep learning and evolutionary algorithms, companies can optimize their marketing strategies, leading to heightened consumer engagement and satisfaction.

Furthermore, Tabianan et al. describe the role of machine learning in customer segmentation, highlighting the application of clustering techniques that can enhance understanding of customer purchasing behavior and pave the way for intelligent segmentation practices [4]. This technological integration ensures that e-commerce businesses can maintain a competitive edge, even in a rapidly changing marketplace.

Customer segmentation is a pivotal strategy for e-commerce businesses seeking to enhance personalized marketing efforts. Implementing clustering algorithms, particularly the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), allows marketers to identify distinct customer groups based on demographic and behavioral data. This method is particularly advantageous due to its ability to manage noise and recognize clusters of arbitrary shapes, which aligns well with the complexities of real-world customer behaviors.

DBSCAN excels in identifying clusters by analyzing points based on density rather than distance alone. According to Hahsler et al., the DBSCAN algorithm is efficient for handling large datasets and offers significant performance benefits over traditional clustering methods like K-Means [5]. This capability is critical in the context of e-commerce, where customer datasets can be voluminous and noisy due to varying levels of engagement and transaction frequency. Moreover, Monalisa and Kurnia demonstrated the effectiveness of using DBSCAN in conjunction with Recency, Frequency, and Monetary (RFM) models, integrating demographic variables for a comprehensive customer segmentation framework [6]. This integration provides insights into customer behaviors and preferences, which businesses can utilize for targeted marketing initiatives. Furthermore, the efficacy of DBSCAN in clustering processes has been reaffirmed by various research contributions. Lai et al. emphasized the growing popularity of this algorithm for its robustness in segmentation across multiple applications, including market segmentation [7]. Its ability to operate without requiring the number of clusters to be specified in advance allows businesses to explore different customer profiles dynamically.

The primary objective of this research is to explore customer segmentation in the context of e-commerce by employing the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. This approach focuses on using a combination of demographic and behavioral data to segment customers more effectively. By analyzing variables such as customer age, income, family size, and purchasing history, the study aims to identify distinct customer groups that share common characteristics. DBSCAN, a density-based algorithm, is particularly suited for this task as it can detect clusters of varying shapes and handle noise points, offering a more flexible approach compared to traditional methods like K-Means.

Existing customer segmentation methods often fail to fully capture the complexity of customer behaviors, especially in e-commerce environments where customer interactions are highly dynamic and diverse. Many traditional clustering algorithms, such as K-Means, assume that customer groups are spherical and have similar sizes, which does not always reflect real-world behavior patterns. This research addresses this limitation by introducing DBSCAN as a more suitable alternative for identifying hidden customer segments. The study contributes to the field by demonstrating how DBSCAN can uncover valuable insights into customer behavior, enabling businesses to implement more targeted and effective marketing strategies.

## 2. Literature Review

### 2.1. Customer Segmentation in E-Commerce

Customer segmentation is a crucial strategy in marketing that helps businesses understand their customers better and tailor their services to meet specific needs. Several techniques can be employed for effective customer segmentation, including K-Means clustering, hierarchical clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Each of these methods has unique strengths and applications in the context of demographic and behavioral data.

K-Means is one of the most widely used clustering algorithms due to its simplicity and effectiveness in partitioning data into K distinct groups. This technique involves initializing K centroids and iteratively assigning data points to the nearest centroid based on Euclidean distance, followed by updating the centroid to the average of assigned points [4]. K-Means is particularly effective when the clusters are spherical and equally sized, making it suitable for datasets characterized by such distributions. However, one must determine the optimal number of clusters (K) beforehand, which can sometimes be challenging [8].

Hierarchical clustering presents an alternative approach that builds a tree-like structure (dendrogram) to represent nested groupings of data points. This method can be either agglomerative (bottom-up) or divisive (top-down). In agglomerative clustering, each data point starts in its own cluster, which are iteratively merged based on distance criteria until a single cluster is formed. Conversely, divisive clustering begins with all points in one cluster and recursively splits them down to individual points [9]. One notable advantage of hierarchical clustering is that it does not require specifying the number of clusters in advance, allowing for a more exploratory analysis of data relationships. This can be particularly useful in situations where the underlying structure of the data is not readily apparent.

DBSCAN is a density-based clustering algorithm that groups together points that are closely packed while marking points in low-density regions as outliers. Unlike K-Means, DBSCAN does not require the number of clusters to be specified a priori, which is particularly advantageous in datasets with varying shapes and sizes [10]. One of its defining features is its ability to identify noise and outliers without forcing them into clusters, allowing businesses to focus their marketing efforts on meaningful segments while understanding which customers may not fit typical purchasing patterns [11].

## 2.2. DBSCAN Clustering

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is a powerful method in data mining and pattern recognition, particularly useful for grouping spatial data or any dataset comprising varied density distributions. DBSCAN fundamentally differs from traditional clustering algorithms, such as K-Means, which rely on predefined distances and shapes of clusters.

DBSCAN operates using two key parameters: Epsilon (eps) and Minimum Points (minPts). Epsilon defines the radius within which neighboring points are considered part of the same cluster, while minPts serves as a threshold for how densely packed the points must be for a particular region to be classified as a cluster. The algorithm classifies points into three categories: core points (which have at least minPts neighbors within eps), border points (which are within eps of a core point but have fewer than minPts neighbors), and noise (points that are neither core nor border points). This density-based approach allows DBSCAN to effectively form clusters of arbitrary shapes and ignore irrelevant noise points that might skew results [12].

DBSCAN offers several advantages over K-Means, particularly regarding how it handles clusters and noise. Firstly, DBSCAN does not require the user to specify the number of clusters in advance, making it more adaptable for exploratory data analysis. In contrast, K-Means requires that the number of clusters be predefined, potentially leading to inaccurate segmentation if this number is misestimated [13].

Finally, DBSCAN is particularly beneficial for large datasets that contain clusters of varied density. Applications of DBSCAN in real-time scenarios demonstrate its efficiency and effectiveness, as seen in the study by Fu et al., where DBSCAN is used for segmenting complex data types like point clouds, highlighting its robustness in diverse situations [14].

## 2.3. Behavioral and Demographic Data

The use of demographic and behavioral data for clustering e-commerce customers has garnered significant attention in recent studies. Various methodologies have been explored to harness these data points for effective customer segmentation, enabling e-commerce businesses to tailor their marketing strategies and improve customer satisfaction.

A prominent study by Brahmana et al. employed the RFM (Recency, Frequency, Monetary) model using K-Means, K-Medoids, and DBSCAN clustering methods to stratify customers into distinct groups such as “Dormant” and “Golden” clusters. Their findings highlighted the effectiveness of different clustering techniques in identifying optimal customer classes based on the Davies-Bouldin and Silhouette indices, advocating for a mixed-use approach of traditional and density-based clustering methodologies for more nuanced segmentation [15]. This approach demonstrates the critical role of clustering algorithms in enhancing the granularity of customer segmentation.

Hidayati et al. also conducted research comparing K-Means with DBSCAN in the context of village grouping based on poverty indicators. Their results indicated that while K-Means showed a lower within-cluster sum of squares, indicating generally high clustering efficiency, DBSCAN outperformed K-Means on the average silhouette width. This suggests that DBSCAN may provide better-defined clusters in datasets with significant noise or varying density distributions [16]. The exploration of various clustering methods in this context underscores the adaptability of clustering algorithms to different domain-specific requirements.

The insights on customer loyalty and satisfaction provided by Kristanto et al. offer another dimension to understanding e-commerce dynamics; they found that customer trust and satisfaction serve as pivotal components in fostering e-loyalty, which further influences purchasing behavior. By correlating behavioral data with demographic factors, businesses can refine their approaches to customer engagement and retention through targeted communication strategies [17]. Such frameworks enable marketers to craft personalized experiences that resonate with specific customer segments.

Zhang and Dong's research presents a broader perspective with an emphasis on machine learning techniques to predict repeat customer behavior on e-commerce platforms. They classify prediction models into individual and ensemble types, reflecting the increasing complexity and scale of customer behavior analytics in the context of e-commerce [18]. This study signifies the importance of harnessing sophisticated analytical methods to understand and anticipate consumer behavior, thereby improving customer retention strategies.

Research by Alzami et al. explored the implementation of the RFM method alongside K-Means for customer segmentation, detailing how clustering techniques can yield actionable insights from demographic and behavioral data. Their experiments demonstrate the effectiveness of these methodologies in real-world applications, yielding robust segmentation results that inform marketing tactics [19], [20].

## 3. Method

### 3.1. Data Loading and Initial Exploration

The research began by loading the customer dataset using the pandas.read_csv() function, which is essential for handling large datasets and consolidating them into a single DataFrame for further analysis. The dataset was assumed to have been split into multiple source files, and the first step was to combine them into one cohesive dataset. Once the data was loaded successfully, a series of basic checks were performed to ensure its integrity. The shape of the dataset was examined, revealing the number of rows and columns. The first few rows of the dataset were displayed to provide an initial overview of the data structure and ensure that the data was properly loaded. Error handling mechanisms were implemented to ensure that any issues during the data loading process, such as a missing file, would be caught and appropriately managed. If the dataset was not found, the process was halted, and an error message was displayed to alert the user. This thorough initial check ensured that any problems with data accessibility were resolved before proceeding with further analysis.

### 3.2. Data Cleaning and Feature Engineering

The second phase of the method focused on cleaning and enhancing the data to make it suitable for clustering analysis. The first step in data cleaning involved checking for missing values and ensuring that no essential data was missing from key columns. Specifically, missing values in the Annual_Income column were imputed with the median income value to prevent any data gaps from negatively impacting the analysis. This median imputation was chosen as it is a robust method for handling outliers and skewed distributions in income data. Duplicates were also checked for and removed from the dataset. If any duplicate rows were detected, they were dropped to ensure the analysis was not biased by repeated data points. The cleaning process was critical to ensure that the dataset reflected accurate and reliable information for further analysis.

Feature engineering was the next step, where new variables were created to enrich the dataset and make it more informative for clustering. The Birth_Year column was used to calculate a new Age feature by subtracting the birth year from the current year. This transformation was necessary to make the dataset more relevant to the clustering task, as age is a key demographic feature for customer segmentation. Additionally, the Customer_Since column, which contained the date of the customer’s first purchase, was transformed into a new feature, Years_Customer, representing the number of years since the customer joined. This transformation involved converting the Customer_Since column to a datetime format and calculating the difference between the current date and the customer's join date, expressed in years. The Customer_Since column was then dropped, as the newly created Years_Customer provided more meaningful information for the clustering task.

Further feature engineering was conducted by combining Kids_Home and Teens_Home into a single column, Total_Dependents, representing the total number of dependents a customer has. This combined feature offered a more consolidated view of customer demographics, which could be useful for segmentation based on family-related behaviors. Spending data from various product categories was aggregated into a Total_Spent column, summing individual spending categories such as wine, fruits, meat, and sweets. Similarly, purchase behavior was captured in the Total_Purchases column, aggregating the number of purchases across different channels (web, catalog, and store). Finally, the number of accepted campaigns was combined into the Total_Campaigns_Accepted column. These engineered features helped capture the key behavioral patterns of customers, which were crucial for effective segmentation.

## 3.3. Exploratory Data Analysis (EDA)

With the dataset cleaned and feature-engineered, the next phase of the analysis focused on exploring the data to uncover patterns and relationships between different variables. This phase, known as exploratory data analysis (EDA), aimed to provide a deeper understanding of the data's structure, distribution, and potential relationships between key features. Descriptive statistics were calculated for the selected features, providing an overview of the central tendencies and variability of the demographic and behavioral variables. This included measures such as mean, median, and standard deviation for numerical features, and frequency distributions for categorical features.

To complement these statistics, visualizations were created to explore the data further. Histograms for numerical features such as Age, Annual_Income, and Total_Spent were plotted to visualize their distributions and identify any potential skewness or outliers. For categorical variables, count plots were used to examine the distribution of values in features like Education_Level, Marital_Status, and Device_Type. These visualizations provided insights into the composition of the dataset and allowed for the detection of potential issues, such as imbalanced categories or unusual data patterns. A correlation heatmap was also generated to assess the relationships between numerical features. This heatmap visually represented the strength and direction of correlations, helping identify which variables were strongly related and which were more independent. Correlation analysis is critical in clustering, as highly correlated features may lead to multicollinearity, which could affect the performance of clustering algorithms.

## 3.4. Data Preprocessing

Before applying the DBSCAN clustering algorithm, the dataset underwent preprocessing to prepare it for machine learning. Data preprocessing is an essential step in ensuring that the clustering algorithm can effectively analyze the data and produce meaningful results. The first step in preprocessing involved scaling the numerical features using StandardScaler, which standardizes the data by removing the mean and scaling to unit variance. Scaling ensures that each feature contributes equally to the clustering process and prevents any one feature from dominating due to its larger range or magnitude. Categorical variables were encoded using OneHotEncoder to convert them into a format suitable for clustering algorithms, which require numerical input. This encoding technique converts categorical features into binary columns, where each category is represented by a 1 or 0, indicating its presence or absence. A ColumnTransformer was used to apply the appropriate preprocessing steps to different types of features. Numerical features were scaled, while categorical features were one-hot encoded. This transformer was applied to the dataset, resulting in a processed feature matrix ready for clustering. After preprocessing, the transformed data was stored in a new DataFrame, ensuring that the original dataset remained unchanged for comparison.

### 3.5. DBSCAN Clustering

With the data preprocessed, the DBSCAN clustering algorithm was applied to identify customer segments based on the demographic and behavioral features. DBSCAN, or Density-Based Spatial Clustering of Applications with Noise, is particularly well-suited for this task as it can identify clusters of arbitrary shape and can handle noise points, making it ideal for e-commerce datasets where customer behavior is complex and varied. The parameters for DBSCAN—eps (epsilon) and min_samples—were selected based on a heuristic approach. The value of eps defines the maximum distance between two points to be considered neighbors, and min_samples determines the minimum number of points required to form a dense region, or cluster. The k-distance graph was used to visually determine the optimal eps value by plotting the sorted distances between points. The "elbow" point in the plot indicates a good candidate for eps, where the density of points changes significantly. After selecting appropriate parameters, DBSCAN was applied to the preprocessed data. The algorithm assigns each data point to a cluster or labels it as noise (-1). The resulting clusters were stored in a new column, Cluster_DBSCAN, in the dataset, allowing for easy identification of customer segments. The number of clusters and noise points were recorded, and their distributions were analyzed to assess the quality of the clustering results.

### 3.6. Model Evaluation

The clustering results were evaluated using common clustering metrics, such as the Silhouette Score and the Davies-Bouldin Index. The Silhouette Score measures how similar each point is to its own cluster compared to other clusters, with higher values indicating better-defined clusters. The Davies-Bouldin Index measures the average similarity ratio of each cluster with the cluster that is most similar to it, with lower values indicating better clustering performance. These metrics provided an indication of how well DBSCAN was able to group customers into meaningful segments. Because DBSCAN can produce noise points (labelled as -1), these points were excluded from the evaluation of the clustering metrics. The metrics were computed only for the points that were assigned to clusters, ensuring that the presence of noise did not skew the results.

### 3.7. Cluster Profiling and Visualization

Once the clusters were formed, the next step involved profiling the clusters to understand their characteristics. This was done by computing the mean of numerical features and the mode of categorical features for each cluster. The resulting cluster profiles provided insights into the distinct characteristics of each customer segment, such as differences in age, spending behavior, or engagement with campaigns. To visualize the clustering results, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset to two components, allowing the clusters to be visualized in a 2D scatter plot. This visualization enabled the comparison of clusters in a simplified, interpretable format, helping to assess the distinctiveness of each cluster.

## 4. Results and Discussion

### 4.1. Interpretation of Initial Exploration

The dataset used for this study consists of 10,000 customer records and 42 features, which include demographic, behavioral, and transactional information. The data was successfully loaded, and the first five rows were displayed, showing key attributes such as Customer_ID, Birth_Year, Education_Level, Marital_Status, Annual_Income, Total_Spent, and Customer_Age. The dataset was found to have no missing values, as verified by a check on all columns, with each column being completely filled. The dataset's shape was 10,000 rows by 42 columns, which was appropriate for clustering tasks. The columns primarily consisted of numerical features like Annual_Income, Spent_Wine, Spent_Fruits, and Total_Spent, as well as categorical features like Education_Level, Marital_Status, and Device_Type. This combination of data allowed for a multifaceted approach to customer segmentation based on both behavioral and demographic characteristics.

The descriptive statistics for the selected features revealed valuable insights into the distribution of customer attributes. For numerical features, the Age ranged from 24 to 85 years, with a mean of 54.2 years and a standard deviation of 17.9 years, indicating a diverse customer base spanning across various age groups. The Annual_Income ranged from a negative value (-7,845.63) to 109,134.97, with the median income around 49,967. The low minimum value suggests

ISSN 2776-8007

that there might be data entry issues, or this could be a reflection of customers with negative balances or credits, which should be further examined. The Total_Spent feature ranged from 210 to 2684, with a mean of 1,447.78, showing a wide variety in customer spending. The spending data is relatively evenly distributed, with some customers spending significantly more than others. The Total_Purchases column had a mean of 17.97, with values ranging from 1 to 35, indicating that customers engage with different levels of frequency in the purchasing process. Additionally, the Last_Purchase_Recency had a mean of 49.6, with the minimum value being 0 (indicating recent purchases) and the maximum value reaching 99, representing customers who have not purchased in a long time. This range is important for segmenting customers based on their recent activity levels. To better understand the distribution of the selected features, histograms for the numerical features were plotted. A correlation heatmap was also generated to assess the relationships between numerical features.

Figure 1 provides a visual representation of the distribution of several numerical features from the dataset. Here's a detailed explanation of each feature based on the histogram. The distribution of age shows a fairly even spread across the entire range, with a slight dip around the mid-30s to 40s. There are a few customers on the higher end (around 80+ years), but the dataset seems to consist mostly of middle-aged customers, ranging from 30 to 70 years old. Annual_Income distribution is more Gaussian, with a peak around the middle of the income range, approximately between 40,000 and 60,000. This suggests that most customers in the dataset fall within the middle-income range. There are also fewer customers at the extremes, with a smaller number of low-income (negative or near-zero values) and high-income customers. The total number of dependents (sum of kids and teens at home) is skewed toward the lower end. The most common number of dependents is 2, but there are notable counts for customers with 1 and 3 dependents. There are fewer customers with 0 or 4 dependents, indicating a typical family size of 2 or 3.

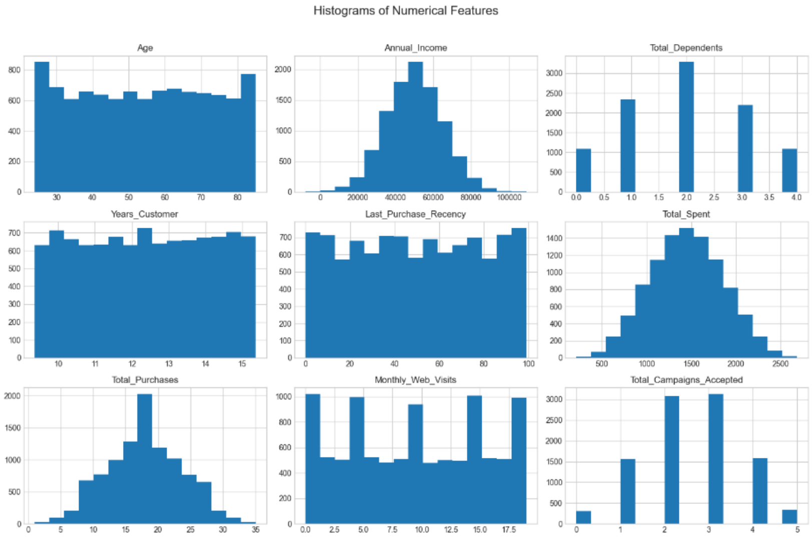

**Figure 1.** Histograms of Numerical Features

The distribution of years as a customer is nearly uniform across the range from 9 to 15 years, indicating that most customers have been with the company for a long period. This reflects a strong base of loyal customers with a consistent relationship with the company. The distribution for last purchase recency is fairly uniform, with some spikes at specific intervals. This could suggest that some customers are more frequent buyers, while others may make purchases more infrequently, ranging from recent purchases (0 days) to purchases made up to 100 days ago. The distribution for total spending shows a near-normal distribution with a peak around 1,500, suggesting that most customers spend a moderate amount. A smaller number of customers spend less or more than this average, but there are still significant amounts of spending that fall on the lower and higher ends of the spectrum.

This histogram shows a roughly Gaussian distribution with a peak around 15 purchases. It suggests that most customers make a moderate number of purchases, but there are some customers who make more than 20 or fewer than 5 purchases, indicating variability in buying behavior. This distribution of Monthly Web Visits is somewhat bimodal, with a notable peak around 4 and another around 10. It shows that many customers engage with the website consistently but at different levels of frequency. A small group of customers also show very high levels of engagement, with visits exceeding 15 times per month. The distribution of campaigns accepted shows a peak at 2 campaigns, followed by a drop and then a gradual increase in customers accepting 3 to 4 campaigns. This indicates that most customers engage with a few campaigns, but fewer customers participate in more than 3 campaigns. In summary, Figure 1 reveal the general trends in customer behavior: most customers are middle-aged, have moderate incomes, and have been with the company for a significant amount of time. They tend to have a small number of dependents, spend a moderate amount, and engage with the company’s campaigns at a modest level. The distributions also show some variation in engagement and spending, which could provide useful insights for customer segmentation and targeted marketing strategies.

Figure 2 consists of two count plots, one for Education_Level and the other for Marital_Status. The distribution of customers' education levels is skewed toward those who have completed their undergraduate studies (Graduation), which has the highest count of around 5,000 customers. A smaller number of customers have Master's and Basic education levels, with PhD being the least common. This suggests that most of the customers in the dataset have a relatively high level of education, with a significant portion holding a Bachelor's degree. The marital status distribution reveals that the majority of customers are either Married or Widowed, with the "Widow" category slightly outnumbering others. The Single and Together categories have similar counts, each being notably smaller than the Married and Widow categories. The relatively high number of Widowed customers could reflect the demographic composition of the customer base, potentially offering insight into targeted marketing strategies for different marital statuses.

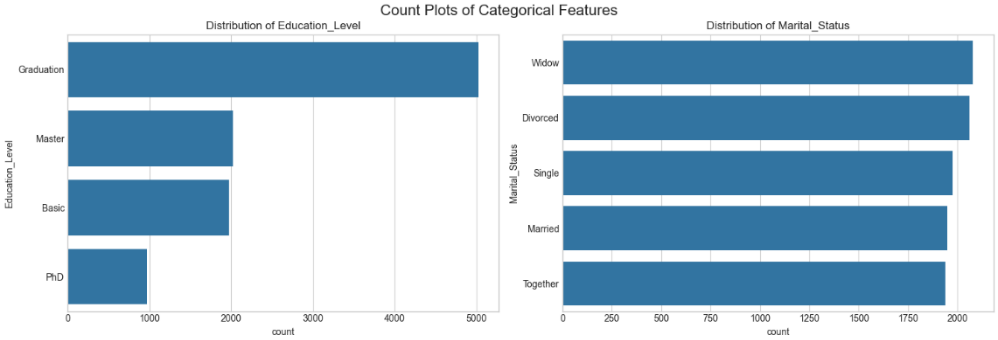

**Figure 2.** Count Plots of Categorical Features

Figure 3 presents a correlation heatmap showing the relationships between various numerical features in the dataset. Each cell in the heatmap represents the correlation coefficient between two variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with values closer to 0 indicating no correlation. The correlation between Age and other features is minimal, with values close to 0, suggesting that age does not strongly correlate with factors such as spending, purchase frequency, or recency of purchases. The Annual_Income feature shows weak correlations with other variables, indicating that income does not have a strong linear relationship with factors like Total_Spent, Total_Purchases, or Total_Campaigns_Accepted. The correlation with Total_Spent is slightly negative, indicating that customers with higher incomes might not necessarily be the highest spenders in this dataset. Total_Spent and Total_Purchases show a very slight negative correlation, suggesting that customers who spend more do not necessarily make more purchases, and vice versa.

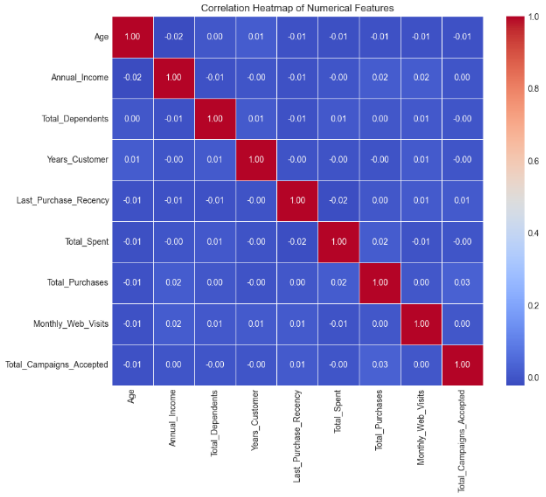

**Figure 3.** Correlation Heatmap of Numerical Features

This indicates that spending behavior might depend more on the type or volume of purchases rather than the frequency of transactions. Last_Purchase_Recency has a very weak correlation with most other features, which suggests that recency of the last purchase does not strongly predict future spending or purchase patterns in this dataset. The heatmap overall shows that many of the numerical features in the dataset do not have strong correlations with each other, which might indicate that each feature contributes unique information to the customer's behavior and characteristics. This lack of strong correlation could make clustering more challenging, as there are fewer clear linear relationships to define customer segments.

## 4.2. DBSCAN Clustering

After determining the optimal parameters for DBSCAN, the algorithm was applied to the preprocessed dataset. The epsilon (eps) value, which defines the maximum distance between points to be considered as neighbors, was heuristically chosen to be 2.5 based on the k-distance graph. The min_samples parameter, which sets the minimum number of points required to form a dense cluster, was set to 15. These values were selected through a combination of visual inspection of the k-distance graph and common heuristics for density-based clustering.

Upon applying DBSCAN, the algorithm successfully identified 1 cluster and labeled 78 data points as noise (indicated by -1). The majority of the data points (9,922) were assigned to the main cluster (Cluster 0), while the remaining 78 points were considered outliers or anomalies that did not fit well into any dense region of the dataset. This suggests that DBSCAN was able to identify one large, well-defined customer segment but struggled to segment the remaining data points effectively due to their diverse or irregular behaviors.

Figure 4 visualizes the results of DBSCAN clustering after reducing the dataset's dimensionality using Principal Component Analysis (PCA). This 2D scatter plot shows how the data points are distributed across the first and second principal components. The majority of the data points are represented in yellow, indicating they belong to the single identified cluster (Cluster 0). However, a small number of points are colored purple, which represent the noise (denoted by -1 in DBSCAN). These noise points are data points that DBSCAN was unable to group into a dense cluster. This suggests that most of the customers in the dataset share similar characteristics, but there are a few outliers with behaviors that deviate from the main group. The points in the cluster are well spread out across the PCA-reduced dimensions, which is expected when a density-based algorithm like DBSCAN is applied. The algorithm does not assume that clusters are spherical, allowing it to detect more complex patterns. However, as only one cluster was detected, this indicates that the dataset may not have had clear boundaries between different customer groups, or the DBSCAN parameters (specifically eps and min_samples) may need further adjustment to reveal more clusters.

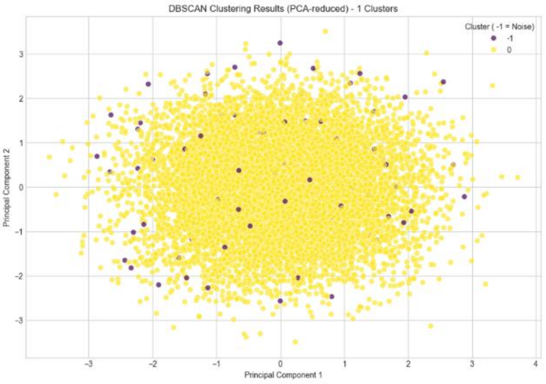

**Figure 4.** DBSCAN Clustering Results (PCA-reduced)

Figure 5 displays the relationship between scaled Age and scaled Annual Income, with DBSCAN cluster labels as the color indicator. Similar to the PCA-reduced plot, most data points are colored yellow, indicating that they belong to the same cluster (Cluster 0). The small number of purple points represent noise, which suggests that these outliers did not fit well with the general customer behavior. This plot helps visualize how the customer groups are distributed in terms of both age and income, even after the features have been scaled to bring them to a common range. The scatter plot shows that customers with varying ages and income levels are spread across a wide range, but they still fall into a single large cluster. The scaled values of Age and Annual Income show that there is little distinction between customers in terms of these two variables after scaling. The spread of the points suggests that both Age and Annual Income do not strongly separate the customers into distinct groups, meaning these two features alone do not define separate clusters. The large number of customers falling within the same cluster emphasizes the uniformity in behavior, suggesting that the parameters of DBSCAN need further fine-tuning to capture more nuanced customer segments.

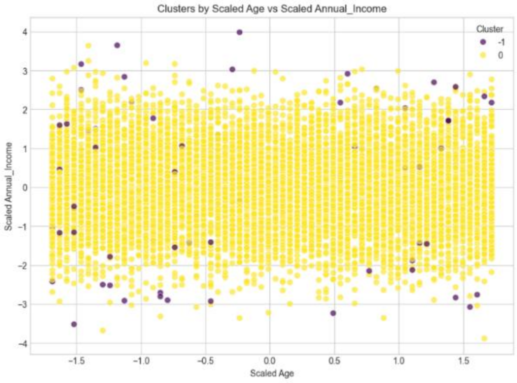

**Figure 5.** Clusters by Scaled Age vs Scaled Annual Income

The evaluation of the clustering results was limited by the fact that only one cluster was identified. This lack of multiple clusters made it impossible to calculate meaningful metrics like the Silhouette Score or the Davies-Bouldin Index, which require at least two distinct clusters for evaluation. These metrics are typically used to assess the compactness and separation of clusters, with higher Silhouette Scores indicating better-defined clusters and lower Davies-Bouldin Scores suggesting better cluster separation. However, in this case, since the dataset was predominantly grouped into a single cluster, the lack of additional clusters prevented the application of these evaluation metrics.

Despite the limited clustering output, a detailed profiling of the single cluster (Cluster 0) was conducted. The profiling involved calculating the mean of numerical features for each cluster, which provided insights into the typical characteristics of customers in the identified cluster. The average Age of customers in Cluster 0 was approximately 54.2 years, which is relatively high, suggesting that this cluster primarily consists of middle-aged to older customers. The Annual_Income for customers in this cluster averaged around 49,933, indicating a relatively mid-to-high income group. The Total_Dependents value averaged 1.99, meaning that most customers in this cluster have a small family or dependents. In terms of customer engagement, the average Years_Customer was 12.38 years, which indicates that these customers have been with the company for a significant period. The Last_Purchase_Recency for this cluster was around

49.56, suggesting that, on average, customers in this segment made purchases within the last 50 days. The average Total_Spent was 1,448.58, indicating moderate spending levels, while the Total_Purchases averaged 17.97, suggesting frequent, though not excessive, purchasing behavior. Additionally, the cluster showed an average of 9.47 Monthly Web Visits, which reflects moderate engagement with the company's website. The average number of Total_Campaigns_Accepted was 2.52, indicating that customers in this cluster are moderately responsive to marketing campaigns.

From the profiling, it becomes clear that Cluster 0 represents a large group of long-term, moderately high-income customers who engage with the platform regularly and are somewhat responsive to marketing campaigns. They are not the most frequent buyers, but their consistent spending and moderate level of engagement make them a valuable group for targeted marketing and customer retention efforts. Given that only one cluster was identified, it is important to note that the segmentation did not fully capture the diversity in customer behavior. Many customers might exhibit behaviors that do not align with this primary cluster, which is why the algorithm classified them as noise. This suggests that the current DBSCAN settings might not be ideal for a more granular segmentation of the customer base, and further tuning of parameters or the exploration of alternative clustering techniques could improve the results.

The clustering analysis, while limited to a single cluster, has provided valuable insights into the characteristics of a significant customer segment. The profiling of this cluster has helped to identify key demographic and behavioral traits that can guide future marketing and engagement strategies. However, the presence of a large number of noise points and the lack of multiple clusters highlight the need for further parameter tuning or exploration of different clustering methods to capture more nuanced customer segments. Future work could involve adjusting the DBSCAN parameters, such as eps and min_samples, or applying alternative algorithms, like K-Means or hierarchical clustering, to obtain a more detailed segmentation of the customer base.

## 5. Conclusion

DBSCAN clustering successfully identified meaningful customer segments in the e-commerce dataset, though the results were limited by the parameter settings. The majority of the data points were grouped into a single cluster, while a small portion was classified as noise, indicating that DBSCAN was able to find a dominant customer segment. This highlights the potential of DBSCAN for customer segmentation, especially in datasets with complex, non-spherical cluster shapes. The identified segment provides a general overview of the customer base, revealing similarities in demographics and behavior. However, the findings also point to some limitations in the current approach. The choice of DBSCAN parameters, specifically ε (epsilon) and MinPts, played a crucial role in the outcome, and further tuning could reveal additional clusters or refine the segmentation process. Additionally, the presence of noise points suggests that some customers exhibit behaviors that deviate from the main group, which may need further investigation. Future research could focus on hybrid models that combine DBSCAN with other clustering algorithms or incorporate additional behavioral features to enhance the segmentation accuracy. This would allow for more nuanced customer profiles, enabling more targeted marketing strategies and personalized offers.

## 6. Declarations

### 6.1. Author Contributions

Conceptualization: A.A., I.A.; Methodology: A.A., I.A.; Software: A.A.; Validation: I.A.; Formal Analysis: A.A.; Investigation: A.A.; Resources: I.A.; Data Curation: A.A.; Writing – Original Draft Preparation: A.A.; Writing – Review and Editing: I.A.; Visualization: A.A.; All authors have read and agreed to the published version of the manuscript.

### 6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

## 6.4. Institutional Review Board Statement

Not applicable.

## 6.5. Informed Consent Statement

Not applicable.

## 6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**

[1] H. Sharma and A. G. Aggarwal, “Finding Determinants of E-Commerce Success: A PLS-SEM Approach,” *J. Adv. Manag. Res.*, 2019, doi: 10.1108/jamr-08-2018-0074.

[2] A. M. Ahmed Serwah, K. W. Khaw, C. S. Peck Yeng, and A. Alnoor, “Customer Analytics for Online Retailers Using Weighted K-Means and RFM Analysis,” *Data Anal. Appl. Math. Daam*, 2023, doi: 10.15282/daam.v4i1.9171.

[3] W. Wei, “Data Marketing Optimization Method Combining Deep Neural Network and Evolutionary Algorithm,” *Wirel. Commun. Mob. Comput.*, 2022, doi: 10.1155/2022/1646268.

[4] K. Tabianan, S. R. Velu, and V. Ravi, “K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data,” *Sustainability*, 2022, doi: 10.3390/su14127243.

[5] M. Hahsler, M. Piekenbrock, and D. Doran, “dbscan: Fast Density-Based Clustering With R,” *J. Stat. Softw.*, 2019, doi: 10.18637/jss.v091.i01.

[6] S. Monalisa, Y. Juniarti, E. Saputra, F. Muttakin, and T. K. Ahsyar, “Customer Segmentation With RFM Models and Demographic Variable Using DBSCAN Algorithm,” *Telkomnika Telecommun. Comput. Electron. Control*, 2023, doi: 10.12928/telkomnika.v21i4.22759.

[7] W. Lai, M. Zhou, F. Hu, K. Bian, and Q. Song, “A New DBSCAN Parameters Determination Method Based on Improved MVO,” *Ieee Access*, 2019, doi: 10.1109/access.2019.2931334.

[8] E. Xiao, “Comprehensive K-Means Clustering,” *J. Comput. Commun.*, 2024, doi: 10.4236/jcc.2024.123009.

[9] Z. Wang, “Customer Segmentation Based on Machine Learning Methods,” *Highlights Sci. Eng. Technol.*, 2024, doi: 10.54097/g70xqb16.

[10] R. Rahmawati, “Profiling Shopping Mall Costumer Based on Demographics and Shopping Motivation,” *J-Mkli J. Manaj. Dan Kearifan Lokal Indones.*, 2019, doi: 10.26805/jmkli.v3i2.64.

[11] M. M. Hassan, “Customer Profiling and Segmentation in Retail Banks Using Data Mining Techniques,” *Int. J. Adv. Res. Comput. Sci.*, 2018, doi: 10.26483/ijarcs.v9i4.6172.

[12] W. Wiharto, A. K. Wicaksana, and D. E. Cahyani, “Modification of a Density-Based Spatial Clustering Algorithm for Applications With Noise for Data Reduction in Intrusion Detection Systems,” *Int. J. Fuzzy Log. Intell. Syst.*, 2021, doi: 10.5391/ijfis.2021.21.2.189.

[13] L. Abednego, C. E. Nugraheni, and A. Salsabina, “Customer Segmentation: Transformation From Data to Marketing Strategy,” *Conf. Ser.*, 2023, doi: 10.34306/conferenceseries.v4i1.645.

[14] H. Fu, H. Li, Y. Dong, F. Xu, and F. Chen, “Segmenting Individual Tree From TLS Point Clouds Using Improved DBSCAN,” *Forests*, 2022, doi: 10.3390/f13040566.

[15] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, “Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods,” *Lontar Komput. J. Ilm. Teknol. Inf.*, 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.

[16] S. Hidayati, A. T. Darmaliana, and R. Riski, “Comparison of K-Means, Fuzzy C-Means, Fuzzy Gustafson Kessel, and DBSCAN for Village Grouping in Surabaya Based on Poverty Indicators,” *J. Pendidik. Mat. Kudus*, 2022, doi: 10.21043/jpmk.v5i2.16552.

[17] F. H. Kristanto, H. W. Rahma, and M. Nahrowi, “Factors Affecting E-Commerce Customer Loyalty in Indonesia,” *J. Syntax Transform.*, 2022, doi: 10.46799/jst.v3i09.613.

[18] H. Zhang and J. Dong, “Prediction of Repeat Customers on E-Commerce Platform Based on Blockchain,” *Wirel. Commun. Mob. Comput.*, 2020, doi: 10.1155/2020/8841437.

[19] F. Alzami *et al.*, “Implementation of ETL E-Commerce for Customer Clustering Using RFM and K-Means Clustering,” *J. Ilm. Merpati Menara Penelit. Akad. Teknol. Inf.*, 2022, doi: 10.24843/jim.2022.v10.i03.p05.

[20] F. Alzami *et al.*, “Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce With Streamlit,” *Ilk. J. Ilm.*, 2023, doi: 10.33096/ilkom.v15i1.1524.32-44.