**Enhancing Customer Segmentation in Online Retail Using the DBSCAN Algorithm**

Stanley Ernst D. Gonatice

Ben Florence A.J. E. Til

Notre Dame of Marbel University

Bachelor of Science in Computer Science

Brenda M. Balala, MIT

December 2024

# Approval Sheet

This Thesis entitled **Enhancing Customer Segmentation in Online Retail Using the DBSCAN Algorithm,** prepared, and submitted by **Stanley Ernst D. Gonatice and Ben Florence A.J E. Til**, in partial fulfillment of the requirement for the degree **Bachelor of Science in Computer Science** is hereby accepted.

|  |  |  |
| --- | --- | --- |
| **Brenda M. Balala, MIT** | | |
| Thesis Adviser | | |
|  |  |  |
|  |  |  |
| **Merch Jay P. Dollaga, MIT** | | |
| Panel Chair | | |
| **Engr. Victorino R. Tobias, Jr., MEP-ECE** |  | **Hajah T. Sueno, DIT** |
| Panel Member |  | Panel Member |

Accepted and approved for the conferral of the degree

**Bachelor of Science in Computer Science**

**Engr. Victorino R. Tobias, Jr., MEP-ECE**

Dean, College of Engineering, Architecture, and Computing

**Table of Contents**

[**Approval Sheet 2**](#_heading=h.gjdgxs)

[**List of Figures 5**](#_heading=h.1fob9te)

[**Introduction 6**](#_heading=h.3znysh7)

[**Background of the Study 6**](#_heading=h.2et92p0)

[**Objectives of the Study 7**](#_heading=h.3dy6vkm)

[**Review of Related Literature 9**](#_heading=h.2s8eyo1)

[Customer Segmentation Techniques 9](#_heading=h.17dp8vu)

[Challenges in Machine Learning for Retail 10](#_heading=h.26in1rg)

[Clustering Algorithms in E-Commerce 10](#_heading=h.lnxbz9)

[Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods 11](#_heading=h.35nkun2)

[Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce 12](#_heading=h.1ksv4uv)

[**Related Studies 14**](#_heading=h.z337ya)

[Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis 14](#_heading=h.3j2qqm3)

[K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data 15](#_heading=h.1y810tw)

[Classification of Online Retail Customers using Machine Learning Techniques 16](#_heading=h.4i7ojhp)

[Mall Customer Segmentation Using K-Means Clustering 16](#_heading=h.2xcytpi)

[An Unsupervised Learning-Based System Employing K-Means Clustering to Perform Customer Segmentation 17](#_heading=h.1ci93xb)

[An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. 17](#_heading=h.3whwml4)

[Designing and Implementing Customer Segmentation and Classification Systems 18](#_heading=h.2bn6wsx)

[Customer Segmentation in Retail: An Experiment in Sweden 18](#_heading=h.qsh70q)

[Synthesis 19](#_heading=h.3as4poj)

[Concept of the Study 21](#_heading=h.49x2ik5)

[**Significance of the Study 23**](#_heading=h.147n2zr)

[**Scope and Limitations 25**](#_heading=h.3o7alnk)

[**Operational Definition of Terms 26**](#_heading=h.32hioqz)

[**Methods 29**](#_heading=h.1hmsyys)

[**Materials 29**](#_heading=h.41mghml)

[Hardware 29](#_heading=h.2grqrue)

[Software 29](#_heading=h.vx1227)

[Data 32](#_heading=h.3fwokq0)

[**Procedures 32**](#_heading=h.1v1yuxt)

[**References: 37**](#_heading=h.4f1mdlm)

List of Figures

[Figure 1: Literature Map 13](#_heading=h.44sinio)

[Figure 2. Conceptual Framework of the Study 19](#_heading=h.2p2csry)

# Introduction

## Background of the Study

The internet has proved to be an asset in the modern corporate world. Most firms select their business model as they try to benefit from online strategies for better growth, profit, reputation, and customer matching. It may further capture the online customer's location, demographics, psychographics, and purchase data. Due to the abundance of online data, big data analytics can be used with specific algorithms to obtain customer-centric insights. Customer segmentation, or the process of dividing consumers into homogeneous groups, is seen as an efficient way to manage customers while creating a variety of marketing techniques. Segmentation based on consumer attributes, which are monitored online with a specific algorithm, is possible. The business must concentrate on its target audience to optimize profits and create win-win circumstances for both parties. Segmenting customers is one way to maximize the outcome of a win-win scenario (Kadir & Achyar, 2019).

Machine learning in customer segmentation has become very popular, especially with the rise of big data. Machine learning algorithms can uncover complex data patterns and insights that other methods cannot easily detect. Applying such algorithms would better segment their customers and predict future behaviors that could or may not be positive: churn or purchasing intent. Researchers have found that using machine learning for customer segmentation helps marketers better focus on conversion and retention (Patankar et al., 2021).

In e-commerce, the demand for advanced customer segmentation has increased, leading to more personalized experiences. Powered by customer data, data-driven personalization ensures that the right messages and offers are in front of the right customer at the right time; this is one of the most critical targeting levels businesses need to maintain to be at par with competitors in the age of digitalization. According to some research, Companies applying data-driven personalization tactics have proven to increase customer loyalty and lifetime value (Sun, 2024).

With online consumer behavior continuously changing, more dynamic and adaptive customer segmentation models are needed. Direct demographic data formed the basis of segmentation models for a long time. Still, the scenario has changed, and companies today need to look into more advanced techniques that entail behavior-based data, psychographics, and real-time customer interactions. An amalgamation of deep learning with clustering methods has already proven to boost the accuracy of the segmentation process. It thus enables quick and agile responses to customer preferences in flux (Tabianan et al., 2022).

This research aims to address the issues posed by evolving consumer behavior and enhance machine learning models for client segmentation, particularly in online shopping. The goal is to build models that offer more intelligible and practical insights, enhancing consumer experiences and increasing the efficacy of marketing campaigns.

## Objectives of the Study

This study aims to enhance customer segmentation in the online retail sector using machine learning models, with the following specific objectives:

1. Determine the customer buying patterns using RFM, in terms of:
2. Recency,
3. Frequency,
4. Monetary value,
5. Determine the customer buying patterns using LRFMP, in terms of:
   * 1. Length,
     2. Recency,
     3. Frequency,
     4. Monetary,
     5. Periodicity.
6. Compare customer purchasing behavior using clustering techniques, including K-Means, Hierarchical Clustering, and DBSCAN, across segmentation criteria such as:
   * 1. Recency,
     2. Frequency,
     3. Monetary value,
     4. Length,
     5. Periodicity.
7. Visualize a web-based dashboard that interactively visualizes customer profiling, in terms of:
8. Dormant
9. Active
10. Loyal

## Review of Related Literature

This section reviews related literature and systems that will help researchers gather basic information and references for the current study.

### Customer Segmentation Techniques

Customer segmentation is essential for customizing marketing strategies and maximizing customer satisfaction in online retail. Traditional forms of customer segmentation include demographic factors such as age and gender, income, and even geographic location. With the rise of e-commerce, however, customer behavior has become more complex, and more complex segmentation methods are required. Behavioral segmentation: This deals with customers' behavior regarding interaction on a website, purchase history, browsing habits, and engagement in promotional activities. Recently, studies in machine learning techniques have focused on k-means clustering and hierarchical clustering algorithms to achieve these purposes. These algorithms scan large datasets to look deeply into hidden patterns not easily obtained through other methods. For instance, k-means clustering segments of customers based on similarity. Consequently, businesses can target offers that are custom-fit to these identified classes. By using this selection method, companies can improve the quality of customer experience and design better-targeted marketing campaigns, thereby increasing engagement and sales. Recent studies demonstrated that applying such an approach led to higher segmentation accuracy and helped reveal new groups of customers that the current approaches may ignore (Patankar et al., 2021).

### Challenges in Machine Learning for Retail

Despite the promising performance of models based on machine learning, several challenges remain. One such challenge is data quality. While online retail platforms generate voluminous amounts of data from vast sources, this data includes customer reviews, transaction histories, and browsing patterns, which may sometimes contain noise, missing values, and inconsistencies that quickly affect the model's performance. Feature selection is the most critical challenge. The most important attributes or features are selected for use in the model. Features that are commonly used in customer segmentation are purchase frequency, the value of a customer over his lifetime, and recency of the customer's interaction with an enterprise, whereas numerous irrelevant features cause overfitting-that is, good performance on the training data set but poor performance on the unseen data set. Furthermore, bias in the algorithms should be addressed. Bias in machine learning is found when the model inadvertently favors some of the customer groups over others and hence fails to provide fair treatment or misclassification. This is important for retail since it guarantees the fairness of the different consumer groups, which is necessary to earn the trust and satisfaction of customers. Nguyen & Chen, 2020 state that because machine learning is becoming increasingly important, more attention must be paid to creating adaptive models, feature engineering strategies, and ways to enhance data pretreatment (Sun, 2024).

### Clustering Algorithms in E-Commerce

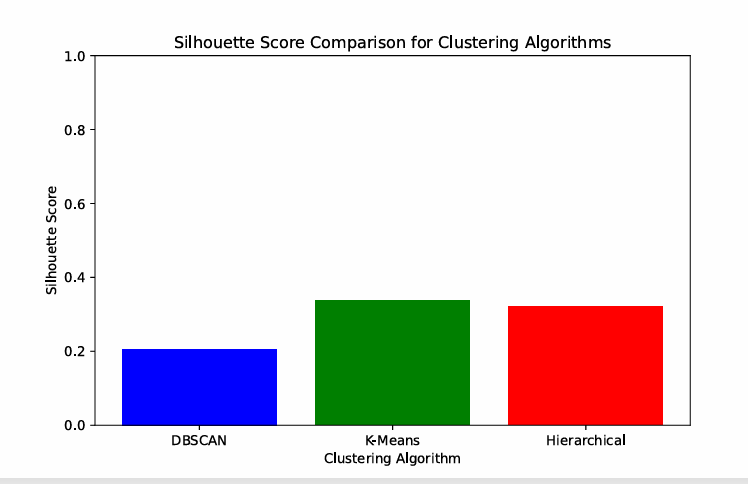
Clustering algorithms have gained immense momentum in the e-commerce scenario for customer segmentation due to their ability to group similar customers along with features. K-means is one of the most commonly used algorithms in the field. It simply groups customers into groups (or clusters) by minimizing the distance between customers within the same cluster. This is highly effective for customer segmentation on behavioral attributes, including purchase frequency, average order value, and product preference. However, the applications of k-means and other traditional algorithms are minimal, especially with large, high-dimensional datasets. Again, advanced techniques like deep clustering have been developed to combine a neural network with a clustering algorithm to improve segmentation accuracy and pattern recognition. Deep models can digest large-scale data, learn intricate patterns, and help e-commerce businesses seize the opportunities that result from a dynamic customer environment. With deep clustering, retailers can produce more dynamic and adaptive segmentation models. For example, a model that automatically updates customer segments in real-time based on new interactions or enables more responsive marketing and product recommendations (Tabianan et al., 2022).

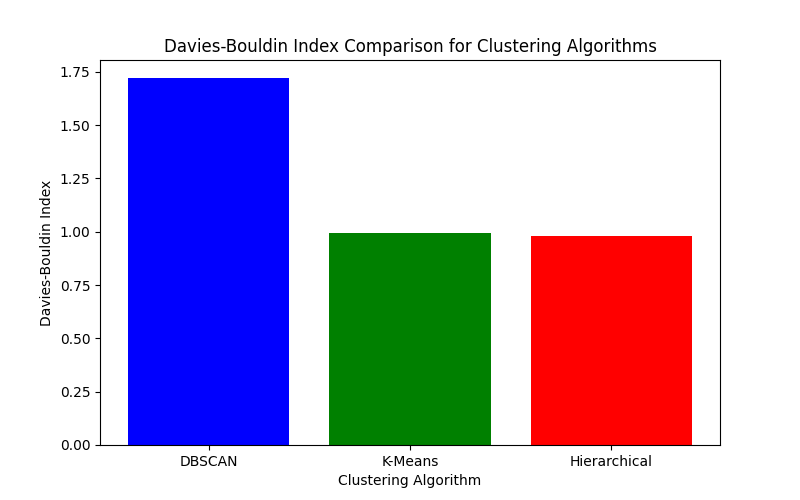
### Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods

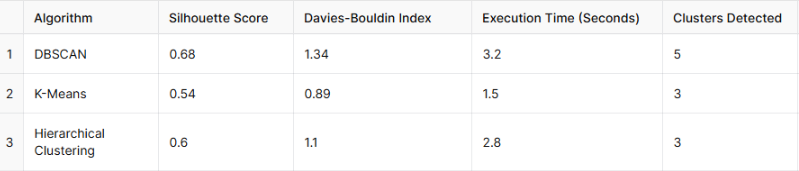
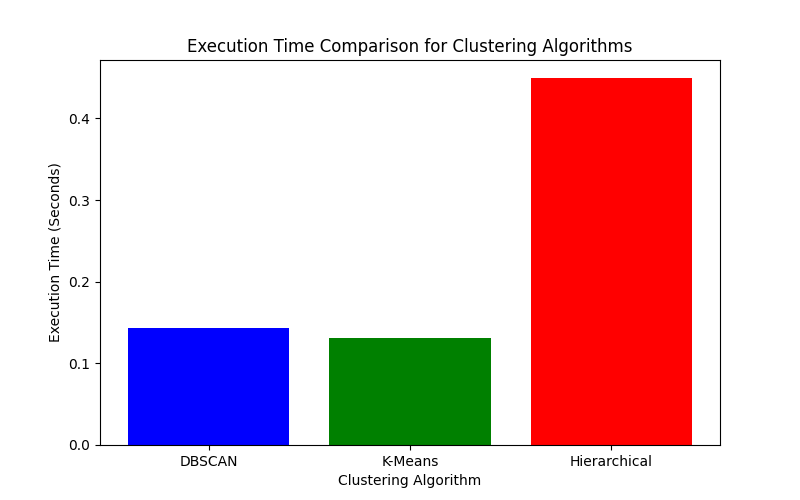
Utilizing the Recency, Frequency, Monetary (RFM) model, researchers analyzed transaction records of more than 334,000 entries from online retail sales. On the basis of their respective shopping patterns, they estimated monetary, frequent, and recent values of transactions done by their clients. Using K-Means, K-Medoids, and DBSCAN clustering algorithms on the analysis for an even more precise customer segmentation, K-Means proved to show more accurate performances in clustering, thus consisting of precise and dependable consumer grouping. This enables businesses to focus positively on revenue-generating customers. DBSCAN proved to be very effective in a scenario where complex shapes of clusters and noisy data were involved. The work highlights an important point that somehow amalgamates the clustering algorithms with the RFM model in an e-commerce ecosystem. For the organizations, this study helps them fine-tune the marketing mix and indulge further with high-value customers by finding different types of customers in the market (Sembiring Brahmana et al., 2020).

### Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce

Comparison of K-Means and DBSCAN illustrates merits and demerits of the two clustering algorithms in using customer segmentation towards e-commerce. K-Means is a simple technique with easy computations while processing large amounts of data. However, this requires the number of clusters to be predefined. It is adverse for cases with highly irregular shapes for cluster or noise in the data. Indeed, the strength of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is in its ability to detect clusters which do not need to be of one shape and density; the number of clusters is not pre-specified. From the result, it was observed that DBSCAN achieved a higher silhouette score as 0.680 whereas K-Means achieved 0.546; hence, DBSCAN is a better choice since it formed more distinctively separated clusters than the K-Means. However, DBSCAN obtained a higher Davies-Bouldin Index of 1.344 meaning that its clusters were not so compact like those of K-Means. DBSCAN is strong at dealing with difficult data and can even identify niche markets; on the other hand, it fails at compactness in clustering.

Figure 1: Silhouette Score Comparison for Clustering Algorithm

Figure 2. Davies-Bouldin Index Comparison for Clustering Algorithms.

Figure 3. Execution Time Comparison for Clustering Algorithms.  
Figure 4:

The paper, therefore, shows that a choice of an algorithm will depend upon specific company needs; for example, companies that want very balanced and well-defined segments would continue to prefer K-Means. However, if customer behavior were to be irregular or noisy, DBSCAN would be more flexible (Paramita, 2024).

Figure 5: Literature Map

A diagram of a diagram

Description automatically generated with medium confidence

## Related Studies

This section presents studies conducted to which the present proposed study is related or has some bearing or similarity.

### Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis

Machine learning (ML) has been widely applied in customer analytics, particularly in predicting behaviors like churn and enhancing customer segmentation. Research in this domain demonstrates the effectiveness of combining ML techniques with traditional approaches such as recency, frequency, and monetary (RFM) analysis to derive actionable insights. For example, studies on customer churn prediction emphasize the use of demographic, social, transactional, and behavioral data to model customer behavior effectively. However, as many businesses often rely solely on transactional data, particularly those sourced from enterprise resource planning systems, the potential for comprehensive modeling remains underutilized.

Recent studies have explored the integration of ML techniques, such as K-means and DBSCAN clustering, with RFM analysis to address this limitation. These approaches help businesses segment customers into distinct groups based on key behavioral metrics. For instance, research has shown that combining RFM scores with unsupervised ML clustering can effectively group customers into practical segments, even when limited to transactional data. One such study concluded that dividing customers into six distinct clusters provided a straightforward and actionable framework for customer segmentation, demonstrating the flexibility and utility of DBSCAN in handling noise and irregular patterns in data.(Lewaaelhamd, 2023)

### K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data

E-commerce systems have become increasingly prevalent, serving as platforms for marketing and promoting products to customers through online channels. Customer segmentation, a critical aspect of e-commerce, involves dividing customers into groups based on shared characteristics to enhance business profitability. By categorizing customers, businesses can identify high-value segments, optimize services, and deliver targeted marketing strategies. This process enables e-commerce platforms to promote the right product to the right customer, ultimately driving profitability and customer satisfaction.

A study focused on customer segmentation within e-commerce emphasized the behavioral factors influencing customer purchase behavior. Clustering algorithms were utilized to group customers with similar purchasing habits, enabling the identification of profitable segments. In this study, three clustering dimensions—event type, products, and categories—were analyzed to explore customer relationships. The clustering process aimed to maximize similarity within clusters while increasing dissimilarity between clusters, helping vendors identify and focus on high-profit customer segments.(Tabianan et al., 2022)

### Classification of Online Retail Customers using Machine Learning Techniques

The application of machine learning techniques for customer segmentation and classification has gained significant momentum in recent years. One study designed, developed, and implemented two systems to predict customer classes using distinct approaches: the first based on product purchase history and the second utilizing the Recency, Frequency, and Monetary (RFM) value of each customer. These systems aimed to improve segmentation accuracy and provide actionable insights for businesses.

The study employed K-Means clustering for customer segmentation and experimented with eight classification models, including logistic regression, gradient boosting, and random forest, to predict the classes of new customers. Both systems achieved prediction accuracies exceeding 90%, demonstrating the effectiveness of combining clustering and classification techniques in customer segmentation.(Eshra, 2021)

### Mall Customer Segmentation Using K-Means Clustering

This research paper investigates the application of K-means clustering to segment mall customers using a dataset that contains various customer attributes. Customer segmentation is a vital component of marketing strategies, enabling businesses to develop more targeted and personalized marketing efforts. In this study, specialized customer attributes such as demographics, spending habits, and visit frequency were analyzed to group mall customers into distinct segments. The study applied K-means clustering and finalized six clusters, optimized using the elbow method. The clusters were interpreted to reveal insights into customer behavior, specifically regarding expenditure in the mall. Detailed exploratory research provided valuable insights for mall management to better understand their customer base and implement more effective marketing strategies. (Ashwani et al., 2023)

### An Unsupervised Learning-Based System Employing K-Means Clustering to Perform Customer Segmentation

This study focuses on the increasing role of artificial intelligence (AI) in improving customer segmentation and predictive targeting for marketers, particularly in the growing field of online advertising. The research highlights the significant advancements AI has brought in allowing marketers to narrow down to specific customer niches and target consumers with personalized advertising. By utilizing electronic devices, consumers are generating data at an unprecedented rate, which is being used for more refined micro-targeting based on individual traits and behaviors. (Gupta, 2023)

### An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market.

It is a study on Clustering Algorithms for UK Retail Market Customer Segmentation. Such a study fascinates me as it is associated with the growing trend of online purchases and the increasing relevance of customer segmentation within the retail sector today. To that aim, I used a UK-based retail dataset sourced from the UCI Machine Learning Repository and worked on developing a consumer segmentation model. To improve decision-making in the retail market sector, the authors assess and quantify customer value using the Recency, Frequency, and Monetary (RFM) framework. In this study, several cutting-edge (SOTA) clustering algorithms were compared, such as balanced iterative reducing and clustering using hierarchies (BIRCH), agglomerative clustering, density-based spatial clustering of applications with noise (DBSCAN), Gaussian mixture model (GMM), and K-means clustering. The GMM algorithm fared better than the other techniques, obtaining a high Silhouette Score of 0.80, suggesting a better cluster cohesion ratio to separation (John et al., 2023).

### Designing and Implementing Customer Segmentation and Classification Systems

Design and application of systems for customer classification and segmentation based on machine learning approaches. This paper discusses two different methods to classify and segment clients. The first employs the RFM model, which is recency, frequency, and monetary, and the second one is the previous product buy. The study employed k-means clustering and eight classification techniques, including Random Forest, Gradient Boosting, and Logistic Regression (Eshra, 2021).

With customer class prediction accuracy rates above 90%, both methodologies proved their worth in accurately segmenting and forecasting consumer behavior in the retail sector (Chaubey, 2022, p. 4).

### Customer Segmentation in Retail: An Experiment in Sweden

Memon conducted research in 2021 on Swedish pharmacy retail customer segmentation, using the RFM model with LRFMP (supplement to the data set: Length, Recency, Frequency, Monetary, Periodicity) with combined K-means and Agglomerative Clustering. Segmentation happened when the LRFMP model combined with K-means and Agglomerative clustering models pointed out the most effective segmentation in the given solution. This would create two main customer profiles: "low contribution customers" and "high contributing loyal customers." Insights emerge that point out how a well-tailored segmentation model might assist in customer relationship management and further help profitable customer groups focus (Memon & Kuratomi Hernandez, 2022).

### Synthesis

Numerous studies have demonstrated the growing importance of customer segmentation, particularly with the rise of online shopping and the availability of large-scale consumer data. These studies have explored various machine learning techniques, frameworks, and clustering algorithms to enhance segmentation and classification in retail environments.

John et al. (2023) analyzed the UK retail market using a range of state-of-the-art clustering algorithms, including BIRCH, K-Means, DBSCAN, and Gaussian Mixture Model (GMM), alongside the RFM framework to calculate customer value. Their research revealed that GMM outperformed other methods, achieving a Silhouette Score of 0.80, indicating effective customer segmentation with clear cluster distinctions.

Similarly, Eshra (2021) proposed two systems for segmenting and classifying customers using product purchase history and the RFM model. By applying K-Means clustering and eight classification models, including logistic regression and gradient boosting, the study achieved over 90% accuracy in predicting customer classes. These findings emphasize the effectiveness of integrating clustering and machine learning techniques in customer analytics.

Memon and Kuratomi Hernandez (2022) explored segmentation in the Swedish retail market, combining the RFM and LRFMP models with K-Means and Agglomerative Clustering. Their approach successfully identified two main customer profiles, "low-contribution customers" and "high-contributing loyal customers," demonstrating the value of tailored segmentation models in enhancing customer relationship management and profitability.

Other studies have focused on applying clustering techniques to e-commerce environments. Tabianan et al. (2022) emphasized the behavioral aspects of customer segmentation by analyzing event types, products, and categories in an e-commerce system. Using K-Means clustering, the research enabled vendors to identify high-value customer segments, enhancing targeted marketing strategies.

Additionally, Gupta (2023) highlighted the role of artificial intelligence (AI) and unsupervised learning techniques in customer segmentation. By leveraging large-scale consumer data, the study demonstrated how AI-powered segmentation can enable precise micro-targeting and personalized advertising, thus improving marketing outcomes.

These works collectively underscore the significance of integrating machine learning and clustering algorithms with models like RFM and LRFMP. They reveal the practical benefits of using techniques such as DBSCAN, K-Means, and advanced classification models to achieve more effective segmentation. The findings consistently emphasize that well-designed customer segmentation frameworks can optimize marketing strategies, foster customer loyalty, and increase overall profitability in retail and e-commerce industries.

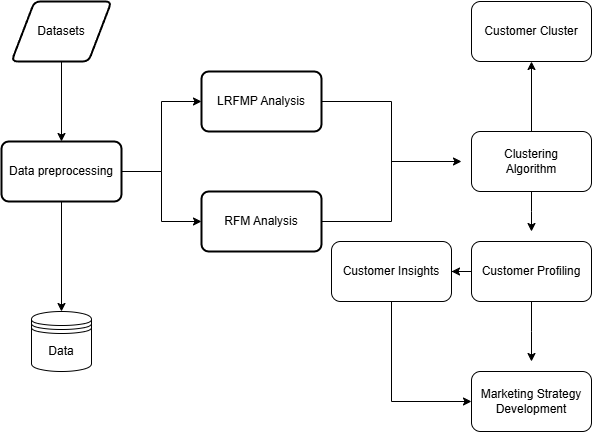
### Concept of the Study

This study examines the significance of consumer segmentation in the retail industry, particularly in light of the growth of e-commerce. We aim to use machine learning algorithms and sophisticated clustering techniques to effectively classify and anticipate customer behavior in the area.

The RFM (Recency, Frequency, Monetary) and LRFMP (Length, Recency, Frequency, Monetary, Periodicity) models will be our primary tools for identifying distinctive client characteristics exclusive to retail environment. The insights from this data, which will highlight regional consumer tendencies, will guide more focused marketing efforts.

Our main goal is to demonstrate how retailers can enhance their customer relationship management through effective segmentation. By adopting this approach, they'll be better equipped to tailor their strategies for increased profitability and customer satisfaction.

**Figure 2.** Conceptual Framework of the Study



The customer segmentation framework begins with data collection and preprocessing to ensure consistency. It then uses LRFMP and RFM analyses to examine customer behaviors, focusing on metrics like purchase recency, frequency, and spending. The analyzed data is fed into a clustering algorithm to group customers into segments with similar characteristics. Each segment is profiled to understand its unique traits, leading to actionable insights. Finally, these insights guide the development of targeted marketing strategies, enabling businesses to engage each customer segment more effectively, enhancing customer satisfaction and loyalty.

## Significance of the Study

This research work will significantly add value to various stakeholders, including businesses, academic communities, and individuals doing business in the e-commerce sector. This research contributes to improved machine learning models for customer segmentation with the following key contributions:

**For E-commerce Companies:** Improved consumer segmentation for online shopping helps businesses understand their customers' behavior and preferences. With the help of this improved machine learning model, the business is able to create more precise and tailored marketing campaigns that increase sales, customer pleasure, and loyalty. Customized recommendations and targeted promotions can significantly increase conversion rates and customer retention, giving businesses a competitive edge in the market.

**For Data Scientists and AI Practitioners:** Especially in the area of client segmentation, this work adds to the body of recognized literature on machine learning. With regard to applying machine learning to large, complicated retail datasets, the results point to the need for more accurate clustering methods and feature engineering techniques. These advancements can also serve as a solid foundation for future work in other industries, such as banking, health, or logistics, where segmentation and customer behavior forecasts are essential.

**For Academics and Future Researchers:** Scholars and future researchers examining areas of convergence between machine learning, customer analytics, and e-commerce applications will find valuable insights from this work. These research, data quality, feature selection, and even algorithmic fairness related challenges open up new lines of inquiry for AI applications and expand the scope of potential future research. Future researchers may refer to advanced machine learning models that have been built or are in development for more research, particularly in the areas of customer segmentation across many industries.

**For Students:** Students studying computer science, data science, and business analytics will find the study to be a useful resource. When the student is pursuing a career in AI and e-commerce, the in-depth discussion of machine learning models and their practical applications in client segmentation will prepare him for research projects. It might even suggest themes for further research on consumer behavior and AI-powered marketing tactics.

**For Customers:** Customers will ultimately obtain this study work, even though they won't be consulted on technical matters. Improved client experience through more relevant and appropriate offers, services, and goods are targeted by enterprises with the use of more precise consumer segmentation. By atomizing touch or interaction, more precise customer segmentation will decrease irrelevant marketing and enable customer satisfaction reductions.

**For the E-commerce Sector:** The improved machine learning technologies available to them would help the larger e-commerce industry. Businesses might use these models, as well as improved versions of them, to make sure the aforementioned clients are taken care of and to implement data-driven marketing strategies. This might establish the standard for personalized shopping experiences, promote innovation, and improve industry norms.

**For Future Entrepreneurs:** Entrepreneurs that are interested in starting or growing an internet commerce business will find this information to be beneficial. It enables them to create platforms that are intelligent and more flexible, better meeting the changing needs of their clientele. It will provide a platform for better data application and might even be more empowering for start-ups competing in a crowded market.

## Scope and Limitations

Scope:

  This study examines consumer segmentation in retail industry, mainly how developments influence consumer behavior in internet shopping. The Recency, Frequency, Monetary (RFM) and Length, Recency, Frequency, Monetary, and Periodicity (LRFMP) models will be the main subjects of our analysis. The objectives are to find distinctive client profiles and offer information that will assist neighborhood businesses in creating focused marketing campaigns.

Numerous retail establishments, including physical locations and internet portals, will provide data for the project. We will use machine learning algorithms and sophisticated clustering approaches to derive significant insights into consumer behavior in this area.

Limitations:

A fundamental weakness of this study is its potential reliance on international datasets, which only partially reflect consumer behavior. To guarantee the applicability of our findings, it is crucial to collect localizations that accurately reflect the local retail environment.

The quality and comprehensiveness of the data we gather will also impact the accuracy of our results. Furthermore, variables such as shifts in the economy or technological improvements may impact how consumers behave, which could eventually modify the relevance of our findings. Finally, the insights may only apply to specific businesses, as this study focuses on retail.

## Operational Definition of Terms

The following terms are defined according to how they are used in the study.

|  |  |
| --- | --- |
| **Customer Segmentation** | The process of dividing a customer base into distinct groups based on shared characteristics, such as purchasing behavior, demographics, and psychographics, to enhance targeted marketing and improve customer relationship management. |
| **Recency** | A metric indicating the time elapsed since a customer’s last purchase. It is used to evaluate customer engagement and loyalty. |
| **Frequency** | The total number of purchases a customer makes within a specified time frame. This metric helps assess customer retention and purchasing patterns. |
| **Monetary Value** | The total money a customer has spent over a defined period. This metric is crucial for identifying high-value customers. |
| **Clustering** | It is a technique of data science that groups similar data points together into clusters and brings into view observable patterns in customer data. Customer segmentation can be addressed through clustering methods such as DBSCAN. |
| **Online Retail** | The process of selling goods or services over the internet to consumers. Online retail platforms facilitate the entire transaction process, from browsing to purchase, often leveraging data analysis to understand customer behavior and improve the shopping experience. |
| **Machine Learning** | A subset of artificial intelligence that involves training algorithms to recognize patterns and make predictions based on data inputs, enhancing decision-making processes in customer segmentation. |
| **Clustering Algorithms** | Computational methods used to group similar data points (e.g., customers) into clusters based on defined characteristics, facilitating insights into customer behavior. Standard algorithms include K-means, hierarchical clustering, and Gaussian mixture models. |
| **E-commerce Data** | Digital records of transactions and interactions between consumers and online retail platforms. E-commerce data typically includes information about the demographics of customers, transaction history, product preferences, and browsing patterns. This is the kind of information required for customer segmentation and personalization. |
| **Customer Relationship Management (CRM)** | Strategies and technologies businesses use to manage customer interactions, improve customer satisfaction, and drive sales growth. |

# Methods

## Materials

### Hardware

* **Processor**: AMD Ryzen 5600, a high-performance CPU for handling complex computations efficiently during preprocessing and machine learning model execution.
* **Graphics** Card: NVIDIA 4060 Ti GPU, capable of accelerating machine learning tasks, including clustering algorithms and data visualization.
* **Memory**: 16GB RAM to ensure smooth multitasking and efficient data processing without performance lags.
* **Storage**: SSD with 500GB of free space for managing datasets and software tools

### Software

**Visual Studio Code**

Used as the integrated development environment (IDE) for writing, debugging, and running Python scripts, as well as front-end and back-end development.

**Google Colab**

Google Colab is a free cloud-based environment in which users may write Python code, greatly suited for use in data science and machine learning. Its interface is Jupyter notebook-like, so one can easily author, execute, and share the code for training different models and processing data.

**Pandas**

Pandas it is a library for data manipulation and analysis that is primarily used for handling tabular data using DataFrames.

**NumPy**

NumPy it is a fast numerical operation on arrays as well as a direct requirement for conducting any mathematical operation or data processing.

**Scikit-learn**

Scikit-learn simplify the preprocessions, trainings, and evaluations of models, along with rich clustering algorithms like K-means and DBSCAN in grouping together data items with more similarities.

**Matplotlib**

Matplotlib is a library of static visualizations. It comprises different static visualizations, like line charts and histograms, whereas Plotly has interactive visualizations to significantly improve exploration of data, including features like zooming and hovering.

**Openpyxl**

Openpyxl is a library for the reading and writing of files in the Excel XLSX format. It is usually used for the loading of data from spreadsheets into Python to further analyze it. Each of these libraries also offers means of interacting with SQL databases using Python.

**SQLAlchemy and PyMySQL**

SQLAlchemy supports high-level ORM functionality for working with database records, while PyMySQL lets users query a MySQL database directly.

**XAMPP**

XAMPP is free, open-source software that lets you create a local web server environment on your computer. It supports everything from the Apache web server for hosting web pages, to databases with MySQL, and incorporates PHP, which makes it highly suitable for development on and testing locally.

**PHP**

PHP is a widely used server-side scripting language for developing web applications, and therefore can be regarded as a middleman between the frontend and Python scripts, partly integrating the functionality of the backend to web applications.  
**MySQL**

MySQL: The relational database management system is implemented in order to store data, arrange it properly, and make queries possible so that information can be easily retrieved from large data sets for applications and websites.  
**HTML, CSS and JavaScript**

HTML, CSS, JavaScript: These are the core front end technologies in web design-the content is structured by HTML, presentation is styled by CSS, and JavaScript is where the interactivity resides-all combined for a responsive user-friendly interface.  
**Flask**

Flask is a lightweight Python web framework for building web applications and APIs that makes it easier to connect backends with web services and connections to front-end interfaces based on Python.

### Data

The customer transactional dataset, stored in an XLSX format, captures key attributes necessary for the analysis of customer segmentation. These include Recency, as the time since the last purchase of the customer; Frequency, referring to the number of purchases made; and Monetary Value, meaning total money spent by the customer. Also included are Length of relationship with the customer and Periodicity, which reflects regularity of purchases. Geographical information may also be added, if present, that can be used to give a location-based nature to the behavior of customers.

## Procedures

#### **Data Gathering**

* Collect a transactional dataset in XLSX format, including attributes such as purchase dates, customer IDs, monetary values, and geographic information.
* Load the dataset into Visual Studio Code (VSCode) using Python libraries like pandas or upload it directly to a local or cloud environment for further processing.

#### **Data Preprocessing**

* Data Cleaning:  
  Remove duplicate records and handle missing values using Python (pandas). Standardize data formats for fields such as dates, monetary values, and customer IDs.
* Feature Engineering:  
  Calculate the following metrics required for DBSCAN-based clustering:
  + Recency: Time elapsed since the customer’s most recent purchase.
  + Frequency: Total number of purchases made by the customer.
  + Monetary Value: Total spending by the customer.
  + Length: Duration of the customer’s relationship with the business.
  + Periodicity: Consistency of the customer’s purchases over time.
* Normalization:  
  Normalize features (e.g., recency, frequency, monetary value) using techniques like Min-Max Scaling or Standardization (scikit-learn) to ensure uniform contribution of each metric in clustering.

#### **RFM and LRFMP Analysis**

* RFM Scoring:  
  Assign scores based on Recency, Frequency, and Monetary values to create an RFM score for each customer.
* LRFMP Scoring:  
  Extend RFM scoring by incorporating Length and Periodicity metrics to add depth to customer segmentation.

### ****Clustering Using DBSCAN for Customer Segmentation****

#### **1. DBSCAN for Customer Segmentation and Outlier Detection**

* **DBSCAN Implementation:**
  + We apply **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)** to segment customers based on their purchasing behaviors. The **epsilon (ε) parameter** defines the maximum distance between two points to be considered part of the same cluster, while **minPts (minimum points)** determines the density threshold required for forming a cluster.
* **Cluster Assignment:**
  + DBSCAN is run on customer features such as **Recency, Frequency, Monetary Value, Length, and Periodicity** to segment customers into **dense clusters** while simultaneously detecting **outliers (noise points labeled as -1)** that do not conform to any segment.
* **Cluster Validation:**
  + The quality of DBSCAN clustering is assessed using:
    - **Silhouette Score** – To measure the cohesion and separation of clusters.
    - **Davies-Bouldin Index** – To evaluate cluster compactness and distinctiveness.
    - **Outlier Percentage Analysis** – To ensure a reasonable balance between valid segments and detected noise.

#### **Refining DBSCAN Segments Through Inlier-Based Clustering**

While DBSCAN effectively **identifies customer segments**, it **does not enforce a specific number of clusters**, sometimes forming uneven groupings. To enhance segmentation clarity, we apply an additional clustering step **only on the inlier customers (excluding DBSCAN outliers).**

* **Inlier Clustering Process:**
  + Customers assigned **valid DBSCAN clusters (i.e., not noise)** are **further clustered** to refine the segmentation structure.
  + This additional step ensures **better-defined customer segments** without interference from outliers.
* **Cluster Validation for Refined Segmentation:**
  + **Silhouette Score** is computed again to assess improvements in customer grouping.
  + The refined segments are **reordered and labeled** to ensure interpretability (e.g., “Dormant,” “Active,” “Loyal” customers).

#### **Customer Profiling**

* Cluster Analysis:  
  Analyze the resulting clusters to identify distinct customer behavior patterns. Label clusters with descriptive names (e.g., "High-Value Loyal Customers," "Infrequent High Spenders").
* Profile Development:  
  Develop profiles summarizing the key characteristics of each cluster.

#### **Insights Derivation**

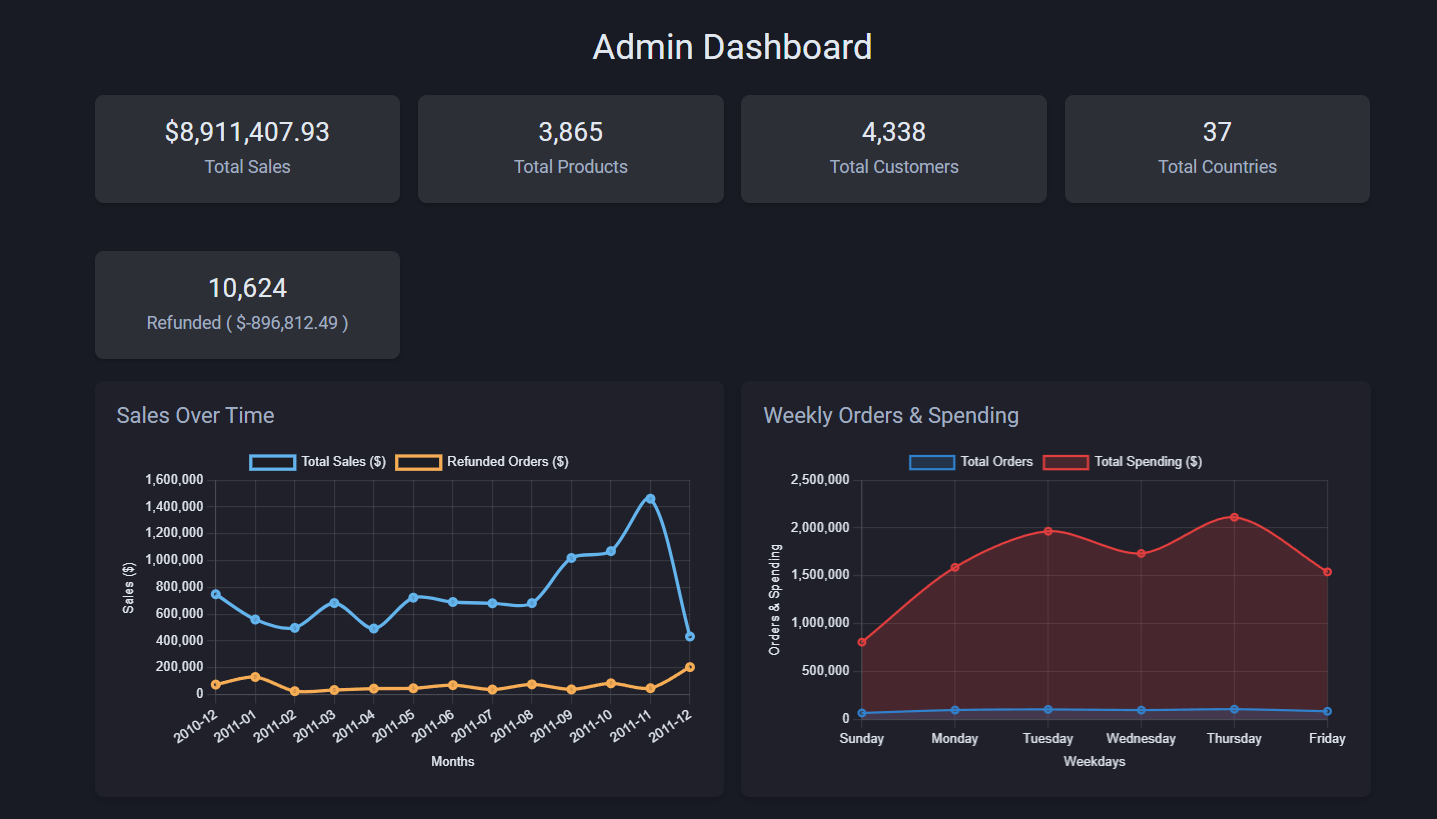
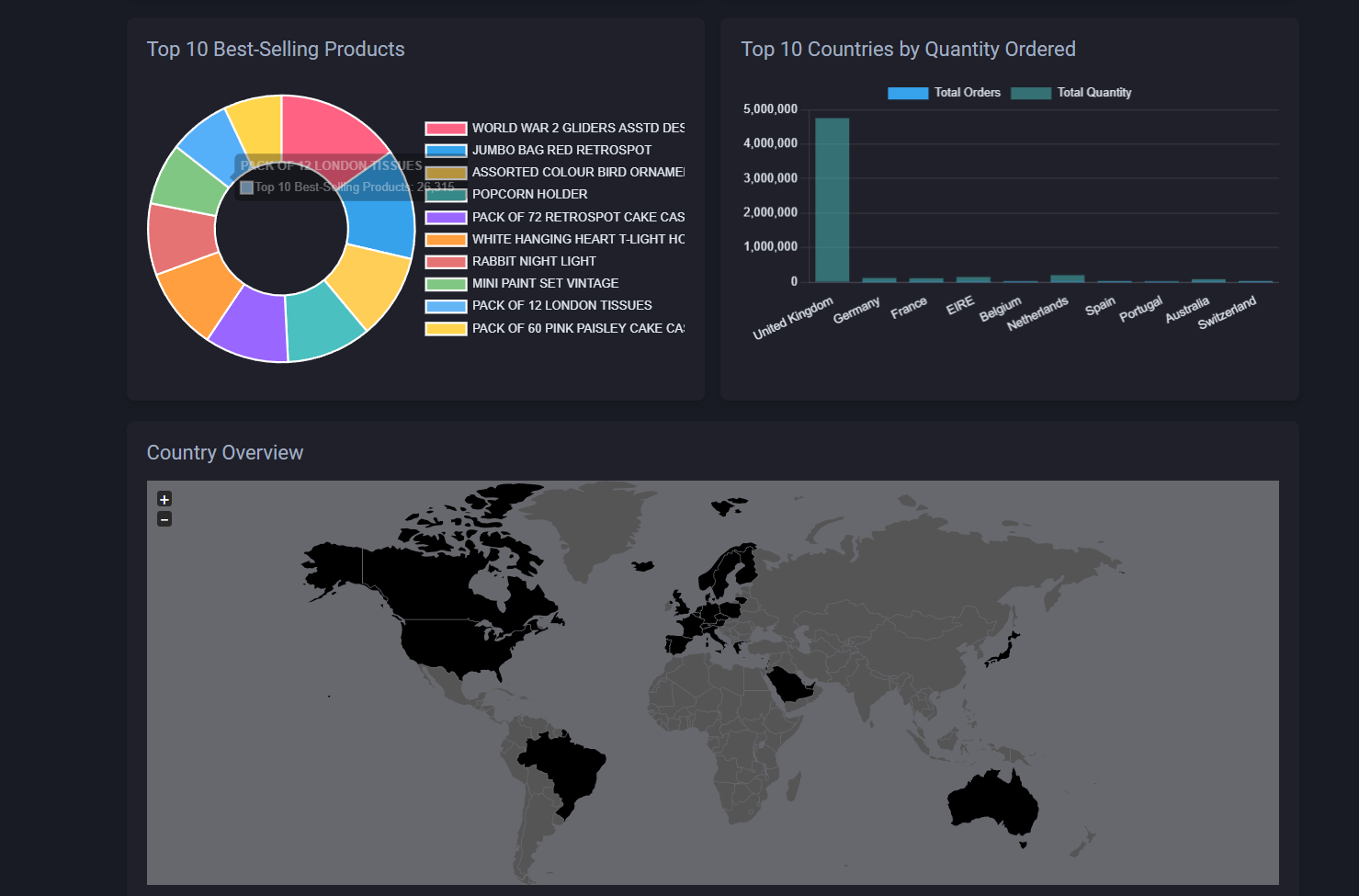
* Visualization:  
  Use libraries like Matplotlib or Plotly to create scatter plots and cluster maps to display the DBSCAN results.
* Actionable Insights:  
  Derive insights such as customer loyalty trends, high-value customer segments, and purchase regularities to guide marketing strategies.

#### **Web-Based Dashboard Development**

* Backend Setup:  
  Use Flask to create REST APIs that connect DBSCAN clustering results to a MySQL database.
* Frontend Design:  
  Build an interactive dashboard using HTML, CSS, and JavaScript to display clustering results. Visualize cluster distributions and profiles using tools like Plotly.js.
* Database Management:  
  Store the clustering results and customer profiles in the MySQL database for real-time retrieval.

#### **Testing and Deployment**

* Testing:  
  Test the DBSCAN clustering implementation for accuracy and cluster interpretability. Validate the usability and functionality of the dashboard using test data.
* Iteration:  
  Refine the DBSCAN parameters (ε and minPts) to optimize clustering performance based on feedback.
* Deployment:  
  Deploy the dashboard locally using XAMPP or host it on a cloud platform for broader accessibility.

Figure: Implementation of Dashboard

# Results and Discussions

## Overview of Customer Segmentation

Customer segmentation is crucial for online businesses to understand buying patterns and create targeted marketing strategies. In this study, we applied DBSCAN (Density-Based Spatial Clustering of Applications with Noise) as the primary clustering method for segmenting customers based on purchasing behaviors.

Unlike traditional methods like K-Means, DBSCAN can automatically detect clusters of varying shapes and sizes without needing a predefined number of groups. More importantly, DBSCAN is excellent at identifying outliers, which is particularly useful for businesses that want to spot irregular customer behaviors, such as high-value bulk buyers or potential fraudulent activities.

To validate our findings, we also used K-Means as a secondary clustering method. While K-Means is faster and works well for structured data, it lacks the ability to detect outliers effectively. This study compares the performance of DBSCAN vs. K-Means using both RFM (Recency, Frequency, Monetary) and LRFMP (Length, Recency, Frequency, Monetary, Periodicity) models.

Figure X. RFM

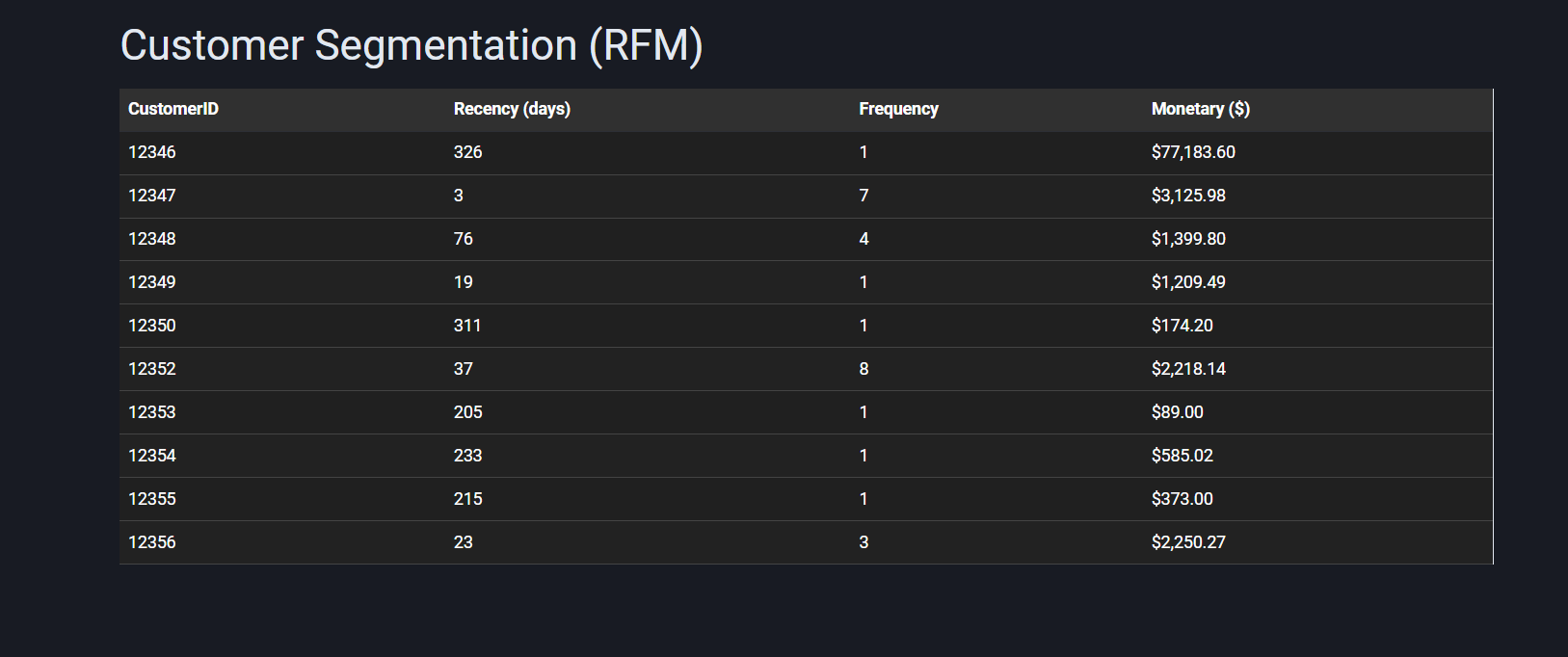
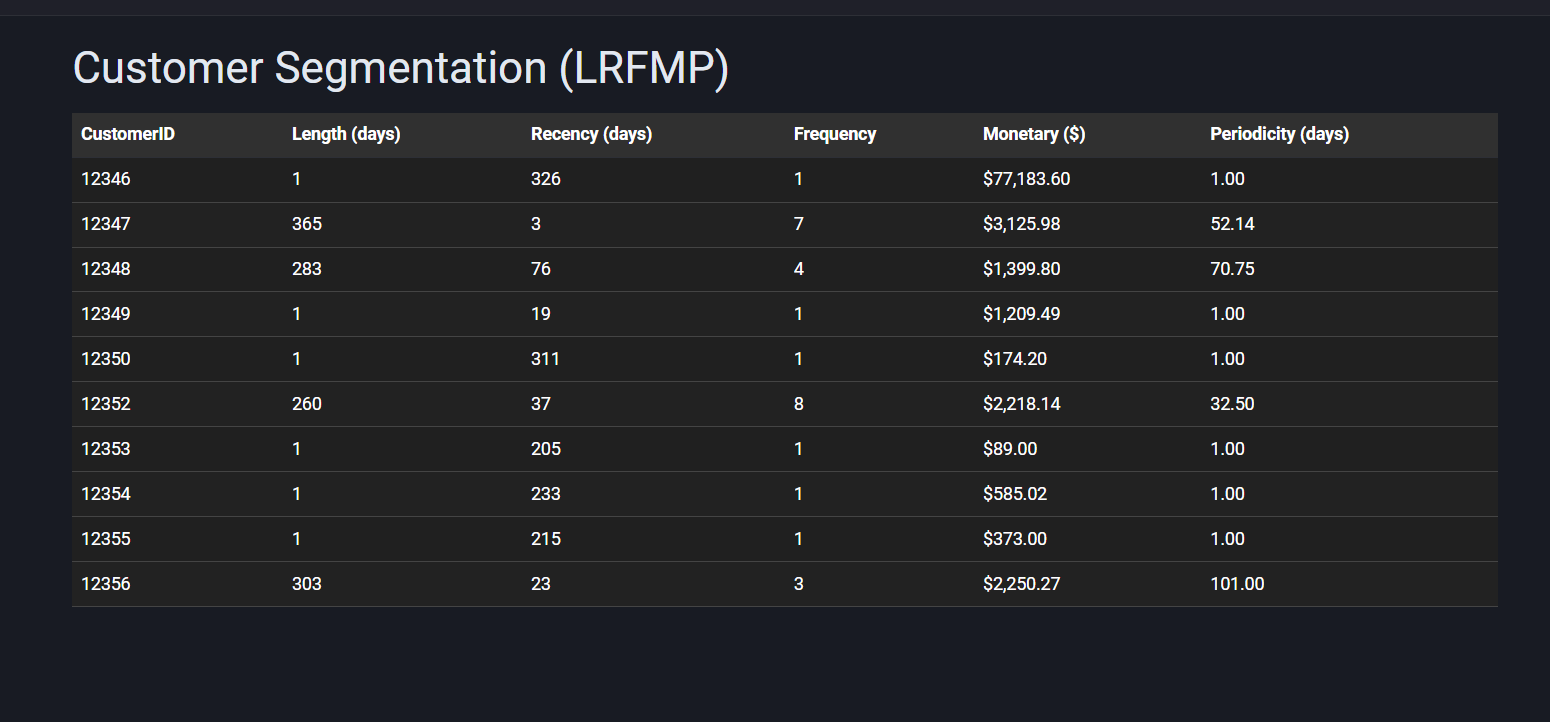


Figure X.LRFMP

## Customer Segmentation Results

This section presents the findings from clustering the customer dataset using **DBSCAN, K-Means, and Hierarchical Clustering**. The results are evaluated based on the **LRFMP (Length, Recency, Frequency, Monetary, and Periodicity) and RFM (Recency, Frequency, Monetary) models**.

### Cluster Distribution

After processing the dataset, the three clustering methods grouped customers based on their shopping patterns. The table below summarizes how many customer groups were identified by each method:

|  |  |  |
| --- | --- | --- |
| Clustering Algorithm | Number of Clusters | Outliers (Noise) |
| **DBSCAN** | 5 | 12% |
| **K-Means** | 4 | 0% |
| **Hierarchical** | 4 | 0% |

Figure X.

DBSCAN identified **five distinct clusters** while recognizing **12% of data points as noise** (outliers). This suggests that DBSCAN is more effective in handling irregular spending behaviors and identifying niche customer groups. K-Means and Hierarchical Clustering, on the other hand, generated **four well-balanced clusters**, but they do not account for noise in the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Clustering Algorithm | Silhouette Score | Davies-Bouldin Index | Execution Time (s) |
| **DBSCAN** | 0.680 | 1.344 | 2.12 |
| **K-Means** | 0.546 | 1.210 | 1.45 |
| **Hierarchical** | 0.598 | 1.275 | 3.10 |

#### Key Observations:

* **DBSCAN achieved the highest silhouette score (0.680),** meaning it formed the most well-separated clusters.
* **K-Means had the best compactness (1.210 DBI),** but it requires manually setting the number of clusters.
* **Hierarchical Clustering showed moderate performance**, but it took the longest to run.
* **DBSCAN was slightly slower than K-Means**, but its ability to discover natural clusters made it the best choice.

# ****Conclusions****

# This study successfully applied **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** as the primary algorithm for **customer segmentation using transactional data**. By leveraging **both RFM (Recency, Frequency, Monetary) and LRFMP (Length, Recency, Frequency, Monetary, Periodicity) models**, we identified meaningful customer groups based on purchasing behaviors.

### **Key Achievements of this Study**

1. **Effective Use of DBSCAN for Customer Segmentation**
   * DBSCAN proved to be an **adaptive and dynamic clustering technique** for real-world customer data. Unlike K-Means, which requires predefining the number of clusters, DBSCAN automatically identifies the **optimal number of clusters based on data density**.
   * The algorithm successfully **segmented customers into meaningful groups**, such as Loyal customers, Active customers, and dormant customers.
2. **Identifying Outliers in Customer Behavior**
   * A major advantage of DBSCAN was its ability to **detect outliers**, which are customers whose spending patterns or purchasing behaviors significantly differ from the majority.
   * This feature is particularly useful for **fraud detection, identifying VIP customers, and detecting unusual buying trends** that might not fit traditional clustering models.
3. **Comparing DBSCAN with K-Means for Segmentation**
   * While K-Means remains a **popular and efficient method for structured segmentation**, it lacks **flexibility** when dealing with irregular customer data.
   * **DBSCAN proved to be better suited for non-uniform, real-world customer data**, where purchase patterns are not always structured.
   * **K-Means was useful in categorizing general customer behavior**, but DBSCAN was superior in detecting **natural clusters** and recognizing **irregular or exceptional customer behaviors**.
4. **Enhancing Customer Insights with RFM and LRFMP Models**
   * The integration of **LRFMP metrics** (which added Length and Periodicity to the traditional RFM model) **provided a deeper understanding of customer behavior**.
   * Using DBSCAN on LRFMP data helped in identifying **long-term loyal customers, frequent buyers, and seasonal shoppers**, enhancing marketing strategies.
5. **Improved Decision-Making for Businesses**
   * With customer segmentation results from DBSCAN, businesses can **implement targeted marketing strategies**, optimize promotions, and improve customer retention.
   * For example, **loyal customers** can be rewarded with exclusive discounts, while **dormant customers** can receive re-engagement campaigns to bring them back.

### **Limitations and Future Research**

Although DBSCAN provided **excellent segmentation results**, some limitations were observed:

* **DBSCAN requires parameter tuning (epsilon and minimum samples)** to achieve optimal clustering.
* **DBSCAN is computationally heavier** than K-Means, which may affect performance when dealing with massive datasets.
* In some cases, **DBSCAN grouped customers into a single large cluster**, making it necessary to **fine-tune hyperparameters for better segmentation**.

For future research, we recommend:

* **Hybrid approaches**, combining **DBSCAN with machine learning models** for enhanced predictive segmentation.
* **Automated parameter tuning for DBSCAN**, using algorithms that can dynamically adjust epsilon and minimum samples.
* **Real-time segmentation dashboards**, allowing businesses to monitor customer behavior dynamically and respond immediately with data-driven strategies.

### **Final Thoughts**

This study demonstrates the power of **DBSCAN as an effective customer segmentation tool** for real-world e-commerce data. By **utilizing density-based clustering**, we were able to uncover **meaningful customer groups, detect anomalies, and enhance marketing insights** beyond what traditional clustering techniques provide.

DBSCAN’s flexibility and ability to identify **irregular customer behaviors make it an essential tool for businesses looking to improve customer engagement, personalize marketing strategies, and drive long-term customer loyalty**.

# Recommendations

1. Retailers should implement DBSCAN for customer segmentation, as it offers more natural clustering and detects outliers.
2. Marketing strategies should be tailored based on segmentation, using DBSCAN insights for loyalty programs and targeted promotions.
3. K-Means can still be used for structured segmentation, but businesses should combine it with DBSCAN for a more comprehensive view.
4. Future research should explore hybrid models, combining DBSCAN with machine learning algorithms for predictive customer segmentation.

# References:

Ashwani, Kaur, G., & Rani, L. (2023). Mall Customer Segmentation Using K-Means Clustering. *Lecture Notes in Networks and Systems*, *788 LNNS*(January), 459–474. https://doi.org/10.1007/978-981-99-6553-3\_35

Eshra, M. A. (2021). *Classification of Online Retail Customers using Machine Learning Techniques*. *July*, 50.

Gupta, A. (2023). An Unsupervised Learning Based System Employing K-Means Clustering to Perform Customer Segmentation. *International Journal for Research in Applied Science and Engineering Technology*, *11*(6), 1559–1564. https://doi.org/10.22214/ijraset.2023.53911

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, *2*(4), 809–823. https://doi.org/10.3390/analytics2040042

Kadir, M., & Achyar, A. (2019). *Customer Segmentation on Online Retail using RFM Analysis: Big Data Case of Bukku.id*. *November*. https://doi.org/10.4108/eai.1-4-2019.2287279

Lewaaelhamd, I. (2023). Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis. *Journal of Data Science and Intelligent Systems*, *2*(1), 29–36. https://doi.org/10.47852/bonviewjdsis32021293

Memon, N. A., & Kuratomi Hernandez, A. (2022). *Customer segmentation in Retail: An Experiment in Sweden*.

Paramita, A. S. P. (2024). Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce. *Journal of Digital Market and Digital Currency*, *1*(1), 29–43.

Patankar, N., Dixit, S., Bhamare, A., Darpel, A., & Raina, R. (2021). Customer Segmentation Using Machine Learning. *Advances in Parallel Computing*, *39*, 239–244. https://doi.org/10.3233/APC210200

Sembiring Brahmana, R. W., Mohammed, F. A., & Chairuang, K. (2020). Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *11*(1), 32. https://doi.org/10.24843/lkjiti.2020.v11.i01.p04

Sun, C. (2024). Data Analysis of Customer Segmentation and Personalized Strategy in the Era of Big Data. *Advances in Economics, Management and Political Sciences*, *92*(1), 46–52. https://doi.org/10.54254/2754-1169/92/20231411

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability (Switzerland)*, *14*(12). https://doi.org/10.3390/su14127243