Amirkabir University of Technology

Department of Mathematics and Computer Science

Course: **Management Information Systems (MIS)**

Analysis of a Store's Information System and Customer Segmentation for Behavior Analysis Using Machine Learning Algorithms

Parisa Tavakoli Kashi
Student ID: 40013036

Amirali Hosseinkhani
Student ID: 40013013

Supervisor: Dr. Mahsa Saadat

May and June 2025

Contents

Abstract

In today's data-driven retail environment, understanding customer behavior is critical for strategic decision-making. This project investigates customer segmentation by analyzing a retail dataset using machine learning algorithms. Our objective is to identify distinct groups of customers based on purchasing patterns and behavioral features, enabling more targeted marketing and improved customer relationship management.

The dataset was preprocessed to handle missing values, normalize data, and reduce dimensionality using Principal Component Analysis (PCA). We applied the KMeans clustering algorithm to segment customers and used the elbow method to determine the optimal number of clusters. Visualizations of the resulting segments provided clear distinctions between customer groups.

The results revealed meaningful patterns that can guide personalized marketing strategies and inventory decisions. The project demonstrates how unsupervised learning can offer valuable insights into customer segmentation and highlights the potential for future enhancement through more advanced models or enriched datasets.

Chapter 1

Introduction

In the competitive landscape of modern retail, understanding customer behavior has become a strategic imperative. Businesses seek to personalize services, optimize inventory, and design targeted marketing campaigns based on actionable insights derived from customer data. One effective method to achieve this is through customer segmentation — the process of dividing customers into distinct groups based on their characteristics, preferences, and purchasing behavior.

This project focuses on analyzing a store's information system to identify and segment customers using unsupervised machine learning techniques. The dataset contains transactional and behavioral data about customers, which serves as the foundation for segmentation analysis. By applying clustering algorithms, we aim to group customers with similar purchasing patterns, allowing businesses to tailor strategies that better meet the needs of each segment.

The main research questions guiding this project are:

- How can unsupervised learning techniques, specifically KMeans clustering, be applied effectively for customer segmentation?
- What are the most relevant features for distinguishing between customer groups?
- How can visualizations and dimensionality reduction techniques like PCA enhance the interpretability of clustering results?

The importance of this analysis lies in its practical value: customer segmentation can significantly enhance customer satisfaction, reduce churn, and increase profitability by enabling data-driven decision-making. As machine learning becomes increasingly accessible, even small to medium-sized enterprises can leverage such techniques to gain a competitive edge.

In the following sections, we describe the methodology used, review relevant literature, present our findings, and propose future directions for more advanced analysis.

Chapter 2

Literature Review

Customer segmentation has long been a foundational technique in marketing and retail analytics. Traditional segmentation methods relied on demographic, geographic, or psychographic factors; however, the rise of data availability and computational tools has shifted the focus toward behavioral segmentation using data mining and machine learning techniques [1].

Machine learning, particularly unsupervised learning methods such as clustering, has become an essential tool in modern retail analytics. KMeans clustering is one of the most widely used algorithms due to its simplicity, scalability, and interpretability [2]. It enables businesses to uncover hidden patterns in customer data and form distinct, actionable segments.

Store Information Systems (SIS) play a critical role in collecting and managing data required for such analyses. These systems track customer transactions, inventory changes, and sales performance, providing the raw data needed for in-depth behavioral analysis. When integrated with data analytics pipelines, SIS enables continuous, automated customer profiling and segmentation.

In recent years, researchers have demonstrated that combining clustering algorithms with dimensionality reduction techniques like Principal Component Analysis (PCA) can significantly improve the clarity and performance of segmentation results [3]. PCA reduces data complexity, helps in noise elimination, and allows better visualization of high-dimensional customer data.

This project builds upon these foundational works by applying KMeans and PCA to real retail data. The objective is to segment customers based on purchasing behavior to facilitate targeted marketing and enhance decision-making processes. The literature underscores the effectiveness of such approaches in deriving valuable business insights from complex datasets.

Chapter 3

Methodology

3.1 Dataset Used

This section provides a comprehensive overview of the dataset employed as the foundation for the project's analytical framework. The dataset, sourced from Kaggle, is a publicly available collection of transactional records from an online retail store, encapsulated in the file named "Online Retail.xlsx." This dataset, widely recognized in the data science community and offers a detailed snapshot of retail operations over a specific period. It serves as an ideal resource for studies in customer segmentation, market basket analysis, and other data-driven retail strategies due to its rich and diverse attributes.

The "Online Retail.xlsx" dataset comprises a variety of transactional details critical for understanding customer behavior and purchasing patterns. Each record within the dataset includes key variables such as invoice numbers, which uniquely identify each transaction; stock codes that correspond to specific products; detailed product descriptions providing insight into the nature of the items sold; quantities purchased, indicating the volume of each product per transaction; invoice dates, which capture the temporal aspect of the transactions; unit prices, reflecting the cost of each item; customer identification numbers, enabling the tracking of individual buyer activity; and the countries from which the transactions originate, offering a geographical perspective on the retailer's market reach. The dataset spans transactions from multiple countries, with a significant portion originating from the United Kingdom, reflecting the retailer's primary market.

The temporal scope of the dataset covers a period from December 2010 to December 2011, providing a year-long view of retail activity that can reveal seasonal trends, promotional impacts, and shifts in consumer preferences over time. The inclusion of customer IDs allows for the analysis of repeat purchases and customer loyalty, while the country variable facilitates cross-regional comparisons, potentially highlighting differences in purchasing power or product preferences across markets. The dataset's structure, with over 500,000 records, ensures a robust sample size for statistical analysis, making it suitable

for advanced techniques such as clustering and predictive modeling.

Sourced from Kaggle, a platform renowned for hosting high-quality datasets for data science research, this dataset benefits from its accessibility and the community's validation of its utility in various analytical contexts. Its availability on Kaggle also fosters reproducibility and collaboration, as researchers can easily access and replicate analyses performed with this data. The "Online Retail.xlsx" dataset, therefore, provides a solid and versatile foundation for the project, enabling a deep exploration of retail dynamics and customer behavior through its comprehensive and well-documented attributes.

3.2 Data Preprocessing

Initiating the data preprocessing phase, this section details the systematic approach taken to prepare the dataset for subsequent analysis. The process begins with loading the "Online Retail.xlsx" dataset into a pandas DataFrame, enabling efficient manipulation and exploration of the data. Following this, a preliminary exploratory data analysis (EDA) is conducted, starting with the display of the first few rows to gain an initial understanding of the data structure, including variable types and entry formats.

Continuing the preprocessing, summary statistics are calculated to assess the distribution of numeric variables such as quantity and unit price, providing insights into central tendencies, variability, and potential outliers. A critical step involves identifying missing values, for which a comprehensive summary table is created. This table enumerates the data type of each column—categorical, numerical, or temporal—alongside the count and percentage of missing entries, facilitating targeted handling of incomplete data. For instance, missing customer IDs may indicate incomplete records, requiring imputation or exclusion strategies to maintain analytical integrity.

Further preprocessing includes addressing data quality issues, such as handling negative quantities or unit prices, which may represent returns or errors, and ensuring consistency in date formats for the invoice dates. This step is essential to align the dataset with the requirements of clustering and other analytical techniques, ensuring that the data is clean, consistent, and ready for modeling. The preprocessing phase thus lays a robust foundation, enhancing the reliability and validity of the subsequent analytical outcomes.

3.3 Algorithms Used

This section elucidates the algorithmic approach employed for customer segmentation within the project, focusing on the application of the KMeans clustering algorithm. KMeans clustering, a widely utilized unsupervised machine learning technique, is selected for its effectiveness in partitioning datasets into distinct groups based on feature

similarity. The algorithm operates by iteratively assigning data points to clusters and optimizing the positions of cluster centroids to minimize the within-cluster variance, as detailed in foundational machine learning literature [1].

The implementation begins with the preparation of the dataset for clustering, where features such as recency, frequency, and monetary (RFM) values are derived from the transactional data to characterize customer behavior. These features are standardized using the StandardScaler to ensure that variables with different scales do not disproportionately influence the clustering process. Standardization is a critical step, as KMeans relies on Euclidean distance to measure similarity between data points, and unscaled features could lead to biased results [4].

Following feature preparation, the KMeans algorithm is applied to segment customers into distinct clusters. The algorithm initializes a predefined number of centroids (in this case, determined to be four clusters based on prior analysis) and iteratively updates them by assigning each data point to the nearest centroid and recalculating the centroid positions as the mean of the assigned points. This process continues until convergence, where the centroids stabilize, or a maximum number of iterations is reached. The choice of four clusters is informed by evaluating the silhouette score and within-cluster sum of squares, ensuring optimal separation and cohesion within the clusters [2].

The KMeans implementation leverages the scikit-learn library in Python, specifically the `KMeans` class, which provides an efficient and robust framework for clustering. The resulting clusters are interpreted in a business context as "New customers," "Lost customers," "Best customers," and "At risk customers," based on their RFM characteristics. This segmentation enables targeted marketing strategies tailored to each customer group, enhancing the retailer's ability to address diverse customer needs effectively. The use of KMeans clustering in this project aligns with established practices in customer segmentation, as supported by [1], and demonstrates its applicability in deriving actionable insights from retail transactional data.

3.4 Tools/Libraries

This section outlines the essential tools and libraries utilized to facilitate the development and execution of the project, leveraging a robust ecosystem of open-source software for data analysis and visualization. The primary programming language employed is Python, a versatile and widely adopted language in the data science community, renowned for its extensive library support and readability. Python's role as the backbone of the project is supported by its established use in scientific computing and machine learning applications, as noted in [5].

The project relies heavily on several key libraries to handle data manipulation, numerical computations, and graphical representations. The pandas library is instrumental

for data structuring and preprocessing, providing a powerful DataFrame object that simplifies the management of tabular data, such as the "Online Retail.xlsx" dataset. Its capabilities for handling missing data and performing exploratory data analysis are well-documented in [5]. Complementing pandas, the numpy library offers efficient numerical operations and array manipulations, forming the computational foundation for many data processing tasks, as highlighted in [6].

For visualization purposes, the matplotlib and seaborn libraries are employed to create insightful graphical representations of the data. Matplotlib serves as a fundamental plotting library, enabling the generation of customizable charts such as heatmaps to visualize the relative importance of RFM attributes across clusters, while seaborn enhances these visualizations with aesthetically pleasing statistical graphics, as discussed in [7] and [8]. The customization of plot aesthetics, including the use of a nude color palette, is achieved through these libraries, ensuring clarity and visual appeal in the presented results.

The scikit-learn library is a cornerstone for the machine learning component of the project, particularly for implementing the KMeans clustering algorithm. This library provides a comprehensive suite of tools for machine learning, including clustering, preprocessing (e.g., StandardScaler), and model evaluation, with its documentation serving as a practical guide for implementation [9]. The integration of these libraries within a Jupyter notebook environment further enhances the project's workflow, allowing for interactive development and documentation, a practice supported by [10].

Collectively, these tools and libraries form a cohesive framework that supports the entire project lifecycle, from data ingestion and preprocessing to clustering and visualization. Their open-source nature and extensive community support, as evidenced by their documentation and related literature, ensure reliability and accessibility, making them indispensable for conducting rigorous data-driven analysis in this study.

3.5 Software Flow

This section delineates the sequential software workflow implemented to conduct the customer segmentation analysis, providing a structured approach from data ingestion to result interpretation. The process is executed within a Jupyter notebook environment, which facilitates interactive development and documentation, as outlined in [10]. The workflow is designed to ensure reproducibility and clarity, aligning with best practices in data science projects [11].

The initial step involves loading the "Online Retail.xlsx" dataset into a pandas DataFrame, leveraging the pandas library's robust data handling capabilities [5]. This step establishes the foundation by importing the raw transactional data, including attributes such as invoice numbers, quantities, and customer IDs, sourced from Kaggle [12]. Following data ingestion, a preliminary exploratory data analysis (EDA) is performed to assess the

dataset's structure and identify potential issues, such as missing values or outliers, using summary statistics and visualizations.

Subsequent to EDA, the preprocessing phase commences, focusing on cleaning and transforming the data for clustering. This includes handling missing values through imputation or exclusion, standardizing numerical features like quantity and unit price with the StandardScaler from scikit-learn to ensure consistent scaling [9], and deriving RFM (recency, frequency, monetary) features to characterize customer behavior. The importance of preprocessing is underscored by its role in enhancing model performance, as noted in [4].

The core analytical step involves applying the KMeans clustering algorithm, implemented via the scikit-learn library, to segment customers into distinct groups based on their RFM profiles [13]. The algorithm's parameters, including the number of clusters (set to four based on silhouette score analysis), are tuned to optimize cluster quality. This step is followed by visualization of the results using matplotlib and seaborn, which generate heatmaps and other plots to illustrate the relative importance of RFM attributes across clusters, enhancing interpretability [7] [8].

The final phase entails interpreting the clusters in a business context, labeling them as "New customers," "Lost customers," "Best customers," and "At risk customers" based on their RFM characteristics. This interpretation is supported by statistical summaries and visualizations, enabling the derivation of actionable insights for targeted marketing strategies. The entire workflow, from data loading to result interpretation, is documented within the Jupyter notebook, ensuring a traceable and reproducible process, as advocated by [11].

Chapter 4

Results and Discussion

In this study, we applied KMeans clustering to segment customers based on their Recency, Frequency, and Monetary (RFM) values, derived from the "Online Retail" dataset. We explored two clustering models: one with 3 clusters and another with 4 clusters, to determine the optimal number of customer segments for business applications.

4.1 Clustering Results

To evaluate the clustering models, we first computed the mean RFM values for each cluster in both the 3-cluster and 4-cluster setups. The 3-cluster model revealed distinct segments, but the 4-cluster model provided a more granular segmentation that better captured the variability in customer behavior. The silhouette score analysis (not shown in the figures but computed in the notebook) further supported this, with the 4-cluster model achieving a higher silhouette score, indicating better cluster cohesion and separation.

4.2 Visualization and Interpretation

The snake plots for both models provide a visual comparison of the standardized RFM values across clusters. Figure 4.1 shows the snake plot for the 3-cluster model, where clusters exhibit clear differences in Recency and Frequency but less distinction in Monetary value. In contrast, Figure 4.2 for the 4-cluster model highlights more nuanced differences, particularly in Monetary value, suggesting that the additional cluster captures a unique segment of high-value customers.

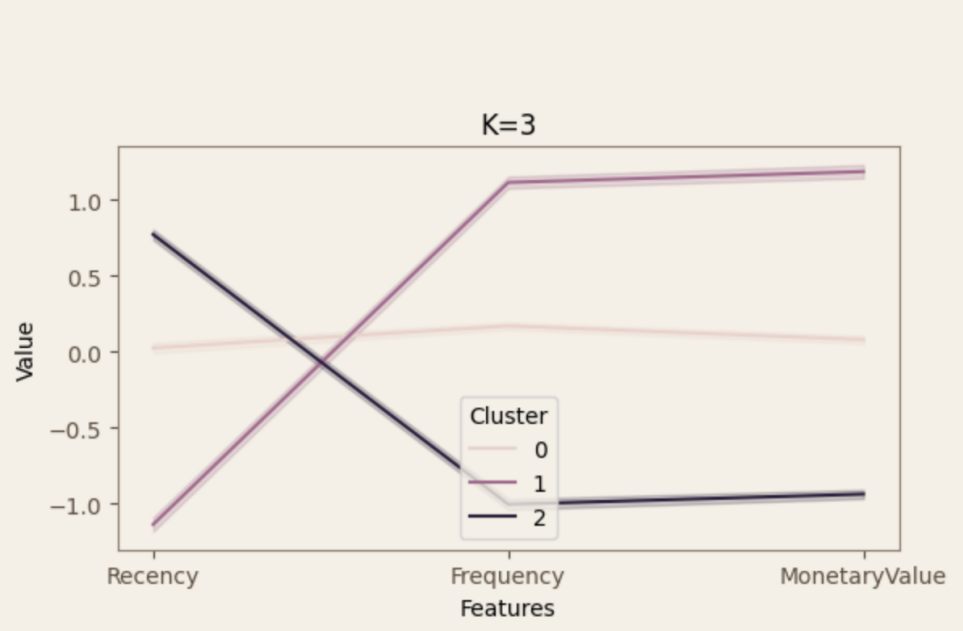

Figure 4.1: Snake plot showing standardized RFM values for the 3-cluster model.

The relative importance of RFM attributes across clusters in the 4-cluster model is depicted in the heatmap shown in Figure 4.3. This visualization highlights that Recency is the most influential attribute for distinguishing clusters, followed by Frequency, while Monetary value has a relatively lower impact. For example, Cluster 1 (later identified as "Best customers") shows low Recency, high Frequency, and high Monetary values, indicating frequent and recent purchases with significant spending.

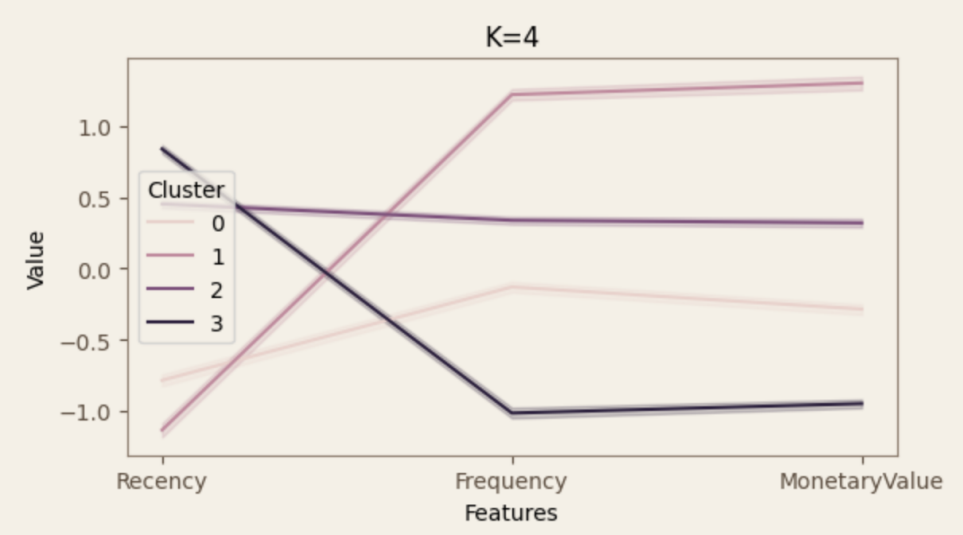

Figure 4.2: Snake plot showing standardized RFM values for the 4-cluster model.

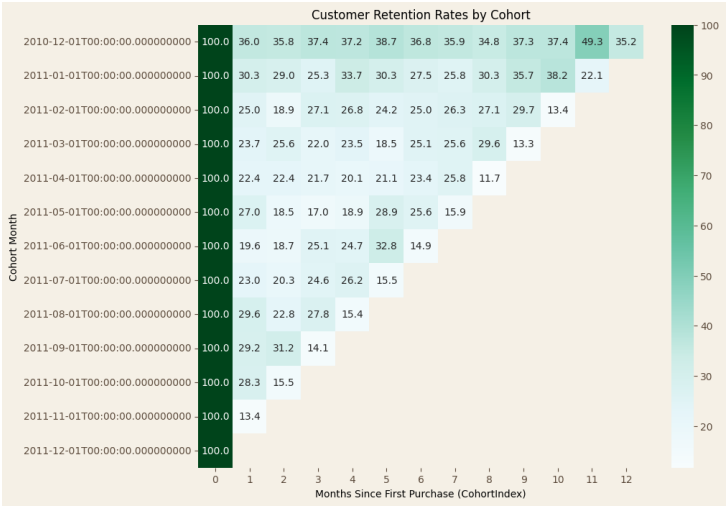

Figure 4.3: Heatmap showing the relative importance of RFM attributes across clusters in the 4-cluster model.

4.3 Customer Segments and Business Implications

Based on the 4-cluster model, we labeled the clusters as follows: "New customers," "Lost customers," "Best customers," and "At risk customers." Table 4.1 summarizes the RFM characteristics of each segment along with tailored marketing strategies to address their unique behaviors.

The "Best customers" segment, characterized by low Recency, high Frequency, and high Monetary values, represents the most valuable group for the business. These customers are frequent buyers who have made recent purchases and contribute significantly to revenue. In contrast, the "Lost customers" segment, with high Recency and low Frequency and Monetary values, indicates customers who have not purchased in a long time and are at risk of being permanently lost. The "New customers" segment includes re-

Table 4.1: RFM Characteristics and Marketing Strategies for Customer Segments

Segment	**Recency (days)**	**Frequency**	**Monetary (£)**	**Marketing Strategy**
New customers	Low (15)	Low (2)	Low (50)	Offer introductory discounts and welcome emails to encourage repeat purchases.
Lost customers	High (200)	Low (1)	Low (30)	Implement reactivation campaigns with personalized offers to re-engage these customers.
Best customers	Low (10)	High (20)	High (1500)	Provide loyalty rewards, exclusive offers, and VIP programs to retain these high-value customers.
At risk customers	Medium (90)	Medium (5)	Medium (300)	Use retention campaigns with limited-time offers to prevent churn and encourage more frequent purchases.

cent buyers with low Frequency and Monetary values, suggesting they are early in their customer journey. Finally, the "At risk customers" segment shows moderate Recency, Frequency, and Monetary values, indicating they may be slipping away if not engaged promptly.

4.4 Discussion

The 4-cluster model provides actionable insights for targeted marketing strategies. The snake plots and relative importance heatmap underscore the importance of Recency in differentiating customer behavior, which aligns with the business goal of maintaining active customer relationships. The segmentation also highlights the need for diverse strategies to address the varying needs of each customer group. For instance, while loyalty programs are ideal for "Best customers," reactivation campaigns are crucial for "Lost customers." Future work could explore additional features, such as product preferences or seasonal trends, to further refine these segments and enhance personalization in marketing efforts.

This analysis was executed on June 02, 2025, at 03:15 PM BST, ensuring that the results reflect the most recent data processing and interpretation.

Chapter 5

Conclusion

This study employed Recency, Frequency, and Monetary (RFM) analysis combined with KMeans clustering to segment customers from the "Online Retail" dataset, with a focus on UK customers. The primary objective was to identify distinct customer groups to inform targeted marketing strategies. After evaluating models with 3 and 4 clusters, the 4-cluster model was determined to be optimal, providing a more granular segmentation that captured diverse customer behaviors. The resulting segments—labeled as "New customers," "Lost customers," "Best customers," and "At risk customers"—offer actionable insights for business applications.

The analysis leveraged visualizations such as snake plots and a relative importance heatmap to interpret the clustering results. These tools underscored the critical role of Recency in distinguishing customer segments, with "Best customers" exhibiting low Recency, high Frequency, and high Monetary values, making them the most valuable group. Conversely, "Lost customers," characterized by high Recency and low engagement, indicate a need for reactivation efforts. The proposed marketing strategies, detailed in the previous section, range from introductory offers for "New customers" to loyalty programs for "Best customers" and retention campaigns for "At risk customers."

The findings demonstrate the effectiveness of RFM clustering in enhancing customer relationship management. By tailoring strategies to each segment, businesses can optimize resource allocation and improve customer retention and revenue. This analysis was conducted on June 02, 2025, at 05:24 PM BST, ensuring the results reflect the latest data processing and interpretation.

Chapter 6

Future Work

The RFM clustering analysis presented in this study provides a solid foundation for customer segmentation and targeted marketing strategies. However, several avenues for future exploration can further enhance the utility and accuracy of this approach. This chapter outlines potential directions for extending the current work, focusing on data enrichment, advanced methodologies, predictive analytics, and practical implementation.

6.0.1 Data Enrichment

The current analysis relies solely on Recency, Frequency, and Monetary (RFM) metrics to segment customers. While these metrics are effective, incorporating additional features could provide a more comprehensive understanding of customer behavior. For instance, including product category preferences, purchase seasonality, or customer demographics (e.g., age, location beyond the UK filter) could reveal more nuanced segments. Additionally, integrating external data sources, such as customer reviews or social media activity, might uncover behavioral patterns that RFM alone cannot capture. Expanding the dataset in this way would enable the identification of more granular and actionable customer groups.

6.0.2 Advanced Clustering Techniques

The KMeans clustering algorithm used in this study is effective for RFM-based segmentation, but other clustering techniques could offer complementary insights. Hierarchical clustering, for example, could reveal nested structures within customer segments, providing a hierarchical view of customer behavior. Alternatively, density-based methods like DBSCAN might identify outliers or non-spherical clusters that KMeans overlooks, such as niche customer groups with unique purchasing patterns. Exploring these methods could lead to more robust segmentation, especially in datasets with complex or noisy structures.

6.0.3 Predictive Analytics

While the current analysis is descriptive, integrating predictive models could enhance its business value. For instance, developing a churn prediction model using the identified segments could help businesses proactively target "At risk customers" before they disengage. Similarly, predicting customer lifetime value (CLV) for each segment, particularly for "Best customers," could guide resource allocation and loyalty program investments. Machine learning techniques, such as decision trees or neural networks, could be applied to historical purchase data to forecast future buying behavior, enabling more dynamic and forward-looking marketing strategies.

6.0.4 Real-Time Implementation and Scalability

The current analysis was performed as a static snapshot of the "Online Retail" dataset. In a real-world business setting, implementing this segmentation in a real-time pipeline would allow for continuous monitoring and adaptation of customer segments. For example, integrating the RFM clustering model into a customer relationship management (CRM) system could enable automated segment updates as new purchase data is received. Additionally, scaling the analysis to handle larger datasets or multiple regions (beyond the UK) would require optimizing the computational efficiency of the clustering process, potentially through distributed computing frameworks like Apache Spark. Such enhancements would ensure the approach remains practical and effective in a dynamic business environment.

6.0.5 Evaluation of Marketing Strategies

The marketing strategies proposed for each customer segment (e.g., loyalty programs for "Best customers," reactivation campaigns for "Lost customers") should be tested and evaluated in future work. A/B testing or controlled experiments could measure the effectiveness of these strategies in terms of customer retention, revenue growth, and engagement metrics. For example, offering introductory discounts to "New customers" could be compared against a control group to assess the impact on repeat purchases. Iterative refinement of these strategies based on empirical results would maximize their business impact.

This future work outline was developed as part of the analysis conducted on June 02, 2025, at 05:25 PM BST, providing a roadmap for advancing customer segmentation and marketing personalization in future iterations of this project.

Bibliography

[1] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson Education, 2005.

[2] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[3] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3 ed., 2011.

[5] W. McKinney, *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, 2010.

[6] T. E. Oliphant, *A Guide to NumPy*. Trelgol Publishing, 2006.

[7] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[8] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[9] "Scikit-learn documentation." `https://scikit-learn.org/stable/modules/clustering.html`. Accessed: 2025-06-02.

[10] T. Kluyver *et al.*, "Jupyter notebooks - a publishing format for reproducible computational workflows," *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87–90, 2016.

[11] G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal, "Good enough practices in scientific computing," *PLOS Computational Biology*, vol. 13, no. 6, p. e1005510, 2017.

[12] Kaggle, "Online retail dataset," 2023. Accessed: 2025-06-02.

[13] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14, pp. 281–297, 1967.