Optimizing Customer Segmentation with Clustering Algorithms: A Comparative Study

[1] Riad Solh Anouar El Mahraoui Amal [2]

Abstract— **In today's competitive market, understanding customer preferences and behaviors is crucial for delivering personalized experiences. Customer segmentation plays a pivotal role in identifying distinct customer groups based on various characteristics such as purchase history, demographics, and behaviors [3]. This study proposes a customer segmentation approach using clustering algorithms, such as K-means and Hierarchical Clustering, to group customers with similar preferences [4]. By applying unsupervised machine learning techniques, the system aims to provide insights that can be used to personalize marketing efforts, enhance customer satisfaction, and drive business growth [7]. The results demonstrate the effectiveness of clustering algorithms in segmenting customers, offering valuable insights for targeted marketing strategies and resource allocation [1].**

This project applies association rule mining and the Apriori algorithm to analyze grocery transactions and discover valuable insights about product relationships. The goal is to uncover frequent itemsets—combinations of items that are often bought together—and generate association rules that predict future purchases. By exploring transaction data from a grocery store, a system is created that not only identifies common purchasing patterns but also supports decision-making for inventory management and marketing strategies [2].

Keywords: Customer Segmentation, Clustering Algorithms, Personalization, K-means, Hierarchical Clustering, Marketing Strategies

I. INTRODUCTION

The retail industry is constantly evolving, and with the rise of big data, understanding customer behavior has become more essential than ever [1]. Retailers are now leveraging data-driven approaches to personalize customer experiences and improve overall business outcomes. One of the most powerful techniques for achieving this is customer segmentation, which involves grouping customers based on shared characteristics such as purchasing habits, demographics, and preferences [3]. Customer segmentation allows businesses to better understand their diverse customer base and tailor their marketing strategies accordingly.

The aim of this paper is to explore the effectiveness of using clustering algorithms for customer segmentation, and how this can drive improvements in customer engagement and business performance.

II. LITERATURE REVIEW

Customer segmentation has long been an essential technique for businesses to understand customer behavior and tailor marketing strategies. The retail industry, in particular, relies heavily on data-driven insights for improving customer satisfaction and sales performance. Several studies have highlighted the importance of segmenting customers to provide personalized experiences, whether it be through targeted promotions, personalized product offerings, or better inventory management [3]. The effectiveness of customer segmentation is further enhanced by the use of clustering algorithms, which automatically group customers based on shared characteristics without the need for prior labeling.

Clustering techniques such as K-means [5] and Hierarchical Clustering [6] are widely used for customer segmentation. K-means is one of the most popular clustering algorithms due to its simplicity and efficiency, especially for large datasets. It divides customers into k clusters based on their feature similarity, minimizing intra-cluster variance [4]. However, K-means has limitations, such as sensitivity to the number of clusters and outliers. Hierarchical Clustering, on the other hand, creates a tree-like structure of clusters, allowing for a more detailed analysis of the data's structure and enabling businesses to choose the level of granularity that best fits their needs [6].

In addition to clustering, recommendation systems have gained significant attention in recent years as they are crucial for enhancing customer experience by providing personalized suggestions. While this study did not focus on building a recommendation system, future work could explore the integration of clustering-based customer segmentation with recommendation models to further improve product suggestions and personalization.

Overall, the literature emphasizes the importance of combining clustering algorithms for customer segmentation with recommendation systems to provide highly personalized and effective solutions. These techniques not only improve customer engagement but also lead to more efficient inventory management, optimized marketing strategies, and ultimately higher sales.

III. METHODOLOGY

This study employs a structured approach to analyze customer transaction data, focusing on data preprocessing, exploratory data analysis, RFM modeling, and clustering techniques to achieve effective customer segmentation. The methodology ensures data integrity, uncovering of behavioral patterns, and the development of meaningful customer clusters.

The first step involves **data collection and cleaning**, where the raw dataset is loaded and processed to remove in-

*This work was not supported by any organization

[1] A. Riad Solh is with Faculty of Sciences, Mohammed V University, 4 Ibn Batouta Avenue, P.O. Box 1014, Rabat, Morocco

[2] A. El Mahraoui is with Faculty of Sciences, Mohammed V University, 4 Ibn Batouta Avenue, P.O. Box 1014, Rabat, Morocco

consistencies. This includes handling missing values, treating canceled transactions to avoid misleading insights, eliminating duplicated rows, and dropping unnecessary columns that do not contribute to the analysis. These steps ensure that the dataset is structured and ready for further processing.

Following data cleaning, **exploratory data analysis (EDA)** is conducted to gain insights into customer purchasing behavior. This involves analyzing the distribution of customers across different countries [Fig: 1], identifying peak shopping days and hours [fig: 2], and performing **correlation analysis** [fig: 4] to examine relationships between key transaction attributes. Understanding these patterns allows for a deeper interpretation of customer behavior and purchasing tendencies.

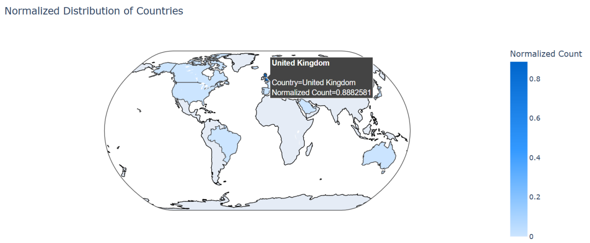

Fig. 1. Normalized Distribution of Countries

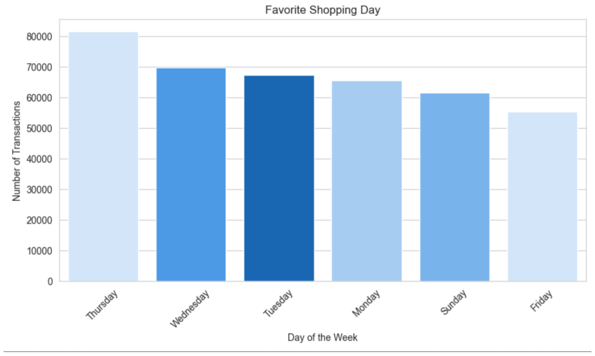

Fig. 2. Favourite Shopping Days

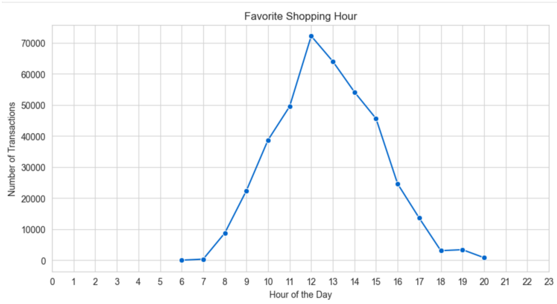

Fig. 3. Favourite Shopping hour

Next, **RFM (Recency, Frequency, and Monetary) modeling** is applied to quantify customer value. Recency measures the time elapsed since a customer's last purchase, frequency accounts for the number of transactions, and monetary reflects the total spending of a customer. These RFM scores are normalized to ensure balanced clustering.

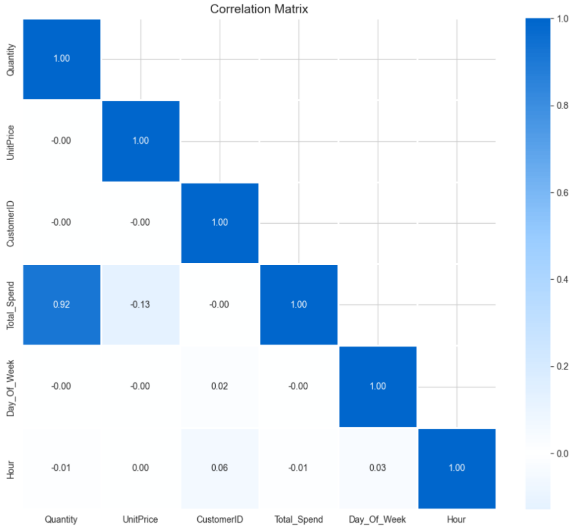

Fig. 4. Correlation Analysis

In the final stage, **clustering techniques** are implemented to segment customers based on their purchasing behavior. The **Elbow Method** is first applied [fig : 5]to determine the optimal number of clusters. The dataset is then scaled to ensure uniformity, followed by **Principal Component Analysis (PCA)** to reduce dimensionality while preserving essential patterns. Finally, **K-Means Clustering** and **Agglomerative Hierarchical Clustering** are performed to segment customers into distinct groups, with cluster quality evaluated using the **silhouette function**.

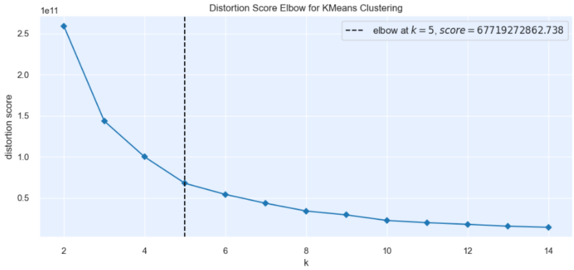

Fig. 5. Elbow score

This methodology provides a robust framework for customer segmentation by combining statistical analysis with machine learning techniques, ensuring actionable insights that businesses can leverage to optimize their marketing strategies.

IV. Proposed Work

The proposed work follows a structured workflow, from data acquisition to final customer segmentation, ensuring systematic and efficient processing. The following sections describe the key components of the system.

A. System Design Overview

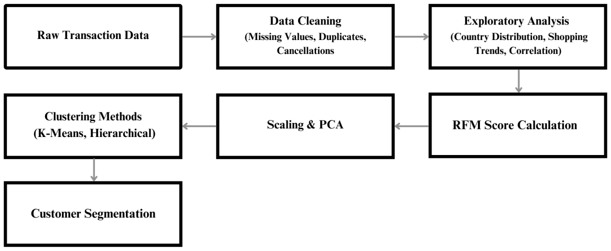
Raw Transaction Data
Data Cleaning (Missing Values, Duplicates, Cancellations
Exploratory Analysis (Country Distribution, Shopping Trends, Correlation)
RFM Score Calculation
Scaling & PCA
Clustering Methods (K-Means, Hierarchical)
Customer Segmentation

Fig. 6. System Design

The system begins with **data collection and preprocessing**, where raw transaction data undergoes a thorough cleaning process, including missing value treatment, cancellation handling, duplicate removal, and column filtering. This step ensures that only high-quality data is used for further analysis.

Subsequently, **exploratory data analysis (EDA)** is performed to identify meaningful trends in customer purchasing behavior. This includes analyzing the **distribution of customers by country**, determining the **most popular shopping days and hours**, and conducting **correlation analysis** to uncover relationships between different variables in the dataset. These insights provide a comprehensive understanding of customer behavior and help in designing effective marketing strategies.

After EDA, **RFM modeling** is applied to compute Recency, Frequency, and Monetary scores for each customer, which serve as inputs for clustering. The dataset is then **scaled**, and **PCA** is applied to enhance clustering performance by reducing dimensionality.

The **clustering stage** involves determining the optimal number of clusters using the **Elbow Method**, followed by the application of **K-Means and Agglomerative Hierarchical Clustering** to segment customers. The final clusters are evaluated using the **silhouette function**, ensuring the segmentation is meaningful and well-defined.

This systematic approach ensures that customer segmentation is data-driven and provides actionable insights for personalized marketing and strategic decision-making.

This system design ensures a logical progression from raw data to meaningful customer segmentation, enabling businesses to make informed decisions based on data-driven insights.

V. Results and Discussion

The results of applying K-means and Agglomerative Hierarchical Clustering are summarized in the table below. The evaluation metrics used to assess the clustering quality include the Silhouette Score and Davies-Bouldin Index. The results from multiple trials for each clustering method are shown.

Model	**Silhouette Score**	**DB Index**
K-means (Trial 1)	0.5816	0.5285
K-means (Trial 2)	0.7972	0.5843
Agglomerative Clustering	0.8830	N/A

TABLE I
Clustering Results for K-means and Agglomerative Clustering

The Silhouette Score for K-means in Trial 1 is 0.5816, indicating that the clustering is somewhat meaningful but could be improved. In Trial 2, the Silhouette Score increases to 0.7972, suggesting a better-defined clustering solution. The Davies-Bouldin Index in Trial 1 is 0.5285, showing that the clusters are relatively well-separated. In Trial 2, the Davies-Bouldin Index is 0.5843, suggesting a slightly less optimal clustering but still with reasonable separation.

For Agglomerative Hierarchical Clustering, the Silhouette Score is 0.8830, significantly higher than K-means, indicating that the clusters are well-separated and meaningful. The Davies-Bouldin Index is not applicable to this method as it is typically used for K-means-based algorithms.

Agglomerative Hierarchical Clustering outperformed K-means in terms of the Silhouette Score, demonstrating better clustering quality with more distinct and well-separated clusters. K-means showed improvement from Trial 1 to Trial 2, but still did not reach the same level of clustering quality as Agglomerative Clustering. The Davies-Bouldin Index indicated that K-means could still benefit from further optimization in both trials. Overall, for this dataset, Agglomerative Hierarchical Clustering appears to be the more suitable clustering technique.

VI. CONCLUSIONS

This study presents a data-driven approach to customer segmentation using transaction data, integrating statistical analysis with machine learning techniques. The methodology involves a systematic process of data preprocessing, exploratory data analysis, RFM modeling, and clustering, enabling businesses to understand customer behavior in depth. By leveraging the Elbow Method, PCA, and clustering techniques such as K-Means and Agglomerative Hierarchical Clustering, customers are effectively grouped based on their purchasing patterns.

The results of this study provide valuable insights that can assist businesses in optimizing marketing strategies, improving customer retention, and increasing revenue. Through customer segmentation, businesses can personalize promotions, enhance customer experience, and implement data-driven decision-making strategies.

Future work could extend this analysis by incorporating additional external factors such as social media interactions, customer demographics, and product categories to refine segmentation further. Additionally, deep learning techniques could be explored to enhance clustering performance and improve predictive accuracy.

This research highlights the significance of data-driven methodologies in business intelligence, demonstrating how structured analytics can drive strategic decision-making and customer relationship management.

APPENDIX

To maintain clarity and focus on the primary objectives, not all result images and code snippets are included in this report. However, the full set of results, including intermediate images, and the complete source code are available in the corresponding GitHub repository

https://github.com/aelmah/Data-Mining

REFERENCES

[1] Ravi S. Bose. Big data and analytics in the retail industry. *Retail Management Review*, 29(3):21–34, 2016.

[2] Shuo Chen, Weiwei Liu, and Ruijun Li. Mining association rules for retail analytics: A data-driven approach. *International Journal of Retail and Distribution Management*, 45(8):909–925, 2017.

[3] Shivendra Gupta, Manish Kumar, and Anjali Verma. Data mining for customer segmentation in retail business. *International Journal of Computer Science and Information Security*, 13(8):12–18, 2015.

[4] Anil K. Jain and Richard C. Dubes. *Data Clustering: A Review*. Springer, 2010.

[5] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[6] F. Murtagh. A survey of hierarchical classification. *Computing Reviews*, 24(5):342–344, 1983.

[7] Xiaoyan Xu, Hongyu Wu, and Shuangyin Liu. A survey on clustering algorithms for customer segmentation. *Journal of Data Science*, 13(2):245–262, 2015.