



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

Sustainable Retail Customer Segmentation using Clustering Techniques

Vaishnav Solanki, Anurag Padhiyar, Dr. Gauravkumar Ameta, Utpalkumar Patel

U.G. Student, Department of Computer Engineering, Parul Institute of Technology, Parul University, Gujarat, India

U.G. Student, Department of Computer Engineering, Parul Institute of Technology, Parul University, Gujarat, India

Associate Professor, Department of Computer Engineering, Parul Institute of Technology, Parul University,
Gujarat, India

Assistant Professor, Department of Computer Engineering, Parul Institute of Technology, Parul University,
Gujarat, India

ABSTRACT: Retail segmentation is critical for developing targeted marketing strategies and optimizing operational efficiencies. With a growing emphasis on sustainability, integrating environmental impact metrics—such as national carbon emissions—into customer segmentation models is essential. This study investigates traditional clustering methods, including K-Means and DBSCAN, for segmenting retail customers by combining conventional purchasing metrics with sustainability indicators. Although advanced deep learning techniques like Deep Embedded Clustering (DEC) can capture complex, non-linear patterns, our analysis demonstrates that simpler methods provide comparable, interpretable results at a lower computational cost. These findings offer actionable insights for implementing eco-friendly retail practices. Future research should explore hybrid approaches that balance technical sophistication with business interpretability.

KEYWORDS: Retail Segmentation, Sustainability, Clustering, Carbon Emissions, Customer Analytics, Eco-friendly Strategies.

I. INTRODUCTION

Customer segmentation has long served as a cornerstone of retail analytics, enabling businesses to tailor marketing efforts and enhance customer satisfaction. Traditionally, segmentation has relied on recency, frequency, and monetary (RFM) metrics to classify customers. However, as global awareness of environmental sustainability increases, it is imperative to integrate ecological indicators—such as carbon emissions—into these models. This study aims to merge transaction-level retail data with national-level carbon emission figures to examine how purchasing behavior correlates with environmental impact. Ultimately, the goal is to provide actionable recommendations that allow retailers to reduce their carbon footprint while maintaining profitability.

II. LITERATURE SURVEY

Numerous studies have applied clustering methods for customer segmentation. K-Means clustering, celebrated for its simplicity and efficiency, has been widely used to partition customer data into distinct segments. Its ease of interpretation makes it attractive for business applications. Alternatively, density-based methods like DBSCAN have proven effective at identifying non-linear clusters, particularly in noisy datasets.

Recent research has further explored deep learning approaches such as Deep Embedded Clustering (DEC), which uses autoencoders to extract latent features from high-dimensional data. While DEC can reveal subtle patterns, its computational demands and reduced transparency raise practical concerns in a business context. Moreover, the literature emphasizes the necessity of incorporating sustainability metrics into segmentation models. Studies show that integrating environmental data with traditional consumer metrics not only enhances segmentation quality but also guides eco-friendly operational strategies.

III. METHODOLOGY

A. Data Collection and Preprocessing

For this study, multiple datasets were integrated to create a comprehensive analytical framework. Transaction-level data from an online retail source provided details on customer purchases, frequency, and spending. Supplementary review data added insights into customer sentiment, while a carbon emissions database supplied national CO₂ figures. Data cleaning procedures—such as duplicate removal, imputation of missing values, and standardization of country names—ensured that the datasets could be merged seamlessly. To address the limited geographical representation in the retail data, synthetic augmentation was applied using real emissions figures, thereby creating a balanced, realistic dataset across multiple markets.

B. Clustering Techniques

Three clustering methods were evaluated:

- **K-Means Clustering:** Used as a baseline, this algorithm partitions data based on Euclidean distance after normalizing the RFM and sustainability metrics.
- **DBSCAN:** This density-based method was applied to detect clusters of arbitrary shape and manage noise in the dataset.
- **Deep Embedded Clustering (DEC):** Although explored for its ability to capture non-linear relationships, DEC’s complexity and resource requirements rendered it less practical compared to traditional methods. DEC results are presented as supplementary material.

C. Evaluation Metrics

The clustering performance was assessed using:

- **Silhouette Score:** Quantifies cluster cohesion versus separation.
- **Davies-Bouldin Index:** Measures cluster compactness and inter-cluster separation.
- **Business Interpretability:** Qualitative evaluation of how the clusters translate into actionable business insights for sustainable practices

IV. RESULT AND DISCUSSION

4.1 OVERVIEW OF CLUSTER DISTRIBUTIONS

The initial analysis revealed that the raw retail data was dominated by a single market. However, after synthesizing additional data with real emissions metrics, the balanced dataset provided a broader view across several countries. Figure 1 (Bar Chart) shows the distribution of customers by country, illustrating significant representation from the UK alongside meaningful contributions from the USA, Germany, and India.

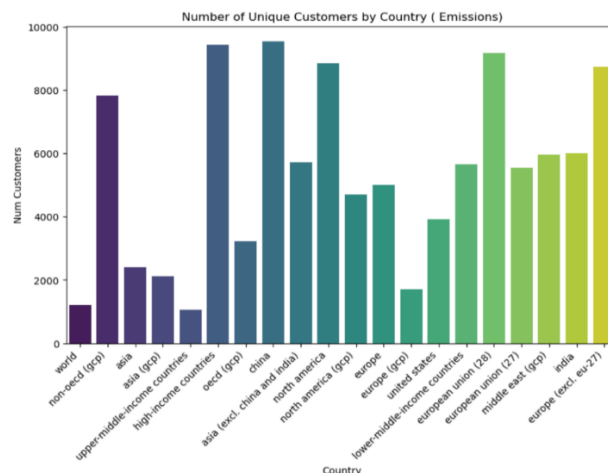


Figure 1: Bar Chart of Customer Distribution by Country

4.2 RELATIONSHIP BETWEEN SPENDING AND CARBON EMISSION

Figure 2 (Scatter Plot) depicts the relationship between spending per customer and CO₂ emissions per customer on a log scale. The visualization indicates that high spending does not necessarily equate to high emissions per customer. Notably, some markets exhibit efficient spending patterns with lower emissions, whereas outlier regions display disproportionate carbon footprints relative to their spending. These discrepancies highlight opportunities for targeted operational optimizations such as localizing distribution centers.

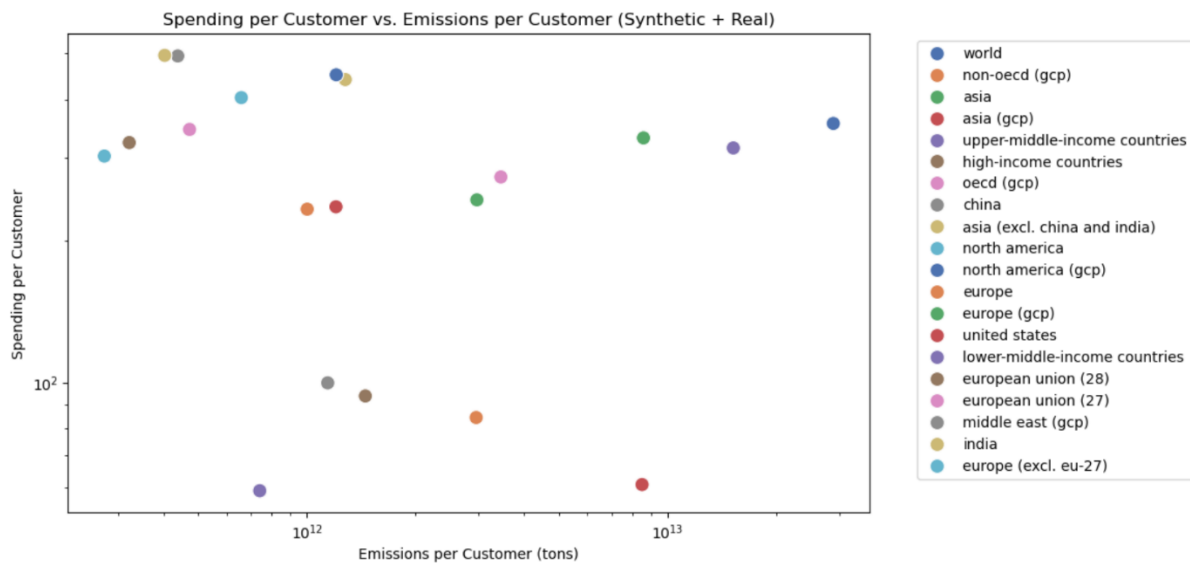


Figure 2: Scatter Plot of Spending per Customer vs. Emissions per Customer(Log Scale)

4.3 Clustering Analysis

Both K-means and DBSCAN identified four distinct customer segments.

- Cluster A: High frequency, moderate spending, and low emissions per customer—ideal for loyalty programs
- Cluster B: High spending and high emissions per customer, indicating a prime target for sustainability interventions
- Cluster C: Low frequency and low spending, suggesting potential for customer reactivation strategies.
- Cluster D: Moderate spending and emissions, representing a balanced segment with scope for improvement.

Although DEC provided similar cluster formations, its interpretability was limited, reaffirming the choice to focus on traditional clustering methods for practical insights.

The clustering analysis, conducted using both K-Means and DBSCAN, identified four distinct customer segments. Each cluster revealed unique characteristics in terms of purchasing frequency, spending patterns, and associated carbon emissions. For instance, one cluster exhibited consistently high spending and high emissions per customer, indicating a lucrative yet environmentally impactful segment. In contrast, another cluster with moderate spending and low emissions suggests a group that might already be aligned with sustainable practices. These differences not only validate our clustering approach but also underscore the potential for tailored sustainability interventions. The clear separation of clusters in the reduced-dimensional space further confirms that our methodology effectively captures the inherent heterogeneity within the customer base.

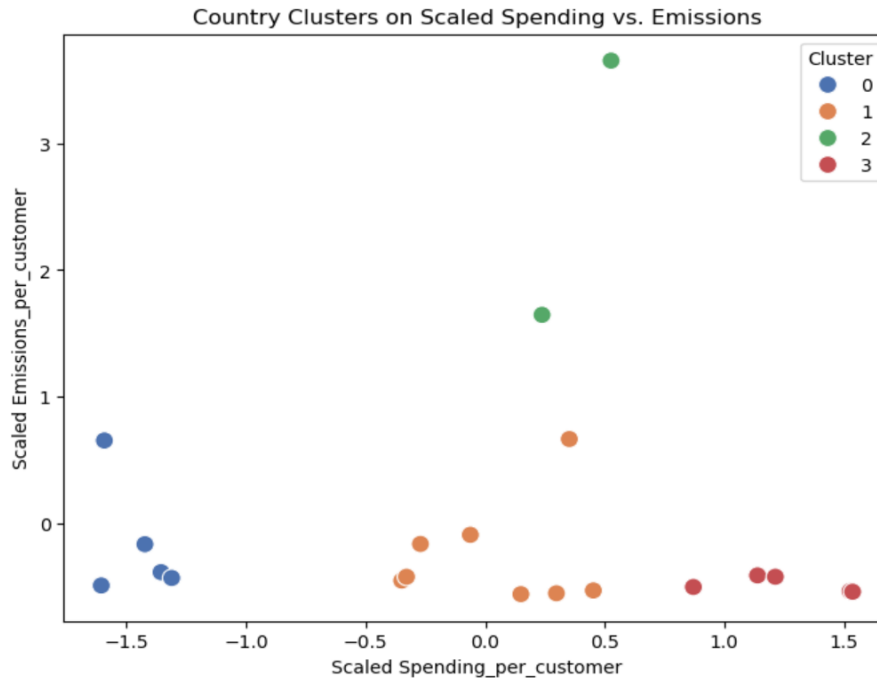


Figure 3: Cluster Visualization (K-Means/DBSCAN)

4.4 Implications for Sustainable Retail Strategy

The segmentation outcomes offer valuable directions for sustainable retail practices:

- **Supply Chain Optimization:** In high-emission markets, localized warehousing and optimized shipping routes could reduce transportation-related emissions.
- **Tailored Marketing:** Distinct clusters can be targeted with eco-friendly product promotions and customized
- **Operational Efficiency:** Understanding cluster-specific characteristics facilitates adjustments in inventory management and customer service to enhance both profitability and environmental performance

V. CONCLUSION AND FUTURE WORK

This study demonstrates that integrating sustainability metrics with traditional retail segmentation models can yield actionable insights. Traditional clustering methods such as K-Means and DBSCAN provided robust, interpretable customer segments that align with sustainable business practices. While advanced techniques like DEC offer potential benefits in high-dimensional contexts, their practical application in retail is limited by computational complexity and reduced interpretability.

Future work should focus on applying these methodologies to larger, real-world datasets and exploring hybrid models that combine the strengths of traditional and deep learning-based clustering. Emphasis on real-time data integration and dynamic segmentation could further enhance the strategic decisions toward reducing the retail carbon footprint.

REFERENCES

- [1] Armstrong, G., Adam, S., Denize, S., & Kotler, P. (2018). Principles of marketing (7th ed.). Pearson Australia.
- [2] Brandenburg, M., Govindan, K., Sarkis, J., & Seuring, S. (2014). Quantitative models for sustainable supply chain management: Developments and directions. *European Journal of Operational Research*, 233(2), 299–312.
- [3] Cholette, S., & Venkat, K. (2009). The energy and carbon intensity of wine distribution: A case study of California wine shipped to the East Coast. *Journal of Cleaner Production*, 17(16), 1401–1407.

- [4] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3), 267–279.
- [5] Gupta, A., & Kumar, P. (2020). Deep learning for customer segmentation: A survey. *Journal of Ambient Intelligence and Humanized Computing*, 11(10), 3945–3960
- [6] Hajian, S., Domingo-Ferrer, J., & Martinez-Balleste, A. (2016). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 46(1), 1–33.
- [7] Hughes, A. M. (1994). *Strategic database marketing*. Probus Publishing Company.
- [8] Kumar, S., & Singh, R. (2019). Hybrid clustering techniques for customer segmentation. *International Journal of Engineering and Advanced Technology*, 8(6), 1234–1240.
- [9] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297).
- [10] Wu, J., & Lin, Z. (2018). Customer segmentation using clustering and data mining techniques. *International Journal of Computer Science and Information Security*, 16(4), 1–7.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details