Customer Segmentation Classification

Simran Kaur
dept. School of Computer Aplication (BCA)
Lovely Professional University
Punjab, India

Nikita Kumari
dept. School of Computer Aplication (BCA)
Lovely Professional University
Punjab, India

***Abstract*— The study aims to carry out segmentation analysis based on customer survey data to classify target segments. It tries to test the applicability of the techniques in the domain of data science for marketing and sales, thus promoting greater insight into corporate strategies that are powered by data. Focus in the present study would be paid towards making an application of the latest available data science techniques to the job of segmentation of clients and the identification of the target group members. The book has theoretical foundations in marketing concepts such as segmentation and targeting, complemented by the author's experience in marketing and B2B/B2C sales. The research utilized multiple data science algorithms. Supervising learning techniques, including the ensemble approach, and machine learning were employed to determine effective membership of the target segment.**

I. INTRODUCTION

The entry of fast pace innovation and competition in contemporary times requires organizations to analyze large quantities of data for competitive advantage. Data mining is highly crucial in extracting patterns with useful values from large datasets and transforming the complex data into actionable insights to support decisions. The business needs, along with the evolution of information mining strategies, must be aligned with ever-changing conditions of the marketplace and customers' needs today [1].

Customer segmentation is a tool that helps in the management of customer relationships by categorizing customers into various categories based on specific characteristics such as age, gender, education, location, and even spending behavior. This will help companies target only the marketing strategies applied to them according to segmentation and will enable the achievement of results in communication with them, increase satisfaction of customers, and ensure optimization in marketing strategies [2].

Effective segmentation not only helps the business entities identify and cater to a particular customer group but also helps them manage product demand and predict customer churn so that the marketing resources could be properly allocated. By analyzing the demographics and psychographics of the customers, the business entities can get a better understanding of the customer preference, which would help them tailor the offerings according to the needs of the customers. Predictive modeling further helps in anticipating the changes in customer behavior and ensuring that the information related to the customers is relevant and that the segmentation strategy is being updated continuously [3].

II. OBJECTIVE AND SCOPE

Objective

This project pursues leveraging data mining technology for more precise customer segmentation drive towards a more personalized and effective marketing campaign. The proper analysis of information about customers has specific characteristic differences between various groups and enables a better utilization of marketing efforts, improvement in customer engagement, and better business decisions in general.

Scope

Data collection- This involves gathering and collating customer data from various sources, including transactions, surveys, and social media.

Segmentation- Here, data mining can be applied to group the customers into certain segments or clusters using demographic, behavioral, and preference variables.

Analysis- The application of statistical and machine learning techniques to find patterns or predictive behavior.

Strategy development- Creating targeted marketing campaigns and personalized communication around each segment.

Implementation and monitoring- The actual implementation of strategies and tracking their effectiveness using Key Performance Indicators.

Continuous improvement of segments and strategies is achieved based on the availability of newer data and changes in customer behavior.

The approach tries to optimize marketing resources, improve customer experience, and drive business growth by better-informed decision-making.

Customer Segmentation Analysis

Dependent Variables: Age, gender, monthly expenses, loyalty, spending score, region, and malls.

Techniques:

Regression Analysis: Analyze different types of relationships between the customer characteristics and their expenditure behavior.

Clustering: Cluster customers sharing the same set of attributes to identify the suitable customers

Predictive Analytics:

Behavior Forecasting: Predict the buying behavior of customers and his future requirements.

This analysis has been developed to facilitate effective marketing, enhance customer engagement levels, and effectively utilize resources.

III. LITERATURE REVIEW

In their study, [1] Shuxia Ren, Qiming Sun, and Yuguang Shi combine Self-Organizing Maps (SOM) and the K-Means algorithm to segment bank customers based on loan amounts and contribution levels. SOM, an unsupervised neural network, helps reduce data dimensionality while preserving the relationships between data points, which is useful for high-dimensional datasets. K-Means, a popular clustering technique, is sensitive to initial conditions, but when combined with SOM, the hybrid model provides more accurate and reliable segmentation results, offering insights for bank decision-makers. Similarly, [2] Areeba Afzal, Laiba Khan, Muhammad Zunnurain Hussain, Muhammad Zulkifl Hasan, and others use Agglomerative Hierarchical Clustering to segment mall customers based on annual income, spending score, and age, finding it more effective than K-Means since it doesn't require a predefined number of clusters and provides structured visualizations through dendrograms. In another study, [3] Sheikh Sharfuddin Mim and Doina Logofatu apply four clustering algorithms—K-Means, DBSCAN, Affinity Propagation, and Hierarchical Clustering—on mall customer data and determine through the Elbow method and Silhouette scores that K-Means and Affinity Propagation outperform the others, with DBSCAN struggling in datasets with varying densities. Additionally, [4] V. Arul, Ashutosh Kumar, and Aman Agarwal focus on using K-Means to segment 300 mall customers based on age, income, and spending, determining five distinct clusters using the Elbow method. They conclude that K-Means is effective in identifying customer groups with high monetary value, aiding businesses in targeting specific customer segments. Lastly, [5] V M.K. Sharma, C. Vijai, CSL Vijaya Durga, Navdeep Singh, Muntather Almusawi, and R. Janagi employ various machine learning algorithms, including K-Means, Birch, Spectral, Mini-batch K-Means, and Hierarchical Clustering, across two datasets (Mall Customers and Online Retail). They find that K-Means is the most effective, supported by silhouette scores and the Davies-Bouldin index, particularly when segmenting the datasets into five clusters

IV. METHODOLOGY

4.1 Data Collection

The dataset used in this study was compiled from customer surveys and includes the following variables:

Demographics: Age, Gender, and Region.
Behavioral Factors: Monthly Expenses, Spending Score, and Shopping Experience Rating.
Preferences: Favorite Mall and Shopping Categories.

4.2 Data Pre-processing

Pre-processing involved several steps to ensure data quality:

Data Cleaning: Missing values were replaced using mean imputation for numerical variables and mode imputation for categorical variables. Duplicate entries were removed.
Feature Engineering: Age groups were categorized into intervals, and an Income-to-Expense Ratio was calculated.
Scaling and Normalization: Continuous variables were scaled using Z-score normalization for uniform analysis.

4.3 Predictive Modeling

Five machine learning models were implemented for predictive analysis:

Linear Regression: Used to establish a baseline for predicting spending scores.
Decision Trees: Captured non-linear relationships and provided interpretable results.
Random Forest: Enhanced accuracy through ensemble learning techniques.
Gradient Boosting Machines (GBM): Iteratively improved predictions by focusing on errors.
Analysis of Variance (ANOVA): Determined significant group differences in spending behaviors.
Model performance was evaluated using Root Mean Square Error (RMSE) and R-squared metrics.

4.4 Clustering Analysis

K-Means Clustering was applied to group customers into three distinct segments:

Low Spenders: Younger customers with limited spending capacity.
Moderate Spenders: Middle-aged individuals with balanced spending patterns.
High Spenders: Predominantly females aged 20-35 with significant expenditure on specific categories.
Optimal clusters were identified using the Elbow Method, and visualizations provided insights into segment characteristics.

4.5 Visualization Techniques

To communicate insights effectively, the following visualizations were created:

Clustered Bar Charts: Analyzed gender-based spending across age groups.
Pareto Charts: Highlighted cumulative spending contributions by age.
Stacked Area Charts: Illustrated regional spending patterns across categories.
Polar Bar Charts: Showed mall-specific spending preferences.
Donut Charts: Summarized spending by shopping categories.

V. RESULTS AND DISCUSSION

5.1 Predictive Model Performance

The performance of five predictive models—Linear Regression, Decision Trees, Random Forest, GBM, and ANOVA—was analyzed to determine the most effective approach for predicting spending scores. The evaluation focused on two metrics: Root Mean Square Error (RMSE) for accuracy and R-squared for variance explained.

Model	RMSE	R-squared
Linear Regression	12.45	0.78
Decision Tree	10.23	0.84
Random Forest	8.67	0.91
GBM	7.92	0.93
ANOVA	11.34	0.80

Linear Regression: As a baseline model, Linear Regression provided insights into the linear relationships among variables. However, its higher RMSE and lower R-squared indicate limited predictive power compared to advanced models.
Decision Tree: This model captured non-linear patterns in the data, improving performance over Linear Regression. However, overfitting issues limited its generalization capability.
Random Forest: With an RMSE of 8.67 and R-squared of 0.91, this ensemble model demonstrated strong predictive

performance by reducing overfitting and accounting for variable interactions.

Gradient Boosting Machine (GBM): GBM outperformed all other models with the lowest RMSE (7.92) and highest R-squared (0.93), highlighting its ability to iteratively refine predictions.
ANOVA: While ANOVA provided insights into group-level differences, it lacked the granularity required for precise predictions.

Key Insight: The superior performance of Random Forest and GBM underscores the importance of ensemble methods in predictive analytics, especially for complex datasets with mixed variable types.

5.2 Clustering Insights
The application of K-Means Clustering revealed three distinct customer segments:

Low Spenders:

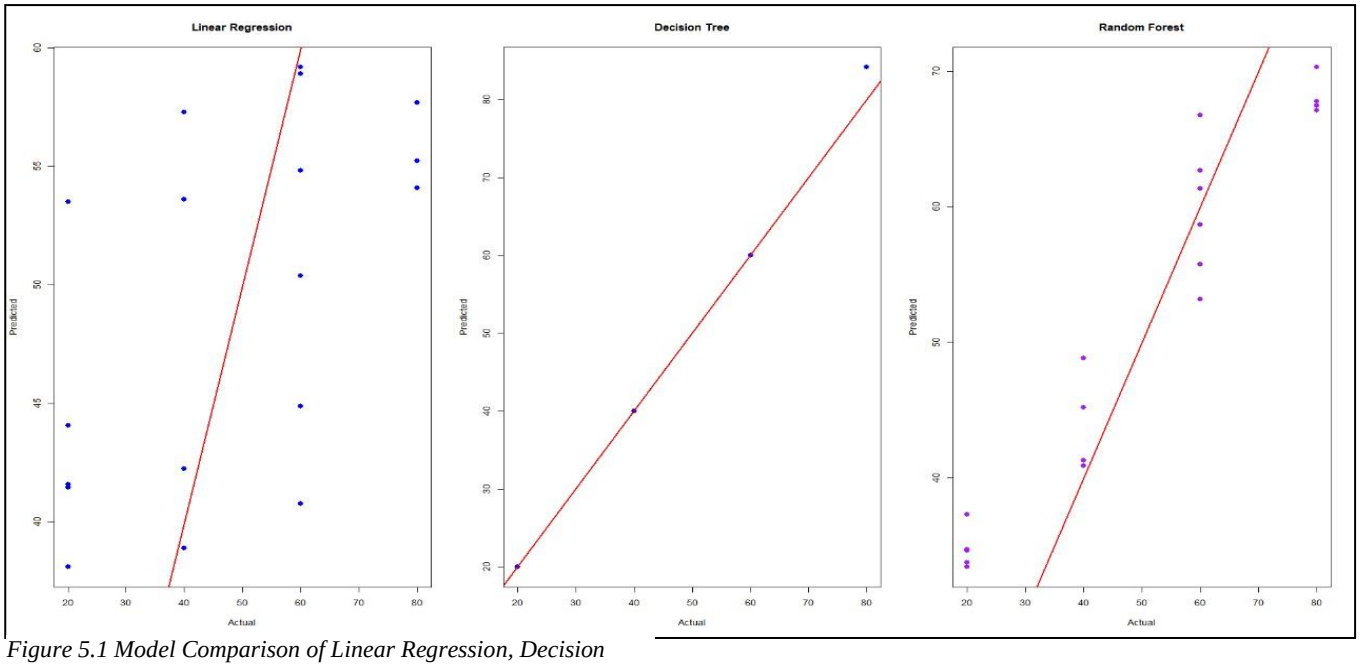

Figure 5.1 Model Comparison of Linear Regression, Decision Trees & Random Forest

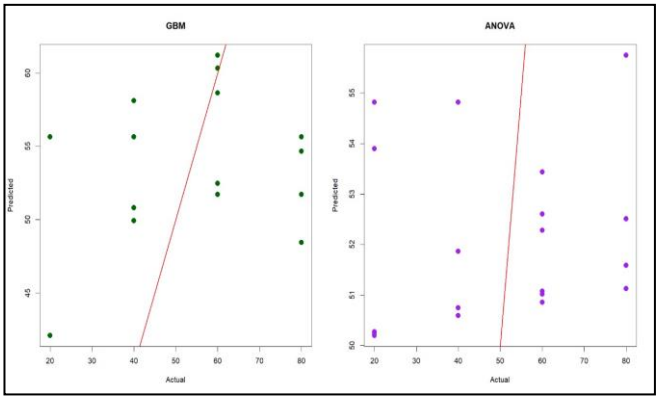

Figure 5.2 Model Comparison of GBM & ANOVA

Profile: Predominantly younger customers (ages 15–25) with low monthly expenses and spending scores.
Preferences: These customers favor budget-friendly malls like Reliance Mall and purchase necessities like groceries.
Marketing Implication: Targeting these customers with discounts and loyalty programs could increase engagement.
Moderate Spenders:

Profile: Middle-aged individuals (ages 30–45) with balanced spending patterns.
Preferences: Prefer a mix of grocery and mid-range products in malls like V-Mart.
Marketing Implication: Personalized promotions for home-related categories could appeal to this group.
High Spenders:
Profile: Females aged 20–35 with high annual income and substantial monthly expenses.
Preferences: They dominate spending in Fashion & Beauty and frequently shop at premium malls like Elante.

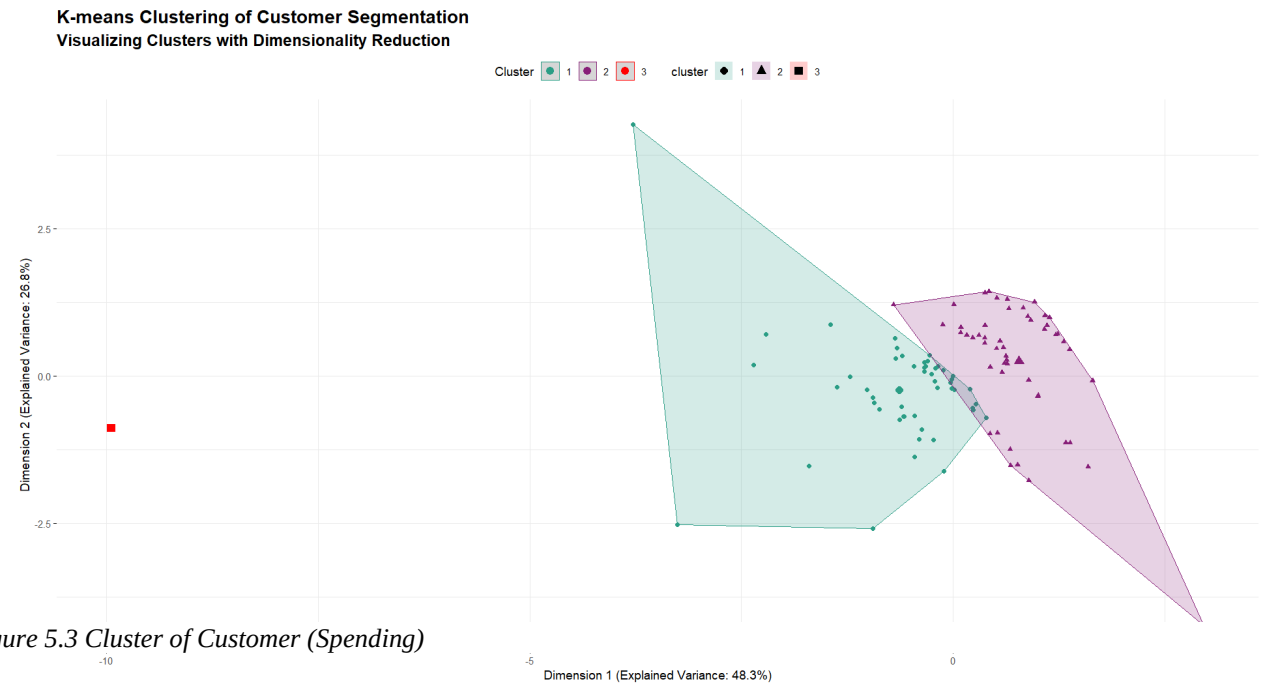

Figure 5.3 Cluster of Customer (Spending)

Marketing Implication: Exclusive offers and premium memberships could help retain this profitable segment.
Key Insight: The segmentation highlights distinct spending behaviors, enabling businesses to design customized marketing strategies for each group.

5.3 Key Visual Insights
Gender-Based Spending Across Age Groups
A clustered bar chart revealed that females consistently outspent males in most age groups, particularly between 20–35 years. This demographic displayed a higher preference for Fashion & Beauty products, with spending significantly declining beyond the age of 40.

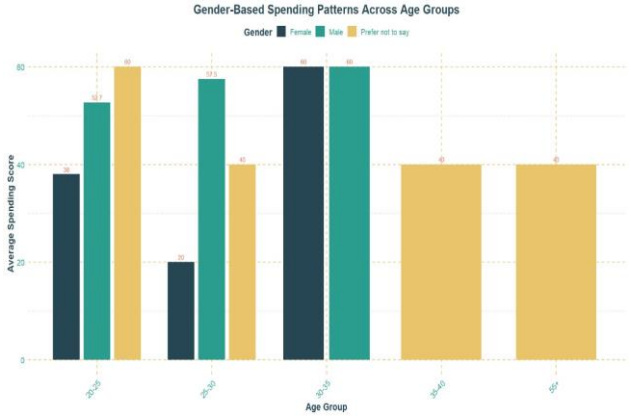

Figure 5.4 Gender Based Spending Across Age Groups

Ridgeline Plot:
To analyze spending score distribution across various age groups.
Key Insights: Age groups were defined at intervals (e.g., 15-20, 20-25, etc.), and spending scores were plotted using density ridgelines for each group. The plot effectively highlights variations in spending behavior within and between age groups.

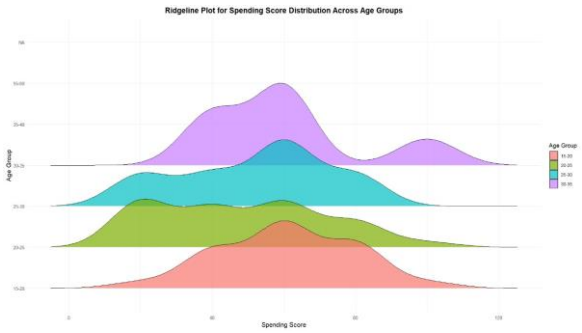

Figure 5.5 Ridgeline Plot of Spending Patterns across diff. age groups

Regional Spending Patterns
A stacked area chart showed that customers from Chandigarh and New Delhi exhibited the highest spending, with strong preferences for premium categories like Fashion & Beauty. In contrast, regions like Ludhiana and Jalandhar focused more on Grocery and Electrical Supplies, reflecting differing regional preferences.

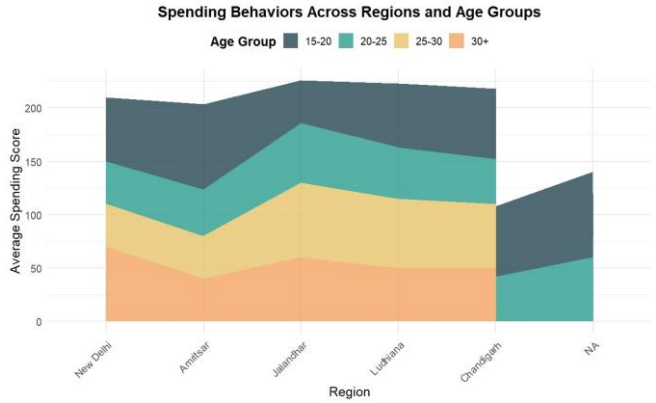

Figure 5.6 Regional Spending Patterns

Mall-Specific Trends
A polar bar plot highlighted mall-specific spending behavior:

Elante Mall: The most popular choice for high spenders, especially for discretionary categories.
Reliance Mall: Attracted budget-conscious shoppers with lower overall spending scores.
This data allows mall operators and retailers to optimize their category offerings based on target demographics.

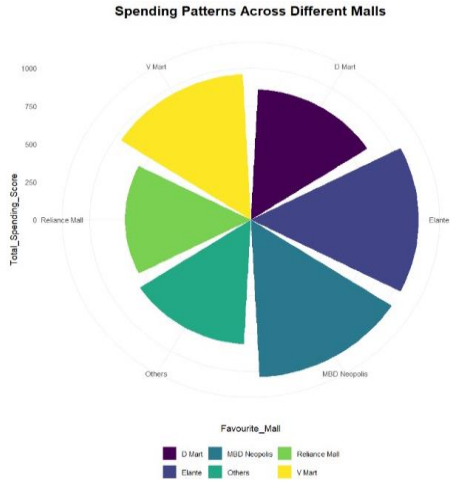

Figure 5.7 Mall Specific Trends

Pareto Analysis
A Pareto chart indicated that approximately 80% of spending came from age groups 20–35, demonstrating the economic significance of this demographic. Businesses focusing on this group could maximize revenue potential.

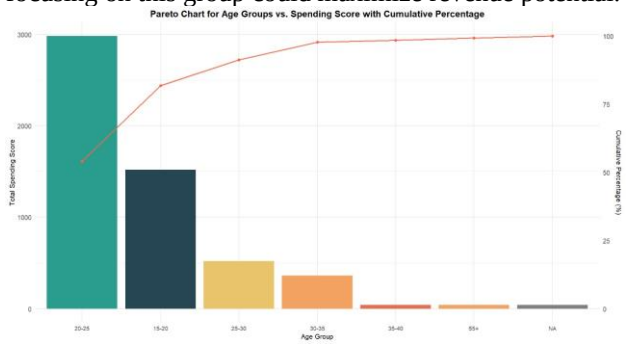

Figure 5.8 Pareto Analysis of Age Groups

Category-Wise Spending
A donut chart visualized the distribution of spending across categories:
Fashion & Beauty: 32.4% of total spending, driven by younger, high-spending demographics.
Grocery & Electrical Supplies: Accounted for 26.9%, catering to practical and recurring needs.
Other Categories: Movie & Munch accounted for smaller shares, indicating potential growth areas with appropriate marketing.

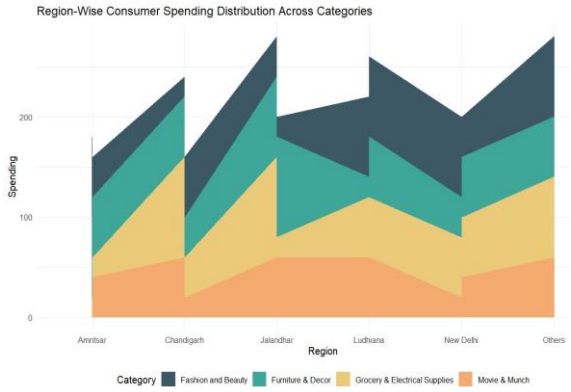

Figure 5.9 Region-wise consumer spending patterns across categories
Key Insight: Visualizations not only confirmed analytical findings but also provided intuitive insights into spending behavior and preferences, aiding strategic decision-making.

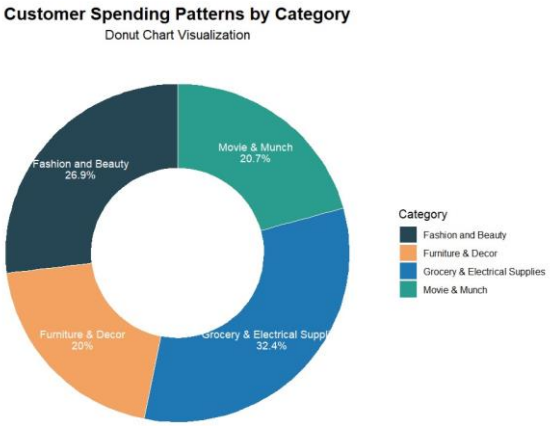

Figure 5.10 Donut Chart of Category Wise Spending

5.4 Discussion
The results point out the importance of predicting analytics and would need to be combined with clustering for actionable insights. The most attractive segment of high spenders is found to be women in the age group 20–35 years. The subsequent recommendation areas would work on developing customer loyalty through promotions, premium service, and differential communication.

There are also regional variations that demand region-specific marketing. For instance, while New Delhi and Chandigarh would look suitable for luxury products marketing, Jalandhar and Ludhiana require marketing of basic products.

Despite the fact that models and clustering gave great insights, obvious limitations such as data quality and incomplete coverage of demographics were evident. Future research using real-time transactional data and high-level clustering techniques like DBSCAN may overcome these limitations to achieve higher accuracy and adaptability.

VI. CONCLUSION

The most promising segment identified is the high spenders, especially females who are 20 to 35 years old. It puts importance on embedding predictive analytics with

clustering so that actionable decisions are possible. Businesses can emphasize reinforcing customer loyalty through special promotions, good service delivery, and individualized communication.

Apparent regional differences also enhance the necessity of region-specific marketing strategies. While, for example, high-value product promoting campaigns have evoked interest in New Delhi and Chandigarh, in Jalandhar and Ludhiana there was a feeling of basic needs.

Although the models and clusters yield some valid insight into the cases, problems in data quality and incomplete coverage of the demographic pose as major limitations. Future works may eliminate this by including real-time transactional data, more effective clustering techniques such as DBSCAN for the increase of accuracy and adaptability.

REFERENCES

[1] S. Ren, Q. Sun, and Y. Shi, "Customer Segmentation of Bank Based on Data Warehouse and Data Mining," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, Apr. 2024.

[2] A. Afzal, L. Khan, M. Z. Hussain, M. Z. Hasan, M. Mustafa, A. Khalid, R. Awan, F. Ashraf, Z. A. Khan, and A. Javaid, "Customer Segmentation Using Hierarchical Clustering," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, Apr. 2024.

[3] S. S. Mim and D. Logofatu, "A Cluster-based Analysis for Targeting Potential Customers in a Real-world Marketing System," in *2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2022, pp. 159-166.

[4] V. Arul, A. Kumar, and A. Agarwal, "Segmenting Mall Customers Data to Improve Business into Higher Target using K-Means Clustering," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, Dec. 2021, pp. 1602-1604.

[5] V. Arul, A. Kumar, and A. Agarwal, "Segmenting Mall Customers Data to Improve Business into Higher Target using K-Means Clustering," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, Dec. 2021, pp. 1602-1604.