Customer Segmentation Using Hierarchical Clustering

Conference Paper · April 2024

DOI: 10.1109/I2CT61223.2024.10543349

10 authors, including:

Areeba Afzal
Information Technology University
1 PUBLICATION **11** CITATIONS
SEE PROFILE

Laiba Khan
Information Technology University
1 PUBLICATION **11** CITATIONS
SEE PROFILE

Muhammad Zunnurain Hussain
Universiti Putra Malaysia
107 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Muhammad Zulkifl Hasan
University of Central Punjab
96 PUBLICATIONS **156** CITATIONS
SEE PROFILE

2024 IEEE 9th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 5-7, 2024

Customer Segmentation Using Hierarchical Clustering

Areeba Afzal[1], Laiba Khan[2], Muhammad Zunnurain Hussain[3], Muhammad Zulkifl Hasan[4], Muzzamil Mustafa[5], Aqsa Khalid[6], Rimsha Awan[7], Farhan Ashraf[8], Zohaib Ahmed Khan[9], Arslan Javaid[10]
[1,2]Department of Computer Engineering Information Technology University Lahore, Punjab, Pakistan
[3]Assistant Professor, Dept. of Computer Science, Bahria University Lahore Campus
[4]Department of Computer Science, Faculty of Information Technology, University of Central Punjab Lahore Pakistan
[5]Department of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan
[6]Information Technology University Lahore, Pakistan
[7]Department of Computer Science National College of Business Administration and Economics, Lahore, Pakistan
[8]Dept. of Computer Science, Bahria University Lahore Campus
[9]Department of Computer Science National College of Business Administration & Economics, Lahore, Pakistan
[10]Department of Computer Science National College of Business Administration and Economics, Lahore, Pakistan
[1]bsce20017@itu.edu.pk, [2]bsce20035@itu.edu.pk, [3]Zunnurain.bulc@bahria.edu.pk, [4]Zulkifl.hasan@ucp.edu.pk,
[5]muzzamil.mustafa@umt.edu.pk , [6] msds19046@itu.edu.pk, [7] rimshaawan.225@gmail.com , [8]farhanashrafali30@gmail.com, [9] zohaibkhanmcitp@gmail.com, [10]Arslanravian97@gmail.com

2024 IEEE 9th International Conference for Convergence in Technology (I2CT) | 979-8-3503-9447-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/I2CT61223.2024.10543349

***Abstract*— In the dynamic landscape of retail, understanding customer behavior is paramount for businesses seeking to optimize marketing strategies and enhance the shopping experience. This study explores the utilization of hierarchical clustering techniques for mall customer segmentation, with a focus on the paper titled 'MALL CUSTOMER SEGMENTATION USING MACHINE LEARNING' as the benchmark. Our dataset encompasses a diverse range of mall customers, spanning demographics and behavioral attributes. Hierarchical clustering systematically groups customers into clusters, revealing distinct segments within the mall's customer base. A comprehensive analysis of these clusters unveils profound insights into customer tendencies, preferences, and purchasing habits. These insights form a solid foundation for tailored marketing campaigns, personalized recommendations, and resource allocation within the mall. The study contributes significantly to customer analytics, providing retailers with a powerful tool to gain a competitive edge in the retail sector. By leveraging hierarchical clustering for mall customer segmentation, businesses can enhance customer satisfaction, drive sales, and foster lasting customer relationships. This paper underscores the importance of data-driven method- ologies in understanding customer behavior and offers a practical framework for businesses to harness the potential of hierarchical clustering for strategic decision-making in the retail industry.**

Keywords: Hierarchical Clustering, Customer Segmentation, Data Mining Techniques, Market Segmentation, Cluster Analysis, Agglomerative Clustering Dendrogram, Machine Learning in Marketing, Consumer Behavior Analysis, Multivariate Data Analysis, Customer Data Clustering, Behavioral Segmentation, Marketing Strategy, Customer Profiling, Predictive Analytics in CRM

I. INTRODUCTION

Understanding customer behavior is a fundamental pursuit for businesses in the dynamic retail sector. The ability to dissect and categorize customers into distinct groups, known as customer segmentation, plays a pivotal role in shaping marketing strategies and enhancing shopping experiences. This research project ventures into the application of hierarchical clustering techniques within the context of mall customers, a diverse and multifaceted consumer demographic. Malls attract a wide range of individuals, each with their unique prefer- ences, demographics, and behaviors. Through the utilization of hierarchical clustering, an advanced data-driven approach, this study aims to unveil valuable insights that can revolutionize how businesses interact with their mall customers.

The essence of this research lies in its potential to uncover hidden patterns and preferences governing customer choices within the mall environment. As businesses grapple with fierce competition in the retail arena, the ability to segment customers effectively and tailor marketing strategies to specific segments holds the promise of a substantial advantage. This paper not only introduces the concept of hierarchical clustering as a potent tool for customer segmentation but also explores its adaptability to diverse mall customer datasets.

In the pages that follow, we delve into the methodology, results, and implications of applying hierarchical clustering to mall customer data. By doing so, we aim to provide businesses in the retail industry with a practical resource to enhance customer satisfaction, drive sales, and foster enduring customer relationships in an ever-evolving retail landscape.

II. BACKGROUND AND RELATED WORK

A. Background

Understanding customer behavior through segmentation is a fundamental strategy across various industries. This process involves categorizing customers into distinct groups based on shared characteristics, enabling businesses to tailor their approaches to meet specific needs and preferences. Malls and shopping centers serve as vibrant hubs attracting a diverse range of customers, each with their unique demographics, shopping habits, and preferences. Yet, the application of hierarchical clustering to segment mall customers remains a relatively unexplored area. This study aims to address this gap by evaluating the suitability and effectiveness of hierarchical clustering in segmenting mall customers. The insights gained will have implications not only for retail but also for various sectors seeking to enhance their understanding of diverse customer bases and optimize their strategies accordingly.

B. Related Work

Customer segmentation and clustering techniques have been extensively studied across various domains, providing

valuable insights into understanding customer behavior and enhancing business strategies. In the context of customer segmentation:

Traditional Segmentation Methods: Traditional methods, including demographic, geographic, and psychographic segmentation, have long been used in marketing. These approaches categorize customers based on factors like age, location, in- come, and lifestyle. While informative, they may oversimplify the complex nature of customer behavior.

K-Means Clustering: K-means clustering is a widely employed technique that groups data points into clusters based on similarity. In the field of customer segmentation, K-means has been used to identify distinct customer groups. However, it assumes spherical clusters and requires specifying the number of clusters beforehand.

DBSCAN and Density-Based Clustering: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identifies clusters based on data density. It is particularly useful for discovering irregularly shaped clusters. In customer segmentation, DBSCAN has been applied to uncover less-defined customer groups.

Hierarchical Clustering in Customer Segmentation: Hierarchical clustering, unlike K-means and DBSCAN, creates a hierarchical structure of clusters, offering a visual representation of the relationships between clusters. Its adaptability to various data shapes and its ability to capture hierarchical relationships make it a promising approach for customer segmentation.

Customer Segmentation in Retail: Numerous studies have explored customer segmentation within the retail sector, focusing on different clustering techniques and data sources. However, the specific application of hierarchical clustering to mall customer data remains an underexplored avenue.

Machine Learning and Predictive Analytics: Beyond clustering, machine learning techniques, such as decision trees, random forests, and neural networks, have been applied to predict customer behavior and preferences. These methods offer predictive capabilities to inform targeted marketing and personalized recommendations.

III. METHODOLOGY

This study advances customer segmentation from k-means to hierarchical clustering. The dataset, sourced from ”Mall__Customers.csv,” underwent preliminary exploration. Univariate clustering was initiated based on ’Annual Income (k)’ using Agglomerative Clustering with three clusters. A dendrogram visually captured the hierarchical structure of in- come segments. Building on this, bivariate clustering incorporated both ’Annual Income (k)’ and ’Spending Score (1-100)’ with five clusters. The Agglomerative Clustering algorithm provided insights into customer segments considering both income and spending. The resulting hierarchical relationships were depicted through a dendrogram.

Multivariate clustering ensued, integrating ’Age,’ ’Annual Income (k),’ ’Spending Score (1-100),’ and ’Genre_Male.’ Categorical features were numerically transformed, and data standardized with StandardScaler. Hierarchical clustering with five clusters and a dendrogram visualization deepened the understanding of complex relationships.

Post-clustering, a detailed analysis of each cluster’s characteristics was conducted. A scatter plot visually portrayed the distribution of clusters in a two-dimensional space, plotting

’Age’ against ’Annual Income (k).’ The derived cluster labels were appended to the original dataset, saved as ’cluster.csv,’ along with additional details like the maximum customer ID.

In summary, this methodology signifies a transition to hier- archical clustering for more nuanced customer segmentation, capturing intricate hierarchical relationships within the data.

IV. EASE OF USE

Here are the key considerations to ensure accessibility and simplicity:

Data Collection and Preprocessing: Streamline data collection procedures to minimize errors and inconsistencies. Develop clear guidelines for data preprocessing, including handling missing values and outliers, and provide automated tools or scripts for data cleaning.

Hierarchical Clustering Tools: Utilize user-friendly software or programming libraries for hierarchical clustering, offering an intuitive interface for researchers or practitioners to apply clustering algorithms without requiring extensive coding ex- pertise.

Visualization: Implement visualization tools that generate dendrograms and cluster visualizations in a straightforward manner. These visual representations should be interpretable for non-technical stakeholders.

Parameter Tuning: Simplify the process of parameter tuning for hierarchical clustering algorithms. Provide clear guidance on choosing appropriate distance metrics and linkage methods, along with automated tools to assist in selection.

Cluster Interpretation: Develop user-friendly dashboards or reports that enable easy interpretation of cluster results. Summarize the characteristics and behaviors of each cluster in a comprehensible format.

Validation Metrics: Include automated validation metrics within the clustering process to assist users in assessing cluster quality. Provide explanations and guidelines on how to interpret these metrics.

Documentation: Offer comprehensive documentation that covers the entire research process, from data collection to interpretation of results. Include step-by-step tutorials, code samples, and examples for reference.

User Support: Establish a support system to assist users with questions or issues they may encounter during the clustering process. Provide contact information for assistance when needed.

Training and Workshops: Consider organizing training sessions or workshops to educate users on the use of hierarchical clustering for mall customer segmentation. These sessions can help bridge knowledge gaps and enhance usability.

Feedback Loop: Encourage users to provide feedback on the tools and processes used for hierarchical clustering.

Regularly update and improve the resources based on user input to enhance ease of use continually.

V. SYSTEM DIAGRAM AND FLOWCHART

Here Fig.1 demonstrates the Flowchart. This flow represents the process of customer segmentation using various Hierarchical Clustering methods. It starts by preparing the data and then employs different clustering techniques (univariate, bivariate, and multivariate) to group customers based on different features. The flow then visualizes these clusters and evaluates their quality using the Silhouette Score. Finally, it presents various visual representations to understand and analyze the segmented clusters before concluding the segmentation pro- cess. The system architecture Fig.2 starts with user interaction through the User Interface, followed by data preprocessing for cleaning and structuring data. The chosen distance metric in the Distance Metric Selector influences the Linkage Matrix generation utilized by the Hierarchical Clustering Engine to form clusters. Evaluation, visualization, labeling, and the final output, then summarize and present the clustered data for user interpretation and further analysis.

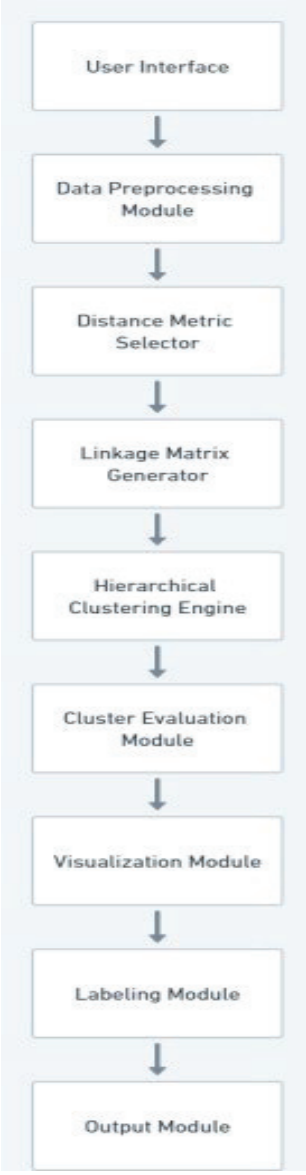


Fig. 1. Flowchart

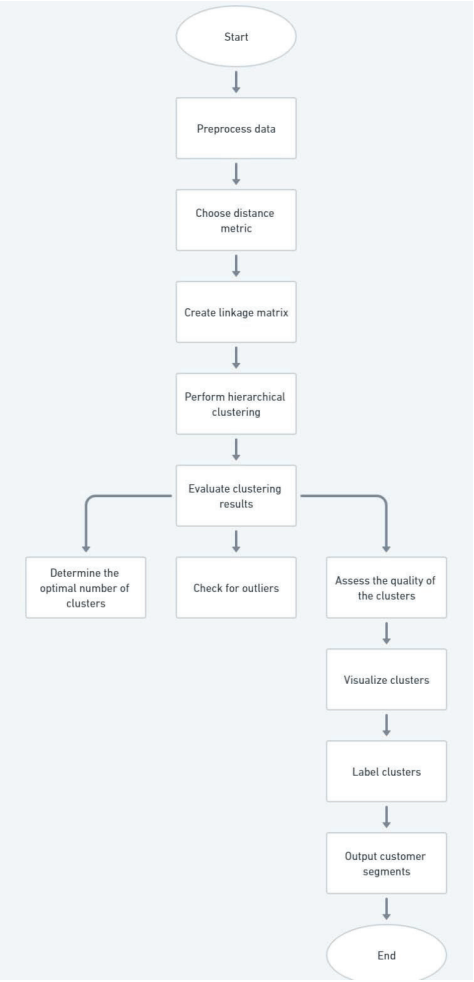


Fig. 2. System Architecture

VI. RESULTS AND EVALUATION

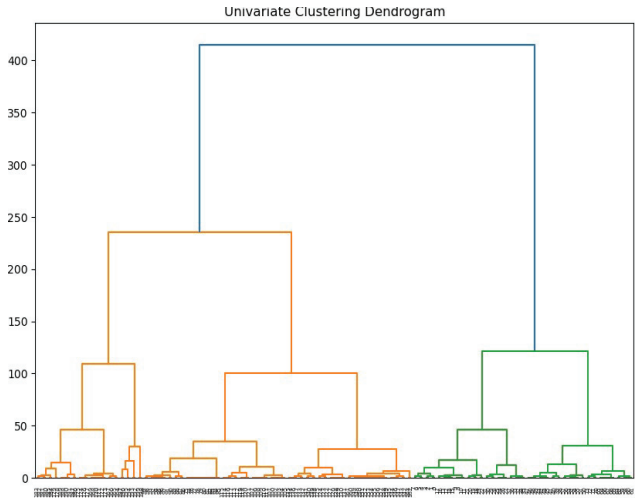
In this section, we present the results and evaluations derived from our hierarchical clustering approach applied to customer segmentation. Our analysis progresses through univariate, bivariate, and multivariate clustering, providing a comprehensive understanding of customer behavior within the dataset.

A. Data Loading and Exploration

The initial stage of our investigation involved loading the customer data from the file ”Mall Customers.cs” into a Pandas DataFrame. This facilitated a comprehensive exploration of the dataset’s structure and attributes. The dataset encompasses essential information such as age, annual income, spending score, and gender. By loading the data into a DataFrame, we laid the foundation for subsequent analyses aimed at unrav- eling patterns and relationships among customer features. The exploration primarily involved univariate and multivariate clus- tering analyses, employing hierarchical clustering methods.

B. Univariate Clustering - Annual Income

Univariate clustering focused on customer behavior exclusively tied to annual income, utilizing the Agglomerative Clustering algorithm with three clusters. This step aimed to distill patterns in income distribution, offering a foundational grasp of income-based customer segments. The primary objective was to identify distinct groups with similar income levels.



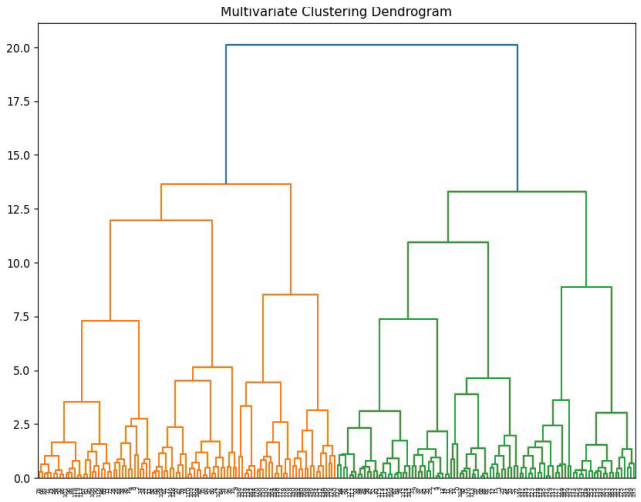


Fig. 3. . Univariate Clustering Dendogram

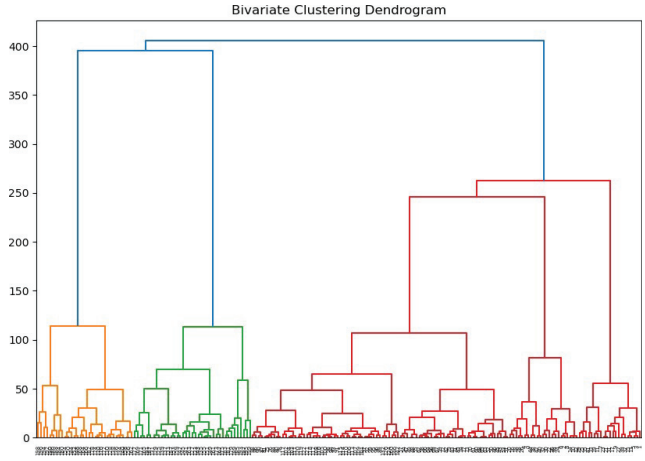


Fig. 4. Bivariate Clustering Dendogram

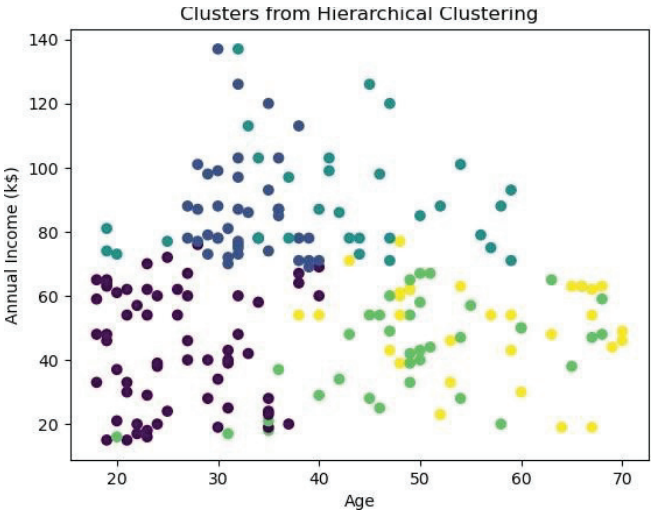


Fig. 5. Multivariate Clustering Dendogram

Visualizing Univariate Clustering:

A dendrogram is generated to visually represent the hierarchical relationships among individuals based on their annual income. Utilizing the 'ward' linkage method, this visualization provides a clear depiction of the hierarchical clustering structure, aiding in the interpretation of income segments (see Fig.3).

C. *Bivariate Clustering - Income and Spending Score*

Building upon univariate clustering, bivariate clustering is performed by considering both 'Annual Income (k)' and

'Spending Score (1-100).' The Agglomerative Clustering algorithm with five clusters is employed to identify more nuanced patterns by incorporating spending behavior along with income.

Visualizing Bivariate Clustering:

Similar to the univariate clustering, a dendrogram is created to visually interpret the hierarchical relationships between income and spending score clusters. This visualization provides insights into how customers cluster based on both income and spending behavior (see Fig.4).

D. *Multivariate Clustering*

A more comprehensive approach is taken by considering multiple features 'Age,' 'Annual Income (k),' 'Spending Score (1-100),' and 'Genre_Male.' These features are pre-processed, scaled, and used in the Agglomerative Clustering algorithm with five clusters, aiming to capture complex relationships among multiple dimensions.

Visualizing Multivariate Clustering:

The hierarchical structure resulting from multivariate clustering is visually represented through another dendrogram as shown in Fig.5. This visualization aids in understanding the complex relationships among the selected features, offering a holistic view of customer segmentation based on various dimensions (see Fig.5).

E. *Analyzing and Visualizing Cluster from Results*

A count is conducted to determine the number of individuals in each cluster resulting from the multivariate analysis. Subsequently, a scatter plot is generated, plotting 'Age' against

'Annual Income (k),' providing a two-dimensional visualization of the clusters. This visual representation allows for an intuitive interpretation of the segmentation results as in Fig.6.

F. Dataset Augmentation and Export

The cluster labels derived from the multivariate analysis were appended to the original dataset, resulting in a new dataset named 'cluster.csv.'Additional details, such as the maximum customer ID, were also extracted and documented.

G. Silhouette Score for Multivariate Clustering Evaluation

To assess the efficacy of our multivariate clustering, we em ployed the Silhouette Score, a metric gauging the compactness and separation of clusters. The calculated Silhouette Score for the multivariate clustering yielded a value for multivariate clustering: 0.28699413201651747.

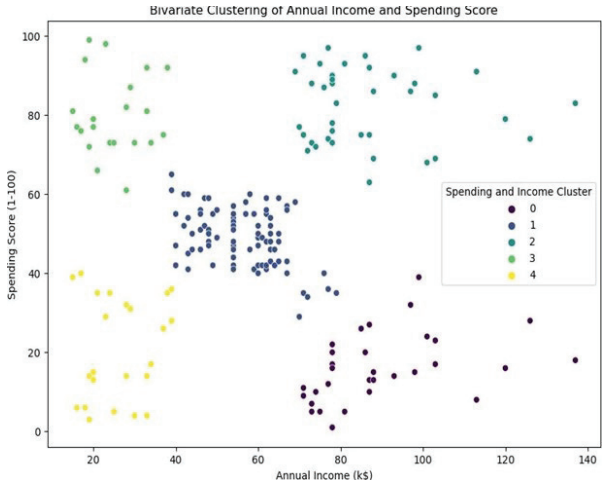


Fig. 6. Clusters from Hierarchical Clustering

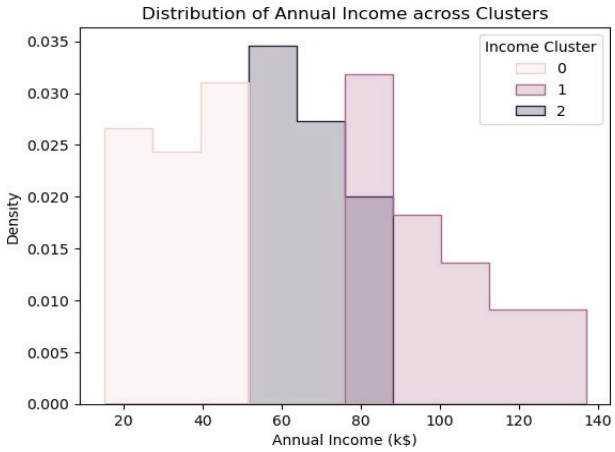


Fig. 7. Univariate Clustering

H. Histogram and Scatterplot Visualisations

To enhance the interpretability of our results, univariate and bivariate visualizations were generated. The histogram illustrates the distribution of annual income across clusters derived from univariate clustering [see Figure 6]. Additionally, the scatter plot visualizes the bivariate clustering of annual income and spending score, providing a clearer representation of the segmented clusters (see Fig.7 and Fig.8).

I. Principal Component Analysis (PCA)

In order to visualize the multivariate clustering in a reduced two-dimensional space, Principal Component Analysis (PCA) was applied. The 2D PCA plot visually represents the clustering patterns, offering insights into the relationships among customers in the context of the selected features (see Fig.9).

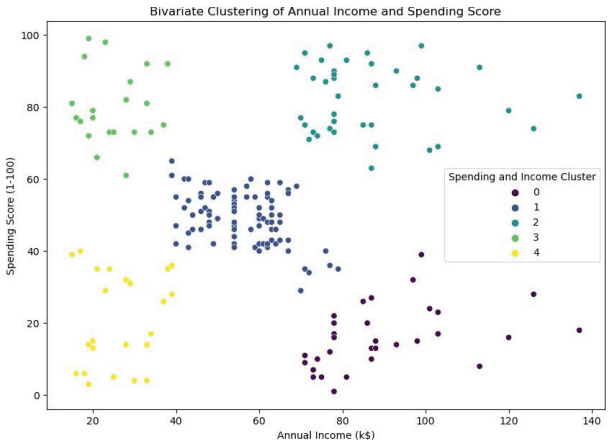


Fig. 8. Bivariate Clustering

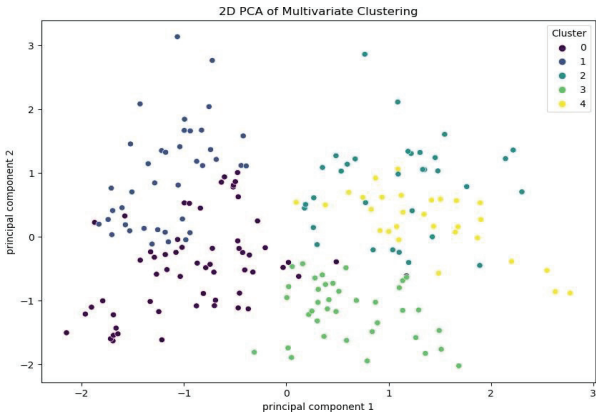


Fig. 9. 2D PCA

VII. CONCLUSION AND FUTURE WORK

In conclusion, the application of hierarchical clustering for customer segmentation, as demonstrated in this study and benchmarked against 'MALL CUSTOMER SEGMENTATION USING MACHINE LEARNING,' has provided valueable insights into diverse customer behaviours. The hierarchical approach, spanning univariate, bivariate, and multivariate analyses, has proven effective in capturing hierarchical structures and revealing relationships among distinct customer segments. By dissecting customer attributes such as annual income, spending score, age, and gender, the hierarchical model, supported by visualizations like dendrograms and scatter plots, has facilitated a nuanced understanding of the data. The incorporation of the Silhouette Score, in conjunction with visualizations, has further enriched our understanding of the identified clusters. These segmentation outcomes directly inform the refinement of marketing strategies and the tailoring of services to meet the specific needs of diverse customer groups, ultimately enhancing satisfaction and loyalty.

Looking ahead, potential areas for future exploration include optimizing feature selection and engineering, exploring alternative algorithms, and incorporating temporal elements, ensuring continual responsiveness to the dynamic landscape of customer preferences and behaviours in the evolving retail industry.

REFERENCES

[1] Chongkolnee Rungruang, Pakwan Riyapan, Arthit Intarasit, Khan- chit Chuarkham, Jirapond Muangprathub,RFM model customer segmentation based on hierarchical approach using CA,Expert Sys- tems with Applications,Volume 237, Part B,2024,121449,ISSN 0957-174,https://doi.org/10.1016/j.eswa.2023.121449.

[2] Sukanlaya Sawang, Chia-Chi Lee, Cindy Yunhsin Chou, Nanjangud Vishwanath Vighnesh, Deepak Chandrashekar, Understanding post-pandemic market segmentation through perceived risk, behavioural intention, and emotional wellbeing of consumers, Journal of Re- tailing and Consumer Services, Volume 75,2023,103482,ISSN 0969-6989,https://doi.org/10.1016/j.jretconser.2023.103482.

[3] Hidenori Komatsu, Osamu Kimura,Customer segmentation based on smart meter data analytics: Behavioral similarities with manual categorization for building types,Energy and Buildings, Volume 283,2023,112831,ISSN 0378-7788,https://doi.org/10.1016/j.enbuild.2023.112831.

[4] D. Teslenko, A. Sorokina, K. Smelyakov and O. Filipov, "Comparative Analysis of the Applicability of Five Clustering Algorithms for Market Segmentation," 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2023, pp. 1-6, doi: 10.1109/eStream59056.2023.10134796.

[5] Luo, L., Li, B., Fan, X. et al. Dynamic customer segmentation via hierar- chical fragmentation-coagulation processes. Mach Learn 112, 281–310 (2023). https://doi.org/10.1007/s10994-022-06276-8

[6] S. Jeena, A. Chaudhary and A. Thakur, "Implementation & Anal- ysis of Online Retail Dataset Using Clustering Algorithms," 2023 4th International Conference on Intelligent Engineering and Man- agement (ICIEM), London, United Kingdom, 2023, pp. 16, doi: 10.1109/ICIEM59379.2023.10166552.

[7] Jie Yu and Xikun Zhang. 2023. Research on Online Learning User Classification Based on Hierarchical Clustering. In Proceedings of the 2023 6th International Conference on Big Data and Education (ICBDE '23). Association for Computing Machinery, New York, NY, USA, 100–105. https://doi.org/10.1145/3608218.3608222

[8] Phan Duy Hung, Nguyen Thi Thuy Lien, and Nguyen Duc Ngoc. 2019. Customer Segmentation Using Hierarchical Agglomerative Clustering. In Proceedings of the 2nd International Conference on Information Science and Systems (ICISS '19). Association for Computing Machinery, New York, NY, USA, 33–37. https://doi.org/10.1145/3322645.3322677