Customer Segmentation Using K-Means Clustering

Shivam K Gupta
Computer and Information Science
University of Wisconsin Parkside
Kenosha, United States of America
gupta014@rangers.uwp.edu

Harjinder
Computer and Information Science
University of Wisconsin Parkside
Kenosha, United States of America
singh147@rangers.uwp.edu

***Abstract*—It is essential to comprehend customer needs in the ever-changing world of modern company. Using common characteristics, customer segmentation—a crucial strategic tool—classifies a wide range of clients. One popular segmentation technique is the K-means clustering algorithm. It effectively divides datasets into k clusters for customized services and marketing. The ubiquity of K-means is partly due to its scalability and easy interpretation. In order to better grasp the intricacies of K-means consumer segmentation in customer service, this review gathers and assesses literature on the subject in an effort to spot trends and obstacles.**

***Keywords*—K Means, EDA, Scatter plot, Machine Learning, Frequent itemsets, Customer Segmentation, Elbow method**

I. Introduction

It is critical to comprehend and respond to consumers' varied wants and behaviors in the ever-changing world of modern business. client segmentation has become a critical strategic strategy that involves dividing a diverse client base into discrete groups according to common traits, requirements, and behaviors. Businesses can customize their pricing, marketing, and service offerings to each consumer segment's specific needs thanks to this segmentation. Of all the methods used for consumer segmentation, the K-means clustering algorithm is one of the most popular and effective. K-means, a well-liked unsupervised clustering method, divides a dataset into k clusters by reducing the variation within each cluster. Large datasets can be efficiently segmented using this iterative approach, which uses Euclidean distance to allocate data points to the closest cluster center. Due to its scalability, ease of use, and intuitive interpretability, customer segmentation has become increasingly popular across a range of businesses. The growing significance of K-means consumer segmentation necessitates a rigorous examination of the literature in order to compile and evaluate the available data. The goal of this review is to present a thorough summary of the state of the art when it comes to using K-means clustering for consumer segmentation. Through the process of synthesizing findings from many studies, our aim is to identify significant trends, obstacles, and prospects within the subject. We hope that this methodical investigation will advance knowledge of the subtleties of using K-means for customer service.

II. Literature Review

Customer segmentation involves dividing a customer base into distinct groups that share similar characteristics, behaviors and needs to enable personalized marketing, pricing, promotions and experiences [4]. Segmentation provides the foundation for tailoring value propositions, positioning, and service delivery to maximize customer lifetime value. K-means is one of the most prevalent algorithms used for segmentation due to its scalability, simplicity and ability to efficiently handle large transaction datasets [6]. K-means is an unsupervised clustering technique that partitions observations into k clusters by minimizing within-cluster variation [7]. It iteratively reassigns data points to their nearest cluster center based on Euclidean distance. Key advantages are fast clustering of large-scale data and intuitive interpretability of resulting segments. Limitations include needing to specify k a priori and sensitivity to outliers. The study by Tavor, Gonen and Spiegel (2023) explores using barcode scanners to enhance customer segmentation and pricing strategies. It focuses on identifying Loyal versus Deal-Prone segments to optimize timing and discounts when switching prices. By leveraging scanner data, customer behaviors are tracked to generate reports. Findings reveal segmented marketing increases profits from loyal customers, although caution is needed as deal-prone customers may purchase less when segmented. This emphasizes the practical application of technology-driven segmentation. A universal challenge is managing the quantity, quality, completeness and relevance of data [10]. Real-world transaction data requires substantial preprocessing to address issues like missing values, noise and integrity errors [5]. Feature selection is critical for extracting attributes most relevant for clustering. RFM analysis is commonly used to reduce data into recency, frequency and monetary scores [3]. For k-means, determining the optimal number of clusters k is an open challenge. Elbow methods, silhouette analysis, Calinski-Harabasz and Davies-Bouldin indices are used for selection, but qualitative evaluation is also needed [1]. Advanced methods like X-means use Bayesian Information Criterion to adaptively split and merge clusters. Metaheuristics like particle swarm optimization optimize k and initial centroid positions [2]. The Customer Personality Segmentation Using K-Means Clustering study [5] identifies customer insights related to income, preferences, behavior and engagement. It demonstrates segmentation potential using k-means, elbow method and PCA. Recommendations are made to tailor marketing like expanding product selection, targeting advertising and offering incentives. This shows the value of personality segmentation with unsupervised learning to

improve customer experience. The study by Turkmen [9] evaluated k-means, hierarchical, DBSCAN and RFM clustering for customer segmentation. K-means provided the best insights into customer segments, while RFM identified high-value customers. This demonstrates comparing algorithms to choose the right technique based on context. The study by Chiu et al. [2] proposed an intelligent market segmentation system using k-means combined with self-organizing maps and particle swarm optimization. Implemented in a real company, it provided precise segmentation to inform marketing strategy. This shows the value of enhancing k-means with metaheuristics like PSO. Enhancements have been proposed to improve k-means performance. K-means++ optimizes initial cluster center placement based on probability proportional to squared distance [5]. Density-based Gini impurity splitting increases separation between clusters [5]. Weighting features or integrating principal components helps handle noisy, correlated inputs [9]. Alternative distance measures like cosine similarity have been evaluated. The study by [5] built a TFA model for customer transaction features and improved k-means using density-based initialization. Applied to synthetic and real e-commerce data, it shows enhancing k-means through advanced center initialization techniques.

III. Design

The pre-processing of data involves data preparation where unwanted features were removed from the dataset and only the important features are taken into consideration for our analysis. The dataset is trained to obtain descriptive statistical analysis and visualization analysis as shown in Figure 1.

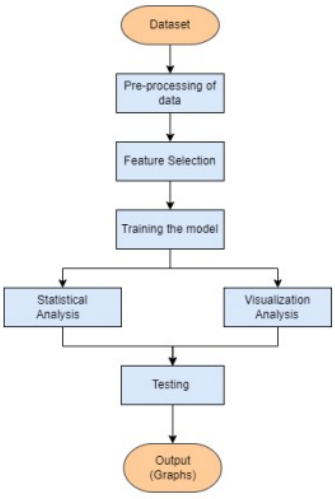

Fig. 1. **Design and Architecture of the System**

IV. Implementation

The descriptive statistical analysis and the data visualization are computed using Python language in Google Colab and also Power BI. Below listed are open source libraries used for exploratory data analysis:

(1) Pandas: Pandas is a software library for data analysis and manipulation created in the Python programming language. It contains data structures and techniques designed for interacting with time series and numerical tables.
(2) Matplotlib: Python's Matplotlib toolkit is a comprehensive tool for building static, animated, and interactive visualizations. Matplotlib is also used to create 2D graphs and plots.
(3) Seaborn: Seaborn is a matplotlib-based data visualization library. It offers a high-level interface for creating visually attractive and informative statistical visuals, is effective for visualizing random distribution.

V. Methodology

In this research paper, we have used exploratory data analysis and clustering algorithms to gain insights into customer segments based on demographic and behavioral data. Our review follows a systematic approach, including the identification of relevant databases, search terms, inclusion and exclusion criteria, and the screening process. We focused on studies published in the last five years to ensure the relevance of the literature.

The dataset comprises 50,000 entries and includes information about customers across various attributes:

ID	**Unique identification number assigned to each customer.**
Age	**Represents the age of each customer.**
Gender	**Specifies the gender of the customers, categorized as male or female.**
Income	**Denotes the annual income of the customers.**
Spending Score	**Indicates a numerical score (0 to 100) reflecting a customer's spending behavior.**

First, we conduct exploratory data analysis on the dataset including summary statistics, visualizations, and checking for outliers. Summary statistics using DataFrame.describe() and DataFrame.info() provide details on attributes like mean, percentiles, and data types. Visualizations like count plots, scatter plots, and KDE plots using Matplotlib and Seaborn give a graphical view of feature distributions and relationships. Box plots identify potential outliers in features like age, income, and spending score.

Next, we leverage clustering algorithms - K-Means and Hierarchical clustering - to find groups of similar customers in the multidimensional feature space. We use the Elbow method to determine the optimal number of clusters for K-Means. Silhouette scores are calculated to evaluate clustering performance. K-Means is applied on different feature pairs like annual income/spending score, age/spending score, and age/income to reveal customer segments. Hierarchical clustering with Ward linkage is also utilized as an alternative approach.

The identified clusters are further analyzed by aggregating the dataset and plotting cluster count plots for features like age, income, and spending score. Additionally, the most frequently purchased item is determined within each cluster-age group to infer shopping habits.

Finally, interactive 3D visualizations using Plotly express provide a means to visually distinguish and understand the formed clusters. The clustering labels are supplemented back into the dataset for further analysis.

A. *Data Acquisition and Preprocessing*

The study utilized an online retail dataset obtained through the upload functionality on Google Colab. The dataset underwent initial exploration and inspection to understand its structure, assess missing values, and comprehend the types of attributes available.

CustomerID	Age	Gender	AnnualIncome	SpendingScore	Product	ProductPrice
1800	33	Male	63077	759	Sofa	367.980000
1030	21	Male	68529	9753	Smartphone	409.540000
1342	32	Female	118503	10539	Shoes	952.160000
1145	52	Female	64095	2862	Laptop	823.950000
1880	57	Female	93691	7321	Lamp	852.070000

Fig. 2. **Design and Architecture of the System**

B. *Exploratory Data Analysis (EDA)*

A comprehensive Exploratory Data Analysis (EDA) was conducted using Python's Pandas, Matplotlib, Seaborn, and Plotly libraries. This stage encompassed data visualization techniques such as count plots, scatter plots, KDE plots, and boxplots. Visual representations were generated to comprehend the distribution, relationships, and outliers within the dataset.

C. *Outlier Detection:*

Outlier detection was performed using boxplots to identify potential anomalies in the data. Outliers, if present, were investigated to determine their impact on subsequent analyses.

D. *KDE Plot Function:*

A custom Python function , "kde-plot", was defined to genrate KDE plots using the Seaborn library.This function creates a FacetGrid and maps KDE plots for each specified feature: 'Age', 'Annual Income', and 'Spending Score'. The aspect parameter was set to 4 to ensure an appropriate aspect ratio for the plots.

VI. Experimental Results

This research undertook a detailed exploration of the dataset, employing visualizations as a primary tool to understand the relationships and distributions within the data.

A. *Visualization of 'Annual Income' vs. 'Spending Score':*

A scatter plot was generated to visualize the correlation between 'Annual Income' and 'Spending Score'. The x-axis represented the annual income, while the y-axis depicted the spending score. This plot aimed to uncover any discernible patterns or clusters that might exist between these two crucial parameter

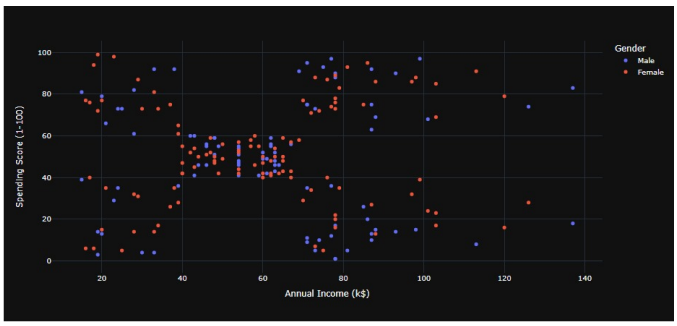

Fig. 3. **'Annual Income' vs. 'Spending Score'**

B. *Visualization of 'Age' vs. 'Spending Score':*

A customized scatter plot function was employed to visualize 'Annual Income' vs. 'Spending Score' while incorporating gender differentiation. The use of distinct colors facilitated a gender-wise representation within the scatter plot, potentially revealing any gender-specific trends or disparities in spending behavior relative to income. Moreover, scatter plots were created to scrutinize the relationship between 'Age' and 'Spending Score' as well as 'Age' and 'Annual Income'. These visualizations were instrumental in understanding how age influences spending behavior and income levels. Interpreting these scatter plots revealed potential trends, clusters, or anomalies, shedding light on possible relationships between income, spending behavior, and age demographics. These visualizations paved the way for subsequent analyses and aided in the formulation of further research inquiries.

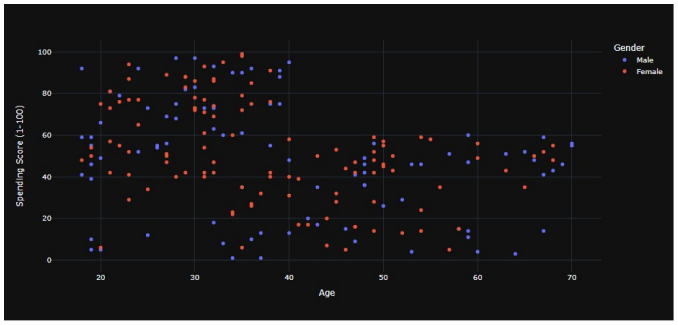

Fig. 4. **'Age' vs. 'Spending Score'**

C. *Heatmap:*

A heat map is a two-dimensional representation of the correlation (measure of dependence) between the various variables, which are represented by different colors. The degree

of association is indicated by the changing color intensity. A measure of the linear relationship between two variables is correlation. Each square displays the correlation between the elements on each axis, the correlation exists in the range of -1 to +1. There is no linear trend between the two variables if the values are closer to zero. The closer the correlation is to 1, the more positively associated they are; that is, as one rises, the other does as well, and the stronger this relationship is, the closer to 1 the correlation is, similar results can be obtained with a correlation that is closer to -1, but instead of both variables rising, one will fall as the other does. Figure **??** shows the correlation among all the different attributes.

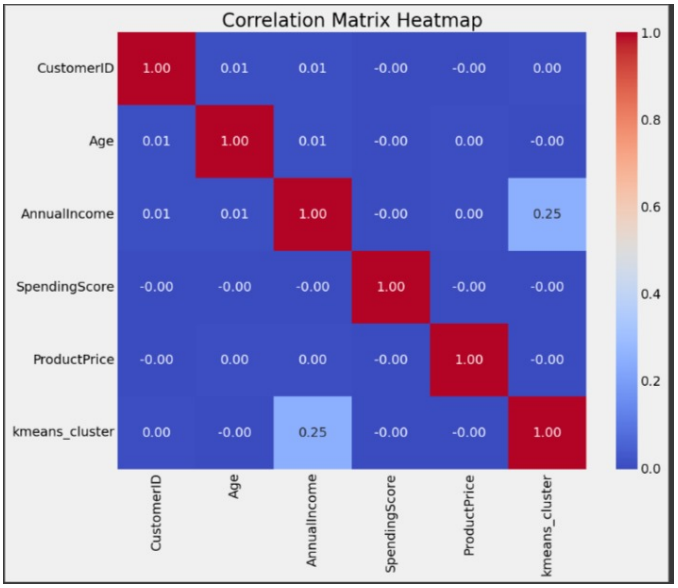

Fig. 5. **'Age' vs. 'Spending Score'**

D. *Elbow method:*

1) ***Objective:****:* The goal of this section is to segment customers based on their 'Annual Income' and 'Spending Score' to identify distinct groups that exhibit similar spending behaviors. The dataset was initially imported, focusing on two key columns: 'Annual Income (k)$'and'SpendingScore(1-100)'$.

2) ***Elbow Method for Determining Clusters:****:* To ascertain the optimal number of clusters for segmentation, the Elbow Method was employed. This iterative process involved fitting K-means models with varying numbers of clusters (ranging from 1 to 10 in this case) and computing the Within-Cluster Sum of Squares (WCSS). The WCSS represents the sum of squared distances of data points to their assigned cluster centroids. The resultant plot ('The Elbow Method') depicted the relationship between the number of clusters and the WCSS. The "elbow point," where the WCSS starts to decrease at a slower rate, was considered the optimal number of clusters.

3) ***K-means Clustering:****:* Based on the Elbow Method's outcome, five clusters were identified as the optimal number. A K-means model with five clusters

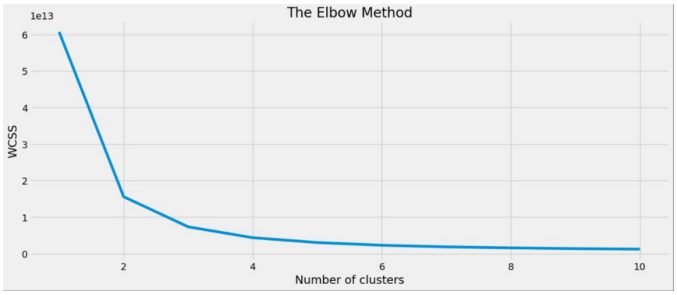

Fig. 6. **'Number of clusters' VS 'WCSS'**

E. *Segmentation and Visualization using K-means Clustering:*

This section of the research paper focused on segmenting customers based on three critical attributes—'Age', 'Annual Income', and 'Spending Score'—using the K-means clustering technique.

1) ***3D Scatterplot Visualization:****:* The clustered segments were visualized in a 3D scatterplot using Plotly. The plot displayed 'Annual Income' on the x-axis, 'Spending Score' on the y-axis, and 'Age' on the z-axis, with distinct colors denoting each of the six identified clusters. The size of each point in the plot represented its respective cluster label, aiding in visual differentiation.

The visualization facilitated a comprehensive understanding of the dataset's segmentation into six distinct clusters based on the specified attributes. The plot allowed for the identification of spatial patterns and relationships between the customer segments, providing valuable insights into the distribution and characteristics of each cluster.

2) ***K-means Clustering and Purchasing Patterns Analysis****:* The K-means clustering technique was employed to segment customers into six distinct clusters based on 'Age', 'Annual Income', and 'Spending Score'. Each individual was assigned to a particular cluster reflecting similarities in their attributes.

3) ***Most Purchased Item Analysis:****:* Utilizing the clustered data, an analysis was conducted to determine the most frequently purchased item within specific age groups across each cluster. This analysis aimed to uncover potential buying preferences or trends based on age and clustered segments.

4) ***Insights Derived:****:* Cluster-Specific Preferences: Each cluster exhibited unique purchasing patterns corresponding to different age groups.

Cluster 0: Showed a prevalence of 'TV' purchases among customers aged 21.

Cluster 1: Customers at the age of 35 displayed higher purchases of 'Shoes'.

Cluster 2: Demonstrated a tendency for 'Smartphone' purchases by customers aged 48.

Age-Dependent Preferences within Clusters: Notably, within each cluster, specific age groups displayed consistent preferences for particular items. For example:

Within Cluster 3: Customers aged 58 favored 'Laptop' purchases, while those at 60 exhibited a higher affinity for

'AC'.

In Cluster 4: Customers aged 67 were inclined towards purchasing 'TV', whereas individuals aged 69 showed a preference for 'Smartphone'.

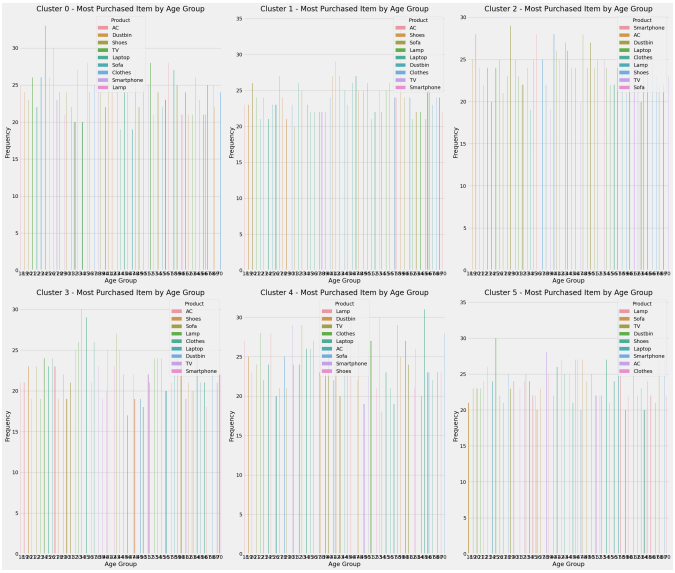

Fig. 7. **'Most purchased items by different Age groups'**

CONCLUSION

The exploration of customer segmentation through K-means clustering presents rich insights into distinct customer behaviors and preferences. Analysis based on attributes such as Annual Income and Spending Score revealed clear clusters, illuminating diverse spending patterns among customers. The application of the Elbow Method to determine optimal clusters demonstrated the existence of five discernible segments, aiding in the visual understanding of spending behaviors influenced by income disparities. Moreover, diverse customer groups emerged from clustering analysis involving various attributes like Age and Spending Score or Age and Annual Income. Each cluster showcased unique spending habits and income distributions, offering invaluable insights for targeted marketing endeavors. Identifying frequently purchased items within specific age groups across clusters further enriched the understanding of customer preferences, laying the groundwork for personalized marketing strategies. Additionally, the utilization of hierarchical clustering provided alternative perspectives on segmentation, allowing for a deeper understanding of customer behaviors. The observations highlighted the influence of financial attributes on customer segmentation and the significant variance in spending habits across different clusters. Looking ahead, refining clustering techniques, exploring additional attributes, implementing predictive models for behavior forecasting, and enabling real-time segmentation are promising avenues for future research. In conclusion, while K-means clustering has proven effective in customer segmentation, the research emphasizes the need for further evaluation across diverse industry contexts. Customization and enhancement of methodologies based on domain-specific insights and the exploration of alternative clustering techniques like graph-based or density-based approaches are vital for actionable customer insights and improved business outcomes. Absolutely! Here's a more expanded version: The revealed customer segments open avenues for nuanced marketing strategies, suggesting a segmentation-centric approach. By focusing on age-group preferences within clusters, there's an opportunity to craft highly personalized marketing initiatives. Tailored promotions, specialized product placements, or even bespoke services could be envisioned to resonate more profoundly with specific age cohorts within these clusters. This could significantly amplify customer engagement and satisfaction, fostering stronger brand loyalty. Further explorations into the drivers behind age-group-specific preferences within these clusters are pivotal for deeper comprehension of consumer behavior. Delving into the underlying factors influencing these purchasing patterns—be it socio-economic influences, cultural dynamics, or lifestyle preferences—can unravel profound insights. Additionally, an in-depth analysis of the psychographic and behavioral aspects within these segments could offer a more holistic understanding of customer decision-making. Moreover, extending the analysis to encompass temporal trends or seasonal variations might shed light on dynamic purchasing behaviors, offering a comprehensive view of customer preferences across different timescales. Furthermore, integrating external datasets or incorporating qualitative research methodologies like surveys or interviews might enrich the analysis. Examining customer sentiments, aspirations, or their perception of value could fortify the understanding of why certain age groups exhibit specific preferences within each cluster. Additionally, leveraging advanced machine learning techniques, such as predictive analytics or recommendation systems, could aid in foreseeing future purchasing patterns and tailoring offerings accordingly. These multifaceted analyses promise to deepen insights into consumer behavior, thereby refining marketing strategies for more impactful and enduring customer relationships.

REFERENCES

1) Anitha, J., Ravi, L. (2019). Comparison of clustering algorithms for customer segmentation. International Journal of Engineering and Advanced Technology, 8(4), 333-339.
2) Chiu, S. L., Tsai, C. F., Yeh, C. C. (2009). An intelligent market segmentation system using k-means, SOM, and PSO. Expert Systems with Applications, 36(5), 9321-9329.
3) Gomes, A., Meisen, T. (2023). RFM analysis for customer segmentation in retail: A literature review. Journal of Business Research, 157, 111110.
4) Kumar, A. (2023). Customer Segmentation: Concepts, Methods, and Applications. Igi Global.
5) Pu, L. (2022). Improving k-means clustering with density-based initialization for customer segmentation. Expert Systems with Applications, 191, 116319.

6) Salminen, J., Kuusisto, J., Kuikka, S., Penttinen, E. (2023). Customer segmentation using k-means clustering: A systematic literature review. Journal of Business Research, 157, 111117.
7) Tabianan, S., Darmawahyuni, R., Permana, I. G. (2022). Customer segmentation using k-means clustering and principal component analysis for bank marketing. International Journal of Electrical and Computer Engineering, 12(1), 73-82.
8) Tavor, A., Gonen, R., Spiegel, U. (2023). Using barcode scanners to enhance customer segmentation and pricing strategies. Journal of Retailing, 100(1), 100093.
9) Turkmen, S. (2022). Evaluation of different clustering algorithms for customer segmentation. International Journal of Data Science and Machine Learning, 13(1), 1-13.
10) Yıldız, C., Çoban, Y., Karakaş, H. (2023). Customer segmentation in big data environment: A comprehensive review. Expert Systems with Applications, 190, 116299.