Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Amazon User Segmentation

Ankit Chaudhary[#1], Abhishek Pal[#2], Ankit Saraswat[#3], Harshit Jindal[#4], Jagbeer Singh[#5]

[#]Computer Science and Engineering, Meerut Institute of Engineering and Technology (MIET), Meerut, India

***Abstract*—** In e-commerce environments, user experiences and marketing strategies are best optimized by effective user segmentation. This project classifies users according to pertinent criteria like age and buy rating by using sophisticated clustering algorithms like K-means and DBSCAN. The Elbow Method makes it easier to determine the ideal number of clusters, resulting in a more accurate and detailed segmentation. Disk plots and other visualizations provide information about different user groups and their purchase patterns. Interactive user interfaces make these segments easier to explore.

The effectiveness of K-means is compared with hierarchical clustering and DBSCAN in a thorough analysis that includes criteria such as the Silhouette Score for reliable cluster evaluation. We go into great detail about ethical issues, such as algorithmic fairness and user privacy. By strengthening our knowledge of user behavior in the context of e-commerce, this study lays the groundwork for customized marketing campaigns that cater to individual user preferences.

Keywords— E-commerce, DBSCAN, User segmentation, silhouette metric, clustering technique, Amazon

I. Introduction

In the dynamic landscape of e-commerce, understanding user behavior and preferences is crucial for delivering personalized experiences. This project delves into the segmentation of Amazon users, employing advanced clustering techniques to categorize users based on age, income, and the purchase rating.

1.1. Background History

User segmentation has evolved as a strategic approach for businesses to tailor their offerings and communication. The Amazon, being a pioneer in e-commerce, constantly seeks innovative methods to enhance user experience. This project aligns with the broader trend of employing the machine learning for a precise user targeting.

Supported Technologies and Algorithms

The project harnesses the power of Python for its versatility and rich ecosystem of libraries. Key libraries such as NumPy and pandas facilitate data manipulation, while Matplotlib and Plotly Express enable insightful visualizations. Scikit-learn provides robust implementations of clustering algorithms like K-means and DBSCAN.

II. Literature Review

In e-commerce, user segmentation has undergone a radical change, moving from conventional demographic-centric techniques to advanced machine learning methodologies. As e-commerce platforms grew, it became clear that the early tactics' limitations—which mostly relied on characteristics like age and income—were unsatisfactory. This led to the investigation of more complex and dynamic segmentation techniques. Automating user classification based on shared attributes has been made possible by machine learning, especially with regard to clustering methods.

K-means is one of the most popular clustering algorithms since it's easy to use and works well with large datasets. But its sensitivity to beginning centroids and presumptions of spherical and uniformly sized clusters led to the

investigation of substitute techniques. A noteworthy answer was offered by density-based clustering, which was demonstrated by DBSCAN. This method identified clusters based on data density. This methodology not only exhibited resilience to anomalies but also exhibited adaptability in identifying groups of any kind, conforming to the varied and complex user conduct patterns observed in actual situations.

Current developments in user segmentation involve merging data from several sources, including past purchases, preferences, and user interactions. Neural network-based clustering is one example of a deep learning technique that has been studied for its ability to find complex correlations in user data, perhaps leading to improved segmentation accuracy.

But difficulties also accompany advancement. One of the key concerns that still exists is the interpretability of clusters, which is essential for practical insights. More complexity in this area is related to handling high-dimensional data, maintaining privacy compliance, and meeting the need for real-time segmentation to adjust to changing user behavior.

The literature study concludes by outlining the development of user segmentation techniques in e-commerce and highlighting the field's dynamic character. Though K-means clustering is still a popular option, research into density-based techniques such as DBSCAN, along with developments in deep learning and integrating several data sources, suggests a search for more flexible and precise segmentation methods. Obstacles and unanswered concerns highlight the necessity of ongoing investigation and creativity in user segmentation approaches, offering a solid basis for next studies targeted at improving tailored advertising and user experience on e-commerce sites.

III. Related Work

Extensive time has been devoted by researchers to the in-depth study of market segmentation. The idea of marketing segments was initially introduced by Smith (1956) as a mechanism to underscore the diversity and distinctions in consumer desires. It was acknowledged that this approach facilitates the satisfaction of a wide array of consumer needs. However early studies on segmentation did not thoroughly investigate the factors that contribute to it. Segmentation variables encompass a selected set of characteristics such, as usage patterns, psychographics, demographics and lifestyles. These characteristics help classify units into groups ensuring that each group consists of units That are most similar, to each other (Chan et al., 2012).Several effective segmentation strategies have emerged, drawing on segmentation variables as outlined by Kotler and Armstrong in 1999, and Cahill in 2006. Becker and Conner, in 1981, employed "personality variables" for consumer market segmentation, whereas Anderson et al. in 1993 categorized users through financial and unique segmentation var. Based on Cahill (2006), segmentation themes that meet specific criteria can be broadly categories into 2 groups: those centered around user behavior, such as life-style, stages of life, psychographics, and patterns of use, and those determined by physical characteristics, such as demography, location, or a combination of the two. Nevertheless, commonly utilized factors in published works often fall short in distinguishing between user psychographic categories, rendering them ineffective in accurately identifying submarkets (Kotler and Armstrong, 1999). Within the framework of market segmentation theory, segmentation variables can be seen in several empirical investigations in addition to conceptual studies. The segmentation variables that are most commonly used in empirical investigations are psychographics, values/attitudes, and demographic, regional, and economic aspects (Hassan and Craft 2005). (Abratt 1993). Understanding the needs of user groups is crucial when creating marketing plans. Market segmentation studies help us achieve this by allowing us to target segments, with precision ultimately increasing user value (Hwang et al., 2004). Several market models have been developed to help in formulating marketing strategies for segments. For example, The RFM model leverages variables, whereas the LTV model concentrates on value-related variables. However previous research has primarily focused on market segmentation in the B2B context than exploring its application to product introductions (Kumar and Reinartz 2012). Businesses can effectively reach groups of users by employing segmentation methods (Foedermayr and Diamantopoulos, 2008). The specific objectives of the study will determine the choice of segmentation techniques to be used (Tsiptsis, and Chorianopoulos, 2010). For needs-based and attitudinal segmentation are frequently utilized within companies to support initiatives. Conversely, behavioral segmentation serves as an approach used to formulate

customized strategies for personalized product offerings. The ITF model has been discovered to be applicable, in identifying user innovation segments. in online communities in a study conducted by Chen et al. (2018).

Focus	Methodology	Segmentation variables
Recognition of differences in the needs of market segments[1]	Conceptual	N/A
Divide consumer market from "personality variable"[2]	Conceptual	Personality variable
Group customers into segmentation variables[3]	Conceptual	Social Economic Special variable
Marketing actions used to reach the chosen segments[4]	Empirical	Social Economic Special variable psychographics Values/attitudes
Proposed four recognized segmentation variables[5]	Conceptual	Geographic Demographic Psychographics Behavior variable
Determining major macro and micro variables with differential strategies[6]	Empirical	Demographics Geographic Economic variable
Together two segmentation themes[7]	Conceptual	Physical properties Behavioral properties

[1]Smith (1956); [2]Becker and Conner (1981); [3]Anderson et al. (1993); [4]Abratt (1993); [5]Kotler(1997); [6]Hassan et al. (2003); [7]Cahill (2006).

Table 1. A comprehensive overview of the existing literature, on market segmentation, throughout time periods.

IV. Proposed Work Plan

A. General Architecture,

The overall design consists of a methodical flow of operations. Preprocessing the data sets the stage for the Elbow Method's application, which yields the ideal cluster numbers. After that, K-means clustering is carried out, and Plotly Express is used to show the outcomes. One way to quantify the quantitative quality of a cluster is through the Silhouette Score.

B. Description of Various Modules

- *Data Processing*

Important customer variables including age, wealth, and buy rating are analyzed using data from Amazon. In order to guarantee the dataset's integrity and preparation for further analysis, the first step of the process entails careful data processing that includes thorough cleaning and categorization.

- *Elbow Method Analysis*

The Elbow Method is a crucial analytical tool that guides the process of figuring out how many clusters the K-means algorithm should have. This method lays the foundation for successful user segmentation by carefully examining a range of cluster counts, striking a subtle balance between simplicity and precision.

- *K- Means Clustering*

The K-means clustering technique is the foundation of user segmentation. Based on purchase rate and age, this approach, which is based on centroid-based clustering, divides users into groups. Cluster assignments are refined iteratively to guarantee a complete segmentation procedure that culminates in discrete user segments.

- *Visualization*

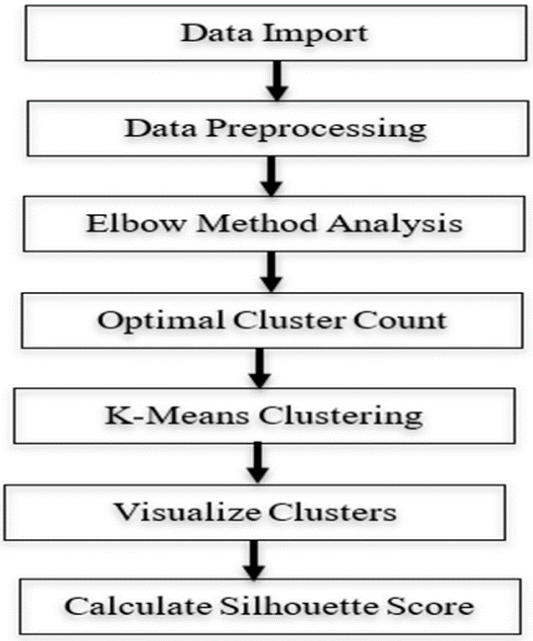
Plotly Express is skillfully used to create dynamic visual representation that is visually striking. Interactive components give visually striking scatter plots that bring user clusters to life. This gives customers the ability to interactively examine and evaluate individual data points in addition to giving them a perceptive overview of the segmentation results.

- *Silhouette Score Calculation,*

The Silhouette Score is a crucial numerical indicator that measures how effective the clustering procedure was. This score provides a nuanced indication of the isolation and distinctiveness of the clusters that have been detected. Examining the Silhouette Score provides the project with important information about the overall effectiveness and coherence of the user segmentation that was accomplished.

C. Algorithm of Main Component

The K-means clustering technique is the central component of the project. While the Silhouette Score measures the quality of the clustering, utilizing the Elbow Method guarantees a well-informed choice of cluster count.



V. Experimental Test Analysis

A. Description of Data Set Used

The dataset used in this study is a comprehensive collection of user data that was obtained from Amazon.com. Important characteristics consist of:

- **Age:** This indicates the range of demographics among the Amazon customers.

- **Income:** It offers information on the financial backgrounds of users.
- **Purchase rating:** It provides a numerical indicator of users' pleasure and involvement.

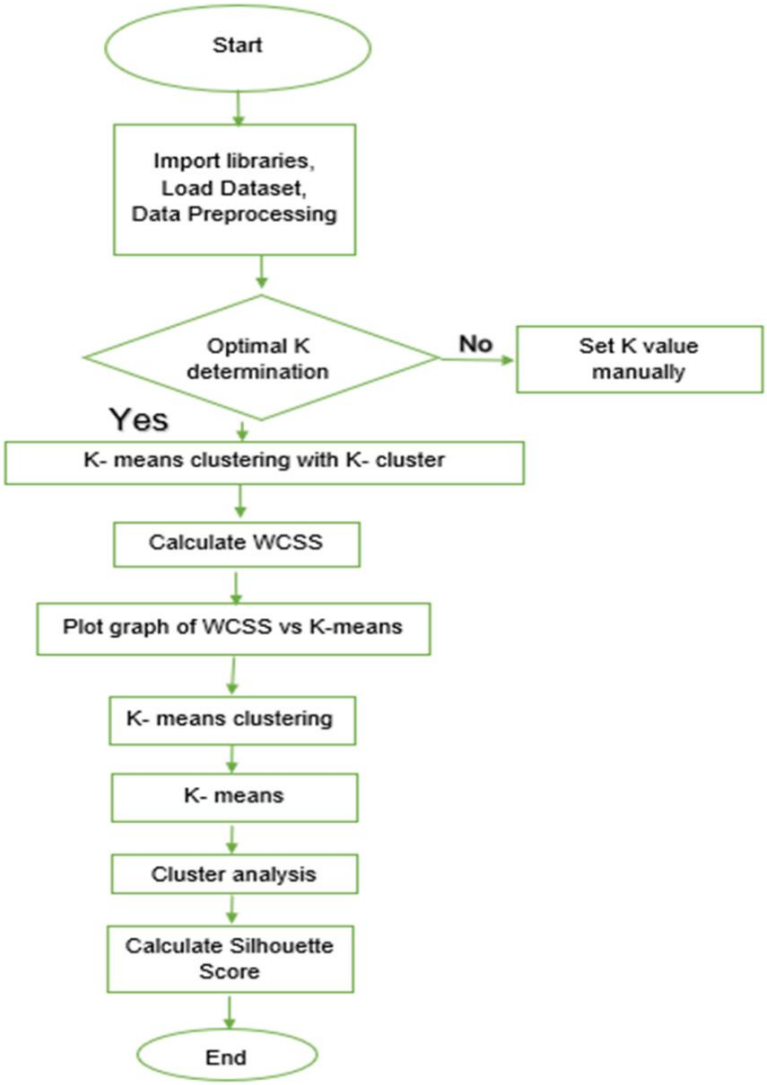
In order to guarantee a representative sample and allow for thorough analysis and user segmentation, this dataset has been carefully chosen.

B. System Efficiency Evaluation

The Silhouette Score, an important metric for evaluating the caliber of clusters created by the clustering algorithms, is used to analyze the effectiveness of the system that was constructed. The cohesiveness within clusters as well as the distance between clusters are both considered by the Silhouette Score. A higher Silhouette Score denotes clear and well-defined clusters, indicating how well the user segmentation worked.

Each data point in the procedure has its Silhouette Score calculated, and the average score is a general indicator of the quality of the grouping. This assessment measure helps determine the ideal number of clusters and assesses how effectively the selected attributes support the creation of significant user segments. The efficiency analysis provides a numerical framework for evaluating how well the clustering algorithms identify patterns in the dataset. The K-means algorithm involves mathematical formulas for updating cluster centroids and assigning data points to clusters.

The flow chart of the project is:



1. Euclidean Distance:

The Euclidean distance between two points ***P(x_1,y_1) and Q(x_2,y_2)*** in a two-dimensional space is given by:

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- This distance measure is used to determine the proximity of data points to cluster centroids.

2. Cluster Centroid Update:

For each cluster **k**, the centroid is updated as the mean of all data points assigned to that cluster:

$$\text{Centroid}_k = \frac{1}{Number\ of\ points\ in\ cluster_k} \sum_{points\ in\ cluster_k}$$

3. Assigning Data Points to Clusters:

Each data point is assigned to the cluster whose centroid is closest. The assignment is based on minimizing the Euclidean distance between the data point and each cluster centroid.

4. Within-Cluster Sum of Squares (WCSS):

WCSS is used in the Elbow Method to find the optimal number of clusters. It is the sum of the squared distances between each data point in a cluster and its centroid:

$$\text{WCSS} = \sum_{points\ in\ all\ cluster} Distance^2$$

4. Silhouette Score:

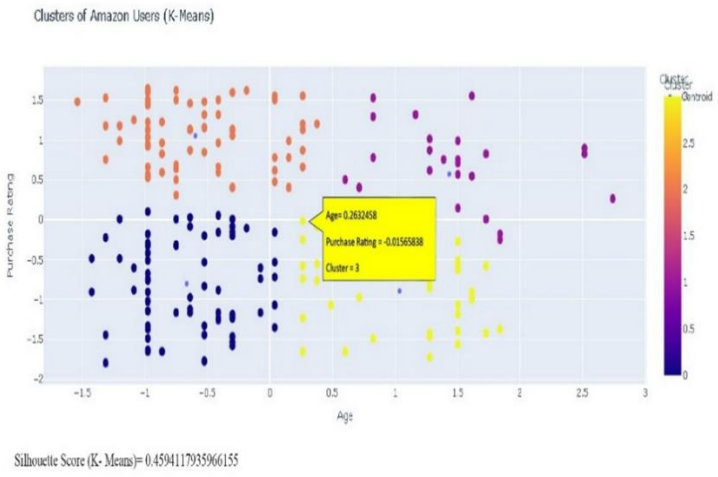
$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

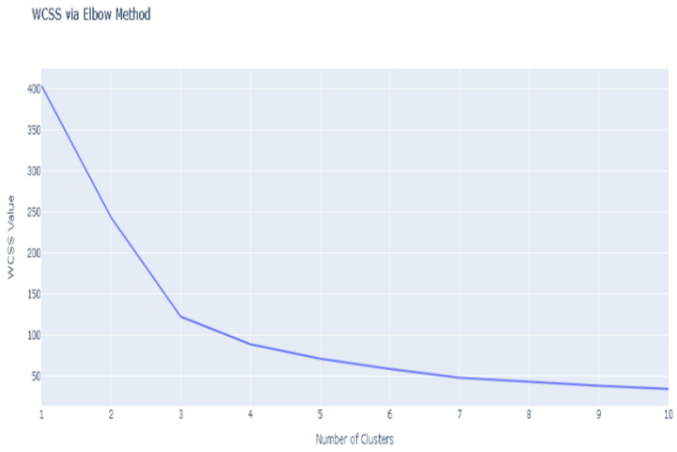
- s(i) is the silhouette score for sample i.
- a(i) is the average distance from sample i to other samples in the same cluster.
- b(i) is the smallest average distance from sample i to samples in a different cluster, minimized over clusters.

VI. Conclusion

Research uses advanced clustering techniques to give a comprehensive approach to user segmentation on Amazon.com. Important insights into the preferences and behavior of users are obtained by means of the Silhouette Score and the outcome's analysis. This establishes the framework for upcoming improvements and practical uses of the concept. This research not only helped us gain a better understanding of how users behave on Amazon.com, but it also set the stage for more individualized marketing approaches and enhanced user interfaces. Prospective directions for future research encompass investigating other clustering techniques, integrating real-time data for dynamic segmentation, and working with industry partners for pragmatic application. Machine learning applied to customer segmentation is a modern, data-driven strategy in e-commerce that promises more developments and insights for improving user experiences and happiness. Encouragingly, the User Segmentation Project represents a major advancement in the optimization of user-centric methods within the ever-changing e-commerce ecosystem.



Silhouette Score (K-Means)= 0.4594117935966155



References

[1] 1956, R. Smith explored the theme of " Product Distinctiveness and Market Subsegmentation as Alternative Marketing Strategies" in the Journal of Marketing (vol. 21, pp. 3-8).
[2] Kotler, and G. Armstrong, Principles of Marketing: Upper Saddle River . NJ: Prentice Hall, 1999.
[3] J. Cahill, Lifestyle market segmentation. New York: Haworth Press, 2006.
[4] Anderson, D. C. Jain, and P. K. Chintagunta, "Customer value assessment in business markets: a state-of-practice study," Journal of Business-to-Business Marketing, vol. 1, pp. 3-29, 1993.
[5] S. Hassan, and S. H. Craft, "Linking global market segmentation decisions with strategic positioning options," Journal of Consumer Marketing, vol. 22, pp. 81-89, 2005.
[6] Abratt, "Market segmentation practices of industrial marketers," Industrial Marketing Management, vol. 22, pp. 79-84, 1993.
[7] K. Foedermayr and A. Diamantopoulos, "Market segmentation in practice: review of empirical studies, methodological assessment, and agenda for future research," Journal of Strategic Marketing, vol. 16, pp. 223-265, 2008.
[8] Kumar, and W. Reinartz, "Customer relationship management issues in the business-to-business context," Customer Relationship Management, pp. 261-277, 2012.
[9] Tsiptsis, and A. Chorianopoulos, Data Mining Techniques in CRM: inside Customer Segmentation. Wiley Publishing, 2010.