Application Of K-Means Clustering For Customer Segmentation In Grocery Stores In Kenya

Edwin Omol[1*], Dorcas Onyangor[2,], Lucy Mburu[3], Paul Abuonji[4]

[1,3,4] School of Technology, KCA University P. O. Box 56808 – 00200 Nairobi, Kenya
[2] School of Business, KCA University P. O. Box 56808 – 00200 Nairobi, Kenya
*Corresponding Author:
Email: omoledwin@gmail.com

***Abstract**.*

The retail industry, particularly in the context of grocery stores, plays a vital role in meeting consumers' daily needs. To optimize marketing strategies and enhance customer satisfaction, understanding customer behavior and preferences is crucial. Customer segmentation, a powerful market research technique, enables businesses to group customers with shared characteristics into distinct segments, allowing targeted and personalized approaches. This article explores the application of the K-means clustering algorithm for customer segmentation in grocery stores within the unique context of Kenya. By leveraging transactional and demographic data from diverse grocery stores across Kenya, the study aims to identify homogeneous customer groups with similar purchasing behaviors and preferences. The data collection process involved obtaining consent from store owners and ensuring data privacy and security. Following data preprocessing, K-means clustering was applied, and various validation techniques were utilized to determine the optimal number of clusters. The results yielded valuable insights into customer segments, aiding the identification of key customer groups and their distinct preferences.

***Keywords:** customer segmentation, market segmentation, k-means clustering algorithm and marketing analytics.*

I. INTRODUCTION

The retail industry is continuously evolving in the wake of digital transformation [1, 30], and grocery stores play a pivotal role in meeting the daily needs of consumers. Understanding customer behavior and preferences is essential for these stores to optimize their marketing strategies, improve customer satisfaction, and enhance overall profitability [2]. Customer segmentation, a powerful technique in market research, enables businesses to group their customers into distinct segments based on shared characteristics, thereby facilitating targeted and personalized approaches [3].This article explores the application of K-means clustering, a popular unsupervised learning algorithm, for customer segmentation in grocery stores within the unique context of Kenya. As the retail landscape in Kenya differs significantly from that of other regions, with its diverse cultural and economic aspects, tailoring segmentation methods to suit the local market becomes imperative. Through this research, we hope to contribute valuable knowledge to the realm of retail analytics and enhance the overall shopping experience for customers in Kenya's grocery retail sector.

II. METHODS

The methodology employed involved a multi-step process combining data collection, data preprocessing, and the application of the K-means clustering algorithm. see figure 1.

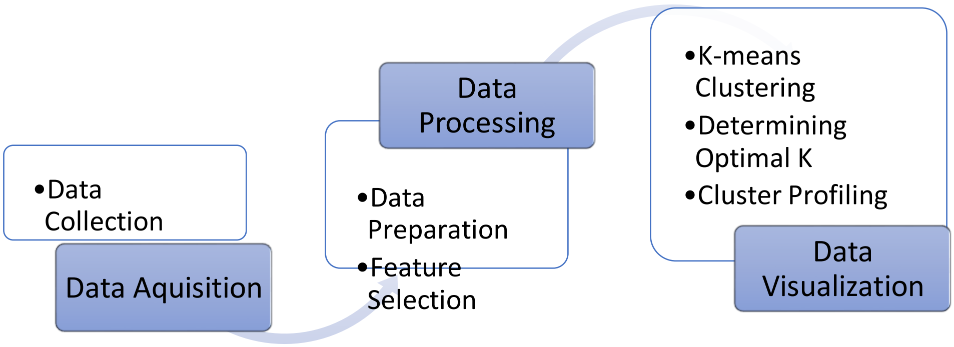


Fig 1. Study Methodology

1. **Data Collection:**

The first step was to collect comprehensive data from Beyond Fruits grocery stores operating in main towns in Kenya. The data included both transactional information and relevant demographic details of customers. Transactional data included purchase histories, item categories bought, and timestamps. Demographic data comprised age, gender, location, annual income, and spending scores.

2. **Data Preprocessing:**

To ensure the quality and suitability of the data, a thorough data preprocessing phase was executed. This involved data cleaning, handling missing values, removing outliers, and normalizing numerical variables. The objective was to create a consistent and reliable dataset for subsequent analysis.

3. **Feature Selection:**

Next, appropriate features were selected from the preprocessed dataset for the customer segmentation task. Feature selection was a critical step in identifying the most relevant attributes that could effectively differentiate customer behavior and preferences.

4. **K-means Clustering:**

The K-means clustering algorithm was then applied to the selected features. K-means is an unsupervised learning algorithm that partitions data into K clusters, where K is a user-defined parameter. The primary goal of this step was to group customers into distinct clusters based on their shared purchasing patterns and demographic characteristics.

5. **Determining Optimal K:**

To determine the optimal number of clusters (K) for the K-means algorithm, the Elbow Method was employed. The goal was to identify the value of K that provided the best trade-off between compactness within clusters and separation between clusters.

6. **Cluster Profiling**:

Once the optimal value of K was established, each cluster was analyzed and profiled to understand the unique characteristics and preferences of customers within the segment. This included examining the age [19], annual income, and spending score.

III. LITERATURE REVIEW

Customer segmentation is a crucial aspect of modern retail analytics, enabling businesses to understand and cater to the diverse needs of their clientele. Over the years, various clustering techniques have been employed to achieve effective customer segmentation, with the K-means clustering algorithm emerging as a prominent method due to its simplicity and scalability [4-7, 9-14, 20-29]. Several studies have demonstrated the effectiveness of K-means clustering in customer segmentation across different industries. Kansal, Tushar, et al. [5] applied K-means clustering to identify customer segments in a Chinese supermarket chain, highlighting the algorithm's ability to distinguish between high and low-value customers based on their purchase histories. Similarly, in the context of e-commerce, Deng and Qianying [25] utilized K-means clustering to segment online shoppers, revealing distinct clusters based on browsing behavior and purchase patterns.In Shirole et al. [7], the segmentation is based on the RFM (Recency, Frequency, Monetary) model and the K-means algorithm applied to transaction data from an online retail store. Four clusters, Class A, Class B, Class C, and Class D, are identified, with Class A generating the highest revenue and Class D generating the least. The segmentation helps uncover customer behavior and preferences. The calculated silhouette index indicates a good level of clustering quality.While K-means clustering has been widely adopted in various retail domains, limited research has been conducted specifically on its application in grocery stores in Kenya. Given the unique cultural and economic factors influencing consumer behavior in the region, understanding local customer segmentation becomes imperative for store owners and marketers to tailor their strategies effectively [17].

In the Kenyan context, Kennedy, Ryan, et al. [13] investigated customer segmentation in the retail sector using traditional clustering techniques, such as hierarchical clustering and partitioning algorithms, others are [16, 18]. Although their study provided valuable insights, the potential of K-means clustering remained unexplored in this specific context.Another relevant aspect of customer segmentation in grocery stores is the incorporation of demographic data. Li, Yue, et al. [11] demonstrated the significance of combining transactional data with demographic attributes in customer segmentation to create more meaningful and actionable clusters. The integration of such data in the K-means clustering process holds promise for enhancing segmentation accuracy and targeting specific customer groups in the Kenyan grocery retail sector.Moreover, advances in K-means clustering extensions have further extended its capabilities. For instance, Zhao, Hong-Hao, et al. [27] proposed an improved version of K-means, incorporating customer lifetime value as an additional feature for customer segmentation, resulting in more robust clusters with higher business value.For instance, in the context of online retail, K-means clustering has been utilized to group customers with similar browsing and purchasing behaviors, enabling the formulation of targeted marketing campaigns [12]. Similarly, in traditional brick-and-mortar stores, K-means clustering has been applied to analyze transactional data, leading to the identification of homogeneous customer groups based on their buying patterns [26-29].

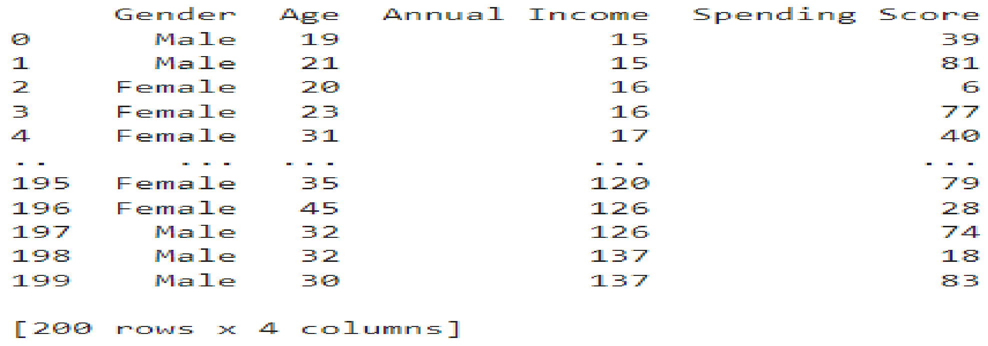
However, while K-means clustering has been widely studied in retail customer segmentation, there is a limited body of literature specifically focusing on grocery stores in the Kenyan context. Given the unique cultural and economic factors influencing consumer behavior in Kenya, the application of customer segmentation techniques in this setting becomes of paramount importance.A notable study by Maneno et al. [4] explored customer segmentation in the Kenyan retail sector using a different clustering algorithm. Their findings indicated significant variations in shopping preferences across different regions of Kenya, highlighting the need for tailored marketing strategies. However, their research did not encompass the application of K-means clustering, which is known for its simplicity and efficiency in large datasets.In a broader context, research on K-means clustering has shown its potential in diverse applications beyond retail. For example, La Cruz, Alexandra, et al. [9] employed K-means clustering for geographic segmentation, partitioning customers based on their proximity to stores, thereby aiding the optimization of store locations and logistics. Moreover, K-means clustering has been extensively used in the healthcare industry to segment patients based on medical records, leading to more personalized treatment approaches [18].To the best of our knowledge, no existing research has exclusively focused on applying K-means clustering for customer segmentation in Kenyan grocery stores. Thus, this study aims to bridge this research gap by investigating the utility of K-means clustering in this unique retail landscape. By leveraging transactional and demographic data from various grocery stores in Kenya, this research seeks to contribute valuable insights into the identification of distinct customer segments and their preferences, empowering store owners and decision-makers to implement targeted marketing strategies and enhance customer satisfaction.

Findings:

The application of K-means clustering for customer segmentation in grocery stores in Kenya yielded valuable insights into the distinct customer segments and their preferences. The analysis of transactional and demographic data from various grocery stores across Kenya resulted in the identification of several homogeneous customer groups based on their purchasing behaviors and characteristics.

Data Extraction

Initially, the study imported essential libraries, including pandas, numpy, matplotlib, plotly express, seaborn, sys, and warnings, required for conducting the clustering process. Subsequently, we proceed to locate and display our dataset. See output below.

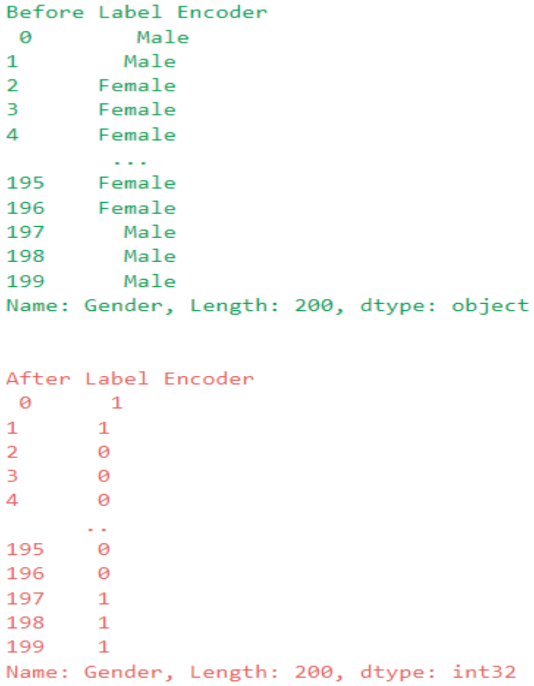


```
     Gender  Age  Annual Income  Spending Score
0      Male   19             15              39
1      Male   21             15              81
2    Female   20             16               6
3    Female   23             16              77
4    Female   31             17              40
..      ...  ...            ...             ...
195  Female   35            120              79
196  Female   45            126              28
197    Male   32            126              74
198    Male   32            137              18
199    Male   30            137              83

[200 rows x 4 columns]
```

Data Labelling

The research integrated essential functions facilitating the conversion of data into numeric formats. Subsequently, Python is directed to display both the coded data and the corresponding numeric representation. See output below.

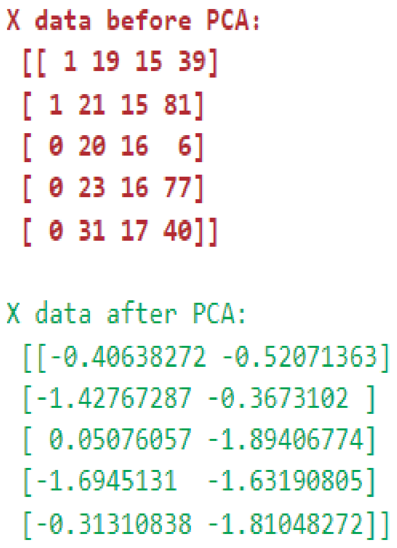


```
Before Label Encoder
 0        Male
1         Male
2       Female
3       Female
4       Female
          ...
195     Female
196     Female
197       Male
198       Male
199       Male
Name: Gender, Length: 200, dtype: object


After Label Encoder
 0      1
1       1
2       0
3       0
4       0
       ..
195     0
196     0
197     1
198     1
199     1
Name: Gender, Length: 200, dtype: int32
```

Data Transformation

In situations involving a substantial number of attributes, Principal Component Analysis (PCA) can be utilized to decrease dimensionality by transforming the attributes into two principal components, facilitating convenient visualization. In this study, the PCA function was applied to compress the data, resulting in the generation of the reduced sample, which was subsequently displayed. See output below.



```
X data before PCA:
 [[ 1 19 15 39]
 [ 1 21 15 81]
 [ 0 20 16  6]
 [ 0 23 16 77]
 [ 0 31 17 40]]

X data after PCA:
 [[-0.40638272 -0.52071363]
 [-1.42767287 -0.3673102 ]
 [ 0.05076057 -1.89406774]
 [-1.6945131  -1.63190805]
 [-0.31310838 -1.81048272]]
```

Number of Clusters

The research employed the Elbow Method, utilizing the Kmeans function from the sklearn.cluster library, to determine the appropriate number of clusters. By conducting training on multiple models with varying numbers of clusters, the study recorded the corresponding within-cluster sum of squares (WCSS) values at each iteration. Subsequently, the plt.figure, plot, title, and label functions were imported to visualize the graph. The optimal number of clusters, determined to be 4, was identified at the specific point where the chart displayed the most significant break or change. see figure 2 and figure 3 below.

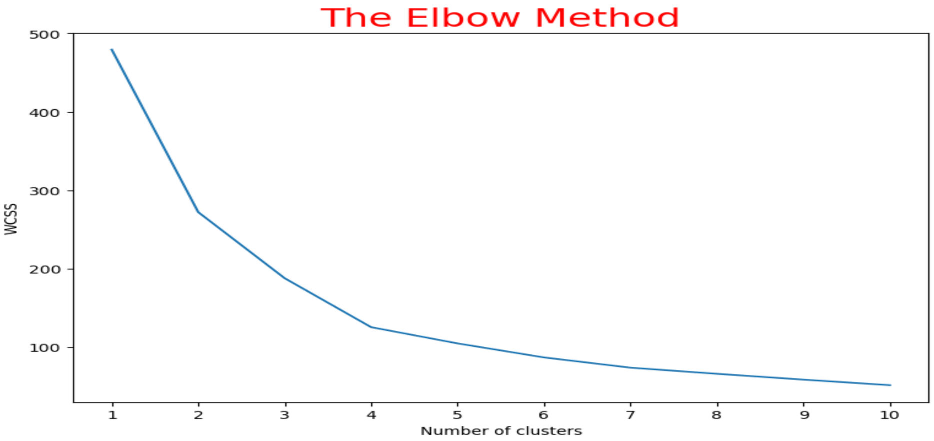


Fig 2. Elbow Method Result

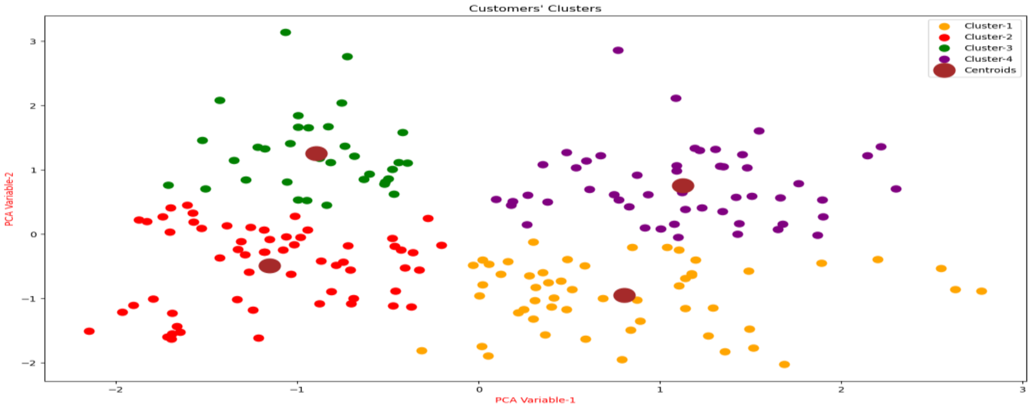
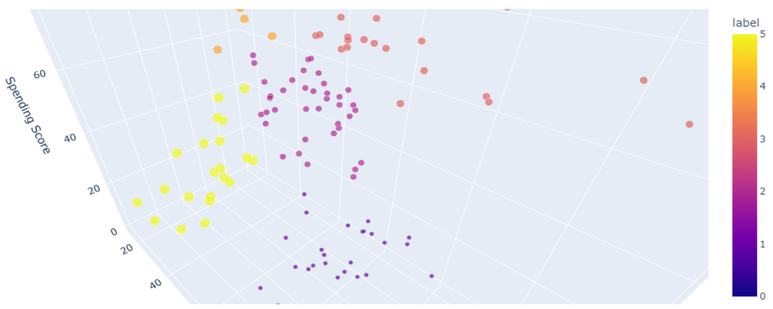


Fig 3. Number of Clusters

Cluster for Age, Annual Income, and Spending Score

To gather the data points for the graph, the research involved training numerous models with different cluster numbers and recording the within-cluster sum of squares (WCSS) values using the inertia_ property at each iteration. The optimal number of clusters, identified as 6, was determined by detecting the most significant break or change on the chart at that specific point. The process involved importing the Kmeans function from the sklearn.cluster library and utilizing the plt.figure, plot, title, and label functions to display the figure 4.

Fig 4. Clusters Visualization



Cluster for Age and Annual Income

In order to obtain the values used in the graph, the study conducted training on multiple models using varying numbers of clusters and recorded the value of the intertia_ property within-cluster sum of squares (WCSS) at each iteration. The optimal number of clusters is determined to be 2, as indicated by the largest break or change observed in the chart at that specific point. The study Imported Kmeans from sklearn.cluster library then imported the function plt.figure,plot,title,label to display the figure 5.

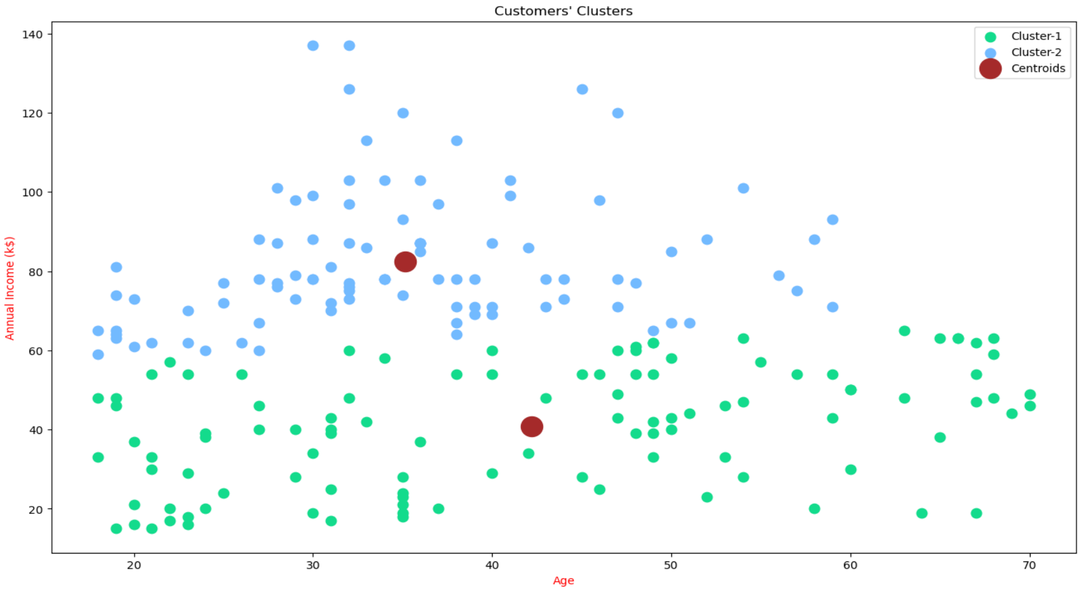


Fig 5. Annual income and Age Cluster

Cluster for Annual Income and Spending Score

To acquire the data points for the graph, the research involved training multiple models with diverse cluster numbers, and at each iteration, it recorded the within-cluster sum of squares (WCSS) values using the inertia_ property. The optimal number of clusters, identified as 5, was determined based on the most substantial break or change observed in the chart at that specific point. To execute these tasks, the study imported Kmeans from the sklearn.cluster library and utilized the functions plt.figure, plot, title, and label to display the figure 6.

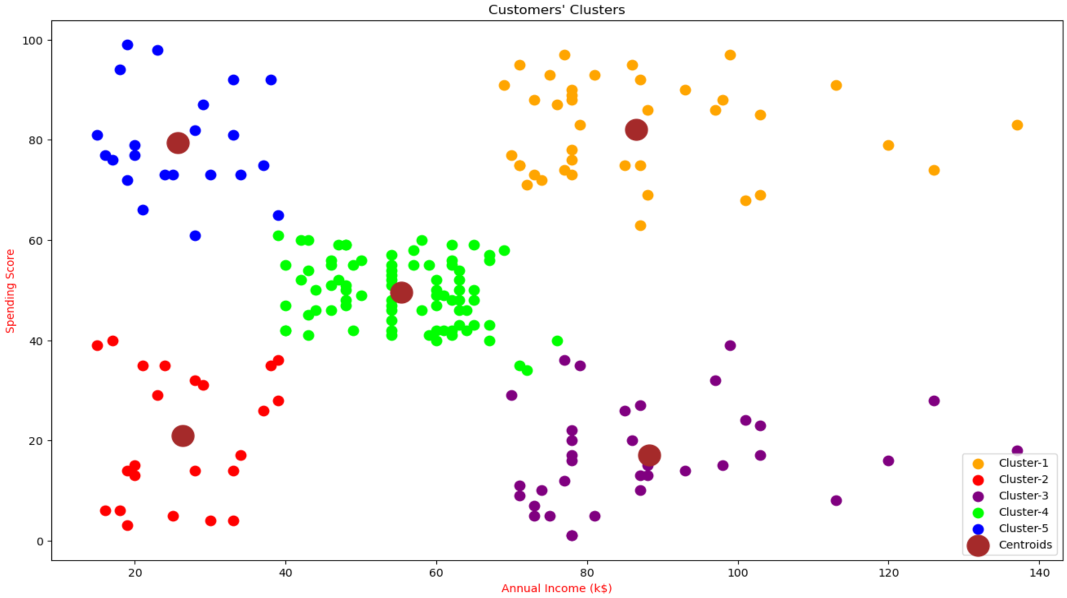


Fig 6. Annual Income and Spending Score Cluster

Cluster for Age and Spending Score

To acquire the data for the graph, the research involved training multiple models with different cluster numbers and recording the within-cluster sum of squares (WCSS) values using the inertia_ property at each iteration. The optimal number of clusters, identified as 5, was determined by detecting the most prominent break or change on the chart at that specific point. The study imported the Kmeans function from the sklearn.cluster library and utilized the plt.figure, plot, title, and label functions to display the figure 7.

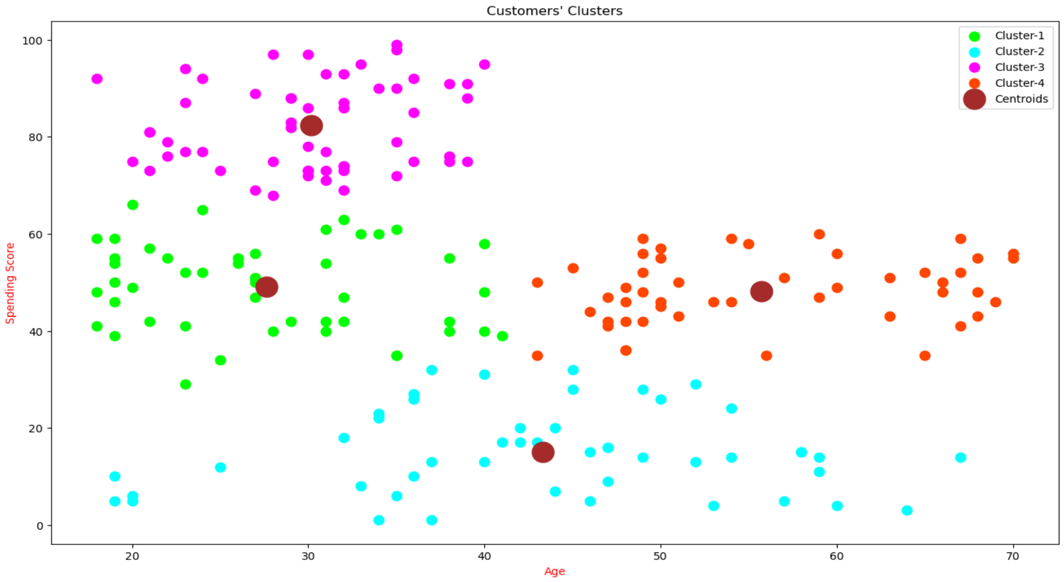


Fig 7. Age and Spending Score Cluster

The results indicate that there is not a significant difference between male and female customers, suggesting that a gender-based targeting strategy is not recommended. However, it is observed that customers in the age range of 20-40 tend to spend more compared to other age groups. Implementing special campaigns to target this age group could potentially increase the supermarket's profits.While targeting middle-income customers (with incomes between 40k-70k dollars) may not lead to significant increases in their spending levels due to their income limitations, attracting more of these customers through campaigns can still contribute to the store's profitability, considering their average spending scores are at a moderate level.The most effective strategy would be to focus on high-income customers. Although some high-income customers already spend a significant amount, there is a considerable portion within this group who have lower spending levels, indicating potential areas of dissatisfaction. By improving service quality, addressing their needs, and enhancing the overall customer experience, it is possible to increase the spending of high-income customers who visit the store but currently spend less. This approach offers the greatest potential for maximizing profits.

Discussion:

The findings of this study have significant implications for grocery stores in Kenya, enabling them to enhance their marketing strategies and improve overall customer experiences. The application of K-means clustering facilitated the creation of targeted approaches tailored to the preferences of different customer segments. By understanding the distinct needs and behaviors of each cluster, store owners and marketers can develop more effective marketing campaigns and personalized promotions.

1. Targeted Marketing Strategies: Armed with the knowledge of customer segments, grocery stores can implement targeted marketing strategies that cater to the specific preferences of each group. For example, marketing promotions can be customized to highlight products that resonate with each segment, increasing the chances of higher engagement and conversion rates.

2. Inventory Management: The identification of customer preferences and purchasing patterns can aid in optimizing inventory management. Stores can stock items in alignment with the demands of each segment, reducing waste and ensuring that popular products are consistently available.

3. Customer Experience Enhancement: By catering to the unique preferences of different customer clusters, grocery stores can offer a more personalized shopping experience. Tailored promotions, discounts, and product recommendations can create a sense of value and satisfaction among customers, potentially leading to increased loyalty and retention.

4. New Market Opportunities: The insights gained from customer segmentation may unveil untapped market opportunities. Understanding the specific needs of different customer groups can inspire the development of new products or services tailored to meet these demands.

Limitations and Future Research:

While the application of K-means clustering provided valuable customer segmentation insights, the study also had its limitations. For instance, the analysis relied heavily on the available data, and there may have been factors not considered that could influence customer behavior. Additionally, the study focused on transactional and demographic data, and future research could explore the inclusion of other data sources such as customer feedback or social media interactions to gain more comprehensive insights.

In conclusion, the application of K-means clustering for customer segmentation in Kenyan grocery stores proved to be a valuable approach. The identified customer segments and their preferences hold significant potential for store owners and decision-makers to optimize marketing strategies, inventory management, and overall customer experiences. By leveraging these findings, grocery stores in Kenya can position themselves competitively in the dynamic retail landscape, fostering stronger customer relationships and driving business growth. Future research in this area may continue to refine customer segmentation techniques and explore additional data sources to gain further insights into consumer behavior in the Kenyan grocery retail sector.

REFERENCES

[1] Omol, Edwin, Lucy Mburu, and Paul Abuonji. "Digital Maturity Action Fields for SMES in Developing Economies." ***Journal of Environmental Science, Computer Science, and Engineering & Technology***, *12*(3), https://doi.org/10.24214/jecet.B.12.3.10114. (2023)

[2] Makara, Isabel. *A Clustering Approach to Market Segmentation Using Integrated Business Data*. Diss. University of Nairobi, 2021.

[3] Makana, Asha P. *Customer Segmentation on Mobile Money Users in Kenya*. Diss. University of Nairobi, 2020.

[4] Maneno, Khamis Mwero, Richard Rimiru, and Calvins Otieno. "Segmentation via principal component analysis for perceptron classification: a case study of Kenyan mobile subscribers." *Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications*. 2020.

[5] Kansal, Tushar, et al. "Customer segmentation using K-means clustering." *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*. IEEE, 2018.

[6] Shirole, Rahul, Laxmiputra Salokhe, and Saraswati Jadhav. "Customer segmentation using rfm model and k-means clustering." ***Int. J. Sci. Res. Sci. Technol*** 8 (2021): 591-597.

[7] Tabianan, Kayalvily, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data." *Sustainability* 14.12 (2022): 7243.

[8] Ong, Ardvin Kester S., et al. "Consumer preference analysis on the attributes of samgyeopsal Korean cuisine and its market segmentation: Integrating conjoint analysis and K-means clustering." *Plos One* 18.2 (2023): e0281948.

[9] La Cruz, Alexandra, et al. "Users Segmentation Based on Google Analytics Income Using K-Means." *Information and Communication Technologies: 9th Conference of Ecuador, TICEC 2021, Guayaquil, Ecuador, November 24–26, 2021, Proceedings 9*. Springer International Publishing, 2021.

[10] Pradana, Musthofa Galih, and Hoang Thi Ha. "Maximizing strategy improvement in mall customer segmentation using k-means clustering." ***Journal of Applied Data Sciences*** 2.1 (2021): 19-25.

[11] Li, Yue, et al. "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm." *Applied Soft Computing* 113 (2021): 107924.

[12] Tabianan, Kayalvily, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data." *Sustainability* 14.12 (2022): 7243.

[13] Kennedy, Ryan, et al. "Multilevel customer segmentation for off-grid solar in developing countries: Evidence from solar home systems in Rwanda and Kenya." *Energy* 186 (2019): 115728.

[14] Matute, Roberto, and Juan Estrada. "Users Segmentation Based on Google Analytics Income Using K-Means." *Information and Communication Technologies: 9th Conference of Ecuador, TICEC 2021, Guayaquil, Ecuador, November 24–26, 2021, Proceedings*. Springer Nature, 2021.

[15] Khan, Riyo Hayat. *LRFS: online shoppers' behavior based efficient customer segmentation model*. Diss. Brac University, 2023.

[16] Salamzadeh, Aidin, et al. "Grocery apps and consumer purchase behavior: application of Gaussian mixture model and multi-layer perceptron algorithm." ***Journal of Risk and Financial Management*** 15.10 (2022): 424.

[17] Omol, Edwin, and Collins Ondiek. "Technological Innovations Utilization Framework: The Complementary Powers of UTAUT, HOT–Fit Framework and; DeLone and McLean IS Model." ***International Journal of Scientific and Research Publications (IJSRP)*** 11.9 (2021): 146-151.

[18] Omonge, Jevans. *A Customer segmentation model using logistic regression: a case of Telkom Kenya*. Diss. Strathmore University, 2021.

[19] Omol, Edwin J., Silvance O. Abeka, and Kelvin K. Omieno. "Relevance of Demographic profile on Acceptance of Mobile Money Payment in enterprise management. A case of MSEs in Kisumu City, Kenya." (2016).

[20] Nandapala, E. Y. L., and K. P. N. Jayasena. "The practical approach in Customers segmentation by using the K-Means Algorithm." *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2020.

[21] Syakur, M. A., et al. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster." *IOP conference series: materials science and engineering*. Vol. 336. IOP Publishing, 2018.

[22] Balakrishnan, PV Sundar, et al. "Comparative performance of the FSCL neural net and K-means algorithm for market segmentation**."** ***European journal of operational research*** 93.2 (1996): 346-357.

[23] Wu, Jun, et al. "An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm." *Mathematical Problems in Engineering* 2020 (2020): 1-7.

[24] Huang, Yong, Mingzhen Zhang, and Yue He. "Research on improved RFM customer segmentation model based on K-Means algorithm." *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*. IEEE, 2020.

[25] Deng, Yulin, and Qianying Gao. "A study on e-commerce customer segmentation management based on improved K-means algorithm." *Information Systems and e-Business Management* 18 (2020): 497-510.

[26] Deng, Yulin, and Qianying Gao. "A study on e-commerce customer segmentation management based on improved K-means algorithm." *Information Systems and e-Business Management* 18 (2020): 497-510.

[27] Zhao, Hong-Hao, et al. "An extended regularized K-means clustering approach for high-dimensional customer segmentation with correlated variables." *Ieee Access* 9 (2021): 48405-48412.

[28] Li, Yue, et al. "Customer segmentation using K-means clustering and the hybrid particle swarm optimization algorithm**."** ***The Computer Journal*** 66.4 (2023): 941-962.

[29] Aryuni, Mediana, Evaristus Didik Madyatmadja, and Eka Miranda. "Customer segmentation in XYZ bank using K-means and K-medoids clustering." *2018 International conference on information management and technology (ICIMTech)*. IEEE, 2018.

[30] Omol, Edwin Juma. "Organizational digital transformation: from evolution to future trends." *Digital Transformation and Society* (2023).