Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Segmentation of Coffee Purchasing Behavior Based on Transaction Time Using the K-Means Algorithm

Yuslia Devitri[1]*, Nining Rahaningsih[2], Irfan Ali[3], Willy Prihartono[4]

[1,2,3,4] *STMIK IKMI Cirebon*

yusliadevitri8@gmail.com[1]*, niningr157@yahoo.co.id[2], irfanaali0.0@gmail.com[3], willyprihartono@gmail.com[4]

Abstract

This studyaims to identify customer behavior patterns based on the time of purchaseof beverages at a coffee shop using the K-Means method.Transaction data includes purchase time, payment type, product name,time category, day, and month. The research stages include data cleaning, time attribute transformation, and numerical feature normalization. The optimal number of clustersis determined through testing k = 2–10 with four evaluation metrics,namely Inertia, Silhouette Score, Davies–Bouldin Index, and Calinski–HarabaszIndex. Based on the validation results, k = 3 was selected because it provided the best balancebetween compactness and cluster separation. The clustering results showedthree main customer groups based on transaction time trends:nighttime buyers with a peak around 10:27 p.m., afternoon to early evening buyerswith a centroid of 7:01 p.m., and morning to noon buyers with a centroid11:13. The frequency distribution indicates that the morning–afternoon buyer groupis the largest, while the early evening–night group is thesmallest. Visualization of scatter plots, boxplots, and time category graphsemphasizes the differences in characteristics between clusters. Overall,this study proves that K-Means is effective in mapping the temporal patternsof customer behavior. These findings can be used to develop time-based marketing strategies, operational arrangements, and product stock management,as well as form the basis for further analysis in the industry.

Keywords: *K-Means; Transaction Time; Coffee Purchasing Pattern; Clustering; Data Mining*

1. Introduction

The coffee industry is growing rapidly, and coffee shops have become social spaces that shape the activities of urban communities. Coffee consumption patterns are also becoming more varied throughout the day, and digital transaction data with timestamps opens up opportunities for time-based purchase behavior analysis. However, many businesses have not yet utilized temporal analysis because it is considered complex.

Various studies confirm that time is an important factor in consumer behavior. Purchasing activities tend to follow daily and weekly rhythms [1], are influenced by physiological conditions, and even have an impact on price sensitivity, which decreases by around 0.5% every hour [2]. Time also influences food and beverage preferences, including an increase in coffee searches at certain hours [3].
In data science, K-Means is widely used for customer segmentation because it can effectively map latent patterns [4]. The development of temporal-based K-Means variants, such as multivariate time-series clustering and K-MDTSC, further strengthens its relevance in time behavior analysis [2]. Public datasets such as Kaggle also facilitate research by providing complete transaction data [5].
However, the use of time analysis in the coffee industry is still low, even though temporal visualizations such as histograms and heatmaps are effective for predicting demand [6]. In addition, there are still few studies that specifically examine coffee purchasing patterns based on transaction times, even though temporal analysis has been proven to reveal recurring demand patterns [7].

This study follows the CRISP-DM methodology, including temporal preprocessing such as timestamp cleaning and normalization, which is important to ensure clustering accuracy [8]. Considering the dynamics of coffee consumption, the potential information in transaction data, and the lack of temporal studies in the coffee industry, time-based purchase pattern analysis is important for scientific development and business strategy optimization.

2. Research Methodology

2.1. Research Object

The research object is coffee purchase transaction data from the Kaggle platform (coffee.csv), which contains information on the time of purchase, product type, price, and purchase quantity. The research focuses on temporal dimension analysis, so the attributes used include hour_of_day (0–23), Time_of_Day (morning, afternoon, evening, night), Weekday, and Month_name. This dataset is used to group purchasing patterns based on transaction time using the K-Means algorithm, in order to identify customer behavior such as shopping time preferences and peak transaction hours [9], [4]. Thus, the research object does not focus on the coffee product itself, but on consumer behavior patterns as reflected in transaction time characteristics.

2.2. Research Process

The research flow in the diagram illustrates the CRISP-DM-based data analysis process, which begins with Business Understanding, namely identifying business needs and problems to be solved through the analysis of customer transaction patterns. The process continues to Data Understanding, which includes data collection, data division according to analysis requirements, and initial visualization to understand the characteristics of transaction time distribution. After that, the data is prepared in the Data Preparation stage through time data cleaning, temporal feature extraction, and encoding and normalization processes so that the data is ready to be used by the clustering algorithm. In the Modeling stage, the study determines the best number of clusters (K), applies the K-Means algorithm, analyzes the distribution of cluster members, and interprets the time profile of each cluster based on the centroid. Next, Evaluation is carried out quantitatively using internal validation metrics and qualitative evaluation through visualization to ensure the quality of the clusters formed. The final stage is Deployment, where the analysis results are translated into operational recommendations that can be used by coffee shops for strategic decision making based on customer transaction time patterns.

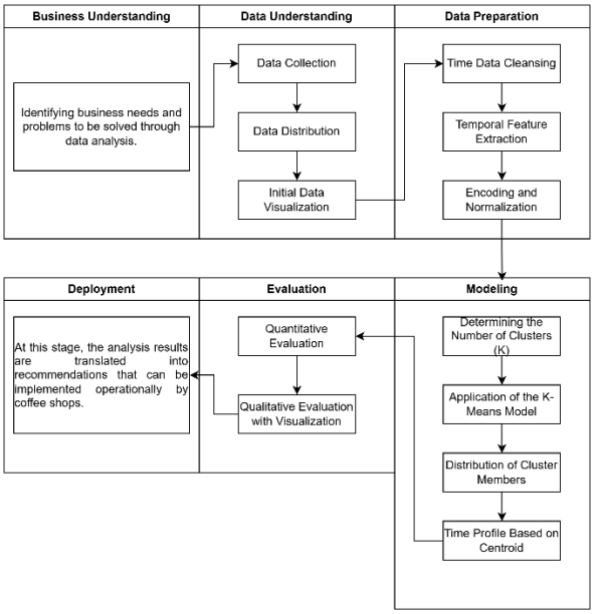

Fig. 1: Research Process

2.3. Business Understanding

The Business Understanding stage aims to understand the operational context of coffee shops and formulate key issues to be resolved through data analysis, particularly related to customer transaction patterns. Business owners need to know the peak transaction hours, daily visit patterns, and the existence of customer groups with specific patterns in order to manage employee numbers, manage stock, develop promotional strategies, and maximize revenue. These issues are then translated into analytical objectives, namely building customer segmentation based on transaction times using unsupervised learning methods, so that the research results not only produce a clustering model, but also provide useful operational insights for coffee shop managers.

2.4. Data Understanding

The Data Understanding stage was conducted using the coffee.csv dataset, which contains coffee sales transaction data along with time attributes such as the date, time, day, and month of the transaction, which are relevant for temporal pattern analysis. The data was obtained from a public repository and imported using Python via the pandas library, then checked for missing values, time format errors, and duplications to ensure data quality. Since the study used unsupervised learning methods, all data were analyzed without separating training and test data, while separation was only performed at the feature engineering level to extract time components such as hour, day, and month. This stage also included initial visualization using transaction hour histograms to understand the distribution of time patterns, which

formed the basis for assessing whether temporal features had strong potential for use in the clustering process. Table 1 displays the first 10 rows of the dataset.

Table 1: Dataset

No	hour_of_day	cash_type	money	coffee_name	Time_of_Day	…	Date	Time
0	10	Card	38.7	Latte	Afternoon		01/03/2024	15:50.5
1	12	Card	38.7	Hot Chocolate	Evening		01/03/2024	19:22.5
2	12	Card	38.7	Hot Chocolate	Evening		01/03/2024	20:18.1
3	16	Card	38.7	Hot Chocolate	Evening		01/03/2024	19:02.8
4	19	Card	38.7	Cocoa	Night		01/03/2024	22:01.8
5	19	Card	33.8	Americano with Milk	Night		01/03/2024	23:15.9
6	10	Card	28.9	Americano	Night		02/03/2024	22:07.0
7	13	Card	38.7	Hot Chocolate	Morning		03/03/2024	09:36.3
8	17	Card	38.7	Cocoa	Morning		03/03/2024	06:40.3
9	17	Card	28.9	Cortado	Morning		03/03/2024	08:45.9

2.5. Data Preparation

The Data Preparation stage is an important process to ensure that data is in optimal condition before being used by the K-Means algorithm. At this stage, the transaction time column is first converted to a datetime format so that temporal features such as hour_of_day, weekday, and month can be extracted to represent customer behavior based on time. Next, categorical variables such as day and month names are converted to numerical format through Label Encoding so that they can be processed by the algorithm. The final stage is normalization using StandardScaler to equalize the scale of all features, because K-Means is very sensitive to differences in value ranges. With this series of steps, the data used becomes more structured and ready to form accurate and representative clusters.

2.6. Modeling

The Modeling stage is carried out by applying the K-Means algorithm to form customer groups based on similarities in transaction patterns. This process begins by determining the optimal number of clusters by testing values of k = 2 to k = 10 using four evaluation metrics, namely Elbow (Inertia/WCSS), Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index, until the best k value is obtained. Once the number of clusters is determined, the K-Means algorithm is trained using the extracted temporal features, where each transaction is grouped based on its proximity to the centroid using Euclidean distance calculated iteratively until the clusters stabilize. Next, the distribution of members in each cluster is analyzed to ensure there are no extreme imbalances that could interfere with interpretation. The final stage is to interpret the inverse scaled centroids to understand the time profile of each cluster, including hourly trends, time categories (morning, afternoon, evening, night), and daily and monthly patterns. The results of this interpretation are then used to describe the behavioral characteristics of customers in each cluster.

2.7. Evaluation

The Evaluation stage is conducted to ensure that the clustering results truly represent customer transaction patterns over time. After all data went through the preprocessing process—from cleaning, time feature extraction, encoding, to normalization—the model was tested by trying k values from 2 to 10 and evaluated using four internal validation metrics, namely Inertia (WCSS), Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. The comparison of metrics shows that k = 3 provides the best balance between cluster separation and cluster structure stability, while larger k values produce patterns that are difficult to interpret. The evaluation continued with centroid analysis to assess the consistency of the emerging time patterns, and the results show that the three clusters have clear temporal characteristics and no significant overlap, with a proportional distribution of members. Overall, the evaluation stage ensured that the K-Means model built had been thoroughly tested and was capable of generating meaningful customer behavior groups based on transaction time patterns.

2.8. Deployment

The Deployment stage focuses on translating the clustering results into practical recommendations that can be used to support coffee shop operations. In this phase, the transaction time patterns found are used to determine the need for additional employees at certain hours, identify periods with high and low transaction intensity, design promotional strategies that match customer visit patterns, and manage raw material stocks based on daily and weekly transaction rhythms. Although this research has not yet produced an automated system, the analytical findings provide a strategic foundation that can be directly utilized in operational decision-making.

3. Results and Discussion

3.1. Data Preparation

The data preprocessing stage is an important step to ensure that the data is in ideal condition before being used in the K-Means algorithm, which is sensitive to format, invalid values, and scale differences. This process includes time data cleaning, temporal feature extraction, as well as encoding and normalization. In the cleaning stage, the Date and Time columns are combined using pd.to_datetime() to form a comprehensive time column. Data with invalid time formats are converted to NaT and removed so as not to interfere with further analysis, ensuring that only temporally valid transaction data is used.

Once the time structure has been cleaned, temporal feature extraction is performed to generate more informative variables, such as hour, minute, hour_frac, weekday, month_name, and Time_of_Day categories (Morning, Afternoon, Evening, Night). These features provide context to consumer behavior patterns and help the K-Means model recognize daily and weekly transaction rhythms more accurately.
The final stage involves encoding and normalization. Categorical features such as weekday, month_name, and Time_of_Day are converted using One-Hot Encoding so they can be processed without creating spurious sequences. All features are then normalized with StandardScaler to equalize the scale, considering that K-Means uses Euclidean distance, which can be distorted if features have different value ranges. The result is a standardized feature matrix (X_scaled).

3.2. Modeling

The Modeling stage begins by determining the most appropriate number of clusters (K) to describe customer transaction patterns over time. The K selection process is carried out by testing k values from 2 to 10 and evaluating them using four internal metrics, namely Elbow (Inertia/WCSS), Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. These four metrics are used to assess the compactness, separation, and stability of the cluster structure. The optimal point is determined when the decline in inertia begins to level off, the silhouette value approaches 1, the DBI is low, and the CH is high—which overall indicates good cluster separation quality. Table 2 shows the determination of the number of clusters (k).

Table 2: Determining the Number of Clusters (k)

Metric	Value
Silhouette Score	0.090839
Davies-Bouldin Index (DBI)	3.174412
Calinski-Harabasz (CH)	104.797598

Once the optimal number of clusters has been determined, the K-Means model is applied to the data that has undergone temporal feature extraction. The algorithm works by grouping transactions based on similarity of time patterns, using Euclidean distance to determine the proximity between data points. This process occurs iteratively until the centroids stabilize and produce three cluster centers that describe dominant time patterns, such as transaction trends in the morning, afternoon, or evening. These centroids serve as “time profiles” that represent the main characteristics of each customer group. Table 3 shows the distribution of the number of cluster members.

Table 3: Distribution of Cluster Members

Cluster	Number of Members
Cluster 0	467
Cluster 1	298
Cluster 2	742

In the next step, we analyze the distribution of members in each cluster and interpret the centroids. Each transaction is labeled with a cluster and the number of members is calculated to ensure that no group is too large or too small, as imbalances can affect the validity of the analysis. The centroids are then converted back from normalized data to the actual time scale so that they can be clearly interpreted. Through this process, researchers can identify the specific characteristics of each cluster, such as the dominant transaction hours, time categories (morning, afternoon, evening, night), and trends on certain days or months. The results of this interpretation produce three main groups, namely the morning, afternoon, and night purchase clusters, which form the basis for customer behavior analysis.

3.3. Evaluation

Visualization of the clustering results is performed to provide a clearer picture of the data grouping patterns based on transaction time, as well as to assess the relative position of each cluster. This stage serves to reinforce the previous numerical analysis results by graphically showing the extent to which clusters are well separated or overlap, so that the cluster structure can be understood more easily and intuitively.

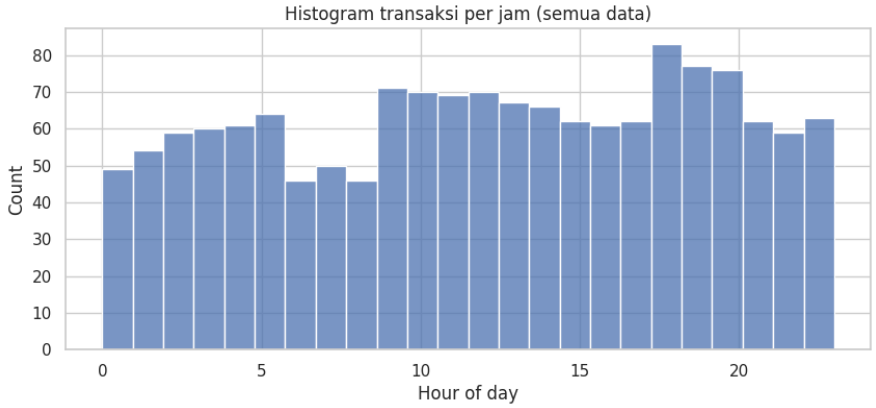

Fig. 2: Transaction Time Histogram (All Data)

The histogram in Figure 2 shows the frequency distribution of transactions based on the time of day, with a pattern that is prominent in two main periods. The peak in transactions occurs at night, particularly between 6 and 8 p.m., when the number of transactions reaches its highest level, possibly influenced by after-work activities. A secondary peak appears between 9 and 11 a.m. as morning activity increases towards noon. Conversely, the lowest number of transactions occurs in the early morning until early morning, especially between 6 and 8 a.m. Overall, the trend shows a gradual increase from midnight to noon, stabilizing at a moderate level during the day, and rising sharply again in the afternoon until evening before slowly declining towards midnight.

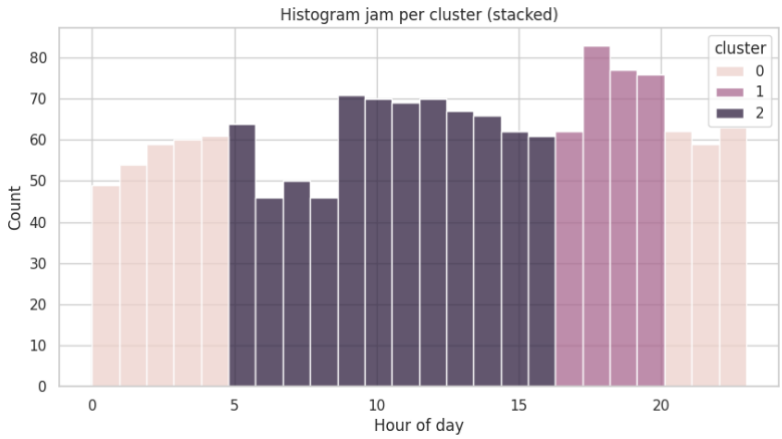

Fig. 3: Hourly Histogram per Cluster

The histogram visualization in Figure 3 shows a very clear distribution of transaction times, with Cluster 2 dominating the main productive hours (06:00–16:00) with a stable activity pattern and a peak at 09:00–11:00. Cluster 1 dominates the evening peak hours (17:00–20:00) with the highest transaction frequency around 18:00, and Cluster 0 operates during marginal hours, namely early morning (00:00–05:00) and late night (21:00–23:00) with lower transaction volumes. This segmentation illustrates three distinct user behavior patterns: routine activity during working hours, peak activity after work, and limited activity outside of peak hours.

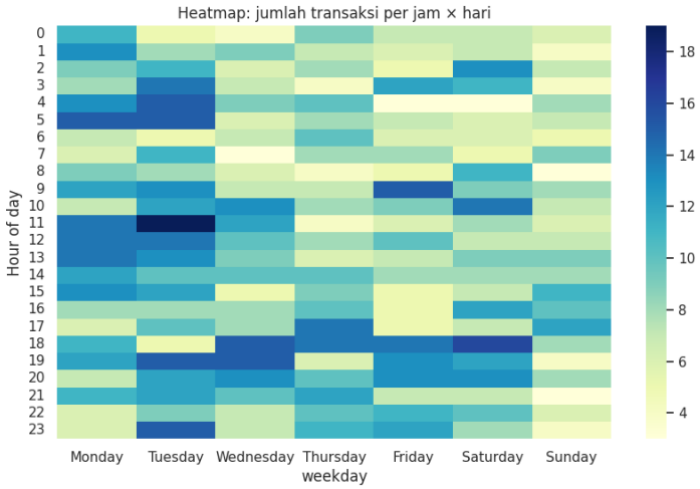

Fig. 4: Heatmap of Distribution by Hour x Day

Figure 4 shows the transaction density pattern based on time and day, with dark blue indicating the highest volume and light yellow the lowest volume. Overall, transactions peaked during two main periods, namely 10:00 a.m.–12:00 p.m. and 6:00 p.m.–8:00 p.m., while the lowest activity occurred in the early hours of the morning (12:00 a.m.–5:00 a.m.). The weekly distribution also varies, with Tuesday, Wednesday, and Saturday appearing to be the busiest days with specific peaks at certain hours, while Sunday, Thursday, and Friday show lower intensity. Several anomalies appear, such as early morning spikes on Tuesdays and Fridays, which are likely influenced by automated processes or other external factors, as well as a shift in patterns on weekends, which tend to be active in the morning and evening.

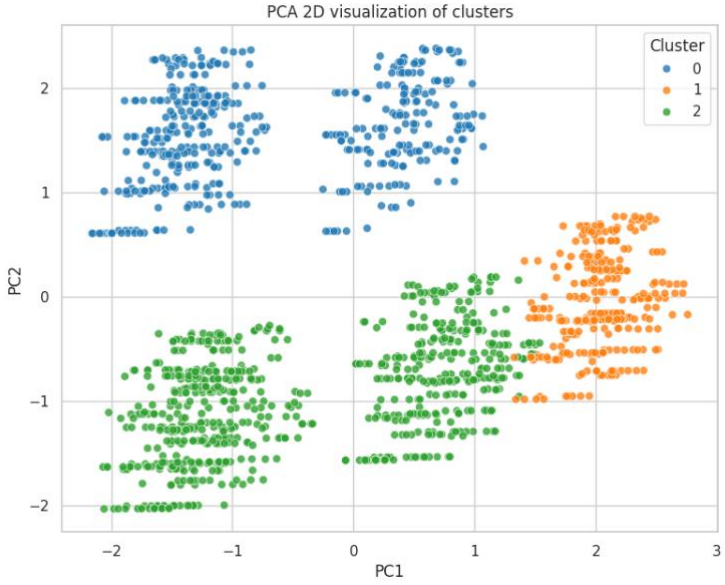

Fig. 5: PCA Visualization

The scatter plot in Figure 5 shows the results of 2D PCA, which visualizes three clusters (0, 1, and 2) formed from the clustering process, with PC1 and PC2 as the two main components representing the greatest variance in the data. The visualization shows a very clear separation of clusters, where PC1 plays a major role in separating Cluster 1 on the high positive side, while Clusters 0 and 2 are in the lower value range; while PC2 separates Cluster 0 in the high positive area and Cluster 2 in the negative area, with Cluster 1 being around the zero point. These positions illustrate the unique characteristics of each cluster, where Cluster 1 has extreme values on PC1, Cluster 0 stands out on positive PC2, and Cluster 2 dominates on negative PC2. Overall, this clear separation shows that the clustering model has successfully formed distinct groups that are consistent with the latent structure in the data.

This discussion highlights transaction patterns based on time and day to answer the research objective, which is to understand user behavior in transactions. The analysis begins with an hourly histogram showing two main peaks at 9:00 a.m.–12:00 p.m. and 5:00 p.m.–8:00 p.m., as well as the lowest activity at 3:00 a.m.–6:00 a.m., reflecting the daily rhythm of users who follow work and leisure patterns. These findings are then explored further through cluster-based segmentation, which clarifies behavioral differences where Cluster 0 is active at night and in the early morning, Cluster 1 is dominant during working hours, and Cluster 2 is more active in the afternoon and evening. This pattern is further reinforced by the hour × day heatmap visualization, which illustrates the consistency of weekly activity, with the highest intensity on weekdays—particularly Tuesdays and Wednesdays—and a shift in activity on weekends, which tends to increase in the afternoon–evening. Overall, this series of analyses shows that transactions follow clear and recurring temporal patterns, both daily and weekly, as well as based on user behavior segments, providing strategic insights for service optimization and time-based operational planning.

4. Conclusion

1. Coffee transaction patterns show a clear temporal structure, with the highest activity in the morning–afternoon and evening–night, and a significant decline in the early hours of the morning.
2. The K-Means method effectively segments purchase times, producing three consistent clusters: night-early morning buyers, working hour buyers, and afternoon-evening buyers for leisure activities.
3. Clustering successfully reveals latent patterns of consumer behavior, while providing important insights for operational optimization, inventory management, and time-based marketing strategies.

Based on the results of the study, several suggestions can be made, namely the need for businesses to adjust their operations during peak transaction times by increasing service capacity, fulfilling raw material requirements, and optimizing workflows. The clustering results can also be used to design more targeted marketing strategies according to the characteristics of each customer group, such as promotions during off-peak hours or special packages for clusters with afternoon-evening activity. Further research is recommended to add analysis variables, evaluate alternative clustering algorithms such as Hierarchical, DBSCAN, or GMM, and develop machine learning-based predictive models to project transaction volumes. In addition, the sales information system needs to be equipped with temporal analytics and real-time dashboard features to support decision making. These recommendations are expected to enrich practical benefits and encourage more comprehensive research on time-based purchasing patterns.

References

[1] D. Giordano, M. Mellia, and T. Cerquitelli, "K-MDTSC: K-multi-dimensional time-series clustering algorithm," *Electronics*, vol. 10, no. 10, p. 1166, 2021.

[2] H. Liu, J. Wang, R. Zhang, and O. Liu, "Does the weather still affect me when I shop at home? The impact of weather on online shopping behavior," *J. Theor. Appl. Electron. Commer. Res.*, vol. 19, no. 3, pp. 2289–2311, 2024.

[3] Z. Liang, "Wake up and search for coffee: Considering the circadian rhythm of consumers on online purchase behavior," *J. Bus. Res.*, vol. 157, p. 113536, 2024.

[4] K. Tabianan, S. Velu, and V. Ravi, "K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, p. 7243, 2022.

[5] D. A. Awaliyah, B. Prasetiyo, R. Muzayanah, and A. D. Lestari, "Optimizing customer segmentation in online retail transactions through the implementation of the K-Means clustering algorithm," *Sci. J. Informatics*, vol. 11, no. 2, 2024.

[6] A. Silva and M. Wang, "Visual analysis of e-commerce user behavior based on time series logs," *J. Sensors*, vol. 2022, p. 4291978, 2022.

[7] R. Autio and et al., "Tensorial principal component analysis in detecting temporal trajectories of purchase patterns in loyalty card data: Retrospective cohort study," *J. Med. Internet Res.*, vol. 25, no. 1, p. e44599, 2023.

[8] C. Fan, M. Chen, and X. Wang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Front. Energy Res.*, vol. 9, 2021.

[9] J. Wen, L. Guillen, T. Abe, and T. Suganuma, "A hierarchy-based system for recognizing customer activity in retail environments," *Sensors*, vol. 21, no. 14, p. 4712, 2021.