Purchasing Behavior Based Consumer Segmentation on TikTok Shop in Indonesia Using K-Means

Wulan Novita Sari[1)], dan Arief Wibowo[1)]

2531600118@student.budiluhur.ac.id, arief.wibowo@budiluhur.ac.id

[1)] Universitas Budi Luhur

ABSTRACT

This study aims to analyze consumer segmentation on TikTok Shop in Indonesia based on purchasing behavior. The rapid growth of social commerce, particularly through TikTok Shop, has changed consumer shopping patterns, creating challenges for businesses in understanding diverse consumer characteristics to design effective marketing strategies.

This study uses a quantitative approach with primary data collected through questionnaires distributed to TikTok Shop users in Indonesia. The variables used include Age (X1), Purchase Frequency (X2), and Type of Product Purchased (X3). The data was analyzed using the K-Means clustering method to classify consumers into homogeneous segments. The results of the study show that TikTok Shop consumers can be grouped into several different segments with different purchasing patterns, spending levels, and product preferences. These findings have practical implications for developing targeted and personalized marketing strategies.

Keywords*: TikTok Shop, Consumer Segmentation, K-Means Clustering, Purchasing Behavior*

INRODUCTION

The development of digital transformation in the modern era has brought significant changes to consumer behavior. The Internet no longer functions solely as a medium for information and communication; instead, it has transformed into a primary platform for conducting digital economic activities known as e-commerce. The role of e-commerce has become increasingly dominant as a means for consumers to fulfill their needs, particularly in Indonesia, which has experienced rapid growth in the number of Internet users and the utilization of digital platforms (Krisnanto, 2025).

This transformation has shifted consumer behavior toward digital purchasing, driven by convenience, accessibility, and efficiency, which significantly influence consumers' online decision-making processes.

Fgure 1 Graph of the Number of E-Commerce Users in Indonesia for the periode 2020-2029

Source: Kementrian Perdagangan, 2024

The COVID-19 pandemic serves as a concrete example of this transformation, during which online shopping and digital payment transactions became increasingly prevalent. Even after the pandemic, these behavioral patterns persisted and evolved into a new lifestyle. As a result, the number of e-commerce users increased significantly from 2020 to 2023, reaching 58.63 million, and is projected to grow to approximately 99.1 million users by 2029 (Kemendag, 2024). This transformation presents opportunities for the government to encourage the optimization of digital platforms as a driver of economic growth.

Based on data from the Indonesia Internet Service Providers Association (APJII), TikTok Shop ranks as the second most accessed e-commerce platform in Indonesia after Shopee, followed by Tokopedia, Lazada, Blibli, and Facebook Marketplace (Databoks, 2025). However, TikTok Shop is relatively new in Indonesia's e-commerce ecosystem and does not yet possess a market share comparable to Shopee, resulting in lower user conversion into active buyers. As a social commerce platform, TikTok Shop integrates entertainment, social interaction, and purchasing functions, influencing consumer engagement and purchase intentions (Annuar, 2023).

This condition requires businesses to adapt marketing strategies to the unique characteristics of social commerce.

In increasingly competitive market conditions, consumer segmentation becomes a strategic approach for identifying consumer groups based on behavioral and demographic characteristics. Through segmentation, companies can better understand differences in consumer needs and preferences, allowing marketing strategies to be designed more effectively. Segmentation based on online shopping behavioral tendencies is particularly relevant, as psychological differences among consumers influence their responses to promotional stimuli and digital content (Budi, 2025).

From a methodological perspective, the K-Means algorithm is a clustering technique that groups data into several clusters based on similarity, with each cluster represented by a centroid (Handoko, 2025). Previous studies have widely applied K-Means in various contexts, Fahrozi et al., (2023) implemented the K-Means algorithm to cluster customer satisfaction data in a health training institution and demonstrated that K-Means was able to group customers into distinct segments based on service evaluation patterns. Their findings indicate that K-Means is suitable for identifying consumer segments that can support managerial decision-making. Numerous studies have applied the K-Means algorithm in various domains, including weather data analysis (F. D. Rahman et al., 2024), Alfitra & Meiyanti, (2025) found that K-Means is more efficient in handling large datasets with numerical attributes. This supports the use of K-Means as an effective clustering method in large-scale behavioral data analysis, The classification of participants in certification programs and science Olympiads with satisfactory accuracy to support data-driven decision making (Purwatiningsih & Habibi, 2024), as well as the clustering of corruption-related criminal cases in Indonesia (Saputro et al., 2020), F. K. Rahman, (2025) applied K-Means clustering to identify machine condition patterns in industrial

processes, while Sonianto & Hartono, (2025) used K-Means to identify students' interests and talents in elementary education. These studies show that K-Means is a flexible and robust method for pattern recognition in different types of data, supporting its application in consumer behavior analysis. These studies demonstrate that K-Means is a flexible and robust method for identifying patterns in large numerical datasets.

However, existing studies predominantly focus on conventional e-commerce platforms or non-commercial domains, while research specifically addressing TikTok Shop users in Indonesia as a social commerce platform remains limited. Most previous research does not incorporate the unique behavioral characteristics associated with social interaction–driven purchasing. Therefore, the originality of this study lies in applying the K-Means clustering method to segment TikTok Shop consumers in Indonesia based on their online shopping behavioral tendencies. This research fills the empirical gap in social commerce studies and provides a data-driven foundation for developing more targeted marketing strategies in emerging digital platforms.

RESEARCH METHODOLOGY

This study employs a quantitative approach using a data mining method based on unsupervised learning. The algorithm applied in this research is K-Means Clustering, which aims to group TikTok Shop consumers based on their purchasing behavior tendencies according to predetermined variables.

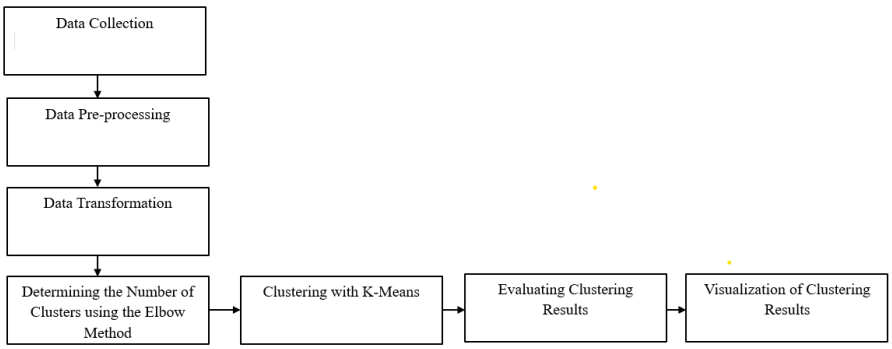
The object of this study is TikTok Shop users in Indonesia who have made at least one purchase through the platform. The population of this research consists of all TikTok Shop users in Indonesia. The sample size in this study is 50 respondents, selected using purposive sampling technique, with criteria that respondents are active TikTok Shop users and have experience in online shopping through the platform. Due to the exploratory nature of this study, a limited sample size was used as an initial segmentation attempt.

Primary data were collected through an online questionnaire distributed to respondents. The variables used include purchase frequency, transaction value, product categories, and intensity of TikTok Shop usage.

Data analysis was conducted using the K-Means Clustering method combined with the Elbow Method. K-Means is used to classify consumers into several clusters based on similarity in purchasing behavior, while the Elbow Method is applied to determine the optimal number of clusters. This combination enables the formation of homogeneous segments within clusters and heterogeneous segments between clusters, ensuring that the resulting consumer segmentation is accurate, interpretable, and relevant for marketing decision making.

The determination of the optimal number of clusters is a critical step in K-Means analysis. Marsa & Ritonga, (2025) used the Elbow Method to identify the appropriate number of clusters in salary data clustering and found that this approach helps balance model simplicity and explanatory power. Similarly, Rafika et al., (2025) demonstrated that combining K-Means with the Elbow Method improves clustering accuracy and interpretability. Therefore, this study adopts the Elbow Method to determine the optimal number of clusters before applying the K-Means algorithm.Research Stages.

Figure 2 Stages of the Research Method



Source: RapidMiner Version 2026 Data Processing Result

The following section describes the stages of the research methodology:

1. Data Collection

   The research instrument consisted of a structured questionnaire designed to collect data on the characteristics of TikTok Shop consumers. The variables collected included respondents' age (X1), purchase frequency (X2), and the type of product most frequently purchased (X3). The questionnaire was distributed to TikTok Shop users, and the collected data were used as primary data for this study.

2. Data Pre-processing

   At this stage, the data were cleaned and prepared to ensure suitability for analysis. The procedures included:

   a. Removing duplicate records.
   b. Handling missing and invalid data.
   c. Converting the "most frequently purchased product type" variable into numerical format.
   d. Anonymizing respondents' identities to ensure data confidentiality

3. Data Transformation

   In the subsequent stage, the data were processed through a transformation procedure, particularly for large-scale datasets. This step aimed to improve the quality of the analysis and facilitate computation.

   Data transformation was performed using Z-Score Standardization, a feature scaling technique that changes each variable value so that it has a mean of zero and a standard deviation of one, so that all features are on a uniform scale for further analysis. The Z-score values can be positive or negative: negative and positive signs represent an observation below or above the mean, respectively (Wongoutong, 2024). The formula for normalization used in this study is (Wongoutong, 2024):

$$X' \frac{X - \mu}{\sigma}$$

Keterangan:

X : Original Data Value

X′ : Mean

σ : Standard deviation of the variable

4. Determination of the Optimal Number of Clusters

   The Elbow Method was employed to determine the optimal number of clusters by calculating the Within-Cluster Sum of Squares (WCSS) and plotting it against the number of clusters. The optimal value of *k* was selected at the "elbow point" of the graph, where the reduction in WCSS begins to diminish.

5. Implementation of the K-Means Clustering Algorithm

   The stages of applying the K-Means algorithm included:

   a. Determining the number of clusters (*k*) based on the Elbow Method.
   b. Automatically initializing the centroids.
   c. Calculating the distance between data points and centroids and assigning each data point to the nearest cluster.
   d. Updating centroid positions iteratively until convergence is achieved.

6. Evaluation of Clustering Results

   This study evaluated the clustering results using the Davies–Bouldin Index (DBI) and the Elbow Method.

   a. Davies–Bouldin Index (DBI)

   The Davies–Bouldin Index was used to assess clustering quality based on the similarity between clusters and the compactness of data points within each cluster. The DBI was calculated after applying the K-Means algorithm with the predefined number of clusters. This evaluation considers both inter-cluster distance and intra cluster distance, aiming to maximize the distance between clusters while minimizing the distance among data points within the same cluster.

   The computation begins with calculating the Sum of Squares Within Cluster (SSW) and the

Sum of Squares Between Clusters (SSB). SSW measures the dispersion of data points within a cluster relative to its centroid, while SSB measures the distance between centroids of different clusters.

The formula SSW used in this study is (Khotimah et al., 2024):

$$\mathrm{SSW} = \frac{1}{\mathrm{m}} \sum_{\mathrm{j=i}}^{\mathrm{mi}} \mathrm{d}\,(\mathrm{x_i C_i})$$

where m is the number of data points in cluster i, x represents the data attributes, C_i is the centroid of cluster i, and d(x, C) denotes the distance between a data point and the centroid.

The formula SSB used in this study is (Khotimah et al., 2024):

$$\mathrm{SSB_{i,j}} = \mathrm{d(C_i, C_j)}$$

where C_i and C_j are the centroids of clusters i and j, respectively.

A large SSB indicates well-separated clusters, whereas a small SSB suggests overlapping clusters.

The ratio for each cluster pair used in this study is (Khotimah et al., 2024):

$$\mathrm{R_{i,j}} = \frac{\mathrm{SSWi + SSWj}}{\mathrm{SSBij}}$$

Where $R_{i,j} \geq 0$and $R_{i,j} = R_{j,i}$. If SSW increases while SSB remains constant, clustering quality deteriorates. Likewise, if SSB decreases while SSW remains constant, clusters become increasingly overlapping.

The formula Davies–Bouldin Index used in this study is (Khotimah et al., 2024):

$$\mathrm{DBI} = \frac{1}{\mathrm{K}} \sum_{\mathrm{i=1}}^{\mathrm{k}} \max(\mathrm{R_i,j})$$

where K is the number of clusters and $R_{i,j}$represents the ratio of intra-cluster dispersion to inter-cluster separation.

b. Elbow Method

The Elbow Method is a commonly used technique for identifying the optimal number of clusters in clustering analysis. It evaluates the WCSS for each potential number of

clusters (k) and plots the results in a graph. The relationship between the number of clusters and the WCSS typically shows a decreasing trend; however, at a certain point, the rate of decrease becomes marginal, forming an "elbow" shape. This point is considered the optimal number of clusters, as it reflects a balance between model complexity and clustering quality.

7. Visualization of Analysis Results

   The clustering results were visualized using several graphical representations, including cluster distribution plots, age distribution per cluster, purchase frequency distribution per cluster, and product type distribution per cluster. These visualizations were used to facilitate the interpretation of patterns and characteristics within each cluster in the analysis of TikTok Shop consumer segmentation.

RESULTS AND DISCUSSION

The results of this study were obtained from the process and main findings of the clustering analysis. The optimal number of clusters was determined using the Elbow Method, using the Within-Cluster Sum of Squares (WCSS) and Davies Bouldin Index (DBI) values, where the elbow point on the graph indicated that the best number of clusters was four. The identification of four optimal clusters in this study occurred because the consumer behavior data formed four distinct groups with similar patterns in key variables such as purchase frequency, transaction value, usage intensity, and product preferences. This is in line with researchers Martí et al., (2021) The Elbow Method and Davies–Bouldin Index (DBI) jointly indicate the most suitable number of clusters by balancing within-cluster cohesion and between-cluster separation; the elbow point at k = 4 suggests that adding additional clusters beyond four yields minimal improvement in explaining variance. The four distinct clusters emerge because TikTok Shop users exhibit heterogeneous purchasing patterns

influenced by their levels of engagement and exposure to interactive content, consistent with findings that social commerce (Laradi et al., 2024).

Through this validation stage, the algorithm that showed the best clustering performance and was most suitable for the data characteristics could be identified.

Data Collection

The first stage of this study involved collecting data as a basis for further analysis. The dataset used was obtained from a questionnaire distributed to 50 respondents, which was converted into numerical data as shown in the following table:

Table 1 TikTok Shop Consumer DataSet in Indonesia

Responden	Age (X1)	Purchase Frequency (X2)	Type of Product Purchased (X3)
R1	2	2	1
R2	2	2	2
R3	1	3	1
R4	1	3	3
R5	3	1	3
R6	2	2	1
R7	3	1	3
R8	1	2	2
R9	3	1	1
R10	1	3	4
R11	3	1	4
R12	2	2	4
R13	2	2	5
R14	1	3	4
R15	2	2	5
R16	2	2	5
R17	2	2	5
R18	3	1	6
R19	3	1	5
R20	2	2	5
R21	1	3	6
R22	1	3	6
R23	2	2	2

R24	2	2	4
R25	1	2	8
R26	3	1	9
R27	3	1	9
R28	2	2	9
R29	3	1	8
R30	1	3	9
R31	2	2	8
R32	1	3	9
R33	2	2	9
R34	1	3	9
R35	2	2	7
R36	1	2	7
R37	1	3	7
R38	3	1	7
R39	3	1	7
R40	3	1	7
R41	1	2	7
R42	1	3	9
R43	3	1	1
R44	3	1	4
R45	1	3	4
R46	1	3	9
R47	2	2	5
R48	2	2	5
R49	3	1	9
R50	2	3	2

Source: Results of RapidMiner Data Processing Version 2026

Data Transformation

The next stage involves data modeling, which serves to facilitate analysis and improve data quality and algorithm performance. One of the steps is the normalization transformation process, which aims to equalize the scale between variables for more accurate and efficient analysis. The normalization method used Z-score standardization, which transforms each variable into a common scale with zero mean and unit variance to eliminate scale differences among attributes. This process is particularly important for distance-based clustering algorithms such as K-Means, as it ensures that all variables contribute equally to the distance computation and prevents

bias caused by variables with larger numerical ranges.

The data were normalized using Z-score standardization to transform all variables into a common scale with zero mean and unit variance, ensuring equal contribution of each variable in distance-based clustering and preventing scale-related bias in the K-Means algorithm.

Table 2 Results of TikTok Shop Consumer Dataset Normalization

Responden	Age (X1)	Purchase Frequency (X2)	Type of Product Purchased (X3)	Age (X1)	Purchase Frequency (X2)	Type of Product Purchased (X3)
R1	2	2	1	-0,4531	-0,6008	-1,0658
R2	2	2	2	-0,4531	-0,6008	-0,0658
R3	1	3	1	-1,4531	0,3992	-1,0658
R4	1	3	3	-1,4531	0,3992	0,9342
R5	3	1	3	0,5469	-1,6008	0,9342
R6	2	2	1	-0,4531	-0,6008	-1,0658
R7	3	1	3	0,5469	-1,6008	0,9342
R8	1	2	2	-1,4531	-0,6008	-0,0658
R9	3	1	1	0,5469	-1,6008	-1,0658
R10	1	3	4	-1,4531	0,3992	1,9342
R11	3	1	4	0,5469	-1,6008	1,9342
R12	2	2	4	-0,4531	-0,6008	1,9342
R13	2	2	5	-0,4531	-0,6008	2,9342
R14	1	3	4	-1,4531	0,3992	1,9342
R15	2	2	5	-0,4531	-0,6008	2,9342
R16	2	2	5	-0,4531	-0,6008	2,9342
R17	2	2	5	-0,4531	-0,6008	2,9342
R18	3	1	6	0,5469	-1,6008	3,9342
R19	3	1	5	0,5469	-1,6008	2,9342
R20	2	2	5	-0,4531	-0,6008	2,9342
R21	1	3	6	-1,4531	0,3992	3,9342
R22	1	3	6	-1,4531	0,3992	3,9342
R23	2	2	2	-0,4531	-0,6008	-0,0658
R24	2	2	4	-0,4531	-0,6008	1,9342
R25	1	2	8	-1,4531	-0,6008	5,9342
R26	3	1	9	0,5469	-1,6008	6,9342
R27	3	1	9	0,5469	-1,6008	6,9342
R28	2	2	9	-0,4531	-0,6008	6,9342
R29	3	1	8	0,5469	-1,6008	5,9342
R30	1	3	9	-1,4531	0,3992	6,9342

R31	2	2	8	-0,4531	-0,6008	5,9342
R32	1	3	9	-1,4531	0,3992	6,9342
R33	2	2	9	-0,4531	-0,6008	6,9342
R34	1	3	9	-1,4531	0,3992	6,9342
R35	2	2	7	-0,4531	-0,6008	4,9342
R36	1	2	7	-1,4531	-0,6008	4,9342
R37	1	3	7	-1,4531	0,3992	4,9342
R38	3	1	7	0,5469	-1,6008	4,9342
R39	3	1	7	0,5469	-1,6008	4,9342
R40	3	1	7	0,5469	-1,6008	4,9342
R41	1	2	7	-1,4531	-0,6008	4,9342
R42	1	3	9	-1,4531	0,3992	6,9342
R43	3	1	1	0,5469	-1,6008	-1,0658
R44	3	1	4	0,5469	-1,6008	1,9342
R45	1	3	4	-1,4531	0,3992	1,9342
R46	1	3	9	-1,4531	0,3992	6,9342
R47	2	2	5	-0,4531	-0,6008	2,9342
R48	2	2	5	-0,4531	-0,6008	2,9342
R49	3	1	9	0,5469	-1,6008	6,9342
R50	2	3	2	-0,4531	0,3992	-0,0658

Source: Results of RapidMiner Data Processing Version 2026

Based on the following table, it can be concluded that the initial data has been successfully normalized using the Z-Score method as explained earlier, as the transformed values are centered around zero with varying positive and negative scores. This indicates that each variable has been standardized to ensure balanced contribution in the clustering process. Thus, the scale differences between the Age (X1), Purchase Frequency (X2), and Type of Product Purchased (X3) attributes no longer affect the analysis, allowing the algorithms in the next stage to perform optimally. Therefore, the normalized data is considered suitable for further processing.

Despite its contributions, this study has several limitations. Due to the exploratory nature of this study, a limited sample size was used as an initial segmentation attempt, which may limit the generalizability of the results.

Elbow Method Implementation

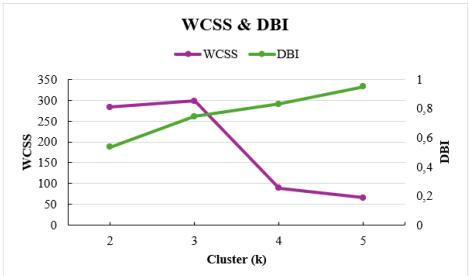
Before the clustering process is carried out, it is necessary to first

determine the most appropriate number of clusters to optimize the analysis results. The number of clusters is determined using the Elbow method by calculating the Within-Cluster Sum of Squares (WCSS) value for several possible cluster numbers. As the number of clusters increases, the WCSS value will decrease until, at a certain point, the rate of decline begins to plateau. This point is known as the elbow point and is used as the basis for determining the optimal number of clusters. The WCSS value will continue to decrease, but at a certain point, the decline begins to slow down. This point is called the elbow point and is considered the optimal number of clusters. In general, the Elbow Method is used to determine the most efficient number of clusters by comparing clustering results at various k values.

The elbow point on the graph indicates a balance between too few and too many clusters, allowing the data structure to be optimally represented without compromising its original meaning. Determining the number of clusters is a crucial step before performing the clustering process using the K-Means algorithm.

By knowing the optimal number of clusters, the clustering process can run more efficiently and produce groupings that represent the true condition of TikTok Shop's consumer segmentation. In this study, the results of the Elbow Method calculation indicate that the elbow point is at k=4, meaning that TikTok Shop's consumer segmentation based on purchasing behavior is most appropriately grouped into four clusters, as shown in the following figure:

Figure 3 Graph of the Relationship between WCSS & DBI in the Elbow Method for Determining the Number of Clusters



Source: Results of RapidMiner Data Processing Version 2026

Based on the graph showing the relationship between Within-Cluster Sum of Square (WCSS) and Davies-Bouldin Index (DBI) against the number of clusters (k), it can be seen that the two evaluation metrics show opposite patterns.

The WCSS value decreases as the number of clusters increases, starting from around 290 at k=2, falling to around 300 at k=3, then decreasing dramatically to around 95 at k=4, and continuing to decrease to around 65 at k=5. The most significant decrease in WCSS occurs between k=3 and k=4, where there is a decrease of around 200 points, indicating an elbow point at k=4. In contrast, the DBI value shows an increasing trend, where at k=2 the value is around 0.55, rising to 0.75 at k=3, continuing to increase to around 0.85 at k=4, and reaching its peak at around 0.95 at k=5.

Based on the elbow method, k=4 is the optimal choice because at that point there is the sharpest decline in WCSS before the curve begins to flatten. Although the DBI value at k=4 is relatively high at around 0.85, the increase in DBI after k=4 is not very significant compared to the substantial decrease in WCSS. This shows that k=4 provides the best balance between compact clusters, as indicated by low WCSS, and acceptable separation between clusters, making it the most appropriate number of clusters for this data grouping. In addition, the opposite trends between WCSS and DBI reflect a trade-off in clustering evaluation. The sharp decrease in WCSS at k=4 indicates that this point captures the main data structure, while the DBI value remains acceptable. Therefore, k=4 represents a balanced and interpretable clustering solution.

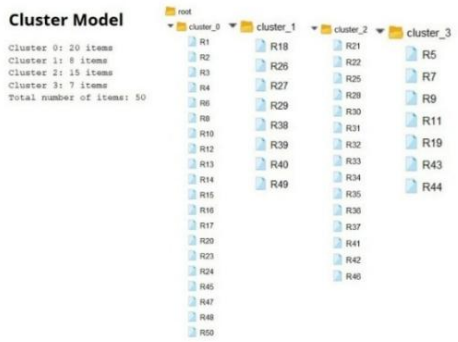
Implementation of K-Means with RapidMiner

In implementing the K-Means algorithm using RapidMiner software, researchers first prepared a dataset in Excel (.xlsx) file format containing data from 50 respondents. The dataset was then imported into

RapidMiner and used as input in the clustering modeling process.

Each attribute was checked to ensure it was in numerical format, and preprocessing steps such as data cleaning and normalization were applied prior to clustering. Subsequently, the K-Means operator was executed by specifying the number of clusters based on the Elbow Method results, allowing the algorithm to group respondents into homogeneous segments based on their purchasing behavior.

Figure 4 Clustering Results in the Rapidminer Application



Source: Results of RapidMiner Data Processing Version 2026

In the figure, the clustering process was performed using the K-Means operator with a set number of four clusters. Next, the Performance operator (Cluster Distance Performance) was used to evaluate the quality of the clustering results. Based on the clustering results, Cluster 0 has a distribution of 20 items, Cluster 1 has a distribution of 8 items, Cluster 2 has a distribution of 15 items, and Cluster 3 has a distribution of 7 items.

The difference in the number of members between clusters indicates that the data is not evenly distributed but rather concentrated in certain groups. Cluster 0 is the largest cluster, so it can be interpreted as the dominant segment in the data population. Conversely, Cluster 3 has the fewest members and can be considered a minority segment with more specific characteristics. These clustering results demonstrate that the K-Means algorithm successfully

identified natural patterns in the data and divided it into distinct groups.

Evaluation of Clustering Results

The evaluation of the number of clusters was performed using a combination of the Elbow and Davies–Bouldin Index methods. The Elbow method showed an elbow point at k=4, marked by a very significant decrease in WCSS compared to the previous k value, as shown in the following table:

Table 3. Calculation Results DBI &WCSS

Cluster	DBI	WCSS
2	0,537272	282,9851
3	0,745884	298,73
4	0,830807	89
5	0,951443	66,01176

Source: Results of RapidMiner Data Processing Version 2026

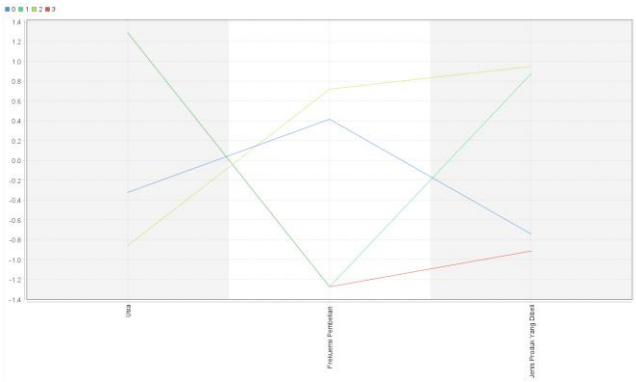
Based on the table, the DBI value shows an increasing trend as the number of clusters increases, from around 0.75 at k=3 to 0.83 at k=4 and increasing again to 0.95 at k=5. Considering that a lower DBI value indicates better cluster quality, the increase in DBI after k=3 indicates that the separation between clusters becomes less optimal when the number of clusters is too large. Conversely, the WCSS value decreases significantly when the number of clusters increases from k=3 to k=4, from around 298.73 to 89. This decrease indicates a substantial increase in cluster compactness. After k=4, the decrease in WCSS becomes relatively small, to around 66.01 at k=5, indicating that the addition of further clusters does not provide a significant improvement in quality. Based on these two relationship indicators, k=4 was selected as the most optimal number of clusters because it provided a significant decrease in WCSS while still maintaining the DBI value within acceptable limits, thus resulting in a balance between cluster compactness and separation between clusters. Therefore, k=4 was selected as the optimal number of clusters, as it provides a

meaningful improvement in compactness while maintaining acceptable cluster separation, resulting in a balanced and interpretable segmentation outcome.

Visualization of Analysis Results

Clustering results are visualized using a cluster profile graph, which shows the standardized average values of the variables age, purchase frequency, and product type purchased in each cluster. This graph is used to facilitate interpretation of differences in characteristics between clusters, as shown in the following figure:

Figure 5: Parallel Coordinates Plot graph



Source: Results of RapidMiner Data Processing Version 2026

The Parallel Coordinates Plot graph shows the differences in characteristics between clusters based on three main variables. Each colored line represents a cluster, and the position of the line on each variable shows the average value of that cluster relative to the overall data. Thus, this graph makes it easier for researchers to visually compare behavior patterns between clusters. Based on the clustering results presented in the image, the following interpretation can be made:

a. Cluster 0 (Blue Line) – Regular Focused Buyer
   This cluster consists of consumers or groups who are relatively younger, most active in terms of transactions (frequent shoppers), but the products they buy tend to be homogeneous or repetitive.

They are most likely buyers of basic/routine necessities

b. Cluster 1 (Green Line) - Selective Senior Buyers
This cluster consists of older consumers or groups who rarely make transactions, but when they do buy, they tend to choose very specific or varied types of products.

c. Cluster 2 (Yellow Line) - Young Heavy Explorers
This cluster consists of active young consumers or market segments who prefer to explore various types of products when shopping.

d. Cluster 3 (Red Line) - Low Engagement Users
This cluster consists of consumers or groups that are the least active. They rarely shop and only purchase a very limited range of products.

CONCLUSION

The application of the K-Means algorithm was able to group respondents into four clusters with different characteristics, which were mainly influenced by the variables of age, purchase frequency, and type of product purchased. The clustering results visualized using a parallel coordinates plot proved effective in revealing complex consumer behavior patterns, allowing differences between segments to be identified more clearly. The visualization shows significant behavioral variations, where young consumers in Cluster 2 tend to be more exploratory towards various types of products, while older consumers in Cluster 1 are more selective. In addition, Cluster 0 has the potential to be targeted with purchase-based promotions, while Cluster 3 requires a specific strategy to increase customer engagement and loyalty.

This study has limitations in that the number of respondents is relatively small and only three variables were used, so further research is recommended to add variables such as income, gender, and purchase motivation.

REFERENCES

Alfitra, A., & Meiyanti, R. (2025). Comparison Of K-Means And K-Medoids Methods In Clustering High Population Density Areas In Bireuen Regency. 9(July), 292–302.

Annuar, N. S. A. & S. N. S. (2023). The Adaptation Of Social Media Marketing Activities In S-Commerce: Tiktok Shop. 15(1), 176–183.

Budi, F. S. (2025). Segmentasi Konsumen Tiktok Shop Berdasarkan Perilaku Pembelian Impulsif Menggunakan K-Means Clustering Segmentation Of Tiktok Shop Consumers Based On Impulse Buying Behavior Using K-Means Clustering. 18(2), 1–9.

Databoks, A. (2025). E-Commerce Yang Sering Diakses Masyarakat Indonesia Pada 2025. Databoks. Https://Databoks.Katadata.Co.Id/Infografik/2025/08/19/E-Commerce-Yang-Sering-Diakses-Masyarakat-Indonesia-Pada-2025

Fahrozi, A. Al, Insani, F., Budianita, E., & Afrianty, I. (2023). Implementasi Algoritma K-Means Dalam Menentukan Clustering Pada Penilaian Kepuasan Pelanggan Di Badan Pelatihan Kesehatan Pekanbaru. 1, 474–492.

Handoko, K. (2025). K-Means Clustering Menggunakan Aplikasi Rapidminer (Efitra & I. K. Sari (Eds.)). Pt. Sonpedia Publishing Indonesia. Https://Www.Google.Co.Id/Books/Edition/K_Means_Clustering_Menggunakan_Aplikasi/Dqbjeqaaqbaj?Hl=Id&Gbpv=0

Kemendag, P. (2024). Perdagangan Digital (E-Commerce) Indonesia Periode 2023.

Khotimah, B. K., Agustina, F., Puspitarini, O. R., Cahyani, A. D. W. I., Kustiyahningsih, Y., & Anamisa, D. R. (2024). Hyperparameters And Centroid Improvements In The K-Medoids Method For Grouping Processed Beef Smes. 1–21.

Krisnanto, K. G. (2025). Pengaruh E-Wom Dan Kualitas Website Terhadap Keputusan Pembelian Dengan Kepercayaan Sebagai Variabel Intervening Pada E-Commerce Tokopedia. Indonesia Economic Journal, 1(2), 609–634.

Laradi, S., Alrawad, M., Lutfi, A., & Agag, G. (2024). Understanding Factors Affecting Social Commerce Purchase Behavior: A Longitudinal Perspective. Journal Of Retailing And Consumer Services, 78. Https://Doi.Org/Https://Doi.Org/10.1016/J.Jretconser.2024.103751

Marsa, A., & Ritonga, H. (2025). Analisis Clustering Gaji Karyawan Menggunakan K-Means Dan Elbow Method Rumah Sakit Xyz. 1652–1660. Https://Doi.Org/10.33364/Algoritma/V.22-2.2915

Martí, R., Martínez-Gavara, A., & Corber, T. (2021). Grasp And Tabu Search For The Generalized Dispersion Problem ☆. 173. Https://Doi.Org/10.1016/J.Eswa.2021.114703

Purwatiningsih, A., & Habibi, M. (2024). Optimalisasi Analisis Data Peserta Olimpiade Sains Nasional Indonesia Menggunakan Algoritma K-Means Clustering. 14(3), 786–793.

Rafika, A., Putri, H., & Wakhidah, N. (2025). Optimization Of K-Means Clusteting With Elbow Method For Identification Of Tb Prone In Central Java. 6(1), 730–736. Https://Doi.Org/10.30596/Jcositte.V6i1.21669

Rahman, F. D., Zulfa, M. I., & Taryana, A. (2024). Clustering Dan Klasifikasi Data Cuaca Kota Cilacap Menggunakan K-Means Dan Random Forest. 1(April), 90–97. Https://Doi.Org/10.61124/Sinta.V1i2.15

Rahman, F. K. (2025). Penerapan Algoritma K-Means Clustering Pada Kondisi Mesin Screw Press.

Saputro, W., Pahlevi, M. R., Wibowo, A., Komputer, M. I., Komputer, F. I., & Luhur, U. B. (2020). Analisis Algoritma K-Means Untuk Klasterisasi Tindak Pidana K-Means Algorithm Analysis For Corruption Criminal Clasterization In Indonesia ' S Law. 3(3), 137–142. Https://Doi.Org/10.33387/Jiko

Sonianto, & Hartono. (2025). Penerapan Algoritma K-Means Untuk Mengidentifikasi Minat Dan Bakat Siswa Sekolah Dasar. 6, 1–9.

Wongoutong, C. (2024). The Impact Of Neglecting Feature Scaling In K-Means Clustering. Plos One. Https://Doi.Org/Https://Doi.Org/10.1371/Journal.Pone.0310839