Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 21, No. 2 September 2024 | DOI https://doi.org/10.33480/techno.v21i2.5796

MULTILEVEL MODAL VALUE ANALYSIS FOR INTERPRETING CATEGORICAL K-MEDOIDS CLUSTERS DATA

Rachmad Fitriyanto [1*]; Ummi Syafiqoh[2]

Information System Department [1]
Informatics Management Department [2]
STMIK PPKIA Tarakanita Rahmawati, Tarakan, Indonesia [1,2]
https://ppkia.ac.id [1,2]
rachmad@ppkia.ac.id [1*], ummi@ppkia.ac.id[2]

(*) Corresponding Author

Abstract *— Consumer segmentation plays a crucial role for business owners in developing their enterprises. K-Medoid is commonly used for segmentation functions due to its low computational complexity. However, K-Medoid has limitations, such as the variability in cluster sizes across different iterations and the challenge of determining the optimal number of clusters. The Davies-Bouldin Index (DBI) is a metric used to evaluate the number of clusters by calculating the ratio between the within-cluster distance and the between-cluster distance. Most segmentation studies typically stop at the formation of clusters without further interpretation, particularly when dealing with categorical data. This study aims to modify the use of K-Medoid and propose a method for interpreting clusters with categorical data. The research began with questionnaire design and the data collecting from 100 respondents, which was normalized in the second stage. Clustering used K-Medoid with variations K values from K=2 to K=10, with each K value tested 10 times. The clustering results were evaluated using the DBI to select the optimal clusters. Data interpretation conducted using modal values, calculated as the ratio of the number of times a specific attribute variable was selected by respondents to the total number of data points in the cluster. Utilization and hierarchical visualization of modal values proposed in this study offer insights into the dominant variables within an attribute and also depict the relationships between attributes based on the ranking of modal values. These advantages facilitate business analysts in labeling clusters for developing consumer-driven business strategies.*

Keywords*: categorical data type, cluster interpretation, davies-bouldin index, k-medoid, modal value, segmentation.*

Intisari *— Segmentasi konsumen memegang peranan penting bagi pemilik usaha untuk pengembangan usaha. K-Medoid umum digunakan untuk fungsi segmentasi karena dikenal dengan kompleksitasnya yang rendah. Meskipun demikian K-Medoid memiliki kelemahan pada ukuran klaster yang selalu berubah-ubah setiap percobaan dan tidak diketahuinya jumlah klaster yang tepat. Davies-Bouldin Index (DBI) adalah nilai yang dapat digunakan untuk evaluasi jumlah klaster dengan menghitung rasio antara jarak data dengan pusat (SSW: klasternya Sum of Square Within Cluster) dan jarak antar pusat klaster (SSB: Sum of Square Between Cluster). Penelitian tentang segmentasi umumnya sebatas segmentasi tanpa dilanjutkan dengan interpretasi klaster, terlebih lagi untuk tipe data kategorikal. Penelitian ini bertujuan untuk memodifikasi penggunaan K-Medoid dan mengusulkan cara menginterpretasi klaster data dengan tipe data kategorikal. Penelitian dimulai dengan perancangan kuesioner dan pengumpulan data sebanyak 100 responden yang dinormalisasi di tahap kedua. klasterisasi menggunakan K-Medoid dengan variasi nilai K=2 sampai K=10. Setiap variasi nilai K diuji sebanyak 10 kali. Hasil klasterisasi dievaluasi menggunakan nilai DBI untuk memilih klaster yang terbaik. Interpretasi data menggunakan nilai modus yang dihitung dari rasio antara jumlah variabel atribut yang dipilih responden dengan jumlah data di klaster. Penggunaan dan Visualisasi hirarki nilai modus yang diusulkan pada penelitian ini mampu memberikan gambaran untuk mengetahui variabel-variabel yang dominan dalam sebuah atribut dan juga mampu menggambarkan relasi antar atribut berdasarkan rangking nilai modus yang ditemukan. Kedua keuntungan tersebut memudahkan bagi analis bisnis untuk pelabelan klaster yang dapat digunakan*

sebagai bahan pertimbangan penyusunan strategi usaha berbasiskan kebutuhan konsumen.

***Kata Kunci**: tipe data kategorikal, interpretasi klaster, indeks davies-bouldin, k-medoid, nilai modal, segmentasi.*

INTRODUCTION

In the rapidly evolving landscape of small business development, understanding and catering to diverse customer needs is paramount for entrepreneurial success. Managing the relationship between a company and its customers is a key factor in ensuring business continuity, particularly within the scope of entrepreneurship. Business owners must be able to gather information about their customers to design appropriate strategies that meet their needs based on customer characteristics (Guerola-Navarro et al., 2024). Each consumer has distinct characteristics that require precise techniques for the company to effectively utilize the data.

Customer segmentation is an appropriate method to define customers (Gomes & Meisen, 2023). Segmentation will group customers based on specific criteria that business owners can use to design appropriate strategies for each customer group (Berahmana et al., 2020). The advantages that segmentation offer, made customer segmentation is crucial for CRM (Customer Relationship Management) purposes. It involves grouping existing and potential customers based on shared characteristics, such as interests, demographics, or buying patterns.

The customers characteristics collected by many methods. Survey in the form of questionnaire is tools to collecting customers data that have many advantages, such as scalability, efficiency and design feasibility. From data preparation perspective, questionnaire offer advantage to prevent missing value by using multiple-choice type question to gather a wide range of data types, relevant to the segmentation process.

Clustering is a widely used method for customer segmentation, frequently found in previous research. Algorithms such as K-Means, K-Medoid, DBSCAN, and AHC are commonly employed due to their low complexity (Abdulhafedh, 2021; Khan et al., 2021; Mufarroha et al., 2022; Nugroho et al., 2024). In these studies, consumer data is processed to form clusters that represent groups of consumers based on the similarity of their characteristics. Various types of diagrams, such as bar charts, pie charts, or scatter plots, were used in previous studies to represent consumer clusters. The interpretation of clusters in these studies was carried out by translating data into descriptive statistical values, which were described independently from one attribute to another. This approach can obscure the relationships between attributes, potentially leading to bias and even inaccurate labeling.

Despite questionnaires and clustering play crucial roles in customer segmentation, previous research has typically concluded with the formation of consumer clusters. This approach is not inherently flawed, as the ultimate goal of clustering is simply to create clusters. However, the intended benefit of customer segmentation—obtaining information that accurately represents each cluster—has not been fully realized. This gap arises because the focus of these studies has been limited to the application of data mining techniques within the context of Knowledge Discovery in Databases (KDD), which ideally should conclude with data interpretation.

Previous studies have generally employed numerical data for clustering. This type of data can be visualized and interpreted using various statistical techniques and diagrams, such as box plots, descriptive statistics, and scatter plot diagrams. However, a different scenario arises when categorical data is used, as it lacks inherent weight or ranking, requiring a different approach for effective interpretation.

To solve these gaps, this study proposes a novel approach to interpreting categorical data within clusters by utilizing modal values. By systematically applying modal values to summarize the most frequent categories in each cluster, the proposed method aims to enhance the interpretability of clustering results. This approach not only provides a clearer understanding of the characteristics that define each customer segment but also makes the results more actionable for non-technical users in marketing and business analytics.

The primary contribution of this research lies in bridging the gap between advanced clustering techniques and practical applications in customer segmentation. By focusing on the interpretation of categorical data, this study seeks to provide a valuable tool for businesses to better understand their customers, leading to more informed decision-making and improved marketing strategies.

MATERIALS AND METHODS

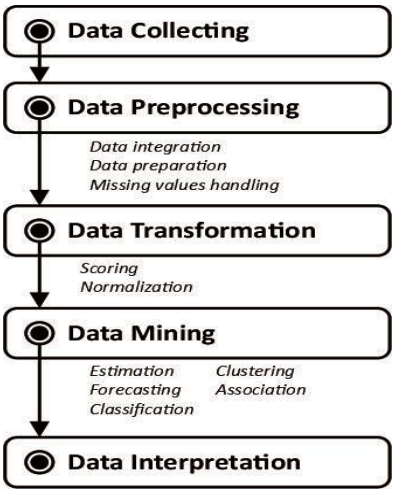
Data mining, a pivotal aspect of contemporary data analysis, involves extracting meaningful patterns and insights from vast datasets. Among the various techniques, clustering is widely used to group similar data points, facilitating better understanding and decision-making. Clustering algorithms like k-means and k-

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 21, No. 2 September 2024 | DOI https://doi.org/10.33480/techno.v21i2.5796

medoids are popular choices for their simplicity and effectiveness. However, notable challenges persist, particularly with the interpretation of categorical data types and the inconsistency in cluster sizes across different trials when using k-medoids.

Categorical data, which refers to variables that can take on a limited number of distinct values, is prevalent in many domains such as social sciences, marketing, and bioinformatics. Traditional clustering methods, such as k-means and hierarchical clustering, are primarily designed for numerical data and often fall short when applied to categorical datasets. This inadequacy arises because these methods rely on distance metrics that are not naturally suited to categorical variables. Although advancements have introduced specialized clustering algorithms for categorical data, the effective interpretation of the resulting clusters remains a significant challenge.

Segmentation is one of the benefits offered from data mining implementation, which is also a part of Knowledge Discovery in Database (KDD). KDD is a framework used to process large amounts of data into information using specific techniques (Duque, 2024; Tattersall, 2021). Figure 1 illustrates the stages within KDD.



Source: (Duque, 2024; Tattersall, 2021).

Figure 1. KDD Stages

The initial stage of KDD is data collection, aligned with the goals of data analysis, followed by data preprocessing, which involves preparing the data for subsequent processing stages. Handling missing values and normalization are two techniques commonly used in the preprocessing stage. The third stage is data transformation, which involves altering the data format according to data analysis requirements. The fourth stage is data mining, where algorithms suitable for the initial goals of KDD and the characteristics of the data are applied. The final stage is evaluation and interpretation. Evaluation in KDD is necessary because there is no single algorithm that is best suited for all data processing tasks, thus requiring an assessment of the results to determine the quality of the data mining application. Data interpretation is the activity of reading and extracting useful information in line with the initial objectives of KDD.

One of the functions offered by algorithms within the scope of data mining is clustering. Clustering provides users with groups of data organized based on similarities among the data points (Ezugwu et al., 2021; Fadhilah et al., 2024; Setiawati et al., 2024). The result of clustering is considered good if the data within a group exhibit strong similarity while the data between groups exhibit weak similarity (Fadhilah et al., 2024; Wegmann et al., 2021). Another characteristic of clustering is that it does not have data labels like classification functions. Clustering only assigns group labels such as C1, K2, A, or other symbols that have no special meaning other than distinguishing one cluster from another.

K-Medoid is one of the widely used clustering algorithms due to its simplicity and low computational complexity (Dwivedi & Bhaiya, 2019; Ezugwu et al., 2021). Additionally, the k-medoids algorithm, valued for its robustness to noise and outliers, minimizes the sum of dissimilarities between points and their respective medoids, making it suitable for various applications. Despite these advantages, k-medoids often yields clusters of varying sizes in different runs, even when applied to the same dataset. This variability, caused by its sensitivity to the initial selection of medoids, leads to instability and unpredictability in the clustering outcomes.

K-Medoid operates using medoid values, which are the central points of data in each cluster. The determination of medoid values is done randomly for each cluster. All medoid values are used to measure the distance to each data point in the dataset. Data points are grouped into specific clusters based on the shortest distance to one of the medoids. Distance measurement can be done using several techniques, one of which is the Euclidean Distance, as shown in Formula 1 (Anggraini et al., 2020; Wahyudi et al., 2023)

$$d_{ij} = \sqrt{\sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_{ij} - c_{ij}\right)^2} \qquad (1)$$

Description:
dij : Distance between 2 data point (Xi and Xj)
Xi : data on the i-th
Xj : data on the j-th
n : the total of data features
p : the total of data point (dataset size)

K-Medoid is applied iteratively, starting with the first iteration to determine the cluster of each data point. In the second and subsequent iterations, the same process as in the first iteration is repeated, ending with the calculation of the Cost value using Formula 2 (Hutagalung et al., 2022; Saragih et al., 2021; Suprianto et al., 2023).

$$c = \sum_{i=1}^{n} d(x_i, m_{xi}) \quad (2)$$

Description:
C : Cost antara iterasi y dan iterasi z
n : the number of data points
d(xi,mxi) : the distance between data point xi and its assigned medoid mxi

The K-Medoid iterations are terminated when the obtained cost is zero.

The Davies-Bouldin Index (DBI) is a metric used to assess the quality of clusters resulting from clustering. The DBI serves as a benchmark for determining the optimal number of clusters in a clustering process. It evaluates the clustering results by comparing two values: the proximity of data points to their cluster center and the distance between cluster centers. (Ghufron et al., 2020; Mughnyanti et al., 2020; Stiadi & Sundani, n.d.; Utomo, 2021). The calculation of the DBI value is performed in five stages (Santoso & Magdalena, 2020):

1) Centroid selection for each cluster conducted by using the average values of each feature data within the respective cluster.
2) Calculate Sum of Square Within Cluster (SSW) value by formula 3.

$$SSW_i = \frac{1}{m}\sum_{j=1}^{m_j} d(x_j, c_i) \quad (3)$$

Description:
SSWi : average distance from each data to the ith centroid in the ith cluster
d(xj,ci) : distance between jth data point to its centroid

3) Calculate Sum of Square Between Cluster (SSB) value by formula 4.

$$SSB_{ij} = d(x_j, c_i) \quad (4)$$

Description:
SSB_{ij} : distance between i^{th} centroid to the j^{th} centroid
X_j : the j^{th} centroid
C_i : the i^{th} centroid

4) Calculate the Ratio ($R_{i,j}$) between SSW dan SSB by formula 5.

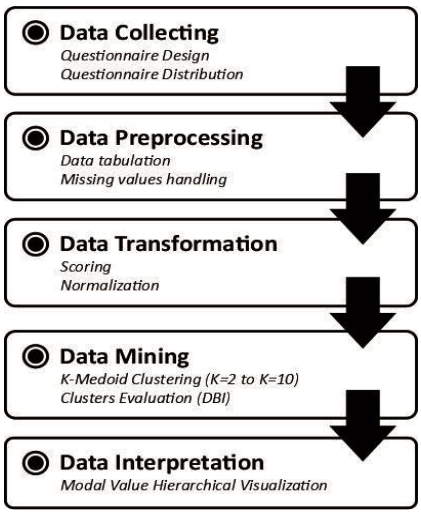
$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{ij}} \quad (5)$$

5) Calculate DBI value by choosing the maximum Rij among all cluster by formula 6.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max(R_{i,j}) \quad (6)$$

The smaller SSW mean the better the quality of the cluster. Similarly, the greater SSB, the better the cluster quality. Therefore, a good DBI value is one that approaches zero (0).

The research methodology is based on the KDD stages, which consist of five stages as illustrated in Figure 2. On the first stage, the questionnaire contains 16 questions representing three categories: individual characteristics, purchasing patterns, and customer satisfaction.



Source: (Research Results, 2024)
Figure 2. Research Stages

Normalization at the preprocessing stage aims to form the same scale for all data attributes. In this research, normalization uses Min-Max Normalization as shown in formula 7. (Henderi, 2021; Raju et al., 2020).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (7)$$

Description:
X' : data after normalizing
X : data before normalizing
X_{max} : Maximum value in attribute
X_{min} : Minimum value in attribute

The number of questions in the questionnaire is designed so that it can be

completed by consumers while waiting for their order to be prepared, with a maximum completion time of 15 minutes. Respondents were selected based on 3 main categories, individual characteristics include the age of respondents ranging from under 17 years old to over 45 years old. Occupations consist of students, civil servants, private employees and entrepreneurs. Another criterion in this category is the distance of residence to the outlet ranging from less than 1 km to 15 km.

The second category of purchasing patterns includes transaction hours starting from 10 a.m. to 10 p.m., the number of cups purchased ranging from 1 cup to 5 cups, cup size (medium - jumbo), frequency of purchase in one week (1x - 3x), favorite drinks and sources of information about the outlet. Table 1,2 and 3 shows the question items based on individual characteristic, purchasing pattern and customer satisfaction.

Table 1 Individual Characteristic Items

Code	Question
P01	Consumer age
P09	Consumer job
P10	Consumers residence to outlet distance

Source: (Research Results, 2024)

Table 2 Purchasing Pattern Items

Code	Question
P02	Purchase time
P03	Numbers of cup
P04	Cup size
P08	Weekly purchasing frequency
P11	Favourite flavor
P15	The information source about product

Source: (Research Results, 2024)

Table 3 Customer Satisfaction Items

Code	Question
P05	Consumer price perception
P06	Consumer promotion experience
P07	Packaging quality
P12	Cup design quality
P13	Price rationality to drink quality
P14	Required facilities
P16	Services Satisfaction

Source: (Research Results, 2024)

Data collection stage, aiming for 100 respondents to complete the questionnaire, which was distributed in softcopy form using Google Forms via mobile phones provided by the seller.

RESULTS AND DISCUSSION

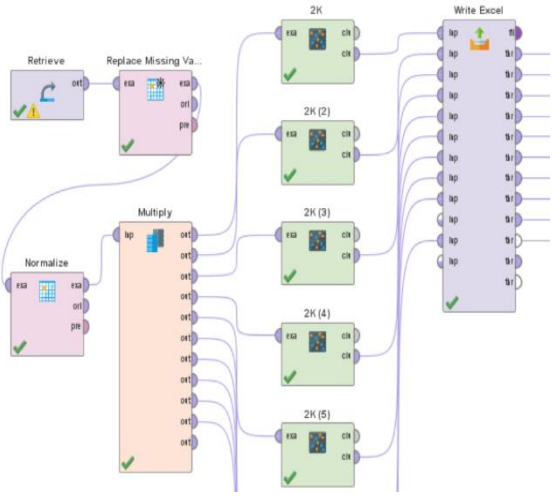
In the data preprocessing stage for the 100 respondent data, no missing values were found, which was successfully prevented through the use of multiple-choice options in the questionnaire. In the data transformation stage, scoring was carried out to convert the answer options from categorical data into numerical values, as illustrated in Table 4

Table 4 Scoring Example

Code	Categorical Data	Numeric Data
P01: Consumers Age	Less than 17 years old	1
	17 years old – 24 years old	2
	25 years old – 34 years old	3
	35 years old – 44 years old	4
	More than 44 years old	5
P09: Consumers Job	Student	1
	College Student	2
	Private Sector	3
	Entrepreneur	4
	Government Employee	5

Source: (Research Results, 2024)

Normalization during the data transformation stage is performed using the "Normalize" operator in the RapidMiner application, as shown in Figure 3.



Source: (Research Results, 2024)

Figure 3. Normalization & Clustering on Rapidminer for K=2

Figure 3 illustrates the normalization and clustering model using K-Medoid in RapidMiner. Data normalization executed twice. The first time using Rapiminer operator as shown on Figure 3. On second time, normalization without automatic tools using 7th formula to verified normalization result from Rapidminer. Clustering was performed for cluster numbers K=2 to K=10. Each cluster number variation was tested by performing clustering experiments 10 times. For example as shown on Figure 3, with K=2, clustering process executed for 10 times marked by operator name as 2K, 2K(1), 2K(2) until 2K(10). This approach was employed to

anticipate the characteristic of K-Medoid, which tends to produce varying cluster sizes with each execution. By varying the value of K from K=2 to K=10, and testing each K value 10 times, a total of 90 clustering results were generated.

After the clustering process was completed, the 90 clustering results were evaluated to determine the optimal number of clusters using the Davies-Bouldin Index (DBI). DBI values computation result for the 90 clustering results are presented in Table 5.

Table 5. DBI Values

Test	K=2	K=3	K=4	...	K=8	9K	10K
1	7.57	4.07	3.85	...	3.00	1.61	1.19
2	7.57	9.21	7.41	...	5.38	2.22	2.29
3	7.57	9.21	6.46	...	3.95	1.93	3.50
4	7.57	3.86	3.73	...	3.55	1.98	2.64
5	7.57	9.21	8.51	...	5.33	1.48	1.22
6	7.57	3.86	6.31	...	3.68	1.68	2.33
7	7.57	7.50	8.37	...	4.52	1.34	2.10
8	7.57	4.07	3.73	...	4.72	2.61	1.73
9	7.57	3.86	8.37	...	3.82	2.96	1.37
10	7.57	7.62	3.73	...	2.22	1.40	2.58

Source: (Research Results, 2024)

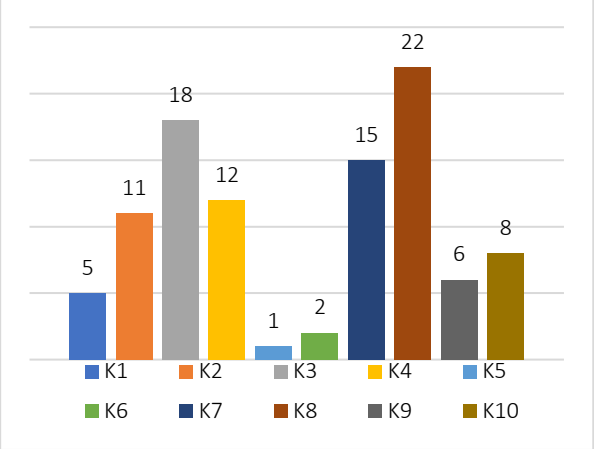
The selection of the clustering results to be used in the data interpretation stage is carried out using the concepts of local minimum and global minimum from metaheuristics. For each variation in the number of clusters, the best DBI value is selected as the local minimum, resulting in 10 DBI values. For example, on K=3 the minimum value is 3.86 from 6th trial clustering. This value choose as local minimum on K=4. The same method implemented on others K value that produce 10 values, as shown on Table 6.

Table 6. Local Minimum & Global Minimum of DBI Values

K	Trial	Local minimum
2	1st	7.5
3	4th	3.86
4	4th	3.72
5	6th	2.12
6	9th	2.10
7	2nd	2.20
8	10th	2.22
9	7th	1.34
10	1st	1.19
Global Minimum		1.19

Source: (Research Results, 2024)

The global minimum represent the best DBI value from 10 best local minimum. As shown on table 6, The best DBI values came from clusterization with 10 clusters on the first trial with 1.19. These clusters result choose to interpreted on the last stages. Figure 4 shown the comparison clusters size.



Source: (Research Results, 2024)

Figure 4. Clusters Distribution

Cluster_8 have 22% members from 100 respondents. Followed by Cluster_7 by 15%. Anomalies in the clustering results are indicated by the presence of two clusters categorized as singleton clusters in Cluster_5 and Cluster_6, which contain only 1 and 2 members, respectively.

Out of the 90 clustering results from the tests with K varying from K=2 to K=10, these two clusters consistently appeared as singleton cluster. The options for handling these two singleton clusters are either to include them in the interpretation stage or to exclude them. In this study, these clusters are deemed outliers based on their data feature characteristics, which are also present in other clusters, made these two clusters lack distinct characteristics. Therefore, these clusters are excluded from the data interpretation process.

The interpretation of data in each cluster begins with identifying the prominent or distinctive data attributes. It should be noted that in the dataset used, all data are categorized as categorical data, so not all descriptive statistical functions can be used for data interpretation purposes. The determination of prominent data features is based on the variability level of each feature. If the variability of a feature is high compared to others, it means the response options for that question in the cluster are evenly distributed or varied. From other perspective, if it is low, it indicates that one value is dominant compared to other responses in that feature. This dominant value becomes a distinctive characteristic shaping the cluster's profile.

For categorical data, the descriptive statistical function that can be used to determine cluster variability is the modal value. Data attributes with the highest modal value indicate that the attribute has a significant influence on the characteristics of a cluster, as exemplified in Table 7, which contains the percentage of modal values from 16 questionnaire questions in Cluster_8.

Table 7. Homogeneity Percentage on Cluster_8

Code	Percentage	Majority Answer
P07	100%	Sufficient (22)
P13	100%	Rational (22)
P15	95%	Friends Info (21)
P16	95%	Satisfied (21)
P04	91%	Jumbo (20)
P11	82%	Non-Boba (18)
P05	82%	Standard (18)
P09	77%	Private Sector (17)
P08	64%	2x - 3x (14)
P03	59%	2 Cups (13)
P06	59%	No (13)
P12	55%	Ordinary (12)
P14	55%	Chair (12)

Source: (Research Results, 2024)

The modal value from each question calculated in percentage and ranked for interpreting cluster. On cluster_8, the biggest modal values found in P07 and P13 by value 100%. In question P07 there are 2 options Sufficient or Not Sufficient. Out of the 22 members in cluster_8, all selected "*Sufficient*" resulting in a 100% ratio for the "*Sufficient*" option compared to "*Not Sufficient*". The percentage of the modal value is calculated using Formula 8.

$$Modal\ Percentage = \frac{total\ of\ chosen\ choice}{total\ of\ cluster\ members} \quad (8)$$

Table 8 shows an example of the percentage calculation of the modal values for questions P07, P15, and P08 in Cluster_8, which consists of 22 data points.

Table 8. Modal Value Percentage Example

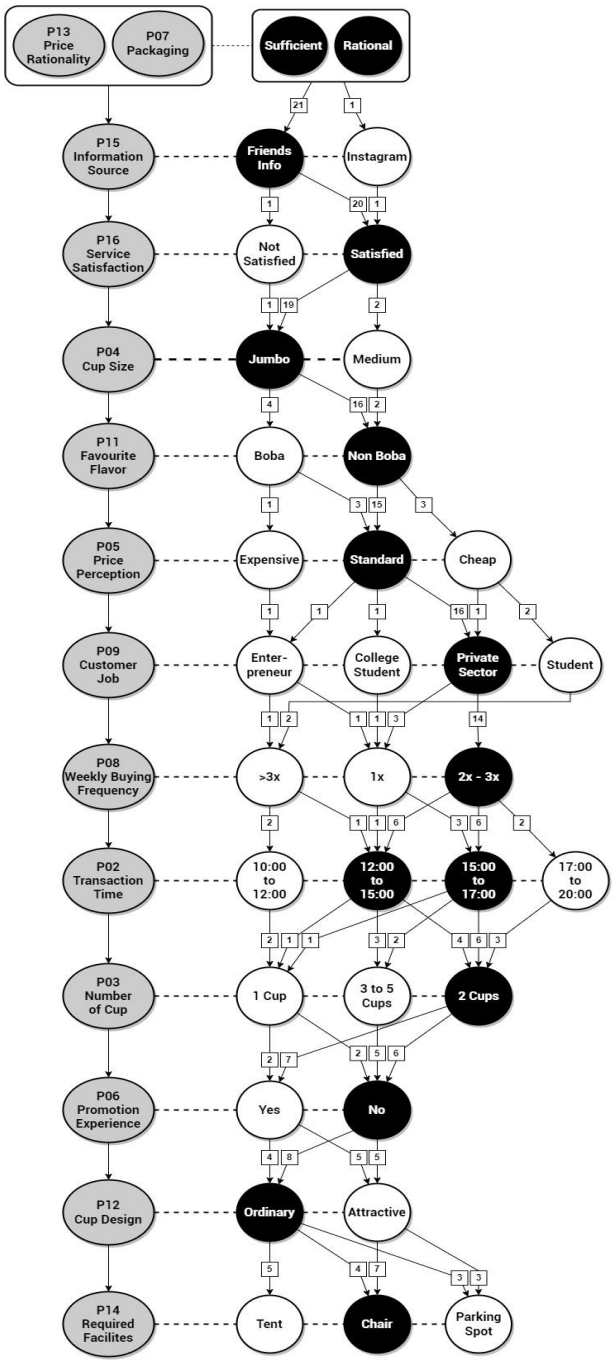
Code	Choices	Total	%
P07	Sufficient	22	100%
	Not Sufficient	0	0%
P15	Friends	21	95%
	Instagram	1	5%
	Facebook	0	0%
P08	1x	5	23%
	2x – 3x	14	64%
	>3x	3	13%

Source: (Research Results, 2024)

Question P07, which inquires about cup quality, offers two answer choices: *Sufficient* and *Not Sufficient*. The questionnaire results indicate that all 22 respondents in Cluster_8 selected "*Sufficient*," meaning 100% of respondents in Cluster_8 believe the cup quality is adequate.

Question P15, regarding the source of product information, provides three options: *Friend, Instagram*, or *Facebook*. Of these options, 21 out of 22 respondents, or 95%, selected "*Friend*," while 1 respondent, or 5%, chose "*Instagram*" as the source of information.

Question P08, concerning the frequency of purchases per week, includes three choices: Once a week (1x), 2-3 times a week (2x-3x) or *more than 3 times a week(>3x)*. The results show that 5 respondents, or 23%, selected "*Once a week*." The second choice, *2-3 times a week*, was selected by 14 respondents, or 64%, while the third option, *more than 3 times a week*, was chosen by 3 respondents, or 13%. Based on modal values, the cluster could explore with hierarchical interpretation as shown on Figure 5.



Source: (Research Results, 2024)

Figure 5. Cluster_8 Hierarchical Visualization

The visualization of Cluster_8 concludes at question P014 due to its modal value homogeneity percentage of 55%, whereas P10 and P01 have homogeneity percentages below 50%. From the diagram, we can label this cluster as "appreciation and promotion required" based on the following reasons:

1) All customers in this cluster have a "rational" opinion about price rationality. This means that the product price is acceptable, and the business owner should retain these customers.
2) 73% (16 out of 22) of customers in Cluster_8 work in the private sector. Consumers in this category have a relatively high purchase frequency, buying 2 to 3 times per week, with an average of 2 jumbo-sized cups purchased per transaction.

However, customers in the private employee category have never received purchase promotions, as indicated by 13 out of 22 consumer responses in this cluster on question P06

For comparison, the visualization of Cluster_3 is shown in Figure 6 generated based on majority answers as shown in Table 9.

Table 9 Homogeneity Percentage on Cluster_3

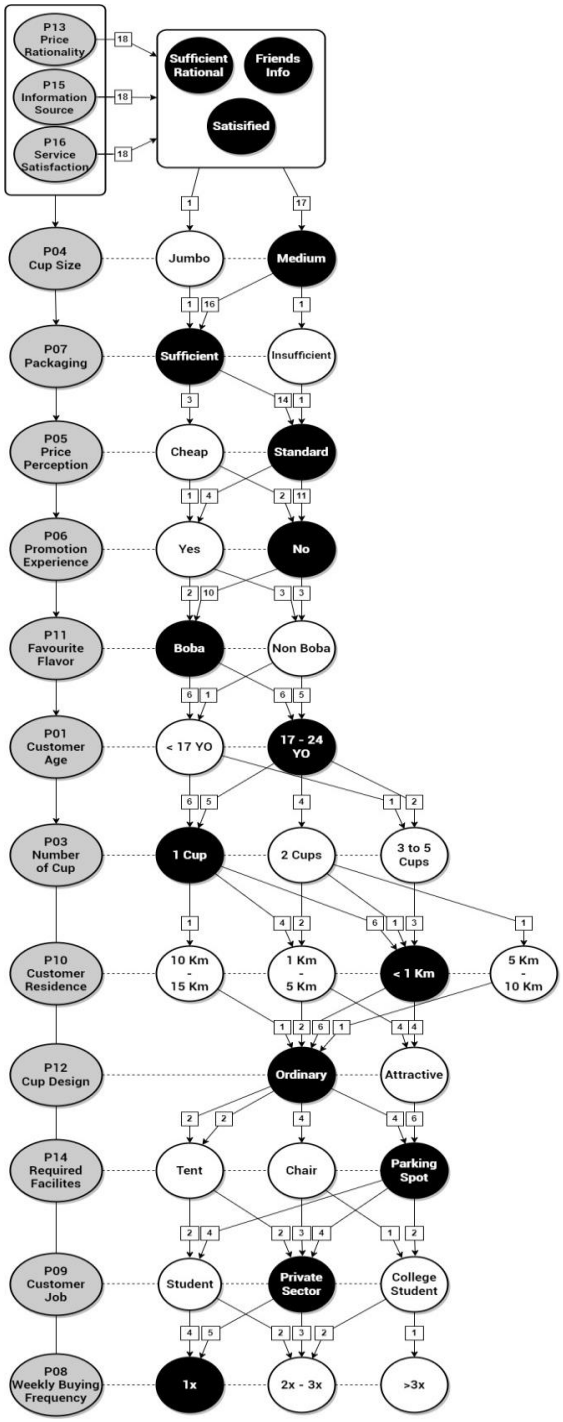
Code	Percentage	Majority Answer
P13		Rational (18)
P15	100%	Friends Info (18)
P16		Satisfied (18)
P4	94%	Medium (17)
P7	94%	Sufficient (17)
P5	83%	Standar (15)
P6	72%	No Promo. Exp. (13)
P11	67%	Boba (12)
P1	61%	17 YO - 24 YO (11)
P3	61%	1 Cup (11)
P10	56%	< 1 Km (10)
P12	56%	Ordinary (10)
P14	56%	Parking Spot (10)
P9	50%	Private Sector (9)
P8	50%	1x (9)

Source: (Research Results, 2024)

Cluster_3 has different characteristics compared to Cluster_8. All customers in this cluster have consistent values regarding information source and service satisfaction, which include price rationality, information source, and service satisfaction. From these three questions, the characteristics of customers in this cluster can be break-down and illustrated as shown in figure 6.

The prominent or distinctive data attributes from Table 7 and figure 6 described as follows:

1) 61% of customers are private sector employees and college students, indicated by customer ages between 17 and 24 years old.
2) These customers' occupations are also indicated by the answers to question P02 about transaction time. 72% of transactions occur between 12 PM and 3 PM and between 3 PM and 5 PM. These time periods coincide with when students finish their classes and when workers leave their jobs.



Source: (Research Results, 2024)

Figure 6. Cluster_3 Hierarchical Visualization

3) Unlike consumers in Cluster_8, the number of products purchased by consumers in this cluster is only one cup per transaction, with a medium cup size.

4) Consumers age in this cluster consist of 2 age categories only, less than 17 years old and between 17 to 24 years old. This in mean we labeled the customers as teenager and young adult.
5) Another characteristic of this cluster that does not appear in Cluster_8 is the distance from consumers' residences to the kiosk. 55% of consumers in this cluster live less than 1 km away from the kiosk.

Based on these four characteristics, consumers in Cluster_3 can be labeled as "*teenager-young adult with promotion required*" due to their average purchase of one medium-sized cup.

The hierarchical visualization of modal values, as shown in Figures 5 and 6, offers advantages over scatter plot diagrams, which only display the shape of the clusters. The modal value visualization in this study enables analysts to identify the dominant factors influencing the level of homogeneity within each cluster. Additionally, it allows analysts to discern the relationships between consumer choices across different questions. For instance, the most common purchase behavior, involving 11 respondents who each bought one cup per transaction, was observed among young adults living within a 1 to 5 km radius of the outlet.

CONCLUSION

Interpretation of clusters with categorical data using modal values can be utilized to identify the data homogeneity that characterizes the cluster. The hierarchical modal value visualization proposed in this study effectively highlights dominant variables within an attribute and illustrates the relationships between attributes based on the ranking of modal values. These advantages facilitate business analysts in labeling clusters, which can then be used as a basis for developing business strategies tailored to consumer needs.

While this study has provided valuable insights into modal value utilizing, there are several areas where future research could further enhance our understanding. Feature selection implementation on attribute with categorical data type would be beneficial to explore the application of cluster interpreting across a larger and more diverse dataset to validate the generalizability of the findings. Moreover, future studies could employ different clustering model to explore the possibility of optimizing the clustering model to prevent singleton clusters. By addressing these areas, future research could contribute to a more comprehensive understanding of cluster interpreting techniques.

REFERENCE

Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, *Vol.3*(1), 12–30. https://doi.org/10.12691/jcd-3-1-3

Anggraini, R. A., Wati, F. F., Shidiq, M. J., Suryadi, A., Fatah, H., & Kholifah, D. N. (2020). Identification of Herbal Plant Based on Leaf Image Using GLCM Feature and K-Means. *Jurnal Techno Nusa Mandiri*, *17*(1), 71–78. https://doi.org/10.33480/techno.v17i1.1087

Berahmana, R. W. B. S., Mohammed, F. A., & Chairuang, K. (2020). Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *11*(1), 32. https://doi.org/10.24843/LKJITI.2020.v11.i01.p04

Duque, J. (2024). Data Mining for Knowledge Management. *Procedia Computer Science*, *239*, 257–264. https://doi.org/10.1016/j.procs.2024.06.170

Dwivedi, S., & Bhaiya, L. K. P. (2019). A Systematic Review on K-Means Clustering Techniques. *International Journal of Scientific Research*, *5*(3).

Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, O. N., José-García, A., & Agushaka, J. O. (2021). Automatic clustering algorithms: A systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, *33*(11), 6247–6306. https://doi.org/10.1007/s00521-020-05395-4

Fadhilah, R., Matdoan, M. Y., Safira, D. A., & Tahalea, S. (2024). Clustering Shrimp Distribution In Indonesia Using The X-Means Clustering Algorithm. *VARIANCE: Journal of Statistics and Its Applications*, *6*(1), 49–54. https://doi.org/10.30598/variancevol6iss1page49-54

Ghufron, Surarso, B., & Gernowo, R. (2020). Implementation of K-Medoids Clustering for High Education Accreditation Data. *Jurnal Ilmiah KURSOR*, *10*(3), 119–128.

Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and E-Business Management*, *21*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

Guerola-Navarro, V., Gil-Gomez, H., Oltra-Badenes, R., & Soto-Acosta, P. (2024). Customer relationship management and its impact on

entrepreneurial marketing: A literature review. *International Entrepreneurship and Management Journal*, *20*(2), 507–547. https://doi.org/10.1007/s11365-022-00800-x

Henderi, H. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *IJIIS: International Journal of Informatics and Information Systems*, *4*(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

Hutagalung, J., Syahril, M., & Sobirin, S. (2022). Implementation of K-Medoids Clustering Method for Indihome Service Package Market Segmentation. *Journal of Computer Networks, Architecture and High Performance Computing*, *4*(2), 137–147. https://doi.org/10.47709/cnahpc.v4i2.1458

Khan, R. H., Dofadar, D. F., & Rabiul Alam, Md. G. (2021). Explainable Customer Segmentation Using K-means Clustering. *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 0639–0643. https://doi.org/10.1109/UEMCON53757.2021.9666609

Mufarroha, F. A., Suzanti, I. O., Satoto, B. D., Syarif, M., & Yunita, I. (2022). K-means and k-medoids clustering methods for customer segmentation in online retail datasets. *2022 IEEE 8th Information Technology International Seminar (ITIS)*, 223–228. https://doi.org/10.1109/ITIS57155.2022.10010135

Mughnyanti, M., Efendi, S., & Zarlis, M. (2020). Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. *IOP Conference Series: Materials Science and Engineering*, *725*(1), 012128. https://doi.org/10.1088/1757-899X/725/1/012128

Nugroho, B. I., Rafhina, A., Ananda, P. S., & Gunawan, G. (2024). Customer segmentation in sales transaction data using k-means clustering algorithm. *Journal of Intelligent Decision Support System (IDSS)*, *7*(2), 130–136. https://doi.org/10.35335/idss.v7i2.236

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 729–735. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Santoso, H., & Magdalena, H. (2020). Improved K-Means Algorithm on Home Industry Data Clustering in the Province of Bangka Belitung. *2020 International Conference on Smart Technology and Applications (ICoSTA)*, 1–6. https://doi.org/10.1109/ICoSTA48221.2020.1570598913

Saragih, R., Irawan, E., & Syahputra, I. (2021). Implementation Of K-Medoids Algorithm in Grouping Diseases in The Community. *Jurnal Teknik Komputer Network*, *1*(1), 1–8.

Setiawati, E., Fernanda, U. D., Agesti, S., Iqbal, M., & Herjho, M. O. A. (2024). Implementation of K-Means, K-Medoid and DBSCAN Algorithms In Obesity Data Clustering. *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, *1*(1), 23–29. https://doi.org/10.57152/ijatis.v1i1.1109

Stiadi, M. J., & Sundani, D. (n.d.). *Performance Measurement of K-Means Clustering Algorithm Using Davies Bouldin Index Method*. *6*(4).

Suprianto, A., Sari, H. L., & Zulfiandry, R. (2023). Perbandingan Algoritma K-Means Dan K-Medoid Dalam Pengelompokan Data Pasien Berdasarkan Rekam Medis di Puskesmas M. Thaha Bengkulu Selatan. *Journal of Science and Social Research*, *VI*(3), 580–586.

Tattersall, M. (2021). Knowledge Mining Methods based on Data Warehouse: A Case Study. *International Journal of Hybrid Innovation Technologies*, *1*(1), 1–14. https://doi.org/10.21742/ijhit.2021.1.1.01

Utomo, W. (2021). The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia. *ILKOM Jurnal Ilmiah*, *13*(1), 31–35. https://doi.org/10.33096/ilkom.v13i1.763.31-35

Wahyudi, E., Meidelfi, D., & Saam, Z. (2023). The Implementation of the K-Medoid Clustering for Grouping Hearing Loss Function on Excessive Smartphone Use. *JOIV : Int. J. Inform. Visualization*, *7(4)*, 2523–2531.

Wegmann, M., Zipperling, D., Hillenbrand, J., & Fleischer, J. (2021). A review of systematic selection of clustering algorithms and their evaluation. *ArXiv, Abs/2106.12792.*