Application of K-Means algorithm for market segmentation (Case study: Lily's cake Pontianak)

Edward Revaldo Danuwinata[1*], Jimmy Tjen[2]

[*1] Department of Digital Business, Faculty of information technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
[2] Department of Informatics, Faculty of information technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.

*Email: 22430154@widyadharma.ac.id

Abstract. In business, effective customer preference and segmentation are essential for decision making. Since data, especially data sales, is a crucial indicator. Data mining is needed to help companies process data into useful resources. Therefore, the author Recommends grouping several attributes from Lily's cake sales data using the K-Means algorithm for clustering, a data mining approach, to determine client preferences during significant holidays. 599 data points from Christmas, Chinese New Year, and Eid al-Fitr sales from 2023 to 2024 will be handled, even though the research participants have never processed and evaluated data before. Using four clusters, the study aims to ascertain the K-Means method's validity for clustering. The findings of this study enhance Lily's cake capacity to connect with various customer segments by allowing them to better customize their product offers to suit customer preferences.

Keywords: business; customer preference; customer segmentation; data mining; clustering

Introduction

In Business, customer segmentation, personalization, behavior analysis, and preference determination are crucial for every business decision making. Data, especially sales data, plays a key role in business, which can provide a lot of insight, from this data company can analyze and find answers to questions such as determining the best-selling product, identifying which products are frequently purchased at the same time, and many others. Numerous businesses are using ways to figure out which products are most suitable for specific customer categories; this is what allows them efficiently offer products to target audiences without directly inquiring about customers. A large amount of data can be processed with data mining, a process that can read and extract information from vast volumes of data, a method for finding patterns, correlations, trends, and anomalies in the database (Gautam et al., 2022). The advancement of technology in data mining offers company significant assistance in enhancing their capacity for making accurate decisions (Hou et al., 2023; Gautam et al., 2022). One of the data mining methods that can help this research is clustering, an unsupervised machine learning method that classifies data based on similar characteristics without requiring prior labels or information (Dalmaijer et al., 2022; Sinaga et al.,2020). Clustering algorithms are commonly used in statistics and computing to perform exploratory data analysis (Oyewole et al., 2023). The benefits of this approach are the ability to identify hidden patterns and structures in large and complex data sets. Facilitating improved decision making (Hou et al., 2023).

K-Means, Dbscan, Prefixed, and hierarchical clustering are just a few of the clustering algorithms that can work effectively on very vast and complicated data sets that would be challenging to analyze using conventional techniques. A popular clustering algorithm called K-Means separates the data into a predetermined number of clusters according to distance or dissimilarity (Ahmed et al., 2020; Wang et al., 2023). While efficient, this technique has disadvantaged including the necessity of defining the number of clusters early on and it is heavily dependent on the random selection of initial centroids (Abdulazeez, 2021).

Workload analysis (Ibáñez et al., 2022), student anxiety identifies (Liu, et al., 2022), and COVID-19 pandemic pattern mapping (Abdullah et al., 2020) are three instances of successful K-Means clustering applications. The potential role of clustering can also help companies understand market niches and customer needs for products, such as product recommendation studies (Santana et al, 2020) and customer segmentation studies (Tabianan et al., 2022; Zhao et al., 2021). The efficacy of K-Means techniques has been validated by several studies across various disciplines.

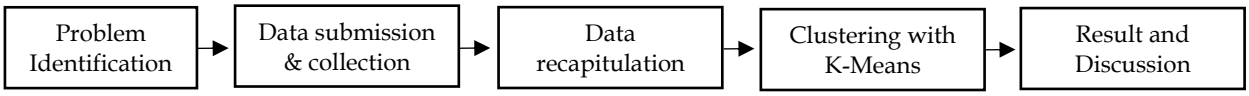
It's a local tradition in Pontianak city to prepare and serve cake products during important celebrations like Christmas, the Chinese New Year, and Eid al-Fitr. After years of being involved in the cake industry, Lily's cake continues to face challenges in sales analysis and understanding customer needs. Even though Lily's cake has been established for approximately 18 years, many components of the company's operations are still done by traditional systems, which makes it challenging to monitor and access data or identify individual customers. Therefore, the research problem in this study is formulated as the application of the K-Means clustering technique to Lily's cake sales data is, thus, the research challenge in this study, with the goal offering insightful information for business decision-making. Knowing the requirements and inclinations of customers in various areas enables. Lily's cake to more effectively target markets and create community-focused plans that promote more inclusive and sustainable local economies by knowing the needs and preferences of its customers in different domiciles and different occupations.

This research builds on these findings by developing a multi-stage clustering-based methodology to investigate purchase patterns, which is then refined using the K-Means clustering algorithm in text context of cake industry demand trends. The rest of research as follows: The methodology section discusses the research steps, data collection, and processing process, and implementation using the K-Means clustering technique applied at Lily's cake; in the Results and discussion section, a comparison of the selected attributes is presented, and the results of clustering using the K-Means method are explained; in the Conclusion section, provide recommendations for the subject and suggest that K-Means clustering be employed for further customer segmentation, and the last draw a conclusion

Methods

The use of the K-Means algorithm to cluster customer preferences and cake sales data during three major holidays at Lily's cake was highlighted in the introduction section. The methods and processes used in this investigation will be methodically described in this part.

Research Framework And Stages



Problem Identification

For the research to be beneficial for Lily's Cakes' future interests, it is important to understand the issues faced by the research subjects before examining the data that is now accessible. As mentioned before, Lily's Cake still uses traditional business procedures, without further processing or analysis of the manually entered data. Better product segmentation based on consumer preferences may be made possible by this data. Designing more specialized product offerings and increasing sales during high demand periods (The three major holidays) requires a deeper understanding of these tastes. With this issue in mind, the objectives of this study are:

1. To Address the product segmentation problem by grouping sales data according to customer preferences during major holiday periods.
2. To tailor the product offering based on customer preferences, thus enhancing the relevance of products offered.

Data Submission And Collection

For this research, data sourced from Lily's cake sales was used as the object of analysis. A total of 599 entries were collected, covering sales during three major holidays: Christmas, Chinese New Year, and Eid al-Fitr. The data, provided through manual records by the business owner, includes transaction details for each cake sale. These holidays were selected due to their significant influence on product demand. However, the reliance on

manual records may lead to poorly structured data, with potential issues such as inconsistent product names and recording errors.

Data Recapitulation

After collecting sales data from manual records, the data was converted into a spreadsheet format, for additional analysis, a total of eleven attributes were inserted in the main spreadsheet, such as customer name, gender, occupation, residence, purchased product, product type, product category, product price, quantity, total purchase. Following that, a selection process was carried out to identify which qualities would be used as key variables in the clustering analysis, ensuring that only relevant and significant features were picked to create accurate and informative clusters. Four attributes were chosen for the clustering procedure and transferred to another spreadsheet file named "convert". These attributes are gender, occupation, domicile, and cake type.

Clustering With K-Means

In data mining, clustering is an unsupervised learning process that divides data into smaller groups based on shared attributes. Since it enables the discovery of patterns in the data without the need for pre-existing labels. Dalmaijer et al. (2022) state that this technique is helpful for discovering subgroups with distinct characteristics even when there is overlap. The methodology employed in this research was the K-Means algorithm, which divides data points from signal processing (Abdulazeez, 2021). After the data is collected. It will be clustered using the K-Means technique. The first stage in the K-Means approach is determining the optimal number of clusters. The silhouette method might help to estimate the optimal number of clusters. The silhouette can determine how close a point is to its centroid (cohesion) in comparison to other clusters (separation). For example, the silhouette for data (i), is displayed:

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{max\{a(i),b(i)\}} & if\ |C_i| > 1 \\ 0 & if\ |C_i| = 1 \end{cases} \quad (1)$$

Notes:

1. $a(i)$ represents the cohesiveness of data (i) or the average distance between the data (i) and other points in the same cluster.

   $$a(i) = \frac{1}{|C_i|-1} \sum_{\substack{j \in C_i \\ i \neq j}} d(i,j) \quad (2)$$

2. $b(i)$ represents the separation of data (i) from data in other clusters.

   $$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i,j) \quad (3)$$

3. Where $|C_j|$ represents the cardinality of cluster (j), or the number of elements included within it.
4. If $|C_i| > 1$ then, the calculation will be carried out using the formula.

   $$s(i) = \left\{ \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \right. \quad (4)$$

   Measures how close a data (i) to its own cluster compared to other clusters, where the silhouette value ranges from -1 to 1, higher values indicate that the data point is more likely to be in its own cluster.
5. If $|C_i| = 1$ then, the silhouette $s(i)$ is set to 0, because the cluster contains just 1 data (i) and there are no other data points in the same clusters to compare the distances, therefore the silhouette value cannot be calculated using the normal formula.

The author attempts to use the numbers 1-4 as the ideal range for testing clusters. As shown in figure 2a, from the 4 numbers tested, there are 2 possible cluster counts that can be used: 2 and 4 clusters. When compared to the complete cluster, the average silhouette value is high, particularly for 2 clusters; however, for

3 clusters, the average silhouette decreases from two clusters. Additionally, there is an increase from 3 clusters to four. In summary, 2 clusters and 4 clusters can produce optimal clustering results. but based on the calculation results 4 clusters are not as good as 2 clusters. 3 Clusters, however, are unsuitable for clustering as they will not offer the best grouping.

The silhouette distribution results for each cluster are then displayed in Figure 2b, where all clusters compared have generally positive outcomes. 1 through 3 clusters have high silhouette values or are almost equal to 1. Despite a few negative values that indicate that some objects may be in less suitable clusters, 4 clusters can also be considered to have good values. According to the above study, it may be inferred that 4 clusters allow for more precise and detailed grouping. As a result, the number of clusters for the test was determined to be 4. Based of the result of the silhouette computation above, which show that 4 clusters are optimal number, the data is separated into 4 clusters.

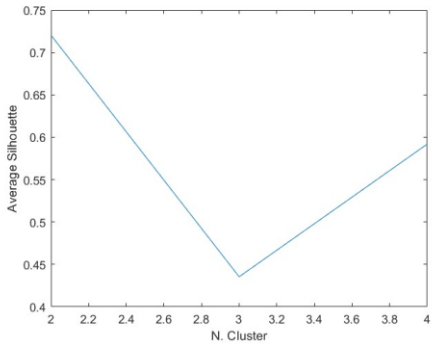


Figure 2a. Silhouette Analysis

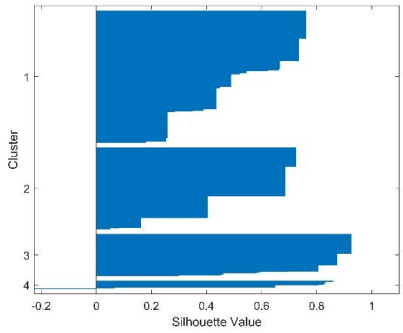


Figure 2b. Silhouette Plot of Clustering Results

The centroids to be utilized can be chosen at random once the optimal number of clusters has been determined, and then the distance to the nearest centroid is measured. Metrics for measuring mathematical distance are crucial for enhancing the K-Means algorithm's output (Ghazal et al., 2021). For this measurement, there are many metrics that can be used, including Minkowski distance, Kosinus distance, Euclidean distance, and Hamming distance.

Result and Discussion

Based on this research, the scatter plot analyzed is based on product category and occupation, as shown in Figure 3a. The result can be taken after the analysis is carried out are that this image only displays 3 clusters namely cluster 2, 3, and 4. Cluster 2, which is dotted at (2,1) and (3,1) is a group of customers who come from self-employed jobs and employees who prefer cookies (y1). This can be said to be relevant to and make sense, because of the culture of Indonesians who like to exchange cakes or cookies during major holidays. The marketing strategy that can be done is to present parcels with several types of cookies coated with wrappers according to the theme of celebrated holiday. Cluster 3, which is dotted at (2,2), (2,3), (3,2), and (3,3), is populated by employees and self-employed individuals who enjoy kue lapis (y2) and kue roll (y3). This can happen because this group may also have the same reason as cluster 2, because there is a tradition that appreciates to exchange souvenirs such as cake to enhance relationships, and there is also a workplace tradition that celebrates holidays together, so kue lapis and kue roll are the option to be served at the event. Cluster 4, which is dotted at (1,1), (1,2), and (1,3), consists of customers characterized as housewives who are interested in various categories of cakes such as cookies, kue lapis, and kue roll. This result is arguably relevant because this group of customers tends to look for cake products for family consumption or dishes to be served during major holidays, for this cluster, the marketing strategy that can be planned is to present hampers that provide all 3 types of cakes so that the products purchased can include all 3 types of cakes to make it look more diverse. Cluster 1 is inconclusive and does not appear in this scatter plot figure.

In Figure 3b, which discusses occupation and cake category, cluster 1, which is dotted at (4,1), (3,2), (3,3), and (6,1), depicts the grouping of customers who reside in Ketapang city (x3) and Sepauk city (x6). This cluster's characteristics include the fact that customers from customers Ketapang city domicile have a wide

interest in cakes and cookie category, as evidenced by their varied preferences among the 3 types of cakes and cookie: cookie(y1), kue lapis (y2), and kue roll (y3), while customers from Sepauk domicile prefer cookies, due to the practicality of serving. As for cluster 2 which is dotted at (1,1), (1,2), (1,3), and (2,3), which shows customers who live in Balai bekuak (x1) show more interest in the 3 categories of cakes available, while the preferences of customers who live in Jakarta tend to buy rolls perhaps reflecting a busier lifestyle, and the importance of practicality. In cluster 3, which is dotted at (5,2), (6,2), (6.3), (7,3), (8,2), (8,3), (9,2), and (9,3) showing a strong preference for kue lapis, kue roll, customers residing in Sampit (x5) showed interest in kue lapis, Sepauk (x6) showed interest in kue lapis dan kue roll, Singkawang (x7) shows interest in roll, Sintang (x8) shows interest in layer and roll, and Tayan (x9) shows interest in kue lapis and kue roll, where it can be analyzed that the characteristics for this cluster are more likely to like wet or baked cakes. And in cluster 4, which is dotted at (2,1), (2,2), (4,1), (4,2), (4,3), is a cluster of customers from big cities who tend to like almost all categories of cakes. Jakarta customers (x2) have a tendency towards cookies, and kue lapis, while Pontianak customers (x4) tend to buy all cake categories.

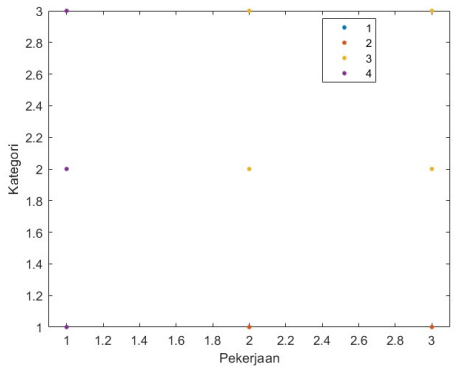


Figure 3a. Scatter Plot of occupation (x) and category (y)

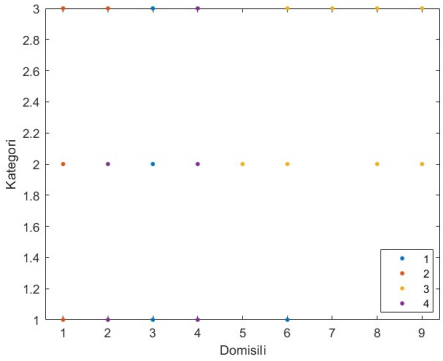


Figure 3b. Scatter Plot of domicile (x) and category (y)

based on research findings and software implementation, Lily's cake can offer unique product offers to specific customer categories and make targeted product offers to customers in particular categories. Lily's cake may benefit from this straightforward yet useful sales data management

Conclusion

This research demonstrates that these 4 clusters are utilized to group the data. After the silhouette approach is used to establish the optimal number of clusters. In the segmentation based on occupation and cake category, cluster 2 comprises of independent contractors and cookie-loving employees. Cluster 3 consists of independent contractors and employees who like kue lapis and kue roll. Clusters 4 has varying tastes for cookies, kue lapis, and kue roll. And cluster 1 cannot be displayed. In terms of customer segmentation by cake category and domicile, cluster 1 revealed that customers from Ketapang city and Sepauk had a diverse range of cake preferences, including kue lapis, kue roll and cookies. Cluster 2 shows that customers from Balai Bekuak were interested in all varieties of cakes and cookies, whereas those from Jakarta chose kue roll. Cluster 3 customers from Sampit, Sepauk, Singkawang, Sintang, and Tayan Particularly enjoyed kue lapis and kue roll. Finally, in cluster 4, Pontianak customers expressed an interest in all 3 varieties of cakes, whereas Jakarta customers preferred cookies and kue lapis.

Given the data, Lily's cake can adjust their product range and marketing techniques to better cater to their customers. Additionally, understanding the interests of customers from various domiciles can aid the development of specialized promotions and product bundles.

References

Abdulazeez, A. M. (2021). PJAEE, 17 (7) (2021) Meta-Heuristic Algorithms for K-means Clustering: A Review META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING: A REVIEW Meta-Heuristic Algorithms for K-means CLUSTERING .

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2022). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Quality and Quantity*.

Ahmed, M., Seraj, R., & Islam, S. M. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics (Switzerland)*.

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics*.

Gautam, N., & Kumar, N. (2022). Customer segmentation using k-means clustering for developing sustainable marketing strategies. *Business Informatics*, 72-82.

Ghazal, T. M., Hussain, M. Z., Said, R. A., Nadeem, A., Hasan, M. K., Ahmad, M., . . . Naseem, M. T. (2021). Performances of k-means clustering algorithm with different distance metrics. *Intelligent Automation and Soft Computing*, 735-742.

Hou, R., Ye, X., Zaki, H. B., & Omar, N. A. (2023). Marketing Decision Support System Based on Data Mining Technology. *Applied Sciences (Switzerland)*.

Ibáñez, S. J., Gómez-Carmona, C. D., & Mancha-Triguero, D. (2022). Individualization of intensity thresholds on external workload demands in women's basketball by k-means clustering: Differences based on the competitive level. *Sensors*.

Liu, F., Yang, D., Liu, Y., Zhang, Q., Chen, S., Li, W., . . . Wang, X. (2022). Use of latent profile analysis and k-means clustering to identify student anxiety profiles. *BMC Psychiatry*.

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *rtificial Intelligence Review*, Data clustering: application and trends.

Pontes, R. V. (2020). Applying K-means Clustering to Create Product Recommendation System Based on Purchase Profiles. *NAVUS-REVISTA DE GESTAO E TECNOLOGIA*.

Santana, R. V., & Pontes, H. L. (2020). Applying K-means Clustering to Create Product Recommendation System Based on Purchase Profiles. *NAVUS-REVISTA DE GESTAO E TECNOLOGIA*.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, 80716-80727.

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability (Switzerland)*.

Wang, X., Shao, Z., Shen, Y., & He, Y. (2023). Research on fast marking method for indicator diagram of pumping well based on K-means clustering. *Heliyon*.

Zhao, H. H., Luo, X. C., & Lu, R. M. (2021). An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation with Correlated Variables. *IEEE Access*.