Enrichment: Journal of Management, 14 (2) (2024)

Published by: Instiute of Computer Science (IOCS)

Enrichment: Journal of Management

Journal homepage: www.enrichment.iocspublisher.org

Customer profiling with k-means clustering and product recommendation with market basket analysis for strategy marketing MSMEs

Ahmad Al Ayubi[1], Hendra Achmadi[2]

Faculty of Economy, Universitas Pelita Harapan, Tangerang, Indonesia

ARTICLE INFO

Article history:

Received Jun 05, 2024
Revised Jun 16, 2024
Accepted Jun 30, 2024

Keywords:

CRISP-DM;
K-Means Clustering;
Market Basket Analysis;
RFM-D Analysis.

ABSTRACT

This research aims to develop a data-driven marketing strategy to increase consumer purchase interest in XYZ Hijab. This small and medium-sized Muslim fashion enterprise experienced a sales decline of 65.2% in 2022 and 2023. Utilizing the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology and employing Python programming language with Google Colab, this study combines RFM-D analysis and K-means clustering for customer segmentation, as well as Market Basket Analysis (MBA) for product bundling strategies. The study uses sales transaction data from December 9, 2023, to January 8, 2024. The analysis results indicate that the optimal RFM-D model uses four customer clusters: Superstar Customers (50.37%), Golden Customers (31.92%), Typical Customers (17.65%), and Dormant Customers (0.06%). The MBA identifies 11 product association rules that can be utilized for bundling strategies. The recommended marketing strategies include exclusive loyalty programs for top customers and tailored promotions for potential and dormant customers. Implementing these strategies will increase customer retention and revenue for XYZ Hijab.

Corresponding Author:

Ahmad Al Ayubi,
Faculty of Economy
Universitas Pelita Harapan
The Plaza Semanggi, Jl. Jend. Sudirman No.50, Jakarta,12930, Indonesia
Ahmadalayubi64@gmail.com

INTRODUCTION

International Data Corporation (IDC) Asia Pacific (International Data Corporation, 2023) projects that the transaction value of e-commerce in Indonesia will reach 118 billion USD by 2027. This growth in transaction value is dominated by digital payment methods. This growth reflects a transformation in consumer preferences in Indonesia, which are increasingly shifting towards the convenience offered by online transactions, as opposed to traditional purchasing approaches.

The projected growth in e-commerce transaction values aligns with Indonesia's increasing number of e-commerce users. Statista (Statista, 2024) projections, the number of users in the e-commerce market in Indonesia is expected to continue rising between 2024 and 2029, reaching a total of 33.5 million users (+51.03 percent). The increasing value of transactions and the use of e-

commerce in Indonesia can lead to a significant rise in the number of e-commerce platforms. This aligns with the fact that e-commerce in Indonesia contributes 52% of the revenue (International Trade Administration, 2024). The increasing number of platforms has led to increasingly fierce competition among various e-commerce platforms to attract sellers, stores, and consumers.

Despite intense competition, Shopee remains the most visited e-commerce platform, according to data from Similar Web cited by Data Boks in 2023. Shopee has managed to maintain its position as the e-commerce platform with the highest number of visits in Indonesia, receiving approximately 2.3 billion visits, reflecting a significant increase of 41.39% since the beginning of the year (Data Boks, 2023). Due to the increasing interest in purchasing Muslim fashion products, the number of sellers using the Shopee platform to market these products has significantly increased, creating more intense competition among sellers or brands operating on the platform. The intense competition has led to a decline in sales for one of the Muslim fashion MSMEs, XYZ Hijab, based on data from 2022 and 2023, showing a decrease in sales of approximately 65.2%.

Therefore, XYZ Hijab strives to increase purchase interest to boost profits and revenue. In pursuit of this goal, the company emphasizes the importance of customer orientation by implementing strategies focused on communication, understanding, and recognition of customer needs. This ensures that customer needs are effectively met (Mintardjo, 2022). By adopting this customer-oriented approach, XYZ Hijab will leverage existing customers to make additional transactions or add more products to their shopping cart, ensuring customer needs are effectively fulfilled and fostering purchase intention.

To meet customer needs, Tsiptsis and Chorianopoulos (Tsiptsis & Chorianopoulos, 2010) emphasize in their book the importance of using data mining for business managers and retailers to extract insights from customers and enhance interactions with consumers in a more personalized manner. Data mining is the process of discovering knowledge in databases. Data mining is crucial in understanding customer behaviour by uncovering hidden patterns in large datasets (Pahwa et al., 2017). By analysing past purchases and predicting future trends, XYZ Hijab can gain deep insights into consumer preferences that are invisible through regular database queries. It involves stages of data selection, data cleaning, data integration, data transformation, data mining, pattern evaluation, and visual knowledge presentation (Supoyo & Prasetyaningrum, 2022). Customer transaction data is one source that can be utilized in data mining to analyze online customers. This transaction data can then serve as the basis for evaluating and setting marketing strategies for the upcoming period (Brick, 2023)

One approach for understanding customer segmentation RFM-D analysis. The RFM-D model, an extension of the RFM model, introduces an additional measure called Diversity; this model evaluates customer value and segmentation based on four main components: Recency (R), Frequency (F), Monetary (M), and Diversity (D) (Smaili & Hachimi, 2023). By integrating Diversity into the RFM model, marketers gain deeper insights into customer behavior, enabling more effective segmentation and marketing strategies. Segmentation based on RFM-D is applied in retail markets to detect customer behaviour patterns because the proposed RFM-D model can improve the quality of customer behaviour prediction so that companies can predict which customers will respond positively (Smaili & Hachimi, 2023). The researcher then conducted an analysis using the clustering method. Clustering is a technique used to group data based on the similarity of characteristics between one data point and another (Shaliha et al., 2021).

One of the clustering methods used by researchers is K-means clustering. The K-means algorithm is a clustering method that requires determining the number of clusters "k" and the objects "n" to be grouped into these clusters beforehand (Mulyo & Heikal, 2022). According to Andre, Widya, and Juanda (Lubis et al., 2023) applying the K-means algorithm in data clustering is based on an iterative process to find the cluster centers, which form the fundamental basis of this algorithm. Therefore, the researcher used the K-means algorithm in the clustering approach. It effectively analyses structured data by reducing the variance within clusters and increasing the

difference between clusters, which is crucial to ensure more accurate empirical conclusions (Bui & Bahtiar, 2024). In its application, the K-means algorithm has advantages and limitations. The advantage of the K-Means algorithm is that it is able to group large objects and can increase the speed of the clustering process (Bui & Bahtiar, 2024). The advantage of applying the K-Mean algorithm is that the wrong initial cluster selection can produce inaccurate results (Ikotun et al., 2023). Therefore, determining the correct number of clusters requires a careful and planned approach.

In addition to applying the K-Means algorithm, this study utilizes the Market Basket Analysis (MBA) method for product bundling strategies. Market Basket Analysis, also known as Association Rule, is a method for discovering relationships between data by analyzing data patterns. This allows researchers to explore and identify existing patterns in the data more deeply (Amna et al., 2023). One algorithm that can be used in the Market Basket Analysis (MBA) method is the Apriori algorithm. The Apriori algorithm aims to find relationships between frequently purchased products in sales transactions; this helps companies discover related product combinations, thereby determining which products should have their stock increased or decreased to enhance sales revenue (Albab & Hidayatullah, 2022).

RESEARCH METHOD

In this study, the researchers used the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology using the pyhton programming language with google colab. CRISP-DM consists of a sequence of steps typically involved in data mining studies (Olson & Delen, 2008). The CRISP-DM process model contributes to data mining as a process, as reflected in its origins (Hahsler et al., 2005). Olson and Delen (2008) argue that the CRISP process provides comprehensive coverage of the business understanding, data understanding, data preparation, model building, evaluation and deployment.

Business Understanding

The first stage of business understanding involves setting business objectives, evaluating the current situation, collecting relevant data, determining mining objectives, and developing a project plan (Olson & Delen, 2008). Collen (2007), business understanding is one of the most important stages in the data mining process. Business understanding in this research is centered around boosting consumer purchasing interest through strategic marketing for the XYZ Hijab. This imperative arises from the company's experiencing a decline in revenue.

Data Understanding

After establishing the business understanding, the next stage is data understanding, which accommodates the data requirements; this process involves initial data collection, data description, data exploration, and data quality verification. The data exploration stage, including summary statistical analysis (including visualisation of categorical variables), is usually conducted in the final stage (Olson & Delen, 2008). Models such as cluster analysis can also be applied in this stage to identify patterns in the data.

Collen (2007) second step in the CRISP-DM analysis process is the data understanding stage. At this stage, data is collected, and the analyst begins to explore and understand the characteristics of the data, including its form, content, and structure. In this study, researchers collected sales data for Dezhaf Hijab on Shopee from 9 December 2023 to 8 January 2024, consisting of a total of 4156 observations across 11 variables.

Data Preparation

After data understanding, the next step is data preparation by selecting, cleaning, building, and formatting the data into the desired form (Olson & Delen, 2008). At this stage, data cleansing and transformation are performed in preparation for data modeling. Deeper data exploration can be conducted at this phase, and additional models can be reused to uncover patterns based on a better understanding of the business. Collen (2007) data preparation involves cleaning and recoding, as well as selecting the necessary training and testing samples. During this stage, merging or re-merging of required datasets or elements is also performed. The aim of this step is to create a dataset that will be used in the subsequent modeling stage of the process.

Model Building

The fourth stage is data mining modeling, which involves visualization (plotting data and establishing relationships) and cluster analysis (identifying suitable variables), which is useful for initial analysis. The first RFM-D Analysis, the RFM-D model extends the traditional RFM (Recency, Frequency, Monetary model by adding the additional dimension of Diversity (D). It is designed to provide a more comprehensive analysis of customer behavior by incorporating the variety of products each customer purchases (Smaili & Hachimi, 2023). Recency is the time interval between the last purchase date and the last date of the statistical period (Imani et al., 2022). A low Recency value characterizes a good customer who visits the company repeatedly (Smaili & Hachimi, 2023). Researchers added the 'Days Since Last Purchas' feature based on this theory as Recency.

The second stage of the researcher examines Frequency. During the period under study, Frequency refers to the number of purchase transactions made by customers. A customer is considered loyal if the Frequency of purchases they make is high enough (Smaili & Hachimi, 2023). Frequency is shows the number of times a customer purchased during the statistical period (Christy et al., 2021). Frequency can also be defined as the total number of transactions made within a specified time (Kumar. V & Reinartz, 2012).From this theory, the researcher created two features that measure the frequency of customer engagement with retailers, namely: a) Total transactions: This feature represents the total number of transactions made by the customer. b) Total Products Purchased: This feature shows the total number of products (quantity) customers purchase across all transactions.

The third researchers examined monetary, this metric shows the overall spending of customers within the specified study period. A higher value for this parameter signifies that the customer has spent more money, thus benefiting the company financially (Smaili & Hachimi, 2023). In this theory, researchers will create two features that represent the monetary aspect of customer transactions: a) Total Spend: This feature represents the total money spent by each customer. It is calculated as the sum of the product of Unit Price and Quantity for all transactions made by the customer. b) Average Transaction Value: This feature calculates Total spending divided by Total Transactions for each customer and shows the average value of customer transactions.

The fourth, researchers examined diversity, which is a matrix for a measure representing the number of products that interest customers during the period under study. The larger the diversity parameter, the more open customers are to new products and providing new offers from the company to potential customers (Smaili & Hachimi, 2023). In this theory, researchers will create a feature, 'Unique Products Purchased,' which represents the number of different products purchased by customers. Understanding the diversity in product purchases can help segment customers based on the diversity of their purchases, which can be an important input in personalising product recommendations.

After completing the RFM-D analysis, researchers conducted a second model building for clustering analysis using the K-Means method to determine the optimal number of clusters. K-means aims to find the optimal clusters in a dataset by dividing it into k groups. K-Means is a

machine learning algorithm used for unsupervised modeling. It is one of the methods that divides data into groups or clusters based on partitioning (Agustino et al., 2022). The optimal number of clusters is determined based on the least-squares error resulting from data partitioning. Researchers determined the cluster using the elbow method. The elbow method finds the dataset's optimal number of 'K'. The elbow method aims to find the smallest 'K' value with a low inertia value (Humaira & Rasyidah, 2020). The elbow method focuses on the percentage of variance as a function of the number of clusters. Determining the number of clusters in the dataset is done by looking at the position of the elbow point (Rao et al., 2018). Researchers used the Yellowbrick Python library to analyze K-Means clustering using elbow methods.

After conducting a K-means clustering analysis, researchers will conduct a market basket analysis and search for item set recommendations. Initially, researchers will create a new data set with the variables 'Days Since Last_Purchase, Total Transactions, Total Products Purchased, Total Spend, Average Transaction Value, Unique Products Purchased, Cluster, InvoiceNo, InvoiceDate, ProductName and Quantity'. For market basket analysis, transaction data will be processed to generate rules that fulfill predefined minimum support and confidence values. The optimal minimum values for support and confidence are determined through several iterations (Liu et al., 2018). The inputs consist of transaction data and the minimum support and confidence values used to identify strict rules that match those parameters. These inputs are then processed using the Apriori algorithm to generate product association rules.

Evaluation

The next phase of the model must be evaluated in the context of the business objectives set in the first phase (business understanding). This will lead to identifying other needs (often through pattern recognition), often returning to the previous phase of CRISP-DM (Olson & Delen, 2008). In the evaluation phase of the project, the models that have been developed will be thoroughly reviewed to determine their accuracy and ability to achieve the goals and objectives identified in the business understanding phase.

Deployment

In the last stage, deployment is the process of transferring the results of data analysis, prediction models, or algorithms from the development and testing environment to the production environment (McCue, 2007). Deployment is the phase where the developed and validated model is applied to actual business operations (Olson & Delen, 2008). The goal is for the analysis or model results to be used directly by end users or operational systems.

RESULTS AND DISCUSSIONS

RFM-D Analysis

The final results of the RFM-D analysis are presented in the figure

	CustomerID	Days_Since_Last_Purchase	Total_Transactions	Total_Products_Purchased	Total_Spend	Average_Transaction_Value	Unique_Products_Purchased
0	[illegible]	0	7	4	354000	354000.0	2
1	[illegible]	28	1	1	138000	138000.0	1
2	[illegible]	4	1	1	138000	138000.0	1
3	[illegible]	5	1	1	167000	167000.0	1
4	[illegible]	10	1	4	432000	432000.0	2
5	[illegible]	0	1	4	902000	902000.0	2
6	[illegible]	25	1	1	166000	166000.0	1
7	[illegible]	10	1	1	133000	133000.0	1
8	[illegible]	28	1	4	466000	466000.0	2
9	[illegible]	10	1	1	138000	138000.0	1

Figure 1 RFM-D Analysis results

Source: Data Processed

The data presented in Figure 1 indicates that the 'Days Since Last Purchase' represents the time interval since the last purchase, ranging from 4 to 28 days. Each customer's 'Total Transactions' recorded only one transaction, but the 'Total Product Purchased' varies from 1 to 4. Significant fluctuations in 'total expenditure' and 'average transaction value' indicate variations in customer spending behavior. For instance, '0d7dnHl0ri', with the highest expenditure of Rp 502,000, may prefer premium products or bulk purchases, while '051500_y', spending Rp 135,000, may be more cautious or opt for more economical products.

Cluster Analysis

The result of K-means clustering with elbow method is presented in Figure 2.

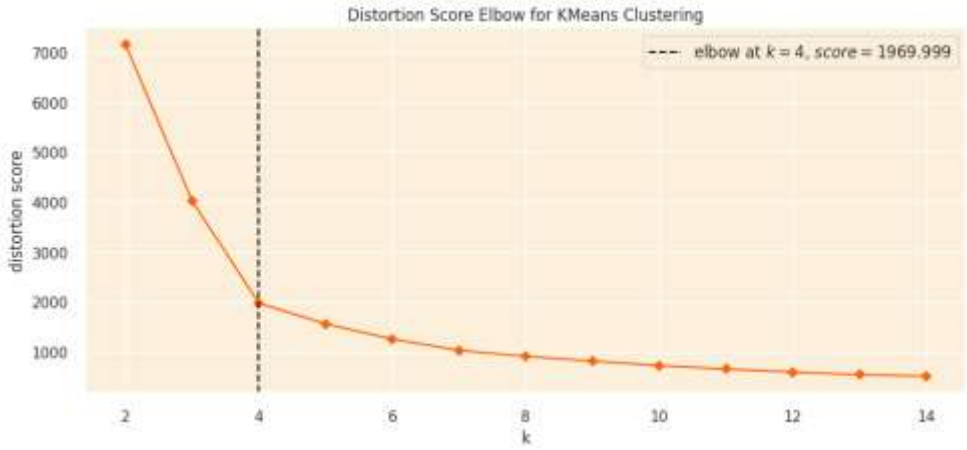

Figure 2 Elbow Method K-means Results
Source: Data Processed

If it has a sub-sectio The Elbow graph presented in Figure 2 shows the values of *k* ranging from 2 to 14, with the distortion score indicated on the vertical axis. From the graph, it is evident that there is a sharp decrease in the distortion score as *k* changes from 2 to 4. At *k* = 4, the graph shows an "elbow" point, indicating a significant reduction in the distortion score to approximately 1969.999. Beyond this point, the decrease in the distortion score becomes slower and relatively stable, with only slight further decreases up to *k* = 14. The flattening of the distortion score after *k* = 4 suggests that adding more clusters does not provide a meaningful improvement in explaining the data's variation.

Based on the analysis of the Elbow graph, *k* = 4 is a good choice for the number of clusters. Choosing this number of clusters allows the model to achieve an efficient balance between accuracy and complexity, ensuring that each cluster is sufficiently homogeneous while avoiding overfitting that might occur with a larger number of clusters. This indicates that four clusters are sufficient to effectively group the data. From the clustering results, the percentage of customers among the four clusters identified in the cluster analysis can be distributed, as presented in Figure 3.

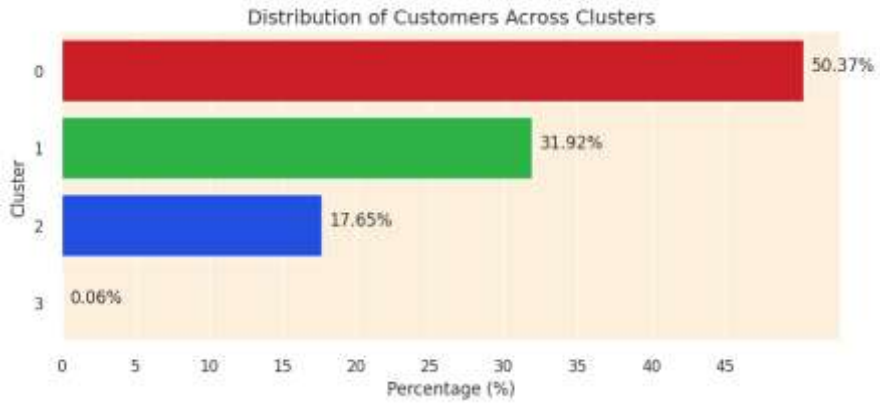

Figure 3 Distribution Customers Across Clusters
Source: Data Processed

From Figure 3 the distribution of customers has an unequal number between clusters: 1) Cluster 0 (Red): The most significant proportion of customers is 50.37% of the total customers. 2) Cluster 1 (Green): Accommodates 31.92% of customers, making it the second largest cluster in numbers. 3) Cluster 2 (Blue): This comprises 17.65% of customers, indicating that the number of customers is smaller than the previous two clusters, but it still represents an important customer segment. 4) Cluster 3 (Purple): Only 0.06% of customers make up the smallest cluster. This cluster may include customers with unique or specific characteristics that are very different from most other customers. This distribution provides essential insights into the heterogeneity within the customer base. Clusters with a higher percentage of customers (Clusters 0 and 1) may require more generic or mass-orientated marketing and service strategies. In contrast, smaller clusters (especially Clusters 3 and 4) may require a more specialized approach or personalization to meet needs and preferences.

Based on the Elbow graph analysis, k = 4 is a good choice for the number of clusters. Choosing this number of clusters allows the model to achieve an efficient balance between accuracy and complexity, ensuring that each cluster is sufficiently homogeneous while avoiding overfitting that may occur with a more significant number of clusters. This suggests that four clusters are sufficient to group the data effectively, providing a solid understanding of the data structure without compromising interpretability or computational efficiency. Researchers conducted a more in-depth evaluation of the clustering quality and adopted several metrics, as presented in Figure 4.

Metric	Value
Number of Observations	3117
Silhouette Score	0.6429978174265455
Calinski Harabasz Score	8813.322220451657
Davies Bouldin Score	0.34343215785480835

Figure 4 Evaluation Matric Results
Source: Data Processed

Based on Figure 4 a comprehensive evaluation of cluster metrics on a dataset comprising 3117 observations is conducted. This evaluation utilizes the Silhouette score, Calinski-Harabasz score, and Davies-Bouldin score, each providing insights into the quality of the cluster division produced by the model. These metrics indicate that the resulting cluster division is adequate, characterizing cohesive and well-separated data division. These results suggest that the clustering

model is reliable for further purposes such as customer segmentation or analysis aiming to develop targeted intervention strategies based on identified cluster characteristics.

Cluster Profiling

In this research, the researchers analyze the characteristics of each cluster to understand the different behaviors and preferences of various customer segments and profile each cluster to identify the key traits that define the customers within each cluster. The cluster profiling results are presented in the figure 5.

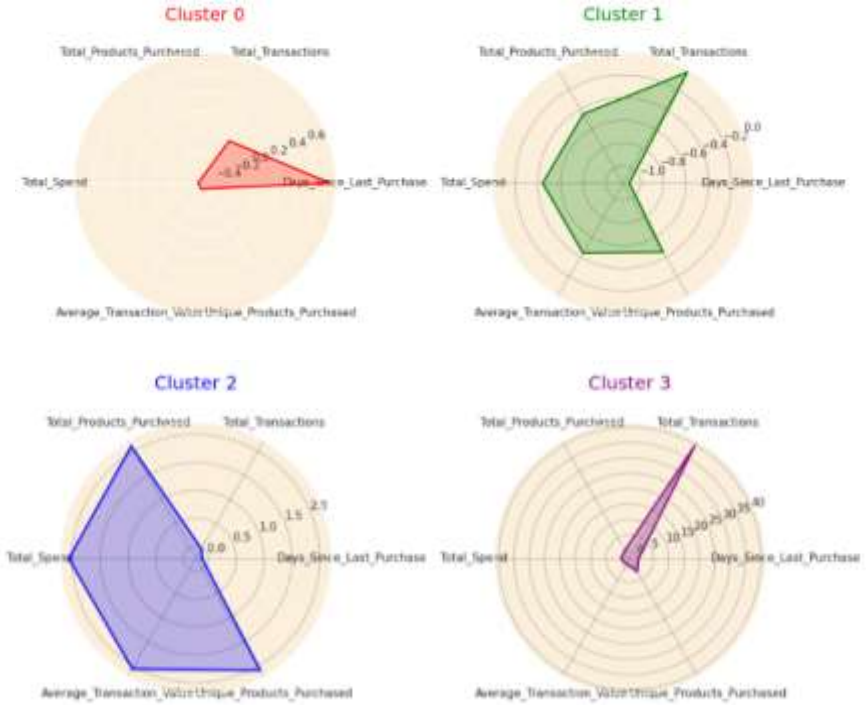

Figure 5 Radar Cluster Segmentation for k=0, k=1, k=2, k=3
Source: Data Processed

Researchers can plot histograms for each feature segmented by cluster labels to validate the profiles identified from the radar charts. These histograms will allow researchers to visually examine the distribution of feature values within each cluster, thus confirming or refining the profiles created based on radar charts.

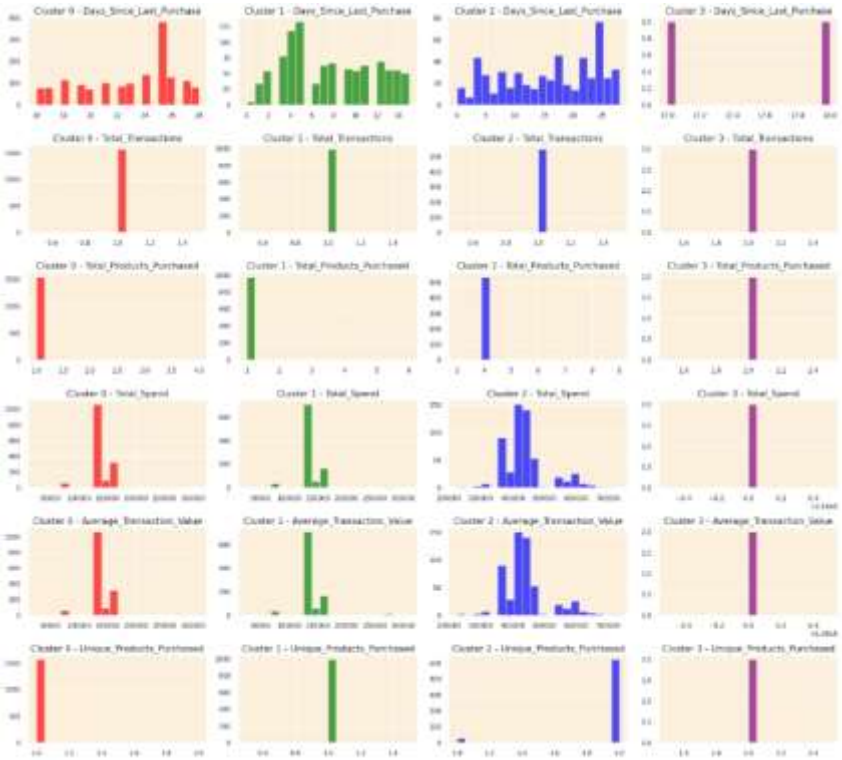

Figure 6 Histogram Cluster Segmentation for *k*=1, *k*=2, *k*=3, *k*=4
Source: Data Processed

Based on Figure 6, the refined profiles for each cluster are as follows: 1) The characteristics of cluster 0 exhibit a profile of "Superstar Customers," known for their high loyalty, high monetary value, and consistent transaction frequency. 2) The characteristics of the cluster 1 this profile is most aligned with the Golden Customers category, which has the highest monetary value after the Superstar category and a high purchase frequency. 3) The characteristics of the cluster 2 is Typical Customers, demonstrating moderate monetary value and transaction numbers. They make large purchases less frequently than Golden or Superstar Customers. 4) The characteristics of the cluster 3 is Dormant Customers, with the lowest frequency and monetary value, indicating a lack of engagement or declining interest during the observed period.

Market Basket Analysis

Next, Researchers will analyze market basket analysis (MBA) to determine what items to sell based on customer segmentation. The goal is to understand each customer segment's purchasing patterns and product preferences, identified through RFM-D and clustering analyses. points. The company has 34 SKU (Stock Keeping Units) across its various products in this study. Therefore, it is almost impossible to identify significant product patterns at the SKU level. To address this issue, purchase patterns need to be identified. Market basket analysis is performed using association rule mining to obtain purchase patterns, resulting in the following item sets from the association rules.

Table 1 Group Association Rules of Items

No	Atecedent	Consequent	Support	Confidence	Cluster
1	Wrap Skirt / Rok lilit Satin	Zada Outer Brukat Pendek / Outer Tile Bordir	0.001818	0.111111	2
2	Flowie Kutubaru / Kutubaru Tile Flowers	Lily Satin Skirt / Rok Satin Span Mermaid	0.003636	0.086957	2

3	Deana Outer Brukat / Outer Brukat / Outer kondangan	Ivona Outer Brukat / Outer Kondangan	0.001818	0.500000	2
4	Alana Series Outer Organza / Outer Kondangan	Inner Kaftan Ikat / Inner Tunik Rok Lilit / Inner Satin	0.001818	1.000000	2
5	Alesha Dress / Silk Dress / Dress Shimer	Flowie Kutubaru / Kutubaru Tile Flowers	0.001818	0.500000	2
6	Inner Satin Blouse / Inner Atasan Satin	Laudya Square | Plain Hijab Polycotton	0.001818	0.111111	2
7	Inner Kaftan Ikat / Inner Tunik Rok Lilit / Inner Satin	Zalina Outer Organza / Outer Organza / Outer kondangan	0.007273	0.102564	2
8	Flowie Kutubaru / Kutubaru Tile Flowers	Wrap Skirt / Rok lilit Satin	0.005455	0.130435	2
9	Inner Kaftan Ikat / Inner Tunik Rok Lilit / Inner Satin	Ivona Outer Brukat / Outer Kondangan	0.018182	0.256410	2
10	Outer Organza Motif / Outer Organza Tille Sapto	Zalina Outer Organza / Outer Organza / Outer kondangan	0.001818	0.076923	2
11	Laudya Square | Plain Hijab Polycotton	Paris Basic Jadul | Paris Jadul	0.000637	0.1	0

Source: Data Processed

Overall, these market basket analysis results reveal varying relationships between different clothing items, with some items having intense and frequent associations while others exhibiting weaker relationships. The support and confidence values help measure the frequency and strength of these associations. Cluster 2 encompasses most rules, indicating that items within this cluster are frequently bought together.

CONCLUSION

RFM-D analysis and k-means clustering results show that choosing (k = 4) is optimal based on the elbow method. Further profiling of this (k = 4) has identified different types of customers based on their behavior and their preferences. The four customer segments are as follows: Superstar Customers, representing 50.37% of the total; Golden Customers, totalling 31.92%; Typical Customers, accounting for 17.65%; and Dormant Customers account for only 0.06%. Thus, there is an imbalance in cluster management, so there are constraints in managing the cluster, namely, high dependence on large segments such as "Superstar Customers." and "Golden Customers" for revenue because in the event of changes in purchasing behavior or loss of loyalty from these segments can experience a significant drop in revenue.

The low number of "Dormant Customers" is 0.06%. This is due to individuals (customers) who have stopped interacting with XYZ Hijab or make infrequent purchases, and their inactivity is caused by various reasons, such as better offers from competitors or lack of personalized engagement according to their preferences. This can be seen from the lowest frequency and monetary values, indicating a lack of engagement or decreased interest during the observation period. So, to increase the number of customers in cluster 3, it is necessary to implement targeted retention strategies such as personalized emails or app (shopee) notifications that feature new products or exclusive offers based on previous purchase history to reawaken their buying interest in addition, incentives such as discounts or free shipping can encourage repeat purchases and gradually turn these dormant customers into more active ones. In addition to Cluster 3, implementation was also carried out on "Superstar Customers" (Cluster 0) by conducting an exclusive loyalty program that offers customer membership tiers that customers can rely on for months or even years and gives customers shopping goals to achieve. The benefits that will accrue

to customers who join as members are coin rewards for product purchases that can be redeemed for vouchers and gifts. In addition, "Gold Customers" (Cluster 1) and "Typical Customers" (Cluster 2) have promotions highlighting various product offerings.

To complete the marketing strategy implementation, the researcher conducted a shopping cart analysis. The shopping cart analysis has also revealed some relationships between the various clothing items, resulting in 11 different product recommendations. Cluster 2 is the most relevant, as most association rules are concentrated in this cluster. This suggests that items in Cluster 2 are often purchased together, highlighting specific buying patterns and affinities among customers within this segment. Leverage shopping cart analysis insights to recommend complementary products and cross-sell opportunities. Implement an interactive campaign that showcases frequently co-purchased product combinations within each cluster. For example, promote Inner Kaftan Ikat / Inner Tunic Wrap Skirt / Inner Satin → Ivona Outer Brukat / Outer Kondangan, aimed at customers in Cluster 2.

The implementation of bundling promotions can be based on association rules with a high level of confidence, such as the rule "Alana Series Organza Outer → Outer Kondangan → Inner Kaftan Ikat / Inner Tunic Skirt Lilit / Inner Satin," which has a confidence level of 1.0. This high level of confidence indicates that purchasing Inner Kaftan Ikat or similar products almost always follows every Alana Series Organza Outer purchase. Therefore, the company can design promotional strategies by offering bundling packages or special discounts for purchasing these two items together.

ACKNOWLEDGEMENTS

Thank you to my lecturers, family, and friends for their support, motivation, and exceptional knowledge while completing this research.

References

Agustino, D. P., Harsemadi, I. G., & Budaya, I. G. B. A. (2022). Edutech Digital Start-Up Customer Profiling Based on RFM Data Model Using K-Means Clustering. *Journal of Information Systems and Informatics*.

Albab, M. U., & Hidayatullah, D. (2022). Penerapan Algoritma Apriori Pada Sistem Informasi Inventori Toko. *Jurnal Media Informatika Budidarma*.

Amna, Wahyuddin, Sudipa, I. G. I., Putra, T. A. E., Wahidin, A. J., Syukrilla, W. A., Wardhani, A. K., Heryana, N., Indriyani, T., & Santoso, L. W. (2023). *Data MInning* (1st ed.). PT Global Eksekutif Teknologi.

Brick. (2023). *Inilah Pentingnya Data Transaksi bagi Perusahaan Bisnis*. Https://Www.Onebrick.Io/Blog/Data-Transaksi.

Bui, M. A., & Bahtiar, A. (2024). Implementasi Metode Algoritma K-Means Clustering Untuk Mengelompokkan Transaksi Penjualan Barang Di Toko Arino. *Jurnal Mahasiswa Teknik Informatika*.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). *RFM Ranking –An Effective Approach to Customer Segmentation*.

Data Boks. (2023, October 11). *5 E-Commerce dengan Pengunjung Terbanyak di Indonesia (Januari-Desember 2023)*. Https://Databoks.Katadata.Co.Id/Datapublish/2024/01/10/5-e-Commerce-Dengan-Pengunjung-Terbanyak-Sepanjang-2023. https://databoks.katadata.co.id/datapublish/2023/10/11/pengunjung-shopee-makin-banyak-bagaimana-e-commerce-lain

Humaira, H., & Rasyidah, R. (2020). *Determining The Appropriate Cluster Number Using Elbow Method for K-Means Algorithm*.

Ikotun, A. M., Ezugw, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, And Advances in The Era Of Big Data. *Information Sciences*.

Imani, A., Abbasi, M., Ahang, F., Ghaffari, H., & Mehdi, M. (2022). *Customer Segmentation to Identify Key Customers Based on RFM Model by Using Data* .

International Data Corporation. (2023). *How Asia Buys and Pays 2023: Tapping into Asia's Regional Commerce Opportunities*. International Data Corporation.

International Trade Administration. (2024, January 9). *Indonesia - Country Commercial Guide*. Https://Www.Trade.Gov/Country-Commercial-Guides/Indonesia-Ecommerce.

Kumar. V, & Reinartz, W. (2012). *Customer Relationship Management: Concept, Strategy, and Tools*. Springer.

Liu, R., Lee, Y., & Mu, H. (2018). *Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from distribution big data of Korean retailing company*.

Lubis, A. H., Utami, W. R., & Lubis, J. H. (2023). *Implementation of K-means Clustering For The Job Provision in Urban Village*.

McCue, C. (2007). *Data Mining and Predictive Analysis Intelligence Gathering and Crime Analysis*. Elsevier.

Mintardjo, B. H. (2022). *Peran Orientasi Pelanggan Departemen Sales dan Marketing dalam Meningkatkan Pendapatan di Petit Boutique Hotel Solo*.

Mulyo, I. A., & Heikal, J. (2022). *Customer Clustering Using The K-Means Clustering Algorithm in Shopping Mall in Indonesia*.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*. Springer Berlin.

Pahwa, B., Taruna, S., & Kasliwal, N. (2017). Role of Data Mining in Analyzing Consumer's Online Buying Behavior. *International Journal of Business and Management Invention*.

Rao, K. R., Garg, S., & Montgomery, J. (2018). *Investigation of unsupervised models for biodiversity assessment*.

Shaliha, K. M., Angelyna, A., Nugraha, A. A., Wahisyam, M. H., & Sandi, T. K. (2021). *Implementasi K-Means Clustering pada Online Retail berdasarkan Recency, Frequency, dan Monetary (Implementation of K-Means Clustering in Online Retail based on Recency, Frequency, and Monetary)*.

Smaili, M. Y., & Hachimi, H. (2023). *New RFM-D Classification Model for Improving Customer Analysis and Response Prediction*.

Statista. (2024, May 14). *Number of Users of E-commerce in Indonesia from 2020 to 2029*. Statista.Com.

Supoyo, A., & Prasetyaningrum, P. T. (2022). *Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Di DIY*.

Tsiptsis, K., & Chorianopoulos, A. (2010). *Data Mining Techniques in CRM: Inside Customer Segmentation*. Wiley.