SCIENCE AND INFORMATION TECHNOLOGY
SINTECH
JOURNAL

SINTECH JOURNAL | ISSN 2598-7305 | E-ISSN 2598-9642
Vol. 8 No 2 – Agustus 2025 | https://s.id/sintechjournal
DOI: https://doi.org/10.31598
Publishing: Prahasta Publisher

Optimization of Customer Segmentation with RFM, K-Means, and FP-Growth for Marketing Strategy

Solehhuddin Aulia[1], Asyiq Nur Muhammad[2], Arief Wibowo[3]

[123] Magister Ilmu Komputer, Fakultas Teknologi Informasi Universitas Budi Luhur,
Jl. Ciledug Raya, RT.10/RW.2, Petukangan Utara, Kec. Pesanggrahan, South Jakarta City, Special Capital Region of Jakarta, Indonesia

e-mail: solehhudinaulia92@gmail.com[1], asyiqnur@outlook.co.id[2],
arief.wibowo@budiluhur.ac.id[3]

Received : August, 2025	Accepted : August, 2025	Published : August 2025

Abstract

Snap Digital Printing has experienced a 6% decline in sales since 2019 and a continuous decline in the number of customers until 2023. The main cause is the company's limited use of customer data, which hinders the identification of behavioral changes, weakens market response, and reduces loyalty. This study aims to evaluate the impact of inadequate customer data management on sales and customer numbers, while analyzing the effectiveness of the RFM model, K-Means algorithm, and FP-Growth in customer segmentation, purchase pattern analysis, and marketing optimization. Data was collected from the Ownshop Snap Digital Printing branch, covering the period from January 2019 to December 2023, with 7,203,059 records before cleaning and 7,029,561 after cleaning. The analysis identified three customer segments: Low Engagement 86%, Active 3%, and VIP 11%. The FP-Growth results showed average Support of 11.36%, 6.73%, and 7.85%; Confidence of 85.70%, 49.72%, and 82.60%; and Lift Ratios of 1.77, 7.67, and 4.92. The findings indicate that data-driven systems strengthen customer-focused marketing strategies, enhance retention and sales, and provide a solid foundation for evidence-based practices in the digital printing industry.

Keywords: *Customer Segmentation, Data Mining, K-Means, RFM, FP-Growth, Purchase Pattern Analysis, Customer Retention.*

Abstrak

Snap Digital Printing mengalami penurunan penjualan sebesar 6% sejak 2019 dan penurunan terus-menerus dalam jumlah pelanggan hingga 2023. Penyebab utamanya adalah penggunaan data pelanggan yang terbatas oleh perusahaan, yang menghambat identifikasi perubahan perilaku, melemahkan respons pasar, dan mengurangi loyalitas. Studi ini bertujuan untuk mengevaluasi dampak pengelolaan data pelanggan yang tidak memadai terhadap penurunan penjualan dan jumlah pelanggan, sambil menganalisis efektivitas model RFM, algoritma K-Means, dan FP-Growth dalam segmentasi pelanggan, analisis pola pembelian, dan optimasi pemasaran. Data diambil dari cabang Ownshop Snap Digital Printing, mencakup periode Januari 2019 hingga Desember 2023, dengan 7.203.059 catatan sebelum pembersihan dan 7.029.561 setelah pembersihan. Analisis mengidentifikasi tiga segmen pelanggan, Low Engagement 86%, Active 3%, dan VIP 11%. Hasil FP-Growth menunjukkan rata-rata Support sebesar 11,36%, 6,73%, dan 7,85%; Confidence sebesar 85,70%, 49,72%, dan 82,60%; serta Lift Ratios sebesar 1,77, 7,67, dan 4,92. Temuan menunjukkan bahwa sistem berbasis data memperkuat strategi pemasaran yang berfokus pada pelanggan, meningkatkan retensi dan penjualan, serta memberikan landasan yang kokoh untuk praktik berbasis bukti di industri percetakan digital.

.

Kata Kunci: *Segmentasi Pelanggan, Data Mining, K-Means, RFM, FP-Growth, Analisis Pola Pembelian, Retensi Pelanggan*

1. INTRODUCTION

Customer segmentation is an important basis for understanding buyer behavior, which plays a role in pricing and demand forecasting to support business decision making [1]. In an increasingly competitive and complex business environment, systematic segmentation can increase customer loyalty and build long-term profitable relationships by expanding the customer database. This process aims to understand customer buying patterns and design appropriate marketing strategies, giving positive benefits to the Company [2].

Snap Digital Printing recognizes the importance of building strong relationships with customers to gain trust and achieve business success. However, despite these efforts, the company has experienced a decline in performance, as evidenced by a 6% drop in sales since 2019 and a decrease in the number of customers during the 2019-2023 period. As shown in Table 1, the number of lost customers defined as customers who did not make repeat transactions within a one-year period shows an increasing trend, from 142,490 customers in 2019 to 134,110 customers in 2022. .

Table 1: Customer's Data

Year	Total Customers	Sales Value in Billions	Total Customer Data Lost
2019	267.455	94.897	142.49
2020	200.663	66.162	104.211
2021	200.309	62.217	106.181
2022	228.482	78.588	134.11
2023	232.833	89.358	-

This research uses RFM and K-Means for customer segmentation, as well as FP-Growth to analyze purchasing patterns, helping Snap Digital Printing design effective data-driven marketing strategies to increase retention and sales.

RFM analysis has been recognized as a highly effective and widely applied method for understanding customer behavior, which enables the development of predictive models relating to such behavior [3]. In other words, RFM can also be used to classify customers based on their purchase history, relying on three main attributes, Recency, Frequency, and Monetary value. [4]. Recency, which measures the time gap between consecutive purchases, is an important metric that influences customer engagement with a brand [5]. Frequently engaged customers tend to be more loyal, indicating a higher level of involvement [6], [7]. The monetary value of a purchase indicates the amount a customer spends, which means customers with large purchases should be treated differently [8].

A review of the literature from 2000 to 2022 shows that K-Means is the simplest customer segmentation method to implement and is widely known and used. Out of the 105 publications analyzed, K-Means was used 41 times (39.0%). The primary objective of this algorithm is to divide a set of data points into k clusters by minimizing the distance between data points. Other widely used segmentation methods include the Hybrid approach 12 times (11.4%), various other approaches 10 times (9.5%), Fuzzy C-Means (FCM) 6 times (5.7%), Latent Class Model 6 times (5.7%), Evolutionary Algorithm 5 times (4.8%), Hierarchical 5 times (4.8%), Self-Organizing Map (SOM) 5 times (4.8%), Rough Set Theory 3 times (2.9%), and Deep Learning, Expectation-Maximization, and Spectral Clustering approaches, each used once (1.0%). In addition to the dominant K-Means method, the author also conducted further analysis by exploring product associations within each customer segment using the FP-Growth algorithm [9]. K-Means divides data into k groups by minimizing the distance between data.

Other than K-Means, this research uses FP-Growth algorithm to analyze product associations in each customer segment, supporting data-driven marketing strategies.

FP-Growth is more efficient in mining frequency patterns with lower memory and CPU usage than Apriori [2].

This research uses FP-Growth to mine sequential patterns after customer

segmentation is established through RFM and K-Means models, with the aim of designing relevant marketing strategies, improving customer retention, and sales.

Several studies use algorithms to improve marketing strategies and customer behavior analysis. One of them is Optimization Product Recommendation Using K-Means, Agglomerative Clustering And Fp-Growth Algorithm [10], The integration of clustering and association rule mining has been shown to improve marketing effectiveness, as evidenced by studies applying FP-Growth and Apriori for food package recommendations [11]. Other studies Clusterization of Agroforestry Farmers using K-Means Cluster Algorithm and Elbow Method [12], By combining K-Means clustering with FP-Growth, researchers are able to discover hidden customer segments and frequent purchasing patterns, which can be leveraged to provide more precise product package recommendations [13].

The application of RFM models and FP-Growth algorithms helps companies understand consumer spending behavior and purchasing patterns, thereby improving the effectiveness of marketing strategies, operational efficiency, and ultimately supporting increased company profitability [14], while FP-Growth identifies product patterns for promotion strategies [15]. The research also divided customers based on transaction behavior with K-Means [16] , and identify high profit customer segments using SPAK and K-Means [17].

This study focuses on the Ownshop Snap Digital Printing branch with data from January 2019 to December 2023 to evaluate the impact of indifference to customer data and the effectiveness of the RFM, K-Means, and FP-Growth models in segmentation and marketing strategies. The limitations of this study include the restricted use of methods to RFM, K-Means, and FP-Growth, evaluation of clusters solely using the Silhouette Coefficient and Davies Bouldin Index (DBI), and data sourced from Ownshop Snap branches with an analysis period limited to 2019–2023. . This research serves as a foundation for Snap Digital Printing to improve its customer behavior-based marketing strategy to support retention and sales, while also contributing academically to the impact of insufficient attention to customer data and the effectiveness of segmentation (RFM, K-Means) and association (Fp-Growth) algorithms in the digital printing business. Additionally, this research offers practical solutions to address customer identification challenges and provides direct benefits to similar companies through the application of analytical methods in data-driven marketing strategies.

2. RESEARCH METHOD

This research uses the Cross Industry Standard Process for Data Mining (CRISP-DM) as a framework.

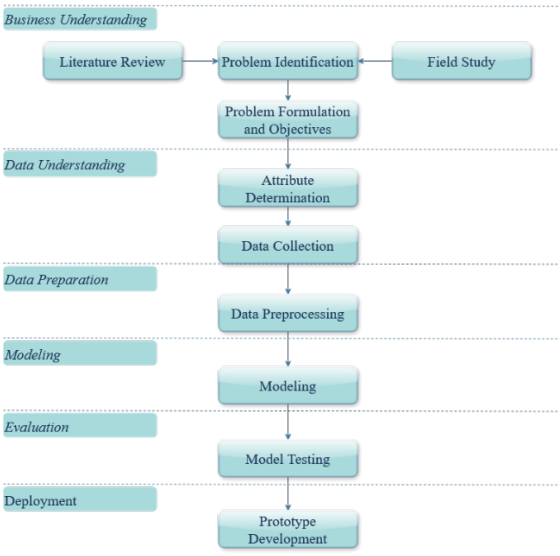

Figure 1. Research Stages

CRISP-DM is a structured, easy-to-use, reliable and commonly applied model independent of industry [18], [19]. The methodology includes six stages described in Figure 1.

2.1. Business Understanding

This stage involves understanding the business situation and available resources. The main goal of this stage is to define specific data mining

objectives and create a comprehensive project plan [18].

2.2. Data Understanding

At this stage, information is collected from various sources and explored to understand its characteristics. Statistical analysis is used to describe the data and check its quality [19].

2.3. Data Preparation

This stage involves selecting relevant data based on predefined criteria. If data quality issues are found, the data will be cleaned. In addition, the attributes required for the data mining model will be built [18].

2.4. Modeling

This stage uses the Recency, Frequency, and Monetary (RFM) model for customer segmentation with the K-Means algorithm [20]. Once the customer segments are formed, the Frequent Pattern-Growth (FP-Growth) algorithm is applied to analyze the association patterns of itemsets or products within each segment [16], [21].

2.5. Evaluation

The results of the research using the RFM model, K-Means algorithm, and FP-Growth are expected to achieve the objectives set in the Business Understanding stage. The evaluation is carried out using the Silhouette Coefficient to evaluate how well the clustering or clusters produced by the K-Means algorithm are. Higher values indicate more distinct and better clusters n [22]. Elbow Method to test clustering and Lift Ratio to measure the recommendation strength of FP-Growth in itemset association, The Davies Bouldin Index (DBI) for this metric will help measure the compactness and separation between clusters. Lower values indicate better-quality clusters, and the Lift Ratio to measure the strength of FP-Growth recommendations in item set association [12]. The data used are sales transactions that meet the Minimum Support value .

2.6. *Deployment*

The final stage is the implementation of the data mining model into the production environment. If this stage is relevant in the project, the implementation of the model will be documented.

3. RESULT AND DISCUSSION

3.1. Business Understanding

This stage included analyzing Snap's branch data for the 2019-2023 period and weekly meetings with Snap Digital Printing's directors every Tuesday. The goal was to identify barriers to stagnant sales through literature review to find relevant theories and evaluate previous steps. The result was clearly defined problem formulation and research objectives.

3.2. Data Understanding

The data used covers sales from January 2019 to December 2023, with a total of 9,841,156 transactions at Ownshop branches in the Jabodetabek area. This data is automatically collected from the branch database to the central server every day.

3.3. Data Preparation

At this stage, data cleaning is carried out to overcome deficiencies such as missing values, duplicates, or incomplete data according to data entry [18]. Data cleansing includes the removal of cell phone numbers with less than 10 digits, in accordance with the 2014 Minister of Communication and Information Technology Regulation. In addition, transactions with Categoryname worth null, Other, or VAT are also removed because they are not products.

The query cleaning process successfully filtered out invalid phone numbers and irrelevant data, with 2,638,097 entries cleaned. The second stage of cleaning reduces noise and enriches the dataset for FP-Growth by retaining only transactions that have more than one item.

After the second stage of the data cleaning process, a query is used to filter out attributes with less than two items [11], transaction data with only one item in a transaction was removed. A total of 173,498 data were cleaned, as shown in Table 2, where the itemset column lists transactions that cannot be used for FP-Growth data mining.

Table 2: Result data after cleaning

Process	Amount of Data
Before Data Cleaning	7.203.059
After Data Cleaning	7.029.561

Table 2 shows the comparison of the number of transactions before and after data cleaning. After removing cell phone numbers less than 10 digits, non-product categories, and transactions with itemsets less than 2, the number of data is reduced from 9,841,156 to 7,029,561.

The attribute selection process for customer segmentation with the RFM model includes three main factors: Recency (last transaction date), Frequency (customer's phone number), and Monetary (total purchase). Additional variables relevant for RFM and K-Means analysis were also considered.

Data transformation involves normalizing values, converting data types, changing formats, and handling blank values to make them ready for further analysis.

Table 3: After RFM Data Cleaning

No	Mobile Number	*Recency*	*Frequency*	*Monetary*
	087884408xxx	16	2	600000.00
2	082129753xxx	1085	7	2493900.00
3	0816775xxx	1512	4	221300.00
4	085945151xxx	416	1	1470000.00
5	085770839xxx	13	3	135900.00
….	….	….	….	….
267992	082243849xxx	567	2	110500.00

After going through several stages of cleaning RFM data, 267,992 data remained after normalizing RFM values, converting data types, changing formats, and handling empty values. The next stage is the modeling phase. The data in Table 3 is the final result of RFM data transformation that is ready to be used for clustering process with K-Means algorithm.

3.4. Modeling

1. *Recency, Frequency, Monetary* (RFM)

Recency, Frequency, and Monetary calculations and analysis are performed to understand the purchasing behavior of each customer. The RFM method groups customers based on similar buying patterns, where Recency identifies active and inactive customers, Frequency reveals spending habits, and Monetary indicates customers with large spending. The RFM value is calculated for each customer based on the cell phone number.

Table 4: Customer RFM

No	Mobile Number	*Recency*	*Frequency*	*Monetary*
	087884408xxx	16	2	600000.00
2	082129753xxx	1085	7	2493900.00
3	0816775xxx	1512	4	221300.00
4	085945151xxx	416	1	1470000.00
5	085770839xxx	13	3	135900.00
….	….	….	….	….
267992	082243849xxx	567	2	110500.00

2. Determination of Cluster Members

The process of determining the members of each cluster involved a total of 267,992 customers who made transactions from January 2019 to December 2023. The test was conducted six times with the number of test clusters ranging from k=2 to k=6, where k represents the number of clusters tested. For example:

1) If k=2, then there are two cluster members: C1 and C2.
2) If k=3, then there are three cluster members: C1, C2, and C3.
3) If k=4, then there are four cluster members: C1, C2, C3, and C4.
4) If k=5, then there are five cluster members: C1, C2, C3, C4, and C5.
5) If k=6, then there are six cluster members: C1, C2, C3, C4, C5, and C6.

After determining the number of clusters to be processed, the next step is to enter the value into the Python program to see the results.

Table 5: Results of 2 Cluster Trial

No	Cluster 0	Cluster 1
1	082122406xxx	081261325 xxx
2	082299282xxx	085348186 xxx
3	085240916xxx	081297129 xxx

4	081284963xxx	087739646 xxx
5	0818774xxx	08170031 xxx
….	….	….
Total	266868 Customers	1124 Customers

Table 6: Results of 3 Cluster Trial

No	Cluster 0	Cluster 1	Cluster 2
1	082122406xxx	081297129 xxx	081284206 xxx
2	081284787xxx	087739646 xxx	081268817 xxx
3	081310097 xxx	081779845 xxx	081290002 xxx
4	087782309 xxx	081261325 xxx	081513931 xxx
5	081319919 xxx	08567286534	08111769 xxx
….	….	….	….
Total	262336 Customers	5285 Customers	371 Customers

Table 7: Results of 4 Cluster Trial

No	Cluster 0	Cluster 1	Cluster 2	Cluster 3
1	087884408 xxx	08567286 xxx	081284206 xxx	082215816 xxx
2	082129753 xxx	085781151 xxx	081261325 xxx	08129738 xxx
3	0816775 xxx	081297831 xxx	081297129 xxx	081802323 xxx
4	085945151 xxx	08119808 xxx	087739646 xxx	085271587 xxx
5	085770839 xxx	0818755 xxx	08111780 xxx	08119782 xxx
….	….	….	….	….
Total	254458 Customers	224 Customers	1242 Customers	12068 Customers

Table 8: Results of 5 Cluster Trial

No	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	087884408 xxx	081261325 xxx	082129753 xxx	08567286 xxx	081284206 xxx
2	0816775 xxx	081297129 xxx	082215816 xxx	085781151 xxx	081385577 xxx
3	085945151 xxx	087739646 xxx	08129738 xxx	081281731 xxx	081382092 xxx
4	085770839 xxx	085348186 xxx	081802323 xxx	082113317 xxx	082112551 xxx
5	087700492 xxx	081779845 xxx	085271587 xxx	082114110 xxx	08111780 xxx
….	….	….	….	….	….
Total	244694 Customers	557 Customers	19705 Customers	152 Customers	2884 Customers

Table 9: Results of 6 Cluster Trial

No	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	087884408 xxx	081261325xxx	081284206 xxx	08129738 xxx	08567286 xxx	082129753 xxx
2	0816775 xxx	081297129xxx	08111780 xxx	081802323xxx	085781151xxx	085945151 xxx
3	085770839 xxx	087739646xxx	087887138xxx	085271587xxx	082113317xxx	08116319 xxx
4	087700492 xxx	085348186xxx	087888316 xxx	08119782 xxx	082114110xxx	082215816 xxx
5	082213782 xxx	081779845xxx	081293682 xxx	081288888xxx	0817878 xxx	085921231 xxx
….	….	….	….	….	….	….
Total	231969 Customers	349 Customers	1244 Customers	5508 Customers	107 Customers	28815 Customers

If the k value is set to 2, the cluster members will be generated as shown in Table 5.
If the k value is set to 3, the cluster members will be generated as shown in Table 6.
If the k value is set to 4, the cluster members will be generated as shown in Table 7.
If the value of k is set to 5, the cluster members will be generated as shown in Table 8.
If the k value is set to 6, the cluster members will be generated as shown in Table 9.

3. Metode Elbow

Based on six tests, the evaluation of clustering results was carried out using the Elbow method. This method aims to determine the optimal number of clusters by observing the percentage decrease between clusters that form a certain elbow point. The most suitable cluster count is observed at the stage where the error reduction curve shifts from steep to gradual, indicating diminishing returns from adding more clusters [12].

Table 10: Sum Sqaired Error (SSE) Calculation Results

Cluster	SSE Value	SSE Difference Value
2	754569.53	
3	600736.78	153832.75
4	483467.87	117268.91
5	398529.85	84938.02
6	329281.72	69248.13

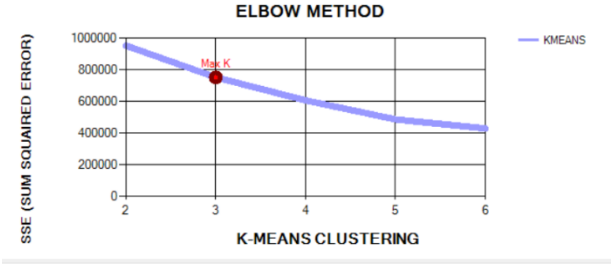

Figure 2. Elbow Method Results

Table 10 and Figure 2 above show.

1) From K=2 to K=3, there is a large decrease in SSE (Sum Sqaired Error), which shows that adding a third cluster provides a significant reduction in error.
2) From K=3 to K=4, the decrease in SSE (Sum Sqaired Error) is still significant, but not as large as the decrease from K=2 to K=3.
3) From K=4 to K=5 and K=5 to K=6, the decrease in SSE (Sum Sqaired Error) gets smaller, indicating that adding more clusters provides less error reduction.

4. Elbow Point

Elbow point is the point where the decrease in Sum Sqaired Error (SSE) starts to decrease significantly. Based on this graph, the Elbow point is located at K=3. This is the point where the decrease in Sum Squared Error (SSE) starts to slow down.

Sum Squared Error (SSE) calculation results in Table 10, Based on this Elbow Method graph, the optimal number of clusters is 3. At this point, adding additional clusters gives a progressively smaller reduction in Sum Squared Error (SSE), indicating that K=3 is the optimal point before a significant decrease in clustering efficiency.

5. Selected Clusters

From the calculation results using the Elbow method, the information obtained shows that the cluster with the optimal value is in cluster 3. Therefore, in this case, the ideal number of clusters is K=3, and this is used as the Default cluster to determine the characteristics of the data.

Table 11: Selected Clusters

No	Cluster 0	Cluster 1	Cluster 2
1	082122406 xxx	081297129 xxx	081284206 xxx
2	081284787 xxx	087739646 xxx	081268817 xxx
3	081310097 xxx	081779845 xxx	081290002 xxx
4	087782309 xxx	081261325 xxx	081513931 xxx
5	081319919 xxx	08567286 xxx	08111769 xxx
....	....	....	....
Total	262336 Customers	5285 Customers	371 Customers

Based on the clustering results with the K-Means algorithm, the cluster characteristics can be explained as follows: Cluster 0 consists of customers with high Recency, low Frequency, and low Monetary, which indicates low engagement and can be categorized as Low Engagement Customers. Cluster 1 includes customers with medium to high Recency and Frequency, as well as high Monetary, which indicates sufficient transaction activity and can

be considered as Active Customers. Cluster 2 contains customers with low Recency, very high Frequency and Monetary, indicating frequent and large purchases, which makes them VIP Customers.

6. FP-Growth

After the formation of customer segments based on clustering using the K-Means algorithm which produces three customer clusters, the next step is to analyze the purchasing patterns or product associations in each cluster using the FP-Growth algorithm.

7. FP-Growth Attribute Selection

Attribute selection for the FP-Growth algorithm includes hp (customer number), cluster (K-Means result), receiptno (receipt number), itemcode (product code), and itemname (product name), as in Table 12.

Table 12: Preferred Attributes

No	Attribute Name	Data Type	Description
1	HP (Telepon Genggam)	Int	As primary each customer data
2	Cluster	Int	K-Means clustering result value
3	Strukno	Varchar	Customer purchase transaction receipt
4	Itemcode	Int	Customer purchase product code
5	Itemname	Varchar	Customer purchase product name

Table 12 shows that the attributes selected for data mining with FP-Growth include HP, Cluster, Receipt Number, Item Code, and Item Name, before proceeding to the algorithm modeling step.

a. Transformation Data

In the data transformation stage, transaction data is converted into boolean tabular form based on the results of K-Means clustering. This process converts transactional data to tabular format according to customer segments, as shown in Table 13.

Table 13: Transaction Data Cluster Ratio

No	Customer Segment	Total Transactions	Total Products	Percentage of Transactions
1	*Low Engagement Customers*	4.295.847	1838	86%
2	*Active Customers*	138.689	1164	3%
3	*VIP Customers*	529.259	1382	11%

Table 13 shows the transaction data cluster ratio by customer segment. The Low Engagement Customers segment recorded 4,295,847 transactions with 1761 products (86% of total transactions). The Active Customers segment had 138,689 transactions with 1134 products (3%), while the VIP Customers segment recorded 529,259 transactions with 1351 products (11%). This data will be converted into a boolean tabular for each segment, to facilitate further analysis based on the clustering results.

8. Support Count

In determining frequent itemsets using the FP-Growth algorithm, this process requires two searches of the database. The first search calculates the Support value of each item and selects those that meet the Minimum Support, which is set at 5% in this study. Next, the itemsets are sorted based on their frequency of occurrence, from largest to smallest.

Table 14: Ranking of Support Count Low Engagement Customers Categories

No	Product Code	Product Name	*Support Count*
1	2010102	Kertas HVS A4 80 gram /500 lbr	806364
2	1020202	Print B/W A4/Q/F (Mesin Fc)	514743
3	1010102	Print Color A4/Q/F	397640
4	1010101	Print Color A3	296706
5	2010244	AC 260 Gr (46x30.5) / Rim 500 Lbr	181136
….	….	….	….
1838	9990020	amplop jaya polos kecil	1

Table 15: Support Count Active Customers Category Ranking

No	Product Code	Product Name	*Support Count*
1	2010102	Kertas HVS A4 80 gram /500 lbr	18356

2	1010101	Print Color A3	9989
3	1020202	Print B/W A4/Q/F (Mesin Fc)	9189
4	1010102	Print Color A4/Q/F	8072
….	….	….	….
1164	4010120	Snowman White Board BIRU /12 pcs	1

Table 16: VIP Customers Support Count Category Ranking

No	Product Code	Product Name	*Support Count*
1	2010102	Kertas HVS A4 80 gram /500 lbr	53172
2	1010101	Print Color A3	45628
3	1010102	Print Color A4/Q/F	30491
4	1020202	Print B/W A4/Q/F (Mesin Fc)	29908
….	….	….	….
1382	2050324	YOYO ID CARD BIRU	1

Tables 14, 15, and 16 show the sorting of products based on the support count for customers (Low Engagement, Active, and VIP) reflecting the frequency of product purchases by each customer category.

9. Itemset Creation

The FP-Growth algorithm utilizes frequent itemsets to directly extract purchase patterns from transaction data. The process involves creating and utilizing itemsets to find frequent purchase patterns.

Table 17: Transaction List of Low Engagement Customers

No	No Receipt	Item
1	-2109100529/INV	1010102, 2010518
2	-21101001288/INV	1020202, 2010102, 4050138
3	-22061003123/INV	1020202, 1030205
4	-22061003261/INV	1010102, 2010102
….	….	….
1082726	WLT-2404301003127/INV	1020202, 2010102

Table 18: Transaction List of Active Customers

No	No Receipt	Item
1	ANT-1907100069/INV	2010102, 3090002
2	ANT-1907100073/INV	1020202, 2010102
3	ANT-1907100230/INV	1020202, 2010102
4	ANT-1907100422/INV	2010244, 2040804, 4050128
….	….	….
37954	WLT-2404301003098/INV	1010102, 2010102

Table 19: VIP Customers Transaction List

No	No Struk	Item
1	ANT-1907100069/INV	2010102, 3090002
2	ANT-1907100073/INV	1020202, 2010102
3	ANT-1907100230/INV	1020202, 2010102
4	ANT-1907100422/INV	2010244, 2040804, 4050128
….	….	….
121703	WLT-2404301003117/INV	1020202, 2010102

Tables 17, 18, and 19 list transactions for low-engagement, active, and VIP customers, including receipt numbers and products purchased in each transaction.

10. FP-Growth Algorithm Calculation Results

At this stage, FP-Growth algorithm is used with Support and Confidence parameters to determine association rules. Customer transactions are classified based on Engagement level (Low, Active, and VIP) to identify buying patterns and preferences, which helps marketing strategies.

Table 20: Calculation Result of Low Engagement Customers

No	*Antecedents*	*Consequents*	*Support* (%)	*Confidence* (%)	Lift Ratio
1	Print Color A3	AC 260 Gr (46x30.5) / Rim 500 Lbr	8.04	37.45	2.85
2	Fotocopy A4/Q/F, Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	6.81	97.43	1.74
3	Print Color A4/Q/F,Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	9.14	97.05	1.74
4	Fotocopy A4/Q/F, Print Color A4/Q/F	Kertas HVS A4 80 gram /500 lbr	3.03	96.75	1.73
5	Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	36.59	96.60	1.73

Table 21: Low Engagement Customers Product Package

No	*Antecedents*	*Consequents*	*Description*
1	Print Color A3	AC 260 Gr (46x30.5) / Rim 500 Lbr	Customers who purchase Print Color A3 have a 37.45% probability of also purchasing AC 260 Gr (46x30.5) / Rim 500 Lbr, indicating a very strong positive relationship.
2	Fotocopy A4/Q/F, Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	Customers who purchase a combination of A4/Q/F Photocopy and A4/Q/F B/W Print (Fc Machine) have a 97.43% probability of also purchasing 80 gram /500 lbr A4 HVS Paper, indicating a very strong positive relationship.
3	Print Color A4/Q/F, Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	Customers who purchase a combination of Print Color A4/Q/F and Print B/W A4/Q/F (Fc Machine) are 97.05% more likely to also purchase HVS A4 Paper 80 grams /500 lbr, indicating a very strong positive relationship.
4	Fotocopy A4/Q/F, Print Color A4/Q/F	Kertas HVS A4 80 gram /500 lbr	Customers who purchase a combination of A4/Q/F Photocopy and A4/Q/F Color Print have a 96.75% likelihood of also purchasing 80 gram /500 lbr A4 HVS Paper, indicating a very strong positive relationship.
5	Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	Customers who purchased Print B/W A4/Q/F (Fc Machine) are 96.60% likely to also purchase HVS A4 Paper 80 grams /500 lbs, indicating a very strong positive relationship.

Based on the table above, the product package analysis shows an average Confidence of 85.70%, with a valid rule if Confidence is above 20%. Lift values range from 1.38 to 2.85, with an average of 1.77, while Support values range from 2.22% to 36.59%, with an average of 11.36%.

Table 22: Active Customer Calculation Results

No	*Antecedents*	*Consequents*	*Support* (%)	*Confidence* (%)	Lift Ratio
1	Print Color Kn A3 (4 Klik) Bb	Box Kartu Nama Besar /24 Pcs	2.32	93.32	26.74
2	Ongkos Potong Kn	Box Kartu Nama Besar /24 Pcs	2.89	83.60	23.95
3	Print B/W A3 Bb (Mesin Fc)	Kertas HVS A3 80 Gram /500 Lbr	2.30	97.60	17.55
4	Ongkos Potong Kn	AC 260 Gr (46x30.5) / Rim 500 Lbr	2.18	62.89	3.38
5	Print B/W A4/Q/F (Mesin Fc), Fotocopy A4/Q/F	Kertas HVS A4 80 Gram /500 Lbr	4.06	97.63	2.77

Table 23: Active Customers Product Package

No	*Antecedents*	*Consequents*	*Description*
1	Print Color Kn A3 (4 Klik) Bb	BOX KARTU NAMA BESAR /24 Pcs	Customers who purchase Print Color A3 have a 37.45% probability of also purchasing AC 260 Gr (46x30.5) / Rim 500 Lbr, indicating a very strong positive relationship.
2	Ongkos Potong Kn	BOX KARTU NAMA BESAR /24 Pcs	Customers who purchase a combination of A4/Q/F Photocopy and A4/Q/F B/W Print (Fc Machine) have a 97.43% probability of also purchasing 80 gram /500 lbr A4 HVS Paper, indicating a very strong positive relationship.

3	Print B/W A3 Bb (Mesin Fc)	Kertas HVS A3 80 Gram /500 Lbr	Customers who purchase a combination of Print Color A4/Q/F and Print B/W A4/Q/F (Fc Machine) are 97.05% more likely to also purchase HVS A4 Paper 80 grams /500 lbr, indicating a very strong positive relationship.
4	Ongkos Potong Kn	AC 260 Gr (46x30.5) / Rim 500 Lbr	Customers who purchase a combination of A4/Q/F Photocopy and A4/Q/F Color Print have a 96.75% likelihood of also purchasing 80 gram /500 lbr A4 HVS Paper, indicating a very strong positive relationship.
5	Print B/W A4/Q/F (Mesin Fc), Fotocopy A4/Q/F	Kertas HVS A4 80 Gram /500 Lbr	Customers who purchased Print B/W A4/Q/F (Fc Machine) are 96.60% likely to also purchase HVS A4 Paper 80 grams /500 lbs, indicating a very strong positive relationship.

Based on Tables 22 and 23, the product association analysis for Active Customers shows significant relationships between products. The average Confidence of the rules found is 49.72%, with rules valid if the Confidence is above 20%. Lift values vary from 1.88 to 26.74, with an average of 7.67, while Support values range from 2.11% to 20.33%, with an average of 6.73%.

Table 24: *VIP Customers* Calculation Results

No	*Antecedents*	*Consequents*	*Support (%)*	*Confidence (%)*	Lift
1	Print Textile Lebar 150	Transfer Paper Kain 85 Gsm 160 / 100 Mtr	2.96	92.97	27.57
2	Print Color A3 Bb	AC 260 Gr (46x30.5) / Rim 500 Lbr	2.13	51.51	3.70
3	Print Color A3	AC 260 Gr (46x30.5) / Rim 500 Lbr	8.18	34.34	2.47
4	Fotocopy A4/Q/F,Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	4.08	97.09	2.32
5	Print Color A4/Q/F, Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	5.65	96.95	2.32

Table 25: VIP Customers Product Package

No	*Antecedents*	*Consequents*	*Description*
1	Print Textile Lebar 150	Transfer Paper Kain 85 Gsm 160 / 100 Mtr	Customers who purchased Print Textile Width 150 have a 92.97% likelihood of also purchasing Transfer Paper Fabric 85 Gsm 160 / 100 Mtr, indicating a very strong positive association with a lift of 27.57.
2	Print Color A3 Bb	AC 260 Gr (46x30.5) / Rim 500 Lbr	Customers who purchased Print Color A3 Bb have a 51.51% likelihood of also purchasing AC 260 Gr (46x30.5) / Rim 500 Lbr, indicating a strong positive association with lift of 3.70.
3	Print Color A3	AC 260 Gr (46x30.5) / Rim 500 Lbr	Customers who purchased Print Color A3 are 34.34% more likely to also purchase AC 260 Gr (46x30.5) / Rim 500 Lbr, indicating a strong positive relationship with a lift of 2.47
4	Fotocopy A4/Q/F, Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	Customers who purchase a combination of A4/Q/F Photocopy and A4/Q/F B/W Print (Fc Machine) have a 97.09% likelihood of also purchasing 80 gram /500 lbr A4 HVS Paper, indicating a very strong positive relationship with a lift of 2.32.
5	Print Color A4/Q/F, Print B/W A4/Q/F (Mesin Fc)	Kertas HVS A4 80 gram /500 lbr	Customers who purchase a combination of Print Color A4/Q/F and Print B/W A4/Q/F (Fc Machine) have a 96.95% probability of also purchasing HVS A4 Paper 80 gram /500 lbr, indicating a very strong positive relationship with lift 2.32.

3.5. Based on the table above, the product association analysis for VIP Customers shows an average Confidence of 82.60%, with a valid rule if Confidence is more than 20%. Lift values vary from 1.82 to 27.57, with an average of 4.92, indicating a positive relationship between products. Support values range from 2.13% to 21.86%, with an average of 7.85%, illustrating the variation in the frequency of product combinations in transactions.***Evaluation***

Based on the evaluation results in the table, cluster 3 has the highest Silhouette Coefficient

value of 0.9478 and the lowest Davies Bouldin Index value of 0.4320, indicating the best cluster separation and highest cluster consistency. However, as the number of clusters increases, the Centroid Distance value decreases, indicating that the clusters become more compact, but the separation quality (DBI) and cluster suitability (Silhouette) tend to decrease. Therefore, the 3-cluster configuration can be considered the most optimal as it provides the best balance between density, separation, and cluster fit compared to other cluster configurations.

Tabel 26 : Cluster evaluation results

Klaster	Centroid Distance	Davies Bouldin Index	Silhouette Coefficient
2	1097111.4528	0.4668	0.8886
3	845737.0756	0.4320	0.9478
4	674810.5877	0.4861	0.8310
5	578535.2757	0.4820	0.7931
6	481107.7505	0.4811	0.7512

3.6. *Deployment*

At this stage, prototype development is carried out with reference to the existing Snap Backend Dashboard. The previous process is part of the Snap Backend Dashboard application development. The following is the interface design of the prototype that has been made.

1. Data Mining Menu

In this menu there are 3 sub tab menus, namely the RFM tab, K-Means tab and FP-Growth tab as shown in Figure 3.

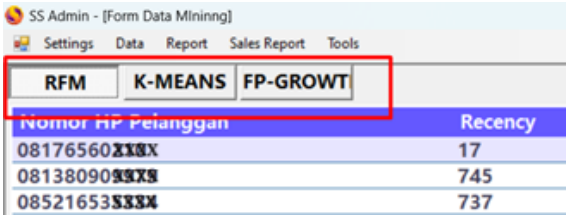

Figure 3. Data Mining Menu

In Figure 3. there is one of the features in a system, namely data processing, by clicking the Association Rules button, the application can display the results of the FP-Growth algorithm calculation process based on each cluster briefly without going through manual calculations.

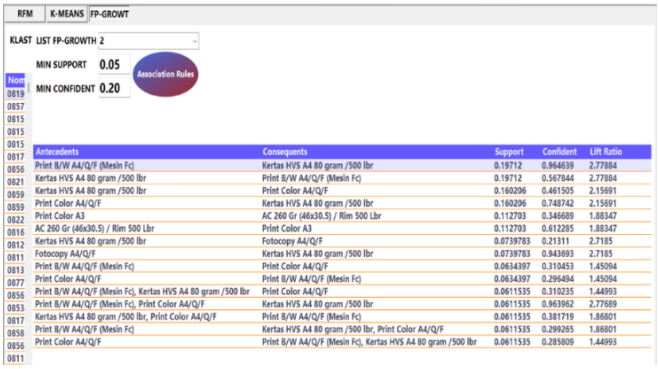

Figure 4. FP-Growth Result Tab

In Figure 4. the data display presented by the application after the admin enters the data and clicks the button according to the available tabs. With the development of the designed Snap dashboard application, admins can easily search for association rules based on customer segments. These rules are used to determine the optimal product combination for sales or promotions.

Cashiers or Customer Service use Snap's Point of Sales (POS) application to process customer transactions efficiently. The first step is to enter the customer's phone number into the system. If the number belongs to a cluster that has been determined through data analysis, the app will display relevant information, as shown in Figure 5.

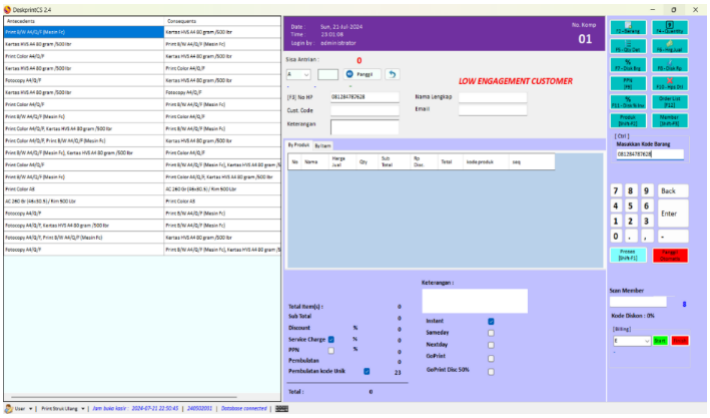

Figure 5. Snap POS App View

In Figure 5. the results of the FP-Growth algorithm calculation show the association of products that are often purchased together. This data helps admins identify buying patterns and make more relevant product recommendations for customers. Product association becomes an important tool in designing the right marketing strategy, increasing sales, and maximizing customer satisfaction by offering products according to their preferences.

4. CONCLUSIONS

Research at Snap Digital Printing using RFM and K-Means analysis, optimized with FP-Growth, identified three customer segments: Low Engagement, Active, and VIP Customers. This segmentation helps understand buying patterns, improve product recommendations, and marketing efficiency. The Elbow Method shows the optimal number of clusters is three, with their respective transaction contributions: Low Engagement (86%), Active (3%), and VIP (11%). FP-Growth analysis results show the average Support for Low Engagement Customers is 11.36%, Active Customers is 6.73%, and VIP Customers is 7.85%. The average Confidence is 85.70%, 49.72%, and 82.60%, respectively, with Lift Ratio of 1.77, 7.67, and 4.92. This model contributes to the development of data mining techniques in marketing. Thus, this study shows that the application of data mining in customer segmentation and purchase pattern analysis can significantly improve customer retention as well as marketing and sales effectiveness. Further research could integrate cutting-edge methods such as deep learning clustering or ensemble methods to improve segmentation accuracy compared to K-Means. Product association analysis can also be developed using graph-based or context-aware recommendation systems as a complement to FP-Growth. Additionally, expanding the data to all branches and using additional evaluation metrics such as the Adjusted Rand Index or Precision@k will make the research results more generalizable and relevant for data-driven marketing strategies.

REFERENCES

[1] T. Tavor, L. D. Gonen, and U. Spiegel, “Customer Segmentation as a Revenue Generator for Profit Purposes,” *Mathematics*, vol. 11, no. 21, Nov. 2023, doi: 10.3390/math11214425.

[2] T. Lakshika and A. Caldera, “Association Rules for Knowledge Discovery From E-News Articles: A Review of Apriori and FP-Growth Algorithms,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 7, no. 5, pp. 178–192, 2022, doi: 10.25046/aj070519.

[3] M. Mehrabioun and B. M. Mahdizadeh, "Customer retention management: A complementary use of data mining and soft systems methodology," *Human Systems Management*, vol. 40, no. 6, pp. 897–916, Dec. 2021, doi: 10.3233/HSM-201075.

[4] O. Akande, E. O. Asani, and B. Dautare, "Customer Segmentation Through RFM Analysis and K-Means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy," *Ceddi Journal of Information System and Technology (JST)*, vol. 3, no. 1, pp. 14–25, Apr. 2024, doi: 10.56134/jst.v3i1.81.

[5] I. Lewaaelhamd, "Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis," *Journal of Data Science and Intelligent Systems*, vol. 2, no. 1, pp. 29–36, Sep. 2023, doi: 10.47852/bonviewJDSIS32021293.

[6] J. Joung and H. Kim, "Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews," *Int J Inf Manage*, vol. 70, p. 102641, Jun. 2023, doi: 10.1016/j.ijinfomgt.2023.102641.

[7] J. Ma, B. R. Nault, and Y. (Paul) Tu, "Customer segmentation, pricing, and lead time decisions: A stochastic-user-equilibrium perspective," *Int J Prod Econ*, vol. 264, p. 108985, Oct. 2023, doi: 10.1016/j.ijpe.2023.108985.

[8] I. Kursan Milaković, "Purchase experience during the COVID-19 pandemic and social cognitive theory: The relevance of consumer vulnerability, resilience, and adaptability for purchase satisfaction and repurchase," *Int J Consum Stud*, vol. 45, no. 6, pp. 1425–1442, Nov. 2021, doi: 10.1111/ijcs.12672.

[9] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Information Systems and e-Business Management*, vol. 21, no. 3, pp. 527–570, Sep. 2023, doi: 10.1007/s10257-023-00640-4.

[10] R. N. Huda, R. Fitriadi, and A. Wibowo, "Optimization Product Recommendation Using K-Means, Agglomerative Clustering And Fp-Growth Algorithm," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 953–960, Jul. 2024, doi: 10.52436/1.jutif.2024.5.4.1901.

[11] C. Satria, A. Anggrawan, and Mayadi, "Recommendation System of Food Package Using Apriori and FP-Growth Data Mining Methods," *Journal of Advances in Information Technology*, vol. 14, no. 3, pp. 454–462, 2023, doi: 10.12720/jait.14.3.454-462.

[12] T. Santoso, A. Darmawan, N. Sari, M. A. F. Syadza, E. C. B. Himawan, and W. A. Rahman, "Clusterization of Agroforestry Farmers using K-Means Cluster Algorithm and Elbow Method," *Jurnal Sylva Lestari*, vol. 11, no. 1, pp. 107–122, Jan. 2023, doi: 10.23960/jsl.v11i1.646.

[13] J. P. B. Saputra, S. A. Rahayu, and T. Hariguna, "Market Basket Analysis Using FP-Growth Algorithm to Design Marketing Strategy by Determining Consumer Purchasing Patterns," *Journal of Applied Data Sciences*, vol. 4, no. 1, pp. 38–49, Jan. 2023, doi: 10.47738/jads.v4i1.83.

[14] D. Dwiputra, A. Mulyo Widodo, H. Akbar, and G. Firmansyah, "Evaluating the Performance of Association Rules in Apriori and FP-Growth Algorithms: Market Basket Analysis to Discover Rules of Item Combinations," *Journal of World Science*, vol. 2, no. 8, pp. 1229–1248, Aug. 2023, doi: 10.58344/jws.v2i8.403.

[15] F. Nuraeni, D. Tresnawati, Y. Handoko Agustin, and G. Fauzi, "Optimization of Market Basket Analysis Using Centroid-Based Clustering Algorithm and FP-Growth Algorithm," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 6, pp. 1581–1590, Dec. 2022, doi: 10.20884/1.jutif.2022.3.6.399.

[16] M. Sarkar, A. R. Puja, and F. R. Chowdhury, "Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis," *Journal of Business and Management Studies*, vol. 6, no. 2, pp. 54–60, Mar. 2024, doi: 10.32996/jbms.2024.6.2.5.

[17] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation

Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022, doi: 10.3390/su14127243.

[18] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[19] S. Panpaeng, P. Phanphaeng, J. Kumnuanta, P. Yommakit, K. Kocento, and P. Wongchompoo, "The application of data mining techniques for predicting education to new undergraduate students at Chiang Mai Rajabhat University," in *2023 IEEE International Conference on Cybernetics and Innovations (ICCI)*, IEEE, Mar. 2023, pp. 1–6. doi: 10.1109/ICCI57424.2023.10112233.

[20] A. Khumaidi, H. Wahyono, R. Darmawan, H. D. Kartika, N. L. Chusna, and M. K. Fauzy, "RFM-AR Model for Customer Segmentation using K-Means Algorithm," *E3S Web of Conferences*, vol. 465, p. 02005, Dec. 2023, doi: 10.1051/e3sconf/202346502005.

[21] R. Cristover, H. Toba, and B. R. Suteja, "Segmentation and Formation of Customer Regression Model Based on Recency, Frequency and Monetary Analysis," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 2, Aug. 2022, doi: 10.28932/jutisi.v8i2.5075.

[22] H. Mulyani, R. A. Setiawan, and H. Fathi, "Optimization of K Value in Clustering Using Silhouette Score (Case Study: Mall Customers Data)," *Journal of Information Technology and Its Utilization*, vol. 6, no. 2, pp. 45–50, Dec. 2023, doi: 10.56873/jitu.6.2.5243.