SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Combination Grouping Techniques and Association Rules for Marketing Analysis based Customer Segmentation

Amir Mahmud Husein[1)*], Dodi Setiawan[2)], Andika Rahmad Kolose Sumangunsong[3)], Andreas Simatupang[4)], Shela Aura Yasmin[5)]
[1,2,3,4,5)] Universitas Prima Indonesia Medan
[1)]amirmahmud@unprimdn.ac.id, [2)]sdodi113@gmail.com, [3)]andikasimangunsong@yahoo.com, [4)]andreassimatupang1998.as@gmail.com, [5)]slla.yasmin27@gmail.com

Submitted : July 21, 2022 | **Accepted** : Aug 2, 2022 | **Published** : Aug 7, 2022

Abstract: Changes in people's transaction behavior using the internet resulted in the exponential growth of e-commerce. With the growth of digital shopping transactions, it is difficult to predict customer segments and patterns using traditional mathematical models. Timely identification of emerging trends from large volumes of data plays a major role in business processes and decision making. This is different from previous research works that apply the RFM model based on K-Means Clustering to find potential customers as an ingredient in determining marketing targets. In this study, a clustering technique approach is proposed to classify customer data which is evaluated using the Davies Bouldin, Calinski Harabasz and Silhouette methods to determine the optimal number of clusters, then the results are used in the Apriori algorithm to find patterns of goods that are often purchased together. Based on the test results on the K-Means Clustering, Spectral Clustering, and Gaussian Mixture Model techniques produced 5 clusters with 76% more accurate the K-Means Clustering method than the other two methods so that it was determined as a method in the RMF model, then the results of customer grouping were used on the Apriori algorithm to find patterns of concurrent product purchases by customers that are expected to be useful in future marketing management.

Keywords: Marketing Targets, RFM Model, K-Means Clustering, Apriori algorithm

INTRODUCTION

The development of internet has changed the way of individuals shop, communicate, disseminate information and conduct business transactions. Changes people's transaction behavior using internet resulted in exponential growth of e-commerce. In addition, e-commerce is not only for buying products online, but also for comparing prices, features, and can also make payments for electricity, water and others. (T. Sai Vijay, S. Prashar, and V. Sahay, 2019). With the growth of today's digital shopping transactions, it becomes difficult to predict customer segments and patterns using traditional mathematical models. Timely identification of emerging trends from large volumes of data plays a major role in business processes and decision making. Huge volumes of data exist, but companies are thirsty for knowledge(O. Piskunova and R. Klochko, 2020). Customer Segmentation(CS) aims to target valuable customers and develop marketing activities by dividing customers into smaller homogeneous groups so that marketing strategies can target each group individually(D. Kamthania, A. Pahwa, and S. S. Madhavan, 2018). Customer Segmentation with clustering algorithm is one of the most widely used algorithms. Clustering technique minimizes intra-cluster distance and maximizes distance between clusters for data segmentation(Y. Cheng, M. Cheng, T. Pang, and S. Liu, 2021), In addition, customer behavior analysis(J. Wu et al. 2020; Dedi, M. I. Dzulhaq, K. W. Sari, S. Ramdhan, R. Tullah, and Sutarman, 2019; . Wu et al., 2020) RFM (Recency Frequency Monetary) model is one of the solutions that need to be considered in marketing analysis.

The application of clustering algorithms has been proposed in many works such as, (S. G. Carbajal, 2021) applying a clustering algorithm for automatic classification of customers can identify several types of customers, and the dynamics of their behavior in the store. Furthermore (J. Wu et al.,) applied K-Means++ for User Value Identification Based on the Enhanced RFM Model and (M. Pohludka and H. Štverková, 2019) to analyze the use of CRM (Customer Relationship Management) systems in small and medium enterprises (SMEs). The application of K-Means Clustering on the RFM model can accurately extract the buying behavior characteristics of each type of customer and create an accurate marketing strategy(J. Wu et al., 2020).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Recently the clustering algorithm approach with Association Rules Mining in marketing, such as(M. Ryu, K. Il Ahn, and K. Lee, 2021) using an item assignment model based on associations between items used to place items in a retail or e-commerce environment to increase sales, then (Y. Cheng, M. Cheng, T. Pang, and S. Liu, 2021) propose a cross-marketing model based on an improved sequential pattern mining algorithm. In the works (W. J. Chen, M. J. Jhou, T. S. Lee, and C. J. Lu, 2021) integrates Association Rules Mining to build a prediction model for basketball results with the aim of providing useful information for operations in the sports market. In recent years the application of clustering algorithms like K-Means, dbscan, K-Means++, Hierarchical clustering and others have proven to be accurate for customer segmentation with different purposes, but several recent works report the application of combined clustering with association rule mining. has produced useful information on marketing strategies, such as (Y. Han, D. Yu, C. Yin, and Q. Zhao, 2020) propose a new and efficient algorithm for finding basic knowledge in the form of Temporary Association Rules(TARs) in the Blast furnace iron manufacture data, then (S. Guney, S. Peker, and C. Turhan, 2020) integrate clustering and mining of association rules to analyze and profile customers in the service Video on Demand (VoD).

Inspired by previous work, in this study we propose a different clustering algorithm (K-Means Clustering, Spectral Clustering, dan Gaussian Mixture Model) on the RFM model to break down the profile of large customers into much smaller segments and to gain insight into the detailed behavior of customers then Association Rules Mining with Apriori algorithm is used to find sets of frequently purchased items together using association rules analysis which is expected to be useful as a recommendation system

LITERATURE REVIEW

In recent years the application of clustering algorithms like K-Means, dbscan, KMeans ++, Hierarchical clustering and others have been shown to be accurate in their application to segmentation of customers with different objectives, but several recent works reporting the application of combined clustering with association rule mining have yielded useful information on marketing strategies, like (Y. Han, D. Yu, C. Yin, and Q. Zhao, 2020) propose a new and efficient algorithm for finding basic knowledge in the form of Temporal Association Rules (TARs) in the Blast furnace iron manufacturing data, then (M. Ryu, K. Il Ahn, and K. Lee, 2021) applying Association Rules Mining to identify useful relationships between large data sets and association rule mining can provide new insights to decision makers, in particular the proposed model is used to place items in retail or e-commerce environments to increase sales based on hybrid genetic algorithms and (S. Guney, S. Peker, and C. Turhan, 2020) integrate clustering and mining of association rules to analyze and profile customers in the service Video on Demand (VoD).

METHOD

In this study, this type of research uses qualitative research, where this qualitative research is descriptive and tends to use analysis by grouping data to find a pattern of what is learned and compare it with the concepts that exist in several sources. The datasets were obtained from survey results in several European countries, namely: United Kingdom, Francis, Jerman, Spanyol.

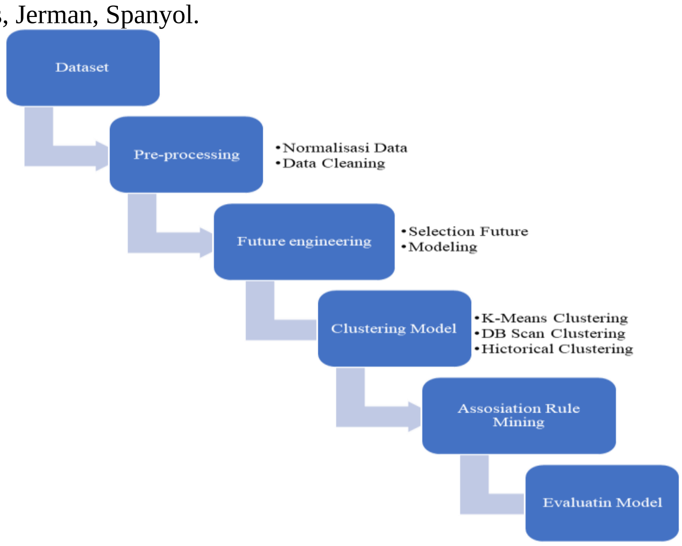


Figure 1. Proposed method

Pre-processing

Data cleaning which includes missing values, smooth noise data, identify and remove outliers, resolve inconsistencies. The process of data integration from several databases and data transformation in the form of normalization and aggression.

Feature Engineering

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Selection of data subsets that are relevant to the problem from the existing set of features, without transforming and combining all features to improve predictability, followed by pre-processing that obtains raw features so that the exact amount of data is not always large.

Clustering Model

The process of grouping data into several clusters or groups so that data in one cluster has a maximum level of similarity and data between clusters has a minimum similarity.

Assosiation Rule Mining

Data mining techniques to find associative rules between a combination of items.

Evaluasi Model

The pattern of information generated from the data playing process needs to be displayed in a form that is easily understood by interested parties.

RESULT

In this study, all experiments and tests used python 3.7, jupyter notebook, pandas, numpy with device specifications Intel Core i5, 8 Ghz, Win 10. The stages of testing were pre-processing, data cleaning, data analysis exploration, application of clustering techniques (K). -Means Clustering, Hierarchical, and DB Scan) and finally the results of the grouping will be used in the a priori algorithm to find product patterns.

4.1. Pre-processing

At this stage it is done to take a quick look at the dataset to be analyzed and see how we can make it better and more efficient for further analysis. First, import the dataset using pandas as shown in table 1.

Table 1 Dataset

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680	France

541909 rows × 8 columns

In table 1 is a summary of the dataset that will be used in this study, the total data is 541,909 rows and 8 columns. In the InvoiceDate column, the format will be changed, besides that the data still has some odd and irregular values in the 'UnitPrice' and 'Quantity' columns. This data will be deleted to prevent it from negatively affecting the analysis. Then in the 'StockCode' column, some transactions are not actually products, but are costs or fees related to post or bank or other transactions that are not actually needed in the data.

Furthermore, for needs analysis will add useful features for future analysis. The features that will be added for now are the FinalPrice of each transaction and the month and day in which the transaction occurs, which can be taken from the 'InvoiceDate' attribute. Then there are also some wrong 'CustomerID' where there are two different countries with the same 'CustomerID'.

This duplicate value will be fixed by grouping the dataframe by 'CustomerID' and if any customer has more than two countries, will replace the incorrect value with the customer country mode value. The results of the overall adjustment of the data will be saved into a file that will be used for further analysis, the results of the preprocessing data can be seen in the table 2.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2 Pre-processing Result

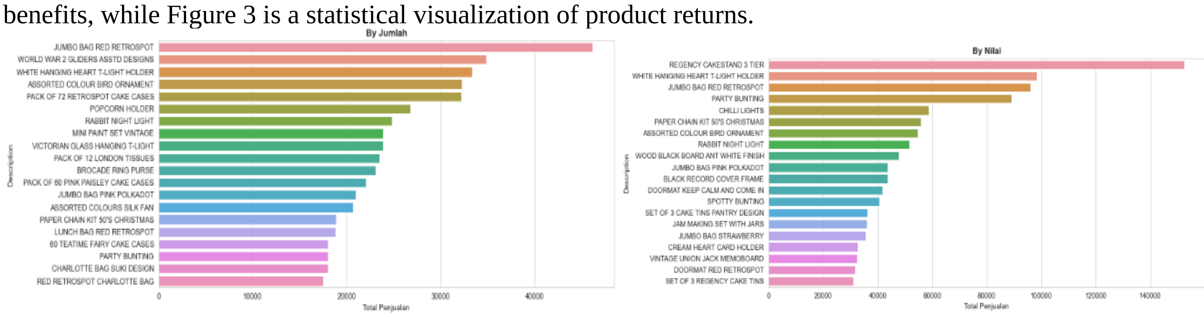
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	FinalPrice	InvoiceMonth	Day of week
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.95	17850	UNITED KINGDOM	17.70	December	Wednesday
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.75	17850	UNITED KINGDOM	22.50	December	Wednesday
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	4.15	17850	UNITED KINGDOM	33.20	December	Wednesday
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.75	17850	UNITED KINGDOM	22.50	December	Wednesday
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART	6	2010-12-01 08:26:00	4.25	17850	UNITED KINGDOM	25.50	December	Wednesday
...	...	...	...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680	FRANCE	10.20	December	Friday
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680	FRANCE	12.60	December	Friday
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680	FRANCE	16.60	December	Friday
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680	FRANCE	16.60	December	Friday
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680	FRANCE	14.85	December	Friday

532820 rows × 11 columns

3.2. Explorasi Data Analisys (EDA)

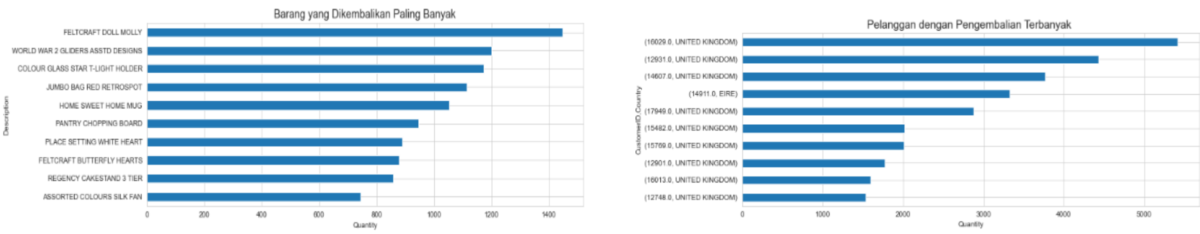
In this section, we describe the analytical data exploration method to visualize the data and create reports that are useful for getting to know and have a clearer vision of the data. For this section, first import the data then convert the index to datetime so that it can work on the time series data, then from the dataset a report is generated showing the featured products sold worldwide.

There are two indicators that can show how much benefit is obtained from each product. Figure 2 (a) shows the top 20 products that are most purchased by customers and 2 (b) is the product that brings the most financial benefits, while Figure 3 is a statistical visualization of product returns.



(a) (b)

Figure 2 Best Selling Products (a) Based Quantity , (b) Based value



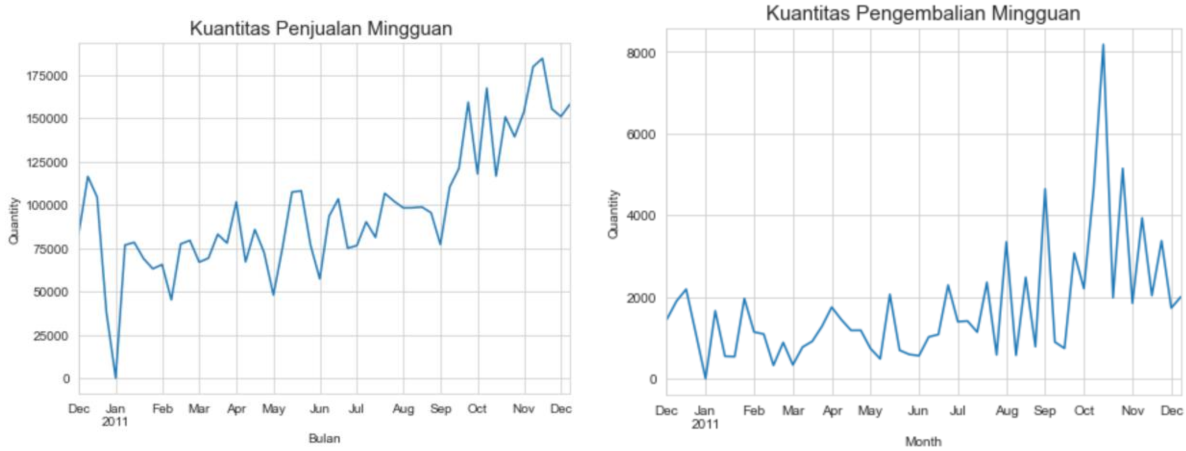
(a) (b)

Figure 3. Product Statistics (a) returned, (b) most customers return products

In the following chart is the trend of sales throughout the year on a weekly basis. In figure 4 (a) is a graph of weekly sales and (b) is weekly returns by customers. After the sudden decline in January, we can see an almost increasing sales trend. For product returns, except for the second week of October, it is almost unchanged but with a slight increase.

*name of corresponding author

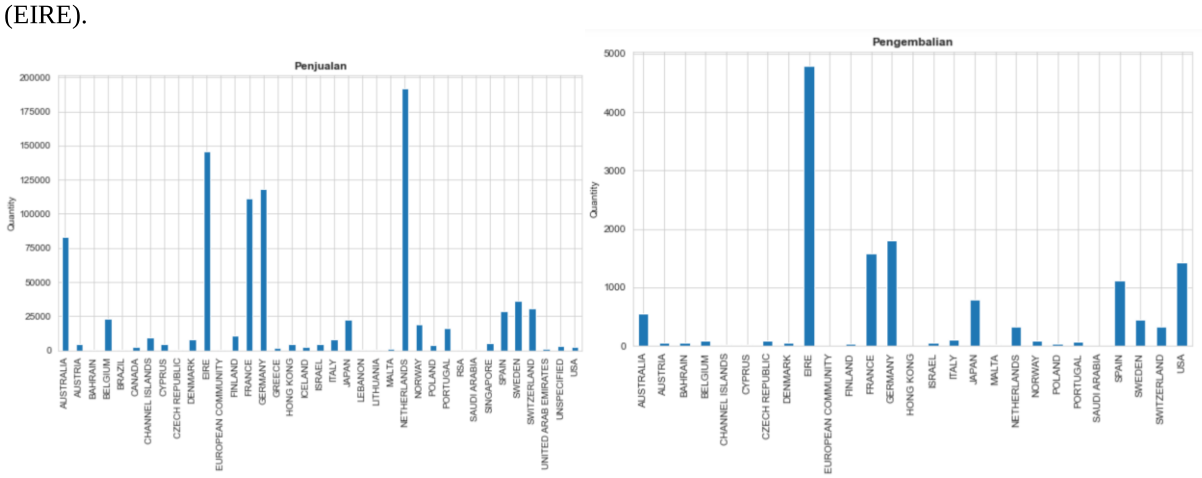
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571
e-ISSN : 2541-2019
p-ISSN : 2541-044X



(a) (b)

Figure 4. Weekly Sales Trends (a) Sales, (b) Returns

The next analysis looks at how many products are sold and returned abroad. Since Great Britain has the majority of sales and will not provide useful information, it is necessary to present it in a better and more informative way. The results of this analysis, as shown in Figure 5, are mostly sold in the Netherlands and mostly returned in Ireland (EIRE).



(a) (b)

Figure 5. Analysis of Sales out (a) Sales, (b) returns

Another statistical analysis that can be used for planning ahead is how many customers are repeat customers, meaning they buy the product more than once. In Figure 6 the first graph plot (on the left) can be seen that almost 70% of customers are repeat customers, while the second plot (on the right) also shows which country's customers repeat transactions the most.

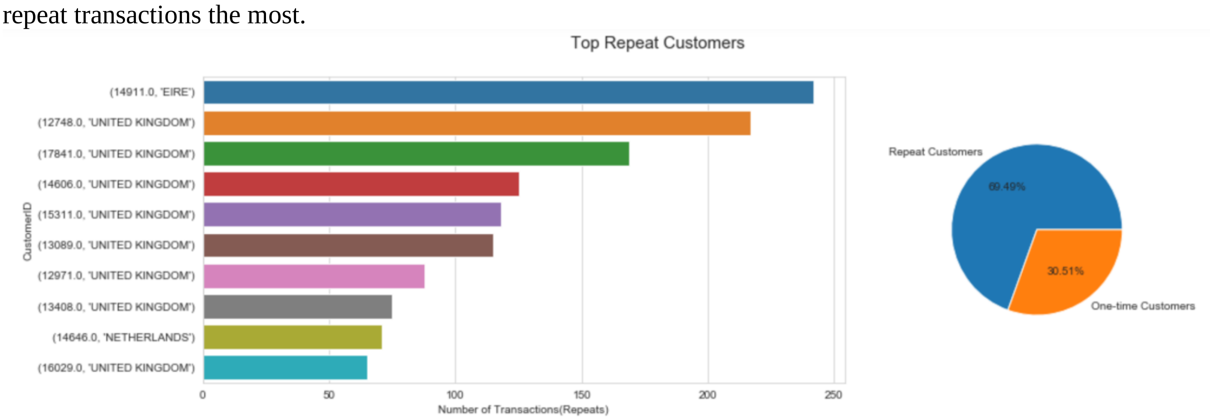


Figure 6. Planning Analysis for the future

3.3. Clustering Technique Analysis

The last section will summarize the results by dividing customers into groups by assigning new customers to clusters based on their behavior. The steps taken for customer segmentation are dividing customers from the UK

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571
e-ISSN : 2541-2019
p-ISSN : 2541-044X

and non-UK customers. The next step is to group the data by customer and choose a way to aggregate the values on it. Before starting the process, scaling the numerical features to avoid the negative effects of feature magnitudes can also speed up the process. Table 3.3 is the result of scaling the dataset into numerical features, the process is carried out by categorizing based on the customer country category and this data can also take into account categorical variables.

In this type of method it is necessary to give the algorithm the right number of clusters. Therefore, in this study 3 methods are proposed to determine the optimal number of clusters, namely Davies Bouldin, Calinski Harabasz and Silhouette. The average value of the three methods will determine how similar the data points to the cluster are compared to other clusters (using Euclidean distance). The results of the comparison of the three methods to the number of clusters can be seen in table 4

Table 3. Feature scale

	CustomerID	UK?	Average Quantity	Average Price	Repeats	Product Variety
0	17850.0	UK	-0.265816	0.467635	3.314710	-0.436451
1	13047.0	UK	-0.230065	0.662160	1.217033	0.506554
2	12583.0	non-UK	0.114924	-0.630673	1.327437	0.634617
3	13748.0	UK	-0.018821	0.688450	0.002589	-0.436451
4	15100.0	UK	-0.164896	6.820805	0.112993	-0.704217
...	...	...	...	...	...	...
4345	13436.0	UK	-0.245888	1.899979	-0.439027	-0.576155
4346	15520.0	UK	0.024086	-0.944941	-0.439027	-0.506303
4347	13298.0	UK	0.766514	0.650674	-0.439027	-0.692575
4348	14569.0	UK	-0.239813	0.838716	-0.439027	-0.599439
4349	12713.0	non-UK	-0.068143	-0.699989	-0.439027	-0.285104

4350 rows × 6 columns

Table 4. Comparison of the Number of Clusters

	Davies_Bouldin_Score	Calinski_Harabasz_Score	Silhouette_Score	n_clusters
KMeans	0.76	1761.33	0.57	[5, 5, 3]
Gaussian Mixture Model	1.52	674.43	0.37	[5, 5, 3]
Spectral Clustering	0.36	544.72	0.83	[3, 5, 5]

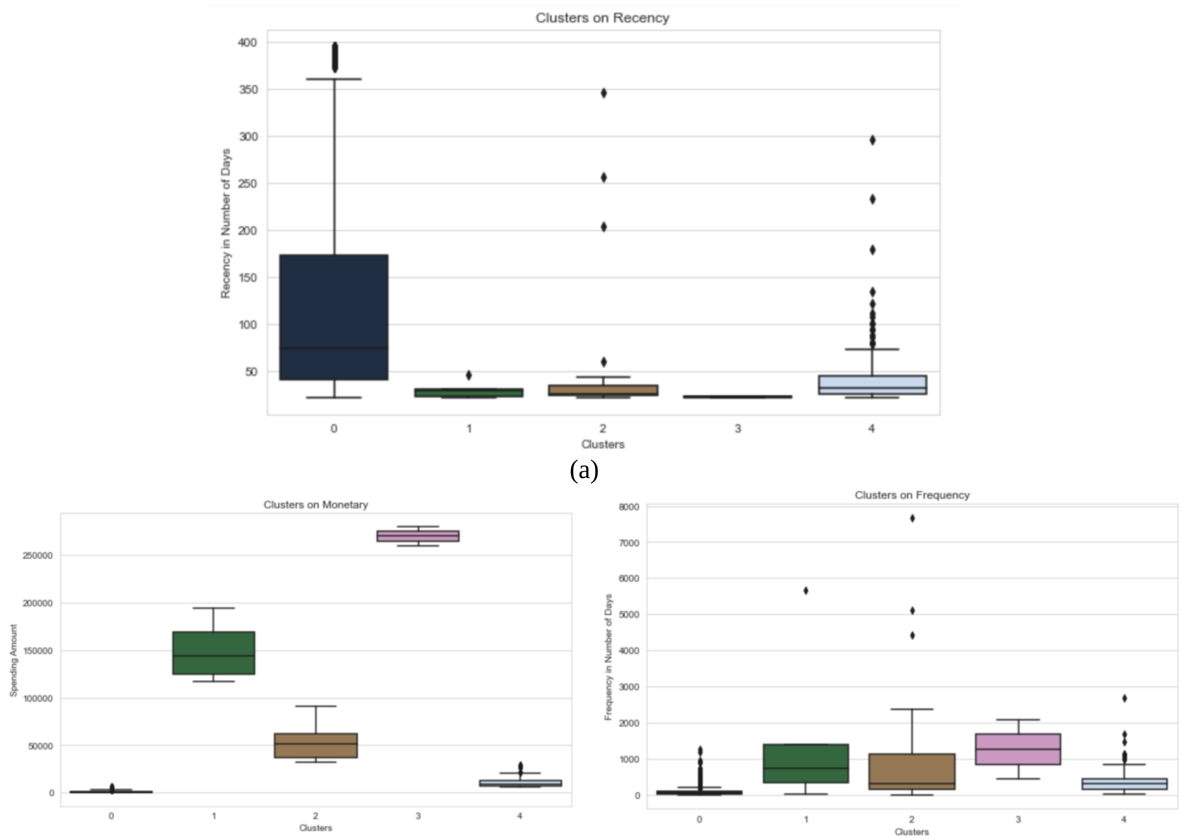
In table 4, it can be seen the results of the comparison of the Davies Bouldin, Calinski Harabasz and Silhouette methods against the three clustering methods. The results of the experiments conducted there are three clustering methods that can determine the number of clusters. The result of the number of clusters in the K-Means Clustering algorithm based on Davies Bouldin is 5 with a value of 76%, Calinski Harabasz 5 and Silhouette are 3, then Gaussian Mixture Model 5.5.3 and finally Spectral Clustering produces 3.5.5 for the three methods.

Based on these results, it was decided to apply the K-Means Clustering algorithm to be applied in the RFM model. The results of applying K-Means Clustering to the RFM model can be seen in Figure 7. The RMF model is useful for breaking up large customer profiles into much smaller segments and for gaining insight into the detailed behavior of customers.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571

e-ISSN : 2541-2019
p-ISSN : 2541-044X



(a)

(b) ©

Figure 7. RMF model cluster results

In the Recency plot, the Cluster0 value has a low recentness, which is not significant. Clusters 1 and 2 have high novelty, indicating that they are more susceptible to marketing. Cluster 3 has high current value or regular customers and cluster 4 is new customers.

In the Monetary box Plot for Cluster 0 has the lowest value indicating that they have the lowest affordability. Cluster 1 is a heavy shopper, then Cluster 2 comes in the middle and 3.4 repeat customers.

On the Plot side of the Frequency box, Cluster 0 has a low frequency, meaning that customers do not frequently make transactions, Cluster 1,2 has a better frequency, indicating that they are generally more satisfied and cluster 3,4 is a regular and new customer.

3.4. Apriori Algorithm Based on K-Means Clustering

In the previous section the results of the K-Means Clustering grouping resulted in 5 customer groups based on the RFM model, then the data from the five customer groups will be applied to the Apriori algorithm to find patterns of products that are often purchased together which are expected to be useful for marketing management in determining the products to be purchased. offer or promote to customers. The following reports the results of product patterns that are often purchased simultaneously, which can be seen in Figure 8.

Cluster 0

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(BAKING SET SPACEBOY DESIGN)	(BAKING SET 9 PIECE RETROSPOT)	0.04	0.06	0.03	0.76	12.56	0.03	3.99
1	(BAKING SET 9 PIECE RETROSPOT)	(BAKING SET SPACEBOY DESIGN)	0.06	0.04	0.03	0.50	12.56	0.03	1.90
2	(PAPER CHAIN KIT 50'S CHRISTMAS)	(PAPER CHAIN KIT VINTAGE CHRISTMAS)	0.07	0.05	0.03	0.44	8.70	0.03	1.70
3	(PAPER CHAIN KIT VINTAGE CHRISTMAS)	(PAPER CHAIN KIT 50'S CHRISTMAS)	0.05	0.07	0.03	0.60	8.70	0.03	2.34

Cluster 0

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(CHARLOTTE BAG PINK POLKADOT)	(CHARLOTTE BAG SUKI DESIGN)	0.06	0.08	0.04	0.54	6.82	0.03	2.00
1	(CHARLOTTE BAG SUKI DESIGN)	(CHARLOTTE BAG PINK POLKADOT)	0.08	0.06	0.04	0.44	6.82	0.03	1.68
2	(CHARLOTTE BAG PINK POLKADOT)	(RED RETROSPOT CHARLOTTE BAG)	0.06	0.09	0.05	0.74	7.92	0.04	3.49
3	(RED RETROSPOT CHARLOTTE BAG)	(CHARLOTTE BAG PINK POLKADOT)	0.09	0.06	0.05	0.51	7.92	0.04	1.93
4	(CHARLOTTE BAG PINK POLKADOT)	(STRAWBERRY CHARLOTTE BAG)	0.06	0.07	0.03	0.50	7.35	0.03	1.87
...	...	...	...	...	...	...	...	...	...
467	(LUNCH BAG CARS BLUE, LUNCH BAG RED RETROSPOT)	(LUNCH BAG BLACK SKULL., LUNCH BAG PINK POLKA...	0.07	0.07	0.03	0.47	6.51	0.03	1.77
468	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG PINK POLKADOT, LUNCH BAG CARS BLUE,...	0.14	0.05	0.03	0.25	5.21	0.03	1.26
469	(LUNCH BAG PINK POLKADOT)	(LUNCH BAG BLACK SKULL., LUNCH BAG CARS BLUE,...	0.13	0.04	0.03	0.27	6.05	0.03	1.31
470	(LUNCH BAG CARS BLUE)	(LUNCH BAG BLACK SKULL., LUNCH BAG PINK POLKA...	0.13	0.05	0.03	0.26	5.00	0.03	1.29
471	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG BLACK SKULL., LUNCH BAG PINK POLKA...	0.17	0.04	0.03	0.21	4.67	0.03	1.21

472 rows × 9 columns

Cluster 1

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ASSORTED COLOUR BIRD ORNAMENT)	(PAINTED METAL PEARS ASSORTED)	0.15	0.04	0.03	0.23	5.59	2.83e-02	1.25
1	(PAINTED METAL PEARS ASSORTED)	(ASSORTED COLOUR BIRD ORNAMENT)	0.04	0.15	0.03	0.82	5.59	2.83e-02	4.79
2	(ASSORTED COLOUR BIRD ORNAMENT)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.15	0.22	0.04	0.27	1.22	6.91e-03	1.06
3	(WHITE HANGING HEART T-LIGHT HOLDER)	(ASSORTED COLOUR BIRD ORNAMENT)	0.22	0.15	0.04	0.18	1.22	6.91e-03	1.04
4	(HOT WATER BOTTLE I AM SO POORLY)	(CHOCOLATE HOT WATER BOTTLE)	0.05	0.07	0.03	0.68	9.36	3.10e-02	2.88
5	(CHOCOLATE HOT WATER BOTTLE)	(HOT WATER BOTTLE I AM SO POORLY)	0.07	0.05	0.03	0.48	9.36	3.10e-02	1.82
6	(HEART OF WICKER LARGE)	(HEART OF WICKER SMALL)	0.10	0.11	0.05	0.54	5.03	4.32e-02	1.96
7	(HEART OF WICKER SMALL)	(HEART OF WICKER LARGE)	0.11	0.10	0.05	0.50	5.03	4.32e-02	1.80
8	(HEART OF WICKER LARGE)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.10	0.22	0.04	0.40	1.85	1.84e-02	1.31
9	(WHITE HANGING HEART T-LIGHT HOLDER)	(HEART OF WICKER LARGE)	0.22	0.10	0.04	0.18	1.85	1.84e-02	1.10
10	(HEART OF WICKER SMALL)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.11	0.22	0.03	0.32	1.46	1.08e-02	1.15

Cluster 2

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(BAKING SET SPACEBOY DESIGN)	(BAKING SET 9 PIECE RETROSPOT)	0.05	0.07	0.03	0.70	10.32	0.03	3.06
1	(BAKING SET 9 PIECE RETROSPOT)	(BAKING SET SPACEBOY DESIGN)	0.07	0.05	0.03	0.52	10.32	0.03	1.96
2	(HAND WARMER OWL DESIGN)	(HAND WARMER RED LOVE HEART)	0.06	0.06	0.04	0.61	10.34	0.03	2.40
3	(HAND WARMER RED LOVE HEART)	(HAND WARMER OWL DESIGN)	0.06	0.06	0.04	0.63	10.34	0.03	2.54
4	(HEART OF WICKER LARGE)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.05	0.14	0.03	0.62	4.49	0.03	2.30
5	(WHITE HANGING HEART T-LIGHT HOLDER)	(HEART OF WICKER LARGE)	0.14	0.05	0.03	0.23	4.49	0.03	1.24
6	(PAPER CHAIN KIT 50'S CHRISTMAS)	(PAPER CHAIN KIT VINTAGE CHRISTMAS)	0.08	0.06	0.04	0.51	8.15	0.04	1.93
7	(PAPER CHAIN KIT VINTAGE CHRISTMAS)	(PAPER CHAIN KIT 50'S CHRISTMAS)	0.06	0.08	0.04	0.66	8.15	0.04	2.67
8	(WHITE HANGING HEART T-LIGHT HOLDER)	(RED HANGING HEART T-LIGHT HOLDER)	0.14	0.05	0.03	0.23	4.31	0.03	1.24
9	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.05	0.14	0.03	0.60	4.31	0.03	2.15

Cluster 3

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11571
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.06	0.08	0.04	0.64	7.91	0.03	2.56
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.08	0.06	0.04	0.49	7.91	0.03	1.85
2	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.09	0.08	0.06	0.68	8.40	0.05	2.87
3	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.08	0.09	0.06	0.74	8.40	0.05	3.48
4	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE IVORY)	0.09	0.05	0.04	0.40	8.71	0.03	1.59
...	...	...	...	...	...	...	...	...	...
63	(REGENCY CAKESTAND 3 TIER, ROSES REGENCY TEACU...	(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY...	0.05	0.06	0.03	0.61	9.95	0.03	2.39
64	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	0.09	0.04	0.03	0.39	10.41	0.03	1.57
65	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...	0.07	0.04	0.03	0.46	10.84	0.03	1.77
66	(REGENCY CAKESTAND 3 TIER)	(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY...	0.18	0.05	0.03	0.19	3.55	0.02	1.17
67	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY...	0.09	0.04	0.03	0.36	9.64	0.03	1.50

68 rows × 9 columns

Cluster 4

Figure 8. Apriori Result Based K-Means Clustering

In Figure 8, the a priori results are set with min_support=0.03, metric = 'lift' and min_threshold=0.5. The data displayed is in accordance with the results of grouping with the number of clusters=5 and each cluster will display data on products that are often purchased together. These a priori results can be used as input for store owners to increase marketing targets in the future.

CONCLUSION

This study discusses the application of the resulting clustering technique used in the a priori algorithm to find products that are often purchased by customers simultaneously. In contrast to the previous work which focused on applying clustering to the RFM model to find potential customers, the potential product that customers buy most often is an integral part of marketing management. Product information is very important to determine marketing targets, such as products that are most often purchased by customers, which can be used as a reference to increase stock and placement adjacent to the product.

Based on test results on Association Rules Mining with the Apriori algorithm to find products that are often purchased by customers based on the RFM model grouped with K-Means Clustering, Exploration data analysis is very useful for understanding data in visual form. The proposed clustering technique was evaluated using Davies Bouldin, Calinski Harabasz and Silhouette to find the number of clusters. The K-Means clustering technique produces 76% accuracy with 5 clusters and is preferred over other techniques because it is faster than the Gaussian Mixture Model and Spectral Clustering. The results of the K-Means clustering have been successfully applied to the Apriori algorithm to find patterns of purchasing goods by customers. The results of the Apriori algorithm can be used by management to determine more optimal marketing targets.

REFERENCES

D. Kamthania, A. Pahwa, and S. S. Madhavan, 2018, “Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business,” J. Comput. Inf. Technol., vol. 26, no. 1, pp. 57–68, doi: 10.20532/cit.2018.1003863.

Dedi, M. I. Dzulhaq, K. W. Sari, S. Ramdhan, R. Tullah, and Sutarman, 2019, “Customer Segmentation Based on RFM Value Using K-Means Algorithm,” Proc. 2019 4th Int. Conf. Informatics Comput. ICIC doi: 10.1109/ICIC47613.2019.8985726.

J. Wu et al., 2021, “ User Value Identification Based on Improved RFM Model and K -Means++ Algorithm for Complex Data Analysis ,” Wirel. Commun. Mob. Comput., vol, pp. 1–8, 2021, doi: 10.1155/2021/9982484.

J. Wu et al., 2020, “An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K -Means Algorithm,” Math. Probl. Eng., , doi: 10.1155/2020/8884227.

J. Wu et al., 2020, “An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K -Means Algorithm,” Math. Probl. Eng., vol. , doi: 10.1155/2020/8884227.

M. Pohludka and H. Štverková, 2019, “The Best Practice of CRM Implementation for Small- and Medium-Sized Enterprises,” Adm. Sci., vol. 9, no. 1, p. 22, Mar, doi: 10.3390/admsci9010022.

M. Ryu, K. Il Ahn, and K. Lee, 2021, “Finding effective item assignment plans with weighted item associations using a hybrid genetic algorithm,” Appl. Sci., vol. 11, no. 5, pp. 1–19, Mar, doi: 10.3390/app11052209.

*name of corresponding author

O. Piskunova and R. Klochko, 2020, “Classification of e-commerce customers based on Data Science techniques,” CEUR Workshop Proc., vol. 2649, pp. 6–20.

S. G. Carbajal, 2021, “Customer segmentation through path reconstruction,” Sensors, vol. 21, no. 6, pp. 1–17, Mar. doi: 10.3390/s21062007.

S. Guney, S. Peker, and C. Turhan, 2020, “A combined approach for customer profiling in video on demand services using clustering and association rule mining,” IEEE Access, vol. 8, pp. 84326–84335, doi: 10.1109/ACCESS.2020.2992064.

T. Sai Vijay, S. Prashar, and V. Sahay, 2019, “The influence of online shopping values and web atmospheric cues on e-loyalty: Mediating role of e-satisfaction,” J. Theor. Appl. Electron. Commer. Res., vol. 14, no. 1, pp. 1–15, doi: 10.4067/S0718-18762019000100102.

W. J. Chen, M. J. Jhou, T. S. Lee, and C. J. Lu, 2021, “Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association,” Entropy, vol. 23, no. 4, p. 477, Apr, doi: 10.3390/e23040477.

Y. Cheng, M. Cheng, T. Pang, and S. Liu, 2021, “Using Clustering Analysis and Association Rule Technology in Cross-Marketing,” Complexity, vol, doi: 10.1155/2021/9979874.

Y. Han, D. Yu, C. Yin, and Q. Zhao, 2020, “Temporal Association Rule Mining and Updating and Their Application to Blast Furnace in the Steel Industry,” Comput. Intell. Neurosci., vol, doi: 10.1155/2020/7467213.

*name of corresponding author