p-ISSN: 2252-3006
e-ISSN: 2685-2411

Implementation Of ETL E-Commerce For Customer Clustering Using RFM And K-Means Clustering

Farrikh Al Zami[a1], Fikri Diva Sambasri, Rifqi Mulya Kiswanto, Rama Aria Megantara, Ahmad Akrom, Ricardus Anggi Pramunendar, Dwi Puji Prabowo, Puri Sulistiyawati
[a]Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
e-mail: [1]alzami@dsn.dinus.ac.id

Abstract

E-commerce is the activity of selling and buying goods through an online system or online. Customer loyalty is important factor in running a Company. To maintain the loyalty, the company can provide several different treatments to its customers so that they can maintain good relations with customers and can increase product purchase revenue. In this study, we are using ETL method with the K-Means clustering algorithm and RFM (Recency, Frequency, Monetary) feature to segment E-Commerce customer. The proposed method obtains the silhouette score is 0.470 in 4 optimum clusters. Thus, we believe that RFM with K-Means Clustering can be used for segmenting the Customer to increase the Customer loyalty.

Keywords: *E-commerce, Customer Segmentation, K-Means, RFM.*

1 Introduction

In the 21st century, internet usage continues to increase and is widely used in various industrial or individual sectors [1]. This can be proven from the results of a survey by the Association of Internet Service Providers in 2018. It is known that 64.8% of internet usage has increased, it is recorded that in Indonesia the Java Island region has a large contribution of internet users, namely 55% [2]. The use of internet technology has various benefits including being able to be used as a means of communicating with other people, seeking information about the current situation, as a learning tool to increase knowledge, and as a means of selling or buying goods. Many industries that use the internet, one of which is the electronic commerce (E-commerce) industry [3] [4].

E-commerce is the activity of selling and buying goods through an online system. Customer to customer (C2C) is one of the e-commerce models, namely a business model where consumers from a marketplace sell products to other consumers [5]. Examples of this business model are Bukalapak, Tokopedia, Shopee, and others.

The thing that needs to be considered in the e-commerce business model is knowing the level of customer loyalty. By knowing the level of customer loyalty, companies can provide some special treatment to customers so that they can maintain good relationships with customers and can increase product sales [6].

To be able to determine the level of customer loyalty, one of the methods used is RFM (Recency, Frequency, Monetary) segmentation [7] [8]. RFM is a method of customer segmentation based on transaction history by dividing customer activities into 3 parameters, namely recency, frequency, and monetary [9] [10] [11]. The recency parameter is the difference between the last day the customer made a transaction and the day the data was analyzed. In this study, the day unit was used. The frequency parameter is the number of transactions made by the customer. The monetary parameter is the total number of orders that have been issued by customers.

The algorithm that is often used for consumer segmentation is clustering [12]. There are several algorithms used for clustering such as the K-Means method [13], the Agglomerative clustering method, and the DBSCAN method. The K-Means method is a data mining clustering method that seeks to group data into one or more groups [14]. The main principle in the K-Means method is to find the center of the cluster that minimizes the total distance of each point to its center.

The previous research that applied the K-Means algorithm and RFM method for customer segmentation was the research conducted by Rahma Wati Br Sembiring Berahmana, Fahd Agodzo Mohammed, and Kankamol Chairuang. Bouldin is 0.33009058 and the result of the Silhouette Index is 0.912671056 with the best number of clusters being 2, namely the Dormant cluster and the Golden cluster [15]. So this study wants to know customer segmentation using the K-Means clustering method by considering the feature construction into RFM on the E-Commerce Olist dataset.

Based on what has been described above, the formulation of the problem in this study is: (1) How to make feature construction from the E-Commerce Olist dataset feature into an RFM feature. (2) How is the performance of the K-Means algorithm in classifying existing customers in the E-Commerce Olist data.

The purpose of this research are: (1) Get the results of feature construction from the E Commerce Olist dataset feature in the form of the RFM feature. (2) Knowing the performance of the K-Means algorithm in classifying existing customers in the E-Commerce Olist data.

2 Research Method / Proposed Method

Here, we are using Extract, Transform and Load Method in executing RFM Analysis. First, we extract the data from data source which consists of 9 tables. Then, we transform the data by merging the tables and perform data understanding. Then, we load the data to Cross-Industry Standard Process for Data Mining (CRISP DM) methods to perform data mining, such as: business understanding, data understanding, data preparation, modeling, evaluation and or deployment. The detailed procedure can be seen as follows:

2.1 Data Source

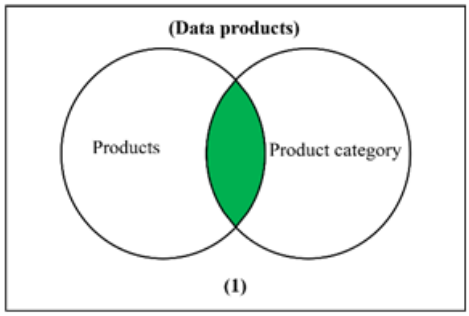
The data used in this research topic is secondary data taken from the Kaggle website (https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce) about performance data in an E-Commerce company in Brazil created by Olist . This data has information on one hundred thousand customer orders from 2016 to 2018. In this study there are 9 types of table data, namely payment type data, product data, purchase review data, purchase data, purchase item data, seller data, location data, product category name, and customer data.

2.2 Data Analysis Technique

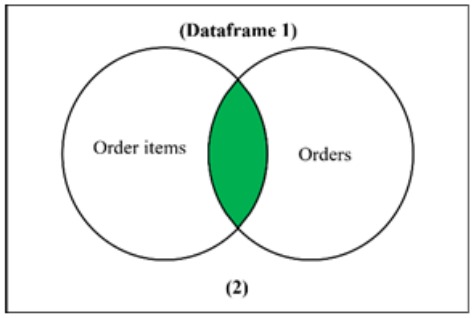
At this stage the author writes down several steps used in analyzing the data held in order to get a solution to the research problem that the author is doing. Data analysis is carried out as follows:

2.2.1 Data Merger

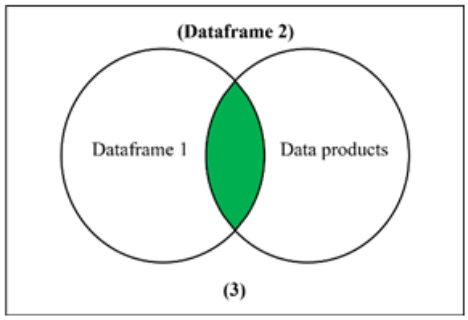
At this stage, the author selects the required data and collects data by merging data from several tables into 1 dataset table with an inner join. The following is the process of joining data:



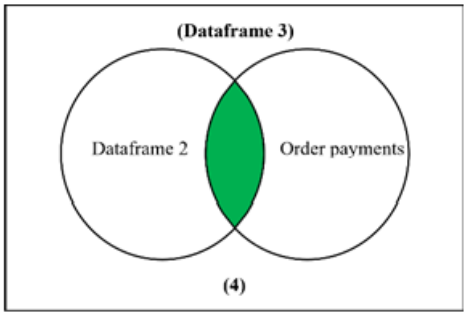
(a) Join Process for Products and Product Category



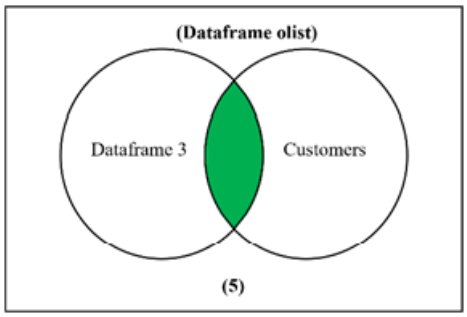
(b) Join Process Order Items and Orders



(c) Join Process from Dataframe 1 to Data Products



(d) Join Process from Dataframe 2 to Order Payments



(e) join process from Dataframe 3 to Customers

Figure 1 Joining Process from Tables

From figure 1, we are joining products and product category resulted in **Data Products table** (fig 1.a), then we are joining order items and orders into **Dataframe 1** (fig 1.b). Data Products and Dataframe 1 are inner join and resulted in **Dataframe 2** (fig 1.c). Dataframe 2 and order payments are joined to become **Dataframe 3** (fig 1.d). Finally, Dataframe 3 and customers are joined then resulted in Dataframe Olist. This Dataframe then used for Clustering for the next step

2.2.2 Data Understanding

In the data understanding process, the authors carry out the Exploratory Data Analysis (EDA) process by conducting descriptive statistics, correlation analysis, and visualization which aims to ensure that the existing data has no problematic data and find patterns and insights that can be used to develop strategies in modeling later.

2.2.3 Data Selection

At the data selection stage, the author selects attributes or features that are in accordance with the method proposed by the author, namely the RFM method. The following data features were selected by the authors:

Table 1 Data Selection

Field	Information
Order_purchase_timestamp	Data that shows the time of purchases made by customers
Customer_unique_id	Customer key/code
Order_status	Customer order status
Price	Price of goods/products
Freight_value	Shipping cost data
Order_id	Order code

p-ISSN: 2252-3006
e-ISSN: 2685-2411

2.2.4 Data Preprocessing

The data that has been selected and then carried out a preprocessing process includes:

2.2.4.1 Data Cleaning

Perform data selection through checking for missing values, checking for duplicate values, checking data types, and removing irrelevant features to be used in the model.

2.2.4.2 Feature Engineering

The feature engineering process is the process of creating new features or adding features.

Table 2 Feature Engineering Results

Field	Information
Customer_unique_id	Customer key/code
Recency	Indicates the distance data of customer purchases which is the time interval of the customer's final order date data until the date of the analysis process.
Frequency	Indicates the data on the number of customer purchases or the number of purchase orders made by the customer.
Monetary	Indicates total customer expenditure data to purchase products.

2.2.4.3 *Standardization*

The standardization process is the process of changing the feature values. The standardization results have a mean value of 0 and a standard deviation of 1.

2.3 Proposed Method

2.3.1 Model Training & Evaluation Model

At this stage, the examiner uses data on the time of purchase in 2016 and 2017 as training data, while for the test data uses data on the time of purchase in 2018. Then the training data will be carried out in the training process using the K-Means clustering algorithm by conducting several trials of the number of clusters. To get the best number of clusters, you can use the evaluation of the Elbow Method and Silhouette Score models. After getting the best number of clusters, a training process will be carried out using a predetermined number of clusters and storing the model for use in the testing phase.

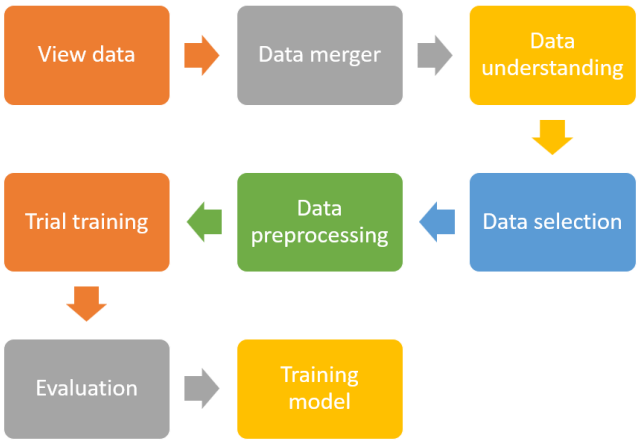


Figure 2 Proposed Method

2.3.2 Testing Stage

At this stage, data on the time of purchase in 2018 was used and made predictions using the previously stored model to get the cluster label.

3 Result and Discussion

3.1 Data Merger

Merge data with the inner join method, according to the data needed to become one dataset. Obtained 115878 records which consist of 22 columns which include 9 features with numeric data type and 13 features with object / string data type. The following is an example of how to combine data and the results of data merging using the following coding:

Table 3 Example of Coding Data Merging Using Python

Coding: Merging data using Python
df_products= df_products.merge(df_category, on='product_category_name')
df_products= df_products.rename(columns={'product_category_name_english':
'product_category_name','product_category_name':'product_category_name_old'})
df_products = df_products.drop(columns=['product_category_name_old'])
df1 = df_order_items.merge(df_orders, on='order_id')
df2 = df1.merge(df_products, on='product_id')
df3 = df2.merge(df_order_payments, on='order_id')
df_olist = df3.merge(df_customer, on='customer_id')

Table 4 Data Merged Result Dataset

Field	**Information**
Order_id	Unique key for order
Order_item_id	Sequence numbers identify the number of items included in the same order
Product_id	Unique key for product
Seller_id	Unique key for seller
Shipping_limit_date	Seller's delivery deadline to handle orders to logistics partners
price	Product price
Freight_value	Product shipping cost price
Customer_id	Unique key for customer
Order_status	Product delivery status
Order_purchase_timestamp	Product purchase time
Order_estimated_delivery_date	Product delivery time
Product_category_name	Product category name
Payment_sequential	The number of payment methods made by the customer

Payment_type	The type of payment used by the customer to buy the product
Payment_installments	The number of installments selected by the customer
Payment_value	Transaction value
Customer_unique_id	Customer unique key
Customer_zip_code_prefix	The first five digits of the customer's postal code
Customer_city	The name of the city the customer comes from
Customer_state	The customer's area comes from
Month_order_purchase	Month of product purchases made by the customer
Year_order_purchase	Year of product purchase made by customer

3.2 Data Understanding

Data understanding is a process to find out information from a data that is owned. To find out the data information, several visualizations were made to gain insight as follows:

3.2.1 Average monthly active users and number of new customers

In this stage, visualization will be made to find out information about the average monthly active users and the number of new customers per year.

Table 5 Average Active Users

No	Years	Average Active Users
1	2016	102.00
2	2017	3589.33
3	2018	5840.22

Table 6 Number of New Customers

No	Years	Number of New Customers
1	2016	306
2	2017	42467
3	2018	51314

p-ISSN: 2252-3006
e-ISSN: 2685-2411

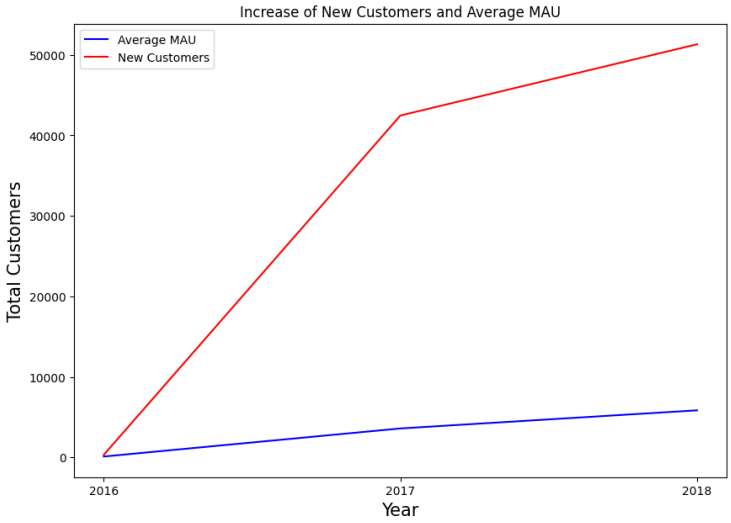


Figure 3 Graph of Average MAU and Number of New Customers

The available data starts from transaction data in September 2016 which causes the results of the analysis in 2016 to have far differences compared to the values in 2017 and 2018. From this analysis, it is seen that monthly customer activity and also the number of new customers have increased.

3.2.2 Number of Customers Making Repeat Orders

At this stage the researcher wants to know data about the number of customers who make repeat orders every year.

Table 7 Number of Customers with Repeat Purchase

No	Year	Total repeat order customers
1	2016	46
2	2017	6454
3	2018	7251

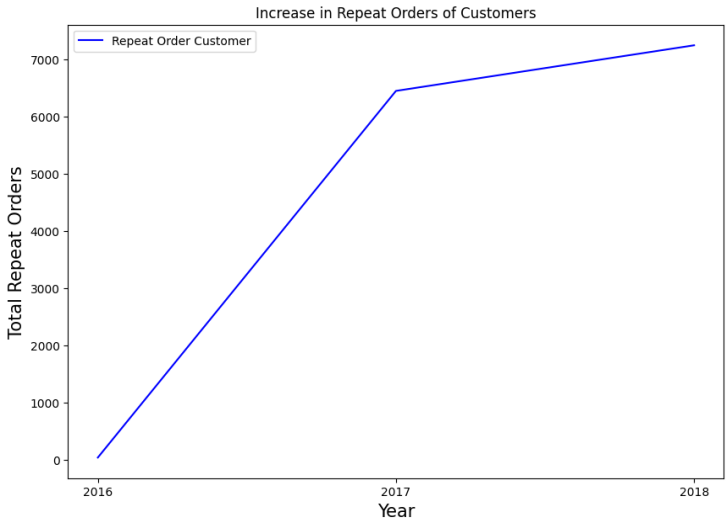


Figure 4 Number of Repeat Orders Grafik

In terms of orders / orders made by customers looks good. It can be seen that the number of customers who make repeat orders has increased from 2016 to 2018.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

3.2.3 Total product category revenue per year

At this stage the researcher wants to find out data information about what product category names have the highest amount of income each year

Table 8 Total Product Revenue per Year

No	Year	Product category name	Total product revenue
1	2016	Furniture décor	7637.49
2	2017	Bed bath table	619833.82
3	2018	Health beauty	885545.71

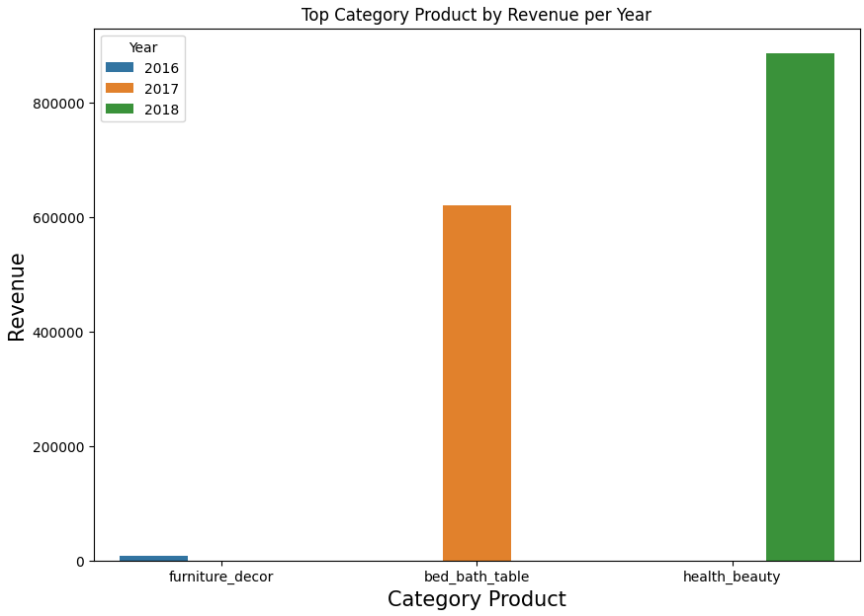


Figure 5 Top Chart of Product Categories Based on Highest Income

From the visualization above, it can be seen that the product category that provides the highest amount of revenue every year is always changing. Viewed from the side of the company's overall income has also increased every year.

3.2.4 Number of product category cancellations per year

At this stage the researcher wants to find out information about the name of the product category that has the highest number of cancellations of product purchases per year.

Table 9 Number of Product Purchase Cancellations

No	Year	Product category name	Number of canceled product orders
1	2016	Toys	3
2	2017	Housewares	32

p-ISSN: 2252-3006
e-ISSN: 2685-2411

3	2018	Health beauty	30

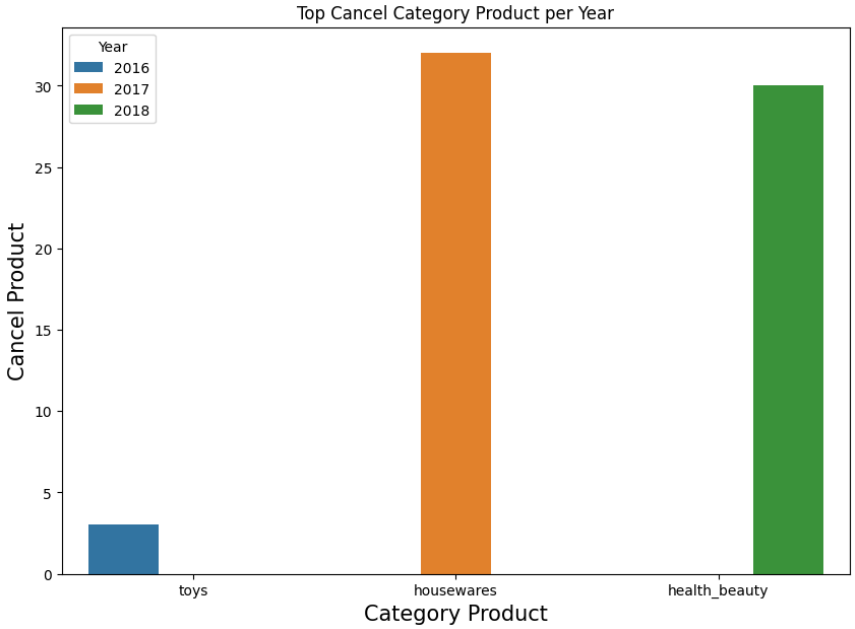


Figure 6 Top Chart of Product Categories by Number of Purchase Cancellations

From the visualization above, each year there are different product categories that experience purchase cancellations. However, there is an interesting thing that in 2018 the health beauty product category was the name of the product category that provided the most revenue as well as the name of the product category that experienced the highest number of product purchase cancellations in 2018.

3.2.5 Favorite payment type per year

At this stage the researcher wants to know the types of payment types that are often used by customers to buy products at E-Commerce companies per year.

Table 10 Number of Users by Payment Type

No	Payment type	Year	Number of users
1	Boleto	2016	70
2	Boleto	2017	10731
3	Boleto	2018	11741
4	Credit card	2016	290
5	Credit card	2017	38151
6	Credit card	2018	47082
7	Debit card	2016	1
8	Debit card	2017	455
9	Debit card	2018	1202
10	Voucher	2016	22
11	Voucher	2017	3220

p-ISSN: 2252-3006
e-ISSN: 2685-2411

12	Voucher	2018	2913

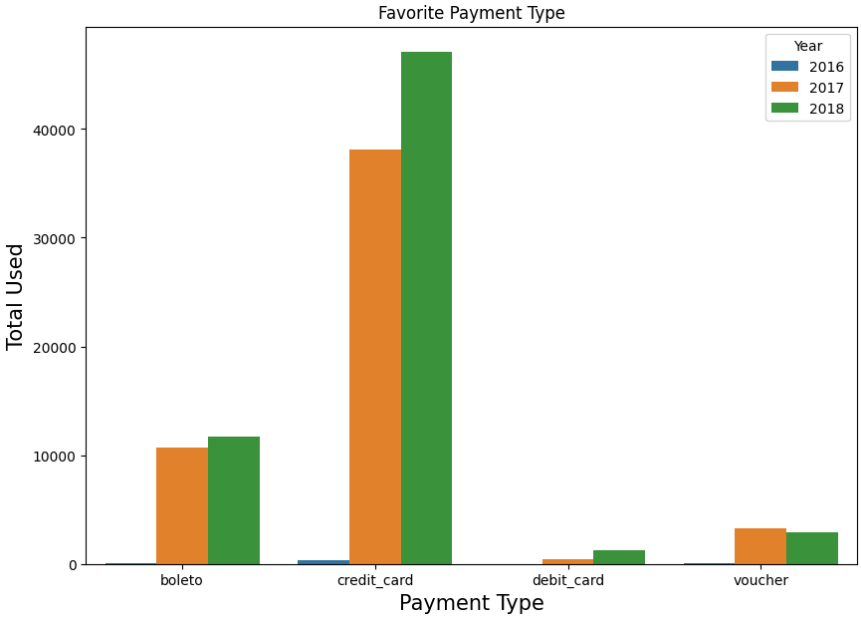


Figure 7 Favorite Payment Type

From the visualization above, it is found that the payment method that is more attractive to customers in the process of purchasing products at an E-Commerce company is a credit card.

3.3 Data Selection

After visualizing the data to be able to understand the content of information in the data held, then selecting the appropriate features to create RFM features. The following is an example of the results of the feature selection used:

	order_purchase_timestamp	customer_unique_id	order_status	price	freight_value	order_id
0	2017-09-13 08:59:02	871766c5855e863f6eccc05f988b23cb	delivered	58.9	13.29	00010242fe8c5a6d1ba2dd792cb16214
1	2017-06-28 11:52:20	0fb8e3eab2d3e79d92bb3fffbb97f188	delivered	55.9	17.96	130898c0987d1801452a8ed92a670612
2	2017-08-01 18:38:42	e7c828d22c0682c1565252deefbe334d	delivered	58.9	16.17	6f8c31653edb8c83e1a739408b5ff750
3	2017-08-10 21:48:40	0bb98ba72dcc08e95f9d8cc434e9a2cc	delivered	58.9	13.29	7d19f4ef4d04461989632411b7e588b9
4	2017-07-27 15:11:51	33449409b16400dbeaf886a5140bf59c	delivered	55.9	26.93	a0f9acf0b6294ed8561e32cde1a966bc

Figure 8 Feature Selection Results

3.4 Data Preprocessing

Data preprocessing is carried out before the modeling process so that the data makes the data more structured, here are some processes for conducting data preprocessing:

3.4.1 Data Cleaning

Data cleaning is the process of cleaning data from data that is incorrect, incomplete, etc.

3.4.2 Feature Engineering

Feature Engineering is the process of making existing features into RFM features. For making RFM, you can use the following coding:

Table 11 Example of Coding for RFM Feature Creation in Python

Coding: RFM using Python

```
data_rfm_olist = olist_train_rfm.groupby(['customer_unique_id']).agg({
    'Invoice Date': lambda x: (collection_date - x.max()).days,
    'order_id': 'count',
    'Total Order Value': 'sum'})
data_rfm_olist = data_rfm_olist.reset_index()
data_rfm_olist.columns = ['customer unique id', 'recency', 'frequency',
'monetary']
```

Table 12 RFM Feature Creation Results

recency	frequency	monetary
297	1	86.22
81	1	43.62
48	1	196.89
303	1	150.12
167	1	29.00

3.4.3 Standardization

Standardization is the process of changing feature values which have a mean value of 0 and a standard deviation of 1.

Table 13 RFM Feature Standardization Results

recency	frequency	monetary
1.610151	-0.268343	-0.326049
-0.594647	-0.268343	-0.486358
-0.931492	-0.268343	0.090416
1.671395	-0.268343	-0.085585
0.283189	-0.268343	-0.541375

3.5 Modelling Process

At this stage the researchers tried to make a model using the K-Means clustering algorithm with the number of clusters from 2 to 9. Then evaluated using a silhouette score to get

p-ISSN: 2252-3006
e-ISSN: 2685-2411

the optimal number of clusters. Here it is found that the number of clusters is 4 clusters with a value of 0.470.

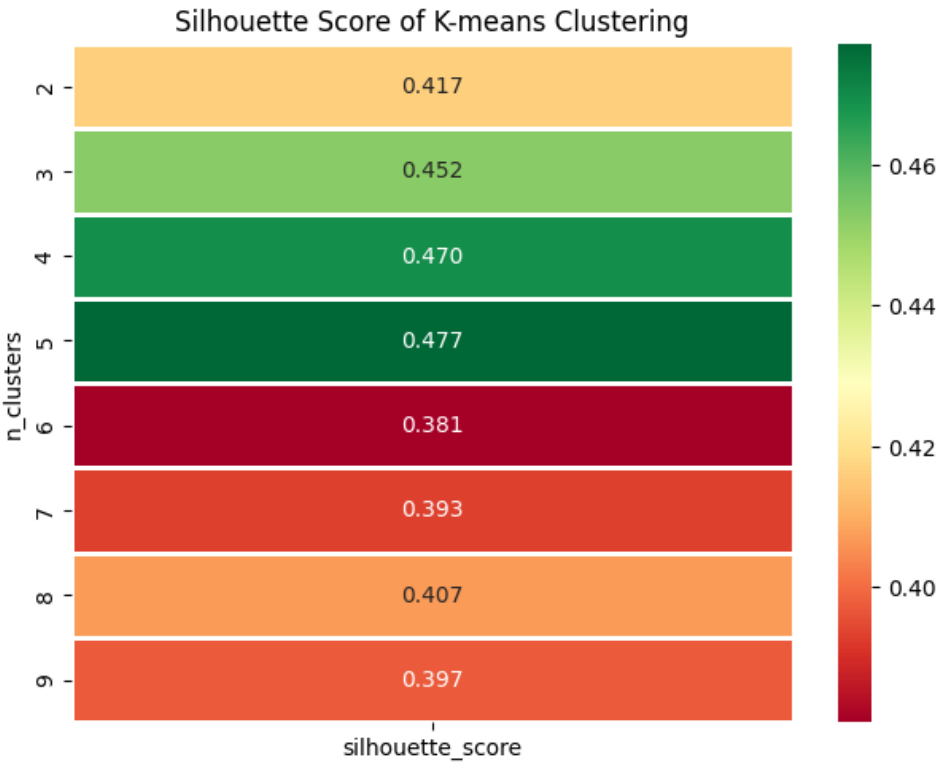


Figure 9 Silhouette Score K-means Clustering

4 Conclusion

Based on the results of research conducted by the author on the Implementation of ETL E-Commerce For Customer Clustering Using RFM And K-Means Clustering, are as follows:

1. The result of feature construction from the E-Commerce Olist dataset feature in the form of the RFM feature has been successfully executed with the findings obtained in the form of the value on the frequency feature at most is 1, so it can be interpreted that most of the customers only make a product purchase once a year.
2. The performance of the K-Means clustering algorithm proves that this algorithm can be applied to segment customers. This is evidenced by the results of the silhouette score evaluation with a value of 0.470 with a cluster value of 4 clusters.

References

[1] S. Palingi and E. C. Limbongan, “Pengaruh Internet Terhadap Industri E Commerce Dan Regulasi Perlindungan Data,” in *Seminar Nasional Riset dan Teknologi (SEMNAS RISTEK)*, 2020.

[2] APJII, “Penetrasi & Profil Perilaku Pengguna Internet Indonesia,” in *Asosiasi Penyelenggara Jasa Internet Indonesia*, 2018.

[3] S. Monalisa, P. Nadya, and R. Novita, “Analysis for Customer Lifetime Value Categorization with RFM Model,” *Procedia Comput Sci*, vol. 161, pp. 834–840, 2019, doi: 10.1016/j.procs.2019.11.190.

[4] G. Liu, “An ecommerce recommendation algorithm based on link prediction,” *Alexandria Engineering Journal*, vol. 61, no. 1, pp. 905–910, Jan. 2022, doi: 10.1016/j.aej.2021.04.081.

[5] Huynh T. K., Le H.-D, Nguyen S. V., and Tran H. M., “Applying Peer-to-Peer Networks for Decentralized Customer-to-Customer Ecommerce Model,” in *International Conference on Future Data and Security Engineering*, 2020.

[6] Latifah N, Widayani A, and Normawati R. A, "Pengaruh Perceived Usefulness Dan Trust Terhadap Kepuasan Konsumen Pada E-Commerce Shopee," *Jurnal Bisnis dan Manajemen*, vol. 14, p. 84, 2020.
[7] R. Heldt, C. S. Silveira, and F. B. Luce, "Predicting customer value per product: From RFM to RFM/P," *J Bus Res*, vol. 127, pp. 444–453, Apr. 2021, doi: 10.1016/j.jbusres.2019.05.001.
[8] M. Song, X. Zhao, H. E, and Z. Ou, "Statistics-based CRM approach via time series segmenting RFM on large scale data," *Knowl Based Syst*, vol. 132, pp. 21–29, Sep. 2017, doi: 10.1016/j.knosys.2017.05.027.
[9] Puspitasari N, Widians J. A, and Setiawan N. B, "Segmentasi Pelanggan Menggunakan Algoritme Bisecting K-Means Berdasarkan Model Recency, Frequency, Dan Monetary (Rfm)," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, pp. 78–79, 2020.
[10] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-based repurchase behavior for customer classification and segmentation," *Journal of Retailing and Consumer Services*, vol. 61, p. 102566, Jul. 2021, doi: 10.1016/j.jretconser.2021.102566.
[11] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1785–1792, May 2022, doi: 10.1016/j.jksuci.2019.12.011.
[12] V. Holý, O. Sokol, and M. Černý, "Clustering retail products based on customer behaviour," *Appl Soft Comput*, vol. 60, pp. 752–762, Nov. 2017, doi: 10.1016/j.asoc.2017.02.004.
[13] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm," *Appl Soft Comput*, vol. 113, p. 107924, Dec. 2021, doi: 10.1016/j.asoc.2021.107924.
[14] Boentarman M, Rostianingsih S, and Setiawan A, "Penerapan Segmentasi Pelanggan Dengan Menggunakan Metode K-Means Clustering Pada Sistem Customer Relationship Management Di PT Titess," *Jurnal Infra*, p. 2, 2021.
[15] Berahmana R, Mohammed F. A, and Chairuang K, "Customer Segmentation Based On RFM Model Using K-Means, K-Medoids, And DBSCAN Methods," *Lontar Komputer*, vol. 11, pp. 32–38, 2022.