Ceddi Journal of Information System and Technology (JST) *ISSN: 2829-808X (print)*
Vol. 3 No. 1 April (2024) *ISSN: 2829-6575 (online)*
https://doi.org/10.56134/jst.v3i1.81

Customer Segmentation through RFM Analysis and K-means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy

Oluwatobi Noah Akande[1*], Emmanuel Oluwatobi Asani[2], Bestman Tobechukwu Dautare[3]

[1]*Computer Science Department, Baze University, Abuja, Nigeria*
[2,3]*Computer Science Department, Landmark University, Kwara State, Nigeria.*

Abstract

Consumer segmentation is a highly effective methodology for organizations to gain a deeper understanding of their consumer base and tailor their strategies accordingly. By categorizing clients based on common attributes, organizations can gain valuable insights into their needs, preferences, and purchasing behaviors. This understanding enables firms to develop targeted marketing strategies and deliver personalized experiences that enhance customer loyalty and boost revenue. Common criteria for consumer segmentation include demographic, psychographic, geographic, and behavioral factors. Methodologies for constructing customer segments typically involve rule-based and cluster-based techniques. Rule-based segmentation uses pre-established rules to assign clients to segments, while cluster-based segmentation uses statistical techniques to identify natural clusters within the customer population. This research focuses on applying the K-Means clustering technique to segment customer behavioral data into categories such as Platinum, Gold, Silver, Bronze, or Bad. The clustering approach demonstrated a high level of accuracy and precision. Implementing an effective customer segmentation strategy can strengthen product offerings, focus marketing communications, and increase client loyalty.

Keywords: Customer Segmentation; Data Analytics; Machine Learning; Consumer Segmentation.

Received: 29 January 2024 Revised: 20 March 2024 Accepted: 29 April 2024

Introduction

Customer segmentation is a strategic method of categorizing customers into groups based on shared characteristics or behaviors (Harahap et al., 2021). This approach enables organizations to effectively engage with each group, optimizing their outreach efforts (Wicaksono, 2019; MR et al., 2024). By utilizing various client attributes, organizations can tailor marketing strategies, forecast trends, develop product plans, design campaigns, and offer relevant items (Putra et al., 2023); (Steiss, 2019). Moreover, integrating personalization techniques in consumer communications enhances the connection between the message and the audience (Rungruang et al., 2024). These attributes may include spending trends, purchasing habits, browsing histories, or other behavioral data (Jaiswal et al., 2023). Behavioral segmentation is a common strategy in Internet advertising, where specific criteria are used to target appropriate clients (Muchardie et al., 2019; Susilo et al., 2024). Consumer behavior is key in this segmentation, covering aspects like buying behavior, usage frequency, consumption rate, and willingness to purchase or use Yuan et al., 2020; Nabella & others, 2022). For example, promotional materials for a product may be displayed to a customer based on their website interactions or visits where the product is available for purchase.

Psychographic segmentation is a consumer categorization approach that classifies clients based on their psychological traits (Rojlertjanya, 2019; Fadilah, 2019). This method considers beliefs, values, lifestyle, social status, opinions, and activities to create buyer personas, guiding marketing and customer success strategies effectively (Hu et al., 2023; Buzzacchi et al., 2023). Unlike behavioral segmentation, which focuses on behavior analysis, psychographic segmentation aims to understand the underlying reasons driving behavior (Talaat et al., 2023). Understanding the psychological attributes of customers allows for customized marketing initiatives tailored to specific target groups, leading to a deeper understanding of their needs and preferences (H. Chen et al., 2019; Silalahi et al., 2023). This approach enables the development of more targeted marketing communications and the delivery of personalized experiences to different customer segments.

*Corresponding author.
E-mail address: oluwatobi.akande@bazeuniversity.edu.ng (Oluwatobi Noah Akande)

Geographic segmentation involves categorizing the target audience based on their locations (Kuo et al., 2023; Awalina & Rahayu, 2023; Ramadhan et al., 2023). This allows advertisers to effectively engage audiences in specific countries, cities, or regions by aligning their messaging with the desires and requirements of these segments (Chen & Gunawan, 2023). By segmenting the audience based on geographic location, from national level to specific postal codes, companies can target individuals living, working, or engaging in commerce in a particular area (Grunig & Repper, 2013). Geographic segmentation is recognized as an effective market segmentation method, used by companies aiming for global expansion or focusing on local markets.

The progress of machine learning has been fueled by the emergence of new learning algorithms and theoretical frameworks, along with the continuous expansion of online data availability and the decreasing costs of computation (Hartoyo et al., 2023). Data-intensive machine-learning methods have been widely used in fields such as science, technology, and business, leading to a shift towards evidence-based decision-making in areas like healthcare, manufacturing, education, financial modeling, policing, and marketing (Jordan & Mitchell, 2015). RFM-analysis has been identified as a highly effective and widely used method for understanding customer behavior, enabling the development of predictive models related to customer behavior (Mehrabioun & Mahdizadeh, 2021). RFM can also be used to categorize clients based on their past purchases, relying on three key customer attributes: the date of purchase, the frequency of purchases, and the monetary value of purchases (Dzulhaq et al., 2019; Ali et al., 2023). Recency, which measures the time between consecutive purchases, is an important metric influencing customer engagement with a brand (Lewaaelhamd, 2024); (Wang, 2022). Customers who engage frequently tend to be more loyal, showing higher levels of involvement Joung & Kim, (2023); Ma et al., (2023). The monetary value of purchases reflects the amount spent by a consumer on a brand within a specific period, suggesting that customers making substantial purchases should be treated differently (Chronopoulos et al., 2020); (Kursan Milaković, 2021). Clustering aims to group a dataset into distinct structures, rather than precisely characterizing unseen samples from the same distribution (Hu et al., 2020); Hu et al., (2023). This study explores a customer segmentation strategy combining RFM analysis and the K-means clustering algorithm to categorize customers into Platinum, Gold, Silver, Bronze, or Bad categories based on their purchasing patterns.

Method

The next section presents an overview of the approach employed in this investigation. The components encompassed in this study are the dataset utilized, the framework applied for implementation, the customer segmentation model provided, and the metrics employed for evaluating performance.

1. Data Collection
   The present work utilized an online retail dataset sourced from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Online+Retail#). This dataset comprises transactional records spanning from 01/12/2010 to 09/12/2011, originating from an E-commerce store situated in the United Kingdom. The dataset comprised a total of 541,909 instances, each consisting of eight attributes: Invoice No, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Figure 1 displays a representative subset of the dataset.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Figure 1. Sample of Online Retail Store Dataset used

2. Data Preprocessing

   The dataset included in this study underwent pre-processing procedures to eliminate abnormalities and extraneous data (Kanoujia & Karuppanan, 2024). The pre-processing steps utilized encompass:

   a. The process of selecting the columns required for analysis from the dataset. As an example, the dataset includes seven selected columns, namely InvoiceNo, Quantity, InvoiceDate, UnitPrice, CustomerID, Description, and StockCode.
   b. The dataset should be processed to eliminate duplicate rows and null values. For example, entries that contained a null customerID were excluded from the analysis.
   c. The identification and removal of transactions with negative values. This denotes the inclusion of orders that have been returned or cancelled, rendering them irrelevant to the current study. It is necessary to obtain data solely from consumers who have engaged in at least one transaction.
   d. Another aspect to consider is the process of verifying and eliminating orders that lack a description.
   e. A new column titled "amount" needs to be generated. This will be employed to determine the result of multiplying the values of Quantity and UnitPrice.

3. The Proposed Costumer Segmentation Model

   The phases involved in the suggested consumer segmentation model were captured using Figure 2. Following the pre-processing of the dataset, the RFM variables were calculated. The frequency values for each client were computed using the parameters InvoiceDate and CustomerID. The InvoiceDate variable was utilized in the calculation of recency, while the newly obtained attribute, amount, was employed in the computation of the Monetary variable.

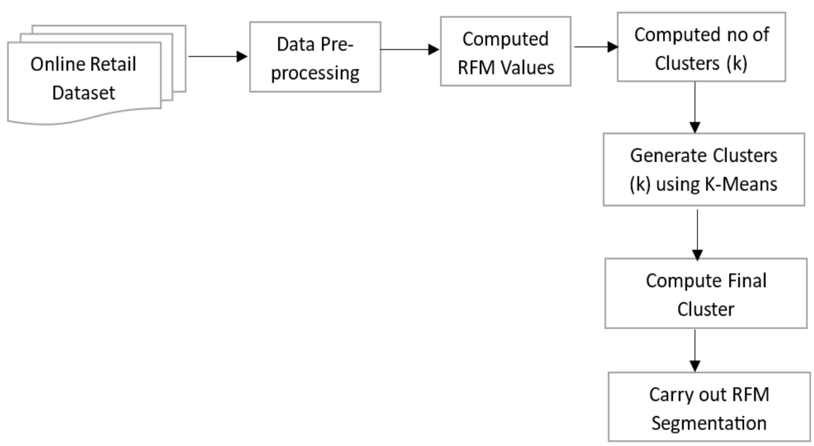

Figure 2. Proposed Research Methodology

The clustering methodology employed in this study is the K-Means algorithm, a distance-based clustering technique that partitions objects into clusters based on their numerical attributes (Ramkumar et al., 2021). Customers are categorized by the utilization of K-means clustering, which is predicated on the identification of the clusters to which they belong (Nandapala & Jayasena, 2020). The elbow method, a heuristic technique, was employed to ascertain the optimal number of clusters within a given dataset (Lezhnina & Kismihók, 2022). The proposed methodology involves plotting the explained variance against the number of clusters and determining the point of inflection in the resulting curve as the optimal number of clusters (Onumanyi et al., 2022). The quantity of potential clusters is ascertained by the value acquired by the elbow method (Mouton et al., 2020).

The customers were categorized and partitioned into distinct categories, employing the clustering outcomes and the recency, frequency, and monetary factors obtained from the dataset (Anitha & Patil, 2022). The RFM values range from 1 to 5 in all three scenarios, with 1 being the lowest value and 5 representing the maximum value (Myburg, 2023). The resultant sum is utilized for the classification of clients within a range of 3 to 15. Customers are categorized into different tiers, namely Platinum, Gold, Silver, Bronze, or Bad, according on their performance. Customers who have been awarded exceptional ratings in all three dimensions are categorized into the Platinum segment, and so on.

4. Model Implementation
   The model implementation utilized various programming tools, including Jupyter Notebooks, Webstorm, MongoDB, and NodeJS. The implementation was conducted on a laptop with an Intel Core i7 processor, running the Windows 7 operating system. The laptop had a clock speed of 1.1 GHz and a RAM capacity of 4 GB. The Python libraries employed in this study encompassed Pandas, Numpy, Seaborn, Matplotlib.pylot, Mpl_toolkit, Sklearn.metrics, Sklearn.clustering, and Sklearn.preprocessing. Pandas was utilized to ingest, analyze, and preprocess the dataset. The Excel file is transformed into data frames, which are two-dimensional arrays that offer enhanced manipulability. The dataset underwent statistical calculations, including the determination of the mean, median, and mode, using the Numpy library (McKinney, 2012). The Seaborn library contains a user-friendly interface, which makes it easier to create statistical visualizations that are both aesthetically pleasing and instructive from a statistical standpoint (Arya et al., 2023). The tool was utilized in order to assist in the visual representation of the findings (Russell et al., 2009; Sarikaya et al., 2018). It was decided to make use of the Matplotlib.pylot module in order to build visualizations in Python that included both static and animated as well as interactive components (Belorkar et al., 2020). Through the employment of the Mpl_toolkit, the construction of 3D graphs, line charts, and mesh charts was made easier. The performance of the classification was tested by utilizing the sklearn.metrics module (Agranovskii & Silukov, 2021), which includes a variety of functions such as the loss function, score function, and utility function. Through the utilization of the Sklearn.clustering module, the K-means clustering method was implemented on the RFM variables. In addition, the Sklearn.preprocessing module was used to convert the raw feature vectors into a format that is better suited for the subsequent estimators. This was accomplished by rearranging the order of the feature vectors.

Results and Discussion

Result

The development of the consumer segmentation system involved two distinct steps. Initially, an RFM (recency, frequency, and monetary) analysis was conducted. Subsequently, the K-Means clustering technique was applied to categorize the clusters into subgroups based on their classifications. In the next phase, a web-based platform will allow users to speculate about the user groups to which others belong, considering the recency, frequency, and monetary aspects gathered in the preliminary study. During the analysis phase, clients were segmented based on specific characteristics, tailored to the firm's needs. This research aims to analyze customer segmentation, focusing on three key factors: recency, monetary value, and frequency. Recency reflects the time elapsed since a customer's last transaction, monetary value indicates the total expenditure, and frequency represents the purchasing regularity.

The segmentation methodology includes RFM analysis, extracting essential aspects from the organization's dataset. Recency was calculated by comparing the most recent date with the corresponding customer ID. Frequency was determined by counting the number of invoices per customer. Monetary value was derived by multiplying the quantity of each product by its price, referencing the customer ID. The study utilized Kaggle notebooks and various Python libraries (refer to Table 1 and Figure 3).

Table 1. Python Libraries and their Uses

Libraries	Uses
Sklearn.metrics	To evaluate the performance of classification models, the sklearn.metrics module provides a range of loss, score, and utility functions.
Sklearn.preprocessing	The library offers a variety of utility functions and transformer classes that facilitate the transformation of raw feature vectors into a more appropriate format for downstream estimators.
Seaborn	The outcomes were depicted utilizing the aforementioned library. The software possesses an interactive user interface that facilitates the creation of graphs that are both instructive and statistical in nature.
Matplotlib.pylot	The library was utilized in the development of Python-based visualizations encompassing static, animated, and interactive elements.
Mpl_toolkit	The library in question utilizes matplotlib as a dependency to generate three-dimensional graphs, line charts, and mesh charts.
Numpy	The library was utilized to do statistical computations on the dataset, encompassing the determination of measures such as the mean, median, and mode, alongside other relevant analyses.

Libraries	Uses
Sklearn.clustering	The library was utilized for the implementation of the K-means clustering algorithm on the RFM variables.
Pandas	The aforementioned library was utilized for the purposes of reading, analyzing, and performing data cleansing on the dataset, as outlined in the given context. The process involves converting the Excel file into data frames, which are structured as two-dimensional arrays, so facilitating easier manipulation.

```
# for data manipulation and analysis
import pandas as pd
import numpy as np

# for plotting
import seaborn as sns
import matplotlib.pyplot as plt
from mpl_toolkits import mplot3d
sns.set_style('darkgrid')

# Silhouette analysis
from sklearn.metrics import silhouette_score

# To perform KMeans clustering
from sklearn.cluster import KMeans

# for scaling
from sklearn.preprocessing import StandardScaler

import warnings
warnings.filterwarnings('ignore')
```

Figure. 3 Intrface that Illustrates the The Libraries used in The Segmentation

The dataset employed in this study is the online retail dataset, accessible from the UCI machine learning repository. The data underwent pre-processing and subsequent analysis in order to extract the key findings from the dataset. Figures 4 and 5 depict the structure of the raw dataset and the pre-processed version respectively.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0.83	14397.0	United Kingdom
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224.69	15498.0	United Kingdom
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10.95	15311.0	United Kingdom
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1.25	17315.0	United Kingdom

Figure 4. An Excerpt from The Dataset

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
469581	576599	23301	GARDENERS KNEELING PAD KEEP CALM	4	2011-11-15 15:06:00	1.65	14544.0
429762	573576	22418	10 COLOUR SPACEBOY PEN	1	2011-10-31 14:09:00	1.63	14096.0
44175	540157	21658	GLASS BEURRE DISH	4	2011-01-05 11:41:00	3.95	15311.0
315505	564729	23200	JUMBO BAG PEARS	2	2011-08-28 12:44:00	2.08	13137.0
355089	567905	72760B	VINTAGE CREAM 3 BASKET CAKE STAND	2	2011-09-22 16:37:00	9.95	12952.0

Figure 5. Sample Data After Pre-Processing

The results of extracting each RFM variable are illustrated in Figure 5. A negative correlation exists between the recency value and its score, indicating that as the recency value increases, the score decreases. Conversely, a positive correlation is observed between the frequency value and its score, meaning that as the frequency value increases, so does the score. Similarly, a positive relationship is noted between the monetary value and its score, indicating that an increase in the monetary value leads to a higher score. The RFM variables were labeled numerically from one (1) to five (5), with one representing the lowest value and five representing the highest. These labels were then categorized based on percentiles (25%, 50%, 75%, and 90%). Customers were classified into five categories—platinum, gold, silver, bronze, and bad—based on their RFM score, which ranged from three (3) to fifteen (15). Figure 6 depicts a scatter plot showing the RFM variables.

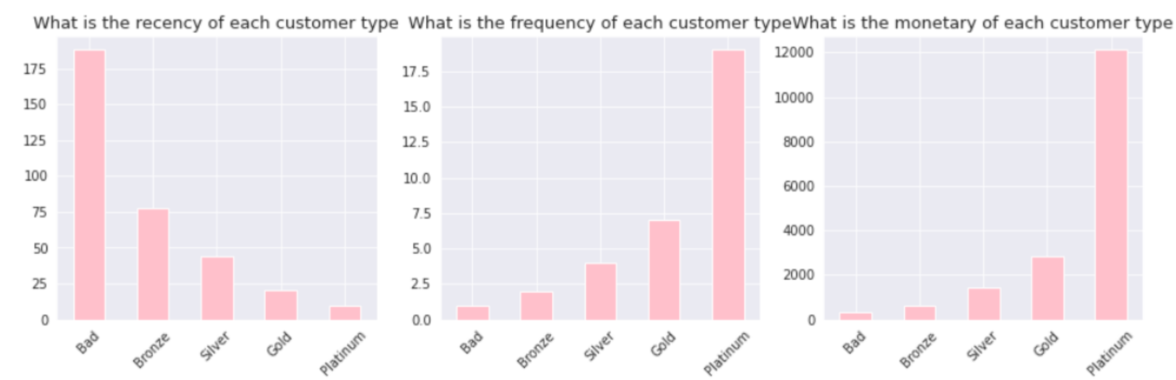

Figure 6. Bar Chart Showing the RFM Values Against the Category

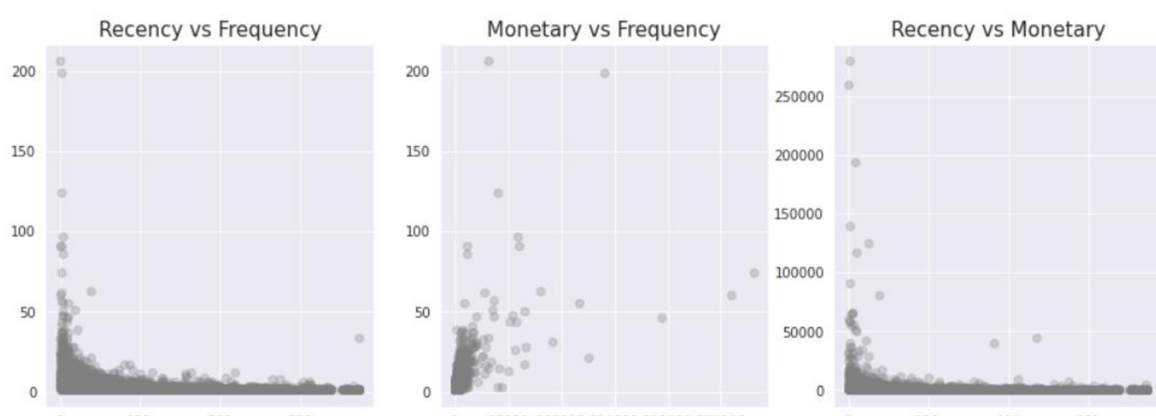

Figure 7. Scatter flot of The Recency Against Other Variables

After conducting the RFM analysis and retrieving the necessary variables, the K-Means clustering technique is used to group the data into clusters and classify clients into distinct groups: Bad, Bronze, Silver, Gold, and Platinum. The optimal value of k, representing the number of clusters, can be determined by applying the elbow curve approach to the dataset. This technique assesses different cluster numbers by evaluating the total within-cluster sum of squares in the dataset. The result of the elbow approach is shown in Figure 8.

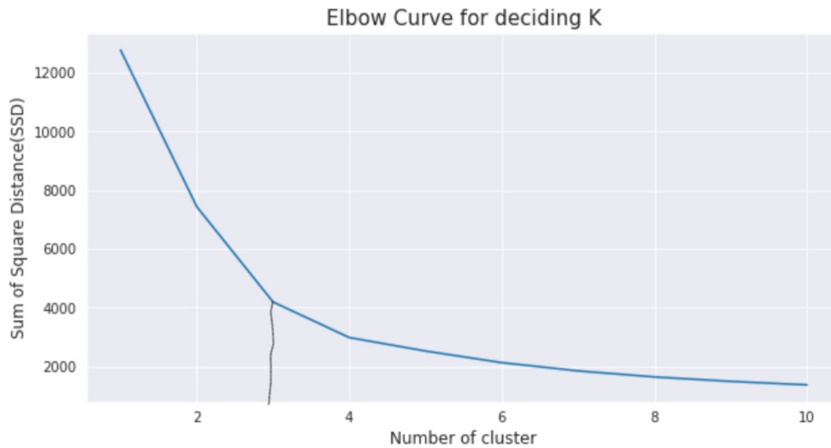

Figure 8. Results of the Elbow Curve Method

In the present scenario, the value assigned to the variable k is three, indicating that the appropriate number of clusters to employ would be three. The values of the RFM variables and their corresponding clusters are depicted in Figure 9. Additionally, Figure 10 illustrates the RFM values for specific consumers and their respective clusters.

	recency	frequency	monetary
clusters			
0	47.0	3.0	1016.0
1	21.0	12.0	5412.0
2	249.0	1.0	434.0

Figure 9. Clusters Based on The RFM Analysis

	recency	frequency	monetary	clusters
0	3	7	4310.00	1
1	76	4	1797.24	0
2	19	1	1757.55	0
3	311	1	334.40	2
4	37	7	1665.74	0

Figure 10. Custumers Clusters Based on Their RFM Variables

Discussion

The model presented in this study was implemented using a web platform to develop a consumer segmentation system. The RFM values obtained were used to categorize clients into segments: platinum, gold, silver, bronze, and bad, where platinum represents the highest tier and bad the lowest. Consumers with cumulative RFM values ranging from three (3) to five (5) are classified as belonging to the bad group, while those with values from twelve (12) to fifteen (15) are categorized as belonging to the prestigious platinum category. Figures 10 to 14 depict the graphical user interface (GUI) of the web platform, displaying the range of categories available. As shown in Figure 10, a customer with an RFM score of four falls into the poor performance category.

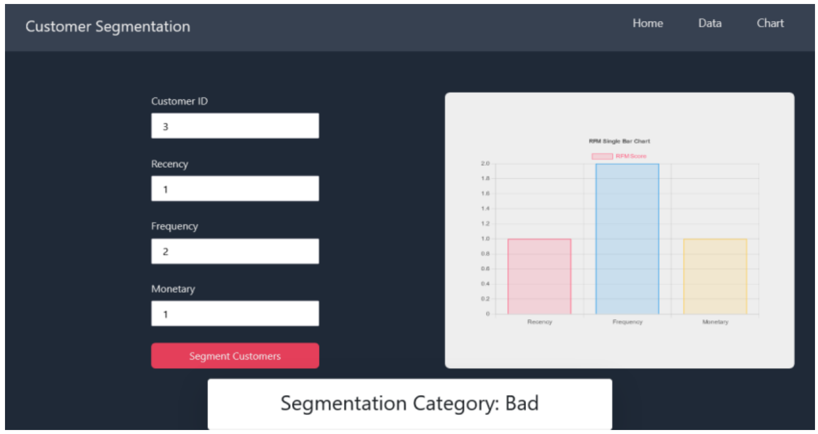

Figure 11. Web Interface Illustration Bad Category Generated from The RFM Scores

Customers with cumulative RFM scores between six (6) and eight (8) were classified as belonging to the bronze level. According to the data presented in Figure 12, the user's cumulative RFM score is six, indicating their classification inside the bronze category.

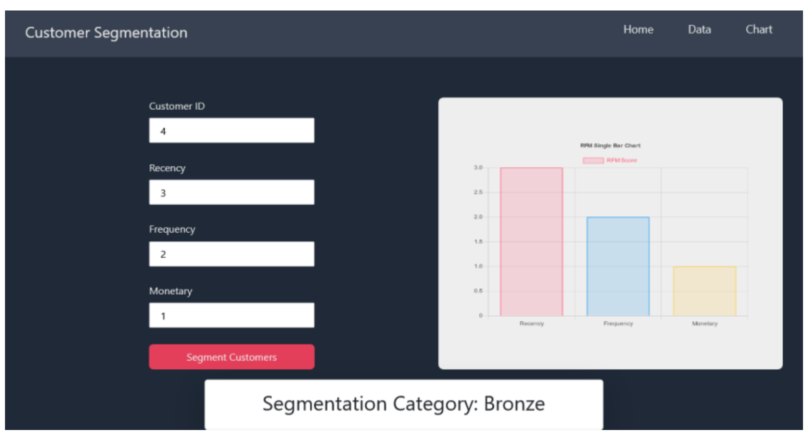

Figure 12. Web Interface Illustrating the Bronze Category Generated from the RFM Scores

In addition, consumers with cumulative RFM score between nine (9) and eleven (11) were classified as belonging to the silver category. According to the data presented in Figure 13, the user possesses a cumulative RFM score of nine, indicating their classification inside the silver category.

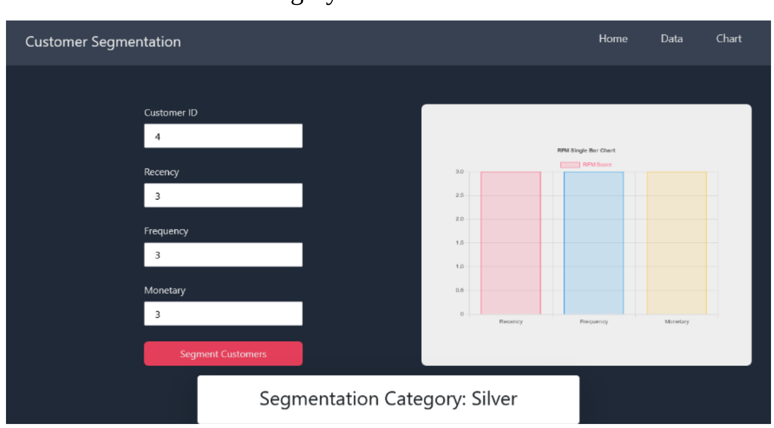

Figure 13. Web Interface Illustrating the Silver Category Generated from the RFM Score

Furthermore, clients whose RFM score ranges from twelve (12) to thirteen (13) are classified as belonging to the gold category. According to the data presented in Figure 14, the user's RFM score is twelve (12), indicating that they fall into the Gold category.

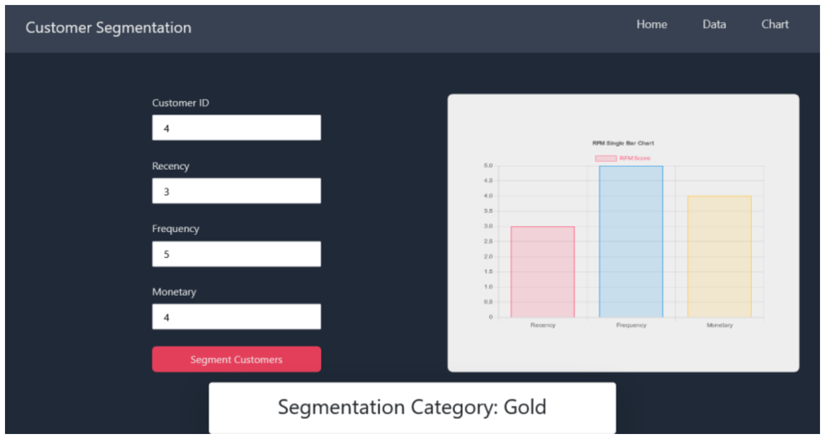

Figure 14. Web Interface Illustrating the Gold Category Generated from the RFM Score

Similarly, customers with RFM score between fourteen (14) and fifteen (15) were classified as belonging to the platinum category. According to the data presented in Figure 15, the user's RFM score totals fifteen (15), indicating that they fall into the platinum category.

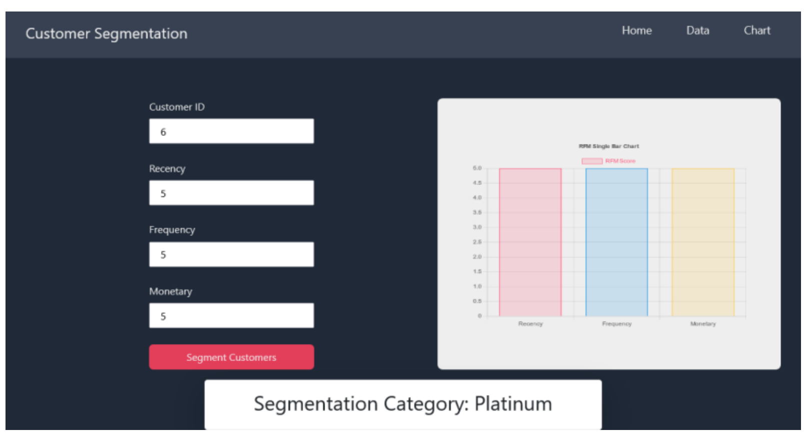

Figure 15. Web Interface Illustrating the Platinum Category Generated from the RFM Score

Based on the depicted web images, it is evident that organizations may conveniently assess the performance of their clients and effectively establish appropriate strategies for engaging with them by leveraging the RFM characteristics.

Conclusions and Suggestions

Conclusions

Understanding customer behavior patterns over time allows for timely adjustments to marketing efforts, ensuring their continued relevance and effectiveness. This study emphasizes the importance of integrating RFM analysis with K-means clustering in customer segmentation to enhance marketing strategies with a more data-centric and efficient approach. By using transactional data to identify discrete client segments, organizations can enhance their understanding of consumer behavior, optimize marketing efforts, and ultimately foster lasting customer loyalty and profitability in a competitive market. The web platform can be expanded to facilitate in-depth analysis of sales, marketing promotions, and the generation of downloadable charts depicting sales data on an hourly, monthly, or yearly basis, as per user specifications. Despite the superior performance of the algorithm in experimental results, there is room for improvement. The current approach requires the input of the number of clusters, and it would be advantageous if the algorithm could autonomously determine the appropriate number. While K-Means has shown promising outcomes, further research could explore alternative metaheuristics or adaptations to enhance computing efficiency and performance. Additionally, future studies could expand the scope to validate findings in datasets with a high number of dimensions and examine the statistical significance of real loss to assess clustering performance.

Suggestions

Tips:
1. State your conclusions clearly and concisely. Be brief and stick to the point;
2. Explain why your study is important to the reader. You should instill in the reader a sense of relevance;
3. Prove to the reader, and the scientific community, that your findings are worthy of note. This means setting your paper in the context of previous work. The implications of your findings should be discussed within a realistic framework, and;
4. Strive for accuracy and originality in your conclusion. If your hypothesis is similar to previous papers, you must establish why your study and your results are original.

References

Agranovskii, A. V, & Silukov, A. P. (2021). Comparative Analysis of Results of Modern Classification Algorithms Usage for Determining the Type of Physical Activity Based on Integrated Sensors Data. *Wave Electronics and Its Application in Information and Telecommunication Systems (WECONF)*, 1–5. https://doi.org/https://doi.org/10.1109/WECONF51603.2021.9470661

Ali, R., Farhat, T., Abdullah, S., & Akram, S. (2023). Customer Segmentation for Targeted Marketing: A Comparative Analysis of Clustering Techniques. *2023 25th International Multitopic Conference (INMIC)*, 1–6. https://doi.org/https://doi.org/10.1109/INMIC60434.2023.10466067

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/https://doi.org/10.1016/j.jksuci.2019.12.011

Arya, R. K., Rajput, A., Singh, S., Mohan, K. M., Choudhry, S., Babu, Y. M., Kumar, P., & Dong, J. (2023). Open-Source Visualization Tools for Antenna Engineers. *IEEE Microwaves, Antennas, and Propagation Conference (MAPCON)*, 1–6. https://doi.org/https://doi.org/10.1109/MAPCON58678.2023.10464133

Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail. *Jurnal Teknologi Dan Informasi*, *13*(2), 122–137. https://doi.org/https://doi.org/10.34010/jati.v13i2.10090

Belorkar, A., Guntuku, S. C., Hora, S., & Kumar, A. (2020). Interactive Data Visualization with Python: Present your data as an effective and compelling story. *Packt Publishing*.

Buzzacchi, L., Milone, F. L., Paolucci, E., & Raguseo, E. (2023). How to react to a shock? Effects of Airbnb hosts' choices and market segmentation at the time of Covid-19. *Information & Management*, *60*(7), 103857. https://doi.org/https://doi.org/10.1016/j.im.2023.103857

Chen, A. H.-L., & Gunawan, S. (2023). Enhancing Retail Transactions: A Data-Driven Recommendation Using Modified RFM Analysis and Association Rules Mining. *Applied Sciences*, *13*(18), 10057. https://doi.org/https://doi.org/10.3390/app131810057

Chen, H., Zhang, L., Chu, X., & Yan, B. (2019). Smartphone customer segmentation based on the usage pattern. *Advanced Engineering Informatics*, *42*, 101000. https://doi.org/https://doi.org/10.1016/j.aei.2019.101000

Chronopoulos, D. K., Lukas, M., & Wilson, J. O. S. (2020). Consumer spending responses to the COVID-19 pandemic: An assessment of Great Britain. *Available at SSRN 3586723*. https://doi.org/https://doi.org/10.2139/ssrn.3586723

Dzulhaq, M. I., Sari, K. W., Ramdhan, S., Tullah, R., & others. (2019). Customer segmentation based on RFM value using K-means algorithm. *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 1–7. https://doi.org/doi: 10.1109/ICIC47613.2019.8985726

Fadilah, A. T. (2019). Implementasi Algoritma K-Means Clustering Untuk Targeting Ads Studi Kasus: Data Pelanggan Asuransi Mobil Ibm Watson Analyti. *Fakultas Sains Dan Teknologi UIN Syarif Hidayatullah Jakarta*.

Grunig, J. E., & Repper, F. C. (2013). Strategic management, publics, and issues. *Excellence in Public Relations and Communication Management*, 117–157. https://doi.org/https://doi.org/10.4324/9780203812303

Harahap, M., Lubis, Y., & Situmorang, Z. (2021). Analisis Pemasaran Bisnis dengan Data Science: Segmentasi Kepribadian Pelanggan berdasarkan Algoritma K-Means Clustering. *Data Sciences Indonesia (DSI)*, *1*(2), 76–88. https://doi.org/https://doi.org/10.47709/dsi.v1i2.1348

Hartoyo, H., Manalu, E., Sumarwan, U., & Nurhayati, P. (2023). Driving success: A segmentation of customer admiration in automotive industry. *Journal of Open Innovation: Technology, Market, and Complexity*, *9*(2), 100031. https://doi.org/https://doi.org/10.1016/j.joitmc.2023.100031

Hu, X., Liu, A., Li, X., Dai, Y., & Nakao, M. (2023). Explainable AI for customer segmentation in product development. *CIRP Annals*, *72*(1), 89–92. https://doi.org/https://doi.org/10.1016/j.cirp.2023.03.004

Hu, X., Shi, Z., Yang, Y., & Chen, L. (2020). Classification method of internet catering customer based on improved RFM model and cluster analysis. *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 28–31. https://doi.org/https://doi.org/10.1109/ICCCBDA49378.2020.9095607

Jaiswal, D., Mohan, A., & Deshmukh, A. K. (2023). Cash rich to cashless market: Segmentation and profiling of Fintech-led-Mobile payment users. *Technological Forecasting and Social Change*, *193*, 122627. https://doi.org/https://doi.org/10.1016/j.techfore.2023.122627

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/https://doi.org/10.1126/science.aaa8415

Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, *70*, 102641. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2023.102641

Kanoujia, S., & Karuppanan, P. (2024). Depression Detection in Speech Using ML and DL Algorithm. *IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, *2*, 1–5. https://doi.org/https://doi.org/10.1109/IATMSI60426.2024.10503510

Kuo, R. J., Alfareza, M. N., & Nguyen, T. P. Q. (2023). Genetic based density peak possibilistic fuzzy c-means algorithms to cluster analysis-a case study on customer segmentation. *Engineering Science and Technology, an International Journal*, *47*, 101525. https://doi.org/https://doi.org/10.1016/j.jestch.2023.101525

Kursan Milaković, I. (2021). Purchase experience during the COVID-19 pandemic and social cognitive theory: The relevance of consumer vulnerability, resilience, and adaptability for purchase satisfaction and repurchase. *International Journal of Consumer Studies*, *45*(6), 1425–1442. https://doi.org/https://doi.org/10.1111/ijcs.12672

Lewaaelhamd, I. (2024). Customer segmentation using machine learning model: an application of RFM analysis. *Journal of Data Science and Intelligent Systems*, *2*(1). https://doi.org/https://doi.org/10.47852/bonviewJDSIS32021293

Lezhnina, O., & Kismihók, G. (2022). Latent class cluster analysis: Selecting the number of clusters. *MethodsX*, *9*, 101747. https://doi.org/https://doi.org/10.1016/j.mex.2022.101747

Ma, J., Nault, B. R., & Tu, Y. P. (2023). Customer segmentation, pricing, and lead time decisions: A stochastic-user-equilibrium perspective. *International Journal of Production Economics*, *264*, 108985. https://doi.org/https://doi.org/10.1016/j.ijpe.2023.108985

McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. *O'Reilly Media, Inc*.

Mehrabioun, M., & Mahdizadeh, B. M. (2021). Customer retention management: A complementary use of data mining and soft systems methodology. *Human Systems Management*, *40*(6), 897–916. https://doi.org/https://doi.org/10.3233/HSM-201075

Mouton, J. P., Ferreira, M., & Helberg, A. S. J. (2020). A comparison of clustering algorithms for automatic modulation classification. *Expert Systems with Applications*, *151*, 113317. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113317

MR, A. P. A. S., Sodik, J., Saripudin, M., Rejeki, S. M., & Firdaus, U. (2024). Kajian untuk Segmentasi Customer Bank dengan Algoritma K-Means. *Karimah Tauhid*, *3*(3), 3899–3906.

Muchardie, B. G., Gunawan, A., & Aditya, B. (2019). E-commerce market segmentation based on the antecedents of customer satisfaction and customer retention. *2019 International Conference on Information Management and Technology (ICIMTech)*, *1*, 103–108. https://doi.org/doi: 10.1109/ICIMTech.2019.8843792

Myburg, M. E. (2023). *Using recency, frequency and monetary variables to predict customer lifetime value with*

XGBoost.

Nabella, R. O., & others. (2022). Analisis Segmentasi Konsumen Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Dan Pengetahuan Data Lainnya Sebagai Rekomendasi Strategi Pemasaran (Studi Kasus: Yayasan Xyz). *Fakultas Teknik*.

Nandapala, E. Y. L., & Jayasena, K. P. N. (2020). The practical approach in Customers segmentation by using the K-Means Algorithm. *IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 344–349. https://doi.org/https://doi.org/10.1109/ICIIS51140.2020.9342639

Onumanyi, A. J., Molokomme, D. N., Isaac, S. J., & Abu-Mahfouz, A. M. (2022). AutoElbow: An automatic elbow detection method for estimating the number of clusters in a dataset. *Applied Sciences*, *12*(15), 7515. https://doi.org/https://doi.org/10.3390/app12157515

Putra, R. F., Zebua, R. S. Y., Budiman, B., Rahayu, P. W., Bangsa, M. T. A., Zulfadhilah, M., Choirina, P., Wahyudi, F., & Andiyan, A. (2023). Data Mining: Algoritma dan Penerapannya. *PT. Sonpedia Publishing Indonesia*.

Ramadhan, H. M. T., Pertiwi, A., Defanka, G. A., & Sari, A. P. (2023). Clustering RFM (Recency, Frequency, Monetary) Publisher Gim Menggunakan Algoritma K-Means. *Prosiding Seminar Nasional Informatika Bela Negara*, *3*, 8–18.

Ramkumar, P., Kalamani, P., Valarmathi, C., & Devi, M. S. (2021). An effective analysis of data clustering using distance-based K-means Algorithm. *Journal of Physics: Conference Series*, *1979*(1), 12015. https://doi.org/https://doi.org/10.1088/1742-6596/1979/1/012015

Rojlertjanya, P. (2019). Customer segmentation based on the RFM analysis model using k-means clustering technique: a case of IT solution and service provider in Thailand. *Bangkok University*.

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, *237*, 121449. https://doi.org/https://doi.org/10.1016/j.eswa.2023.121449

Russell, A. D., Chiu, C.-Y., & Korde, T. (2009). Visual representation of construction management data. *Automation in Construction*, *18*(8), 1045–1062. https://doi.org/https://doi.org/10.1016/j.autcon.2009.05.006

Sarikaya, A., Gleicher, M., & Szafir, D. A. (2018). Design factors for summary visualization in visual analytics. *Computer Graphics Forum*, *37*(3), 145–156. https://doi.org/https://doi.org/10.1111/cgf.13408

Silalahi, H. D., Hasibuan, D., & Sinambela, M. (2023). Segmentasi Pelanggan Online Dengan Metode K-Means Clustering. *METHOTIKA: Jurnal Ilmiah Teknik Informatika*, *3*(1), 13–21.

Steiss, A. W. (2019). Strategic management for public and nonprofit organizations. *Routledge*. https://doi.org/https://doi.org/10.4324/9781482275865

Susilo, D. D., Hilabi, S. S., Priyatna, B., & Novalia, E. (2024). Implementasi Data Mining dalam Pengelompokan Data Pembelian Menggunakan Algoritma K-Means Pada PT. Otomotif 1. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, *13*(1).

Talaat, F. M., Aljadani, A., Alharthi, B., Farsi, M. A., Badawy, M., & Elhosseini, M. (2023). A Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing. *Mathematics*, *11*(18), 3930. https://doi.org/https://doi.org/10.3390/math11183930

Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. *Information Processing & Management*, *59*(6), 103085. https://doi.org/https://doi.org/10.1016/j.ipm.2022.103085

Wicaksono, Y. (2019). Segmentasi Pelanggan Bisnis Dengan Multi Kriteria Menggunakan K-Means. *Indonesian Journal of Business Intelligence (IJUBI)*, *1*(2), 45–53. https://doi.org/https://doi.org/10.21927/ijubi.v1i2.872

Yuan, Y., Dehghanpour, K., Bu, F., & Wang, Z. (2020). A data-driven customer segmentation strategy based on contribution to system peak demand. *IEEE Transactions on Power Systems*, *35*(5), 4026–4035. https://doi.org/https://doi.org/10.1109/TPWRS.2020.2979943