IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Customer Segmentation using Data Science Techniques

Shubhangi Balaji[1] and Ms. Snigdha Shukla[2]
[1]Student, Master of Computer Applications
[2]Assistant Professor, Master of Computer Application
Sadhu Vaswani Institute of Management Studies for Girls, Pune
Savitribai Phule Pune University, Pune, Maharashtra India

Abstract: *Many businesses send the same marketing message to every customer, which wastes budget on customers who are unlikely to respond and under-serves the most valuable ones. This paper applies data science techniques to segment the customers of an online retailer and to link each segment to a practical marketing strategy. Using a sample of 150,000 transaction records from the public Online Retail dataset, each customer's Recency, Frequency and Monetary (RFM) values are calculated, log-transformed and standardised, and customers are grouped with K-Means clustering. The Elbow method, supported by the silhouette score, indicates four clusters. After cleaning, 2,225 customers were segmented and labelled in business terms as Loyal, Big Spenders, New and At Risk. Loyal customers form only 12.2% of the customer base but generate 51.6% of revenue, while At Risk customers make up 42.0% of customers but only 9.9% of revenue. A targeted strategy is attached to each segment, such as rewards for Loyal customers and win-back discounts for At Risk customers, and the segments are presented in a dashboard of charts. The study shows that RFM-based K-Means segmentation, when interpreted in business terms, can turn transaction data into actionable marketing decisions.*

Keywords: Customer segmentation, RFM analysis, K-Means clustering, Elbow method, data science, targeted marketing.

I. INTRODUCTION

Online retail generates a detailed record of every purchase: who bought, what they bought, when and for how much. This data allows a business to understand its customers far better than traditional mass marketing does. Yet many businesses still send the same promotions to all customers. A discount sent to a customer who would have bought anyway reduces profit, while a generic message sent to a customer who is about to leave may not be enough to bring them back.

Customer segmentation addresses this by dividing customers into groups with similar behaviour so that each group can be treated differently. The RFM model is one of the most widely used approaches because it summarises customer value with three simple measures: how recently a customer purchased (Recency), how often they purchase (Frequency) and how much they spend (Monetary). Combined with clustering algorithms such as K-Means (MacQueen, 1967), RFM allows segments to be discovered from data rather than defined by guesswork (Chen et al., 2012).

A. Research Problem

Businesses send the same marketing to all customers, which wastes money and misses opportunities to retain valuable customers and win back inactive ones.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

B. Research Gap

Most studies apply K-Means to RFM data and report technical measures of cluster quality, but they do not explain the resulting segments in business terms or suggest a specific marketing strategy for each segment. Without this final step, segmentation results are difficult for marketing teams to act on.

C. Objectives

The objectives of this study are: (i) to calculate RFM values for each customer from transaction data; (ii) to group customers using K-Means clustering, selecting the number of clusters with the Elbow method; (iii) to label each segment in business terms such as Loyal, Big Spenders, New and At Risk; (iv) to recommend a marketing strategy for each segment; and (v) to present the segments in a dashboard of charts.

II. LITERATURE REVIEW

Recent research (2024–2026) continues to combine RFM analysis with clustering, but differs in the algorithms used, how the number of clusters is chosen, and whether results are turned into business actions. Rungruang et al. (2024) proposed a hierarchical RFM segmentation approach using Formal Concept Analysis to produce structured and interpretable customer groups. Wong et al. (2024) compared K-Means and hierarchical clustering on RFM data from e-commerce transactions after log and Yeo-Johnson transformation, and found that hierarchical clustering performed better, with a silhouette score of 0.47.

Vo et al. (2025) applied RFM-based K-Means to the Online Retail II dataset and focused on choosing the number of clusters automatically with the Elbow method. Syahra et al. (2025) segmented 369 customers of a local retailer with RFM and K-Means, used the Elbow method to select four clusters, and named them Loyal Customers, Potential Loyalists, At-Risk Customers and One-Time Buyers to support customer relationship management. Zaghloul et al. (2025) combined RFM, K-Means and a deep neural network to predict customer churn in online retail.

Ozcan (2026) extended the RFM model with additional behavioural variables and compared K-Means, K-Medoids and fuzzy C-Means using a combined cluster-validity index. Handoko et al. (2026) identified five RFM segments, including VIP, at-risk and new customers, with a silhouette score of 0.4464, and delivered promotion recommendations through a web platform for a small store. Table 1 summarises these studies and the gap that this paper addresses.

Table 1. Summary of recent related studies (2024–2026)

Study	Method	Key Finding	Gap Addressed in This Paper
Rungruang et al. (2024)	Hierarchical RFM segmentation using Formal Concept Analysis	Structured, interpretable RFM segments	No marketing strategy or dashboard attached to segments
Wong et al. (2024)	RFM with K-Means and hierarchical clustering; log and Yeo-Johnson transforms	Hierarchical clustering performed better (silhouette 0.47)	Technical comparison only; segments not linked to marketing actions
Vo et al. (2025)	RFM-based K-Means on Online Retail II; automatic k with Elbow method	Automated choice of the number of clusters	Focus on model building rather than business interpretation
Syahra et al. (2025)	RFM, Min-Max scaling and K-Means; Elbow method (k = 4)	Four CRM segments: Loyal, Potential Loyalists, At-Risk, One-Time	Small local dataset (369 customers); no revenue analysis or dashboard
Zaghloul et al. (2025)	Hybrid RFM, K-Means and deep neural network	Churn prediction for online retail customers	Predicts churn but does not define a strategy for every segment

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Study	Method	Key Finding	Gap Addressed in This Paper
Ozcan (2026)	Extended RFM; K-Means, K-Medoids and fuzzy C-Means	Richer RFM variables and combined validity index	More complex model; this paper keeps a simple, explainable RFM approach
Handoko et al. (2026)	RFM and K-Means on a web platform for a small store	Five segments (silhouette 0.4464) with promotion recommendations	Single small store; this paper uses a larger public dataset and revenue shares

III. METHODOLOGY

The study follows an applied, quantitative design. Figure 1 shows the full workflow, from raw transactions to segments, strategies and a dashboard. The analysis was carried out in Python using pandas and scikit-learn (Pedregosa et al., 2011).

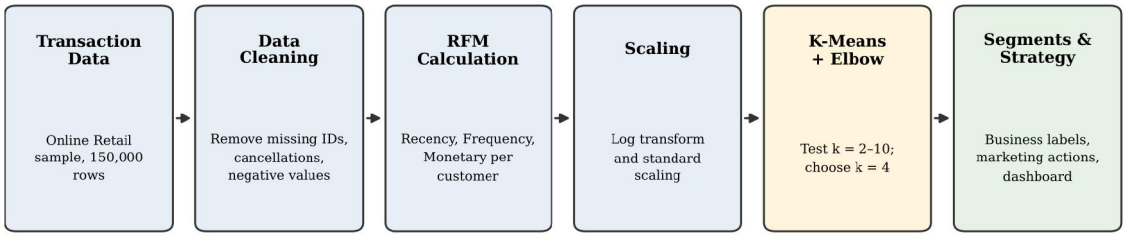


Fig. 1. Methodology workflow for RFM-based customer segmentation.

A. Dataset and Data Preparation

The Online Retail dataset (Chen, 2015), first analysed with RFM by Chen et al. (2012), contains transactions of a UK-based online retailer that sells gift items. As the full file is large, this study used a sample of the first 150,000 transaction lines, covering 1 December 2010 to 8 April 2011. Records without a customer ID (46,477 rows), cancelled invoices (2,597 rows), rows with zero or negative quantity or price (10 rows) and duplicate records were removed, leaving 99,591 valid transaction lines from 2,225 customers in 30 countries (Table 2).

Table 2. Dataset sample and data preparation

Item	Value
Sample size (transaction lines)	150,000
Period covered	1 Dec 2010 – 8 Apr 2011
Rows without customer ID (removed)	46,477
Cancelled invoice rows (removed)	2,597
Rows with zero or negative quantity or price (removed)	10
Valid transaction lines after cleaning	99,591
Unique invoices	5,010
Unique customers	2,225
Countries	30
Total revenue of sample (£)	2,304,280.08

B. RFM Calculation

A snapshot date of 2011-04-09, one day after the last transaction, was used. For each customer, Recency is the number of days between the last purchase and the snapshot date; Frequency is the number of distinct invoices; and Monetary is the total amount spent, calculated as quantity multiplied by unit price and summed across all purchases.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

C. Scaling and K-Means Clustering

RFM values are highly skewed: most customers buy once and spend modestly, while a few buy very often and spend heavily. Each variable was therefore log-transformed and then standardised to zero mean and unit variance, so that no single variable dominates the distance calculation. K-Means then assigns each customer to the nearest of k cluster centres and iteratively updates the centres to minimise the within-cluster sum of squares (MacQueen, 1967).

D. Choosing the Number of Clusters

K-Means was run for k = 2 to 10. The Elbow method plots the within-cluster sum of squares against k and selects the point where adding more clusters gives only small improvements. The silhouette score, which measures how well each customer fits its own cluster compared with the nearest other cluster, was used as a supporting check (Rousseeuw, 1987).

E. Segment Labelling and Strategy

Each cluster was described by its average Recency, Frequency and Monetary values, its number of customers and its share of total revenue. Clusters were then given business labels and matched with a marketing strategy, and the results were combined into a dashboard of charts.

IV. RESULTS AND DISCUSSION

A. Customer Behaviour Overview

Customer behaviour in the sample is highly uneven. The median customer last purchased 38 days before the snapshot date, placed 1 order and spent £408.06. However, the average spend was £1,035.63 and the highest was £81,612.62, and one customer placed 58 orders. This skew confirms the need for log transformation before clustering.

B. Optimal Number of Clusters

Figure 2(a) shows the Elbow curve. The within-cluster sum of squares falls sharply from 3,808 at k = 2 to 2,906 at k = 3 and 2,276 at k = 4, after which the reductions become smaller (1,960 at k = 5 and 1,724 at k = 6). The elbow therefore occurs at k = 4. The silhouette score in Figure 2(b) is highest at k = 2 (0.423), but two groups are too coarse for targeted marketing; k = 4 keeps a score of 0.341, close to k = 3 (0.350) and clearly above k = 5 (0.285). Four clusters were therefore selected as the best balance between statistical quality and business usefulness.

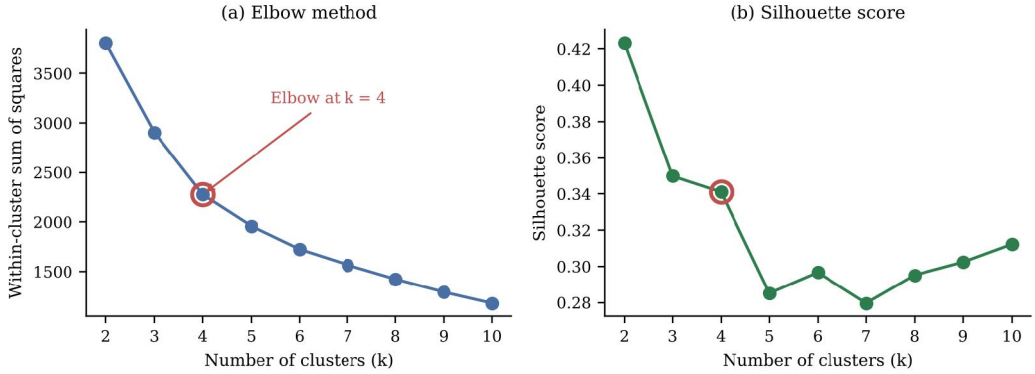


Fig. 2. Selecting the number of clusters: (a) Elbow method and (b) silhouette score for k = 2 to 10.

C. Customer Segments

Table 3 profiles the four clusters and the business label given to each. Figure 3 presents the segments as a dashboard.

Table 3. Profile of customer segments (k = 4)

Segment	Customers	% Customers	Avg. Recency (days)	Avg. Frequency (orders)	Avg. Monetary (£)	% Revenue
Loyal	271	12.2%	13.5	7.55	4,389.96	51.6%
Big Spenders	572	25.7%	48.4	2.22	1,171.44	29.1%

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Segment	Customers	% Customers	Avg. Recency (days)	Avg. Frequency (orders)	Avg. Monetary (£)	% Revenue
New	448	20.1%	9.7	1.57	485.19	9.4%
At Risk	934	42.0%	76.2	1.06	243.22	9.9%

Loyal. 271 customers (12.2%) purchased recently (on average 13.5 days ago), ordered most often (7.55 orders) and spent the most (£4,389.96 on average). This small group generates 51.6% of revenue.

Big Spenders. 572 customers (25.7%) spent well (£1,171.44 on average across 2.22 orders) and contribute 29.1% of revenue, but their last purchase was on average 48.4 days ago, so they need attention to keep them active.

New. 448 customers (20.1%) bought very recently (9.7 days on average) but only 1.57 times, spending £485.19 on average. They are recent, low-frequency buyers who could become regular customers.

At Risk. The largest group, 934 customers (42.0%), bought about once (1.06 orders), spent little (£243.22) and have not returned for 76.2 days on average. They contribute only 9.9% of revenue.

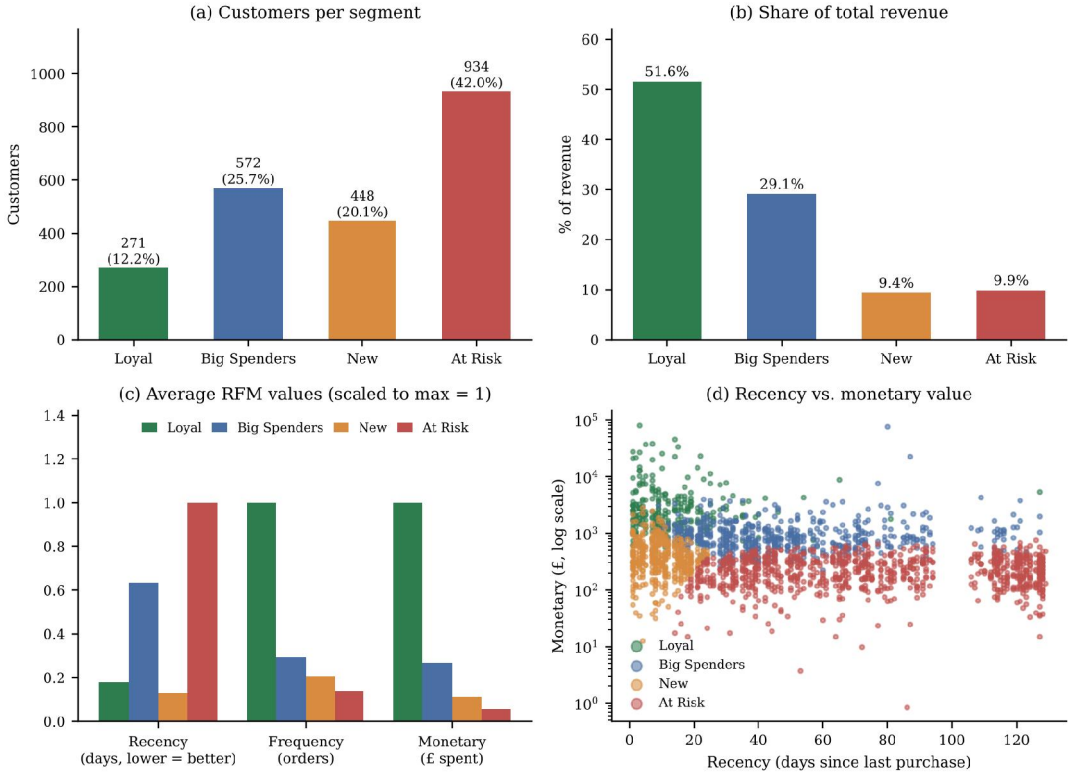


Fig. 3. Customer segmentation dashboard: (a) customers per segment, (b) share of revenue, (c) average RFM values and (d) recency versus monetary value for each customer.

D. Marketing Strategy for Each Segment

Table 4 links each segment to a targeted marketing strategy, so that budget is spent where it has the greatest effect.

Table 4. Recommended marketing strategy for each segment

Segment	Business Meaning	Recommended Strategy
Loyal	Recent, frequent, highest-spending customers	Loyalty rewards and points, early access to new products, VIP service; avoid unnecessary discounts

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40569

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Segment	Business Meaning	Recommended Strategy
Big Spenders	High-value customers whose purchases are slowing	Personalised recommendations, premium bundles, reminder emails before they lapse
New	Recent buyers with few purchases	Welcome series, second-purchase coupon, product guides to build the buying habit
At Risk	One-time, low-value, long-inactive customers	Win-back discount or free-shipping offer, re-engagement email; low-cost channels only

E. Discussion

The results show a strong concentration of value. Loyal customers make up 12.2% of the base but produce 51.6% of revenue, and together with Big Spenders they account for 37.9% of customers but 80.7% of revenue. In contrast, At Risk customers are 42.0% of the base but bring in only 9.9% of revenue. Sending the same promotion to everyone would therefore spend most of the budget on customers who generate the least value, while offering no special recognition to the customers who matter most.

Unlike studies that stop at cluster-quality measures (Wong et al., 2024), this paper converts each cluster into a named segment with a clear action. The silhouette score of 0.341 indicates that segments overlap to some degree, which is common for real customer data; the business labels are therefore best treated as practical guides for marketing rather than exact boundaries.

V. CONCLUSION

This study applied RFM analysis and K-Means clustering to a sample of the Online Retail dataset to replace one-size-fits-all marketing with targeted strategies. The Elbow method identified four clusters, which were labelled as Loyal, Big Spenders, New and At Risk. The segments differ sharply in value: a small Loyal group generates more than half of revenue, while the largest group is inactive and low-value. Linking each segment to a specific strategy, such as rewards for Loyal customers and win-back offers for At Risk customers, makes the segmentation directly useful for marketing decisions, and the dashboard presents the results in a form that managers can understand.

The study has limitations. It used a sample covering about four months rather than the full dataset, so Recency values are limited to this period, and RFM does not capture product preferences or customer demographics. Future work can apply the method to the full dataset, compare K-Means with other clustering methods, add variables such as product categories, and build an interactive dashboard that updates segments automatically as new transactions arrive.

REFERENCES

1. Chen, D. (2015). *Online retail* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33
2. Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17
3. Handoko, K. F., Nugraha, N., & Sujjada, A. (2026). Customer segmentation-based promotion recommendation using RFM and K-means clustering on a web platform. *bit-Tech, 8*(3), 3661–3672. https://doi.org/10.32877/bt.v8i3.3677
4. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
5. Ozcan, T. (2026). Customer segmentation using an extended RFM model and clustering algorithms in e-commerce. *Journal of Theoretical and Applied Electronic Commerce Research, 21*(5), Article 142. https://doi.org/10.3390/jtaer21050142

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
7. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
8. 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*, Article 121449. https://doi.org/10.1016/j.eswa.2023.121449
9. Syahra, Y., Fadlil, A., & Yuliansyah, H. (2025). Customer segmentation using RFM and K-means clustering to support CRM in retail industry. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 9*(3), 1120–1131. https://doi.org/10.33395/sinkron.v9i3.14907
10. Vo, B. G., Van, H. D. S., Van, N. D., & Huynh, H. D. (2025). Customer segmentation: Automatic k-optimization and RFM-based K-means clustering. In *Proceedings of the 2025 10th International Conference on Intelligent Information Technology* (pp. 173–178). Association for Computing Machinery. https://doi.org/10.1145/3731763.3731805
11. Wong, C.-G., Tong, G.-K., & Haw, S.-C. (2024). Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. *Journal of Telecommunications and the Digital Economy, 12*(3), 97–125. https://doi.org/10.18080/jtde.v12n3.978
12. Zaghloul, M., Barakat, S., & Rezk, A. (2025). Enhancing customer retention in Online Retail through churn prediction: A hybrid RFM, K-means, and deep neural network approach. *Expert Systems with Applications, 290*, Article 128465. https://doi.org/10.1016/j.eswa.2025.128465