Unravelling Consumer Segmentation through K-Means Clustering

Introduction

Consumer-centric marketing leverages customer data to deliver personalised, targeted messaging, enhancing the effectiveness of a company's marketing strategies. By detecting similarities and differences among customers, predicting their behaviours, proposing tailored communication to individual customer needs, companies can foster stronger, more valuable relationships. This can, in turn, drive increased satisfaction, loyalty, and advocacy, and thus enhances profitability.

Drawing on the sample of an online retailer, this report uses customer purchase data from 1st December 2010 to 9th December 2011 to enable customer-centric business intelligence. The company in focus is a Europe-based non-store retailer that primarily sells unique gifts for all occasions. A significant portion of its customer base consists of wholesalers.

The key objective of this analysis is to provide the business with deeper insights into its customer base, enabling it to execute more effective customer-centric marketing strategies. Using the Recency, Frequency, and Monetary (RFM) model, customers have been segmented into distinct groups based on their purchasing behaviours. The segmentation was achieved through the k-means clustering algorithm, which identified meaningful patterns across the different consumer segments. The report also outlines the defining characteristics of each customer segment and offers actionable recommendations for the company to enhance its consumer-centric marketing efforts. These insights will help the business tailor its offerings and communications more precisely, ensuring it maximises customer engagement and cultivates long-term relationship with its customer base.

Literature Review

The RFM model has long been recognised as a fundamental tool for customer value analysis and segmentation, offering businesses a systematic way to assess customer engagement and loyalty (Chen et al., 2012). This model evaluates customers on three key dimensions: **Recency (R)**, which refers to the time elapsed since the customer's last purchase; **Frequency (F)**, the number of transactions within a given period; **Monetary (M)**, the total spend within that period. The underlying premise is that customers who purchase more recently, more frequently, and at higher monetary values tend to exhibit greater loyalty and are more likely to make future purchases. As a result, the RFM model is an essential tool in identifying high-value customers and guiding marketing strategies such as retention campaigns, personalised offers, and loyalty programs.

Beyond the RFM model, Yang and Balaji (2003) introduced a pattern-based approach that extends beyond the three traditional dimensions. Their methodology incorporates both temporal and transactional features, offering a more nuanced understanding of how customer purchasing patterns evolve over time. This technique not only considers when and how often customers make purchases, but also factors in variables such as periodicity, shifts in buying habits, and variations in purchase quantities. By analysing these patterns, businesses can gain deeper insights into the likelihood of future purchases, allowing for more proactive customer engagement strategies.

In our analysis, we have integrated both the classical RFM framework and elements of Yang and Balaji's pattern-based approach to develop a hybrid model. Alongside traditional RFM metrics, we include additional transactional variables such as purchase quantities, time intervals between purchases, and product diversity to create a more granular segmentation of customer behaviours. This approach aligns with recent trends in customer segmentation literature, which increasingly emphasise multi-dimensional models that provide a more holistic view of customer engagement. By moving beyond traditional frequency and monetary metrics, we aim to capture the complexity of customer interactions, offering a more predictive and tailored approach to segmentation that can guide effective customer relationship management (CRM) strategies.

Data Wrangling

DATA OVERVIEW

For this report, we will be working with "Assignment1_Dataset.csv", which contains 433,527 transactional data across 8 variables. The variables include order ID, product ID, product description, quantity, order data, unit price, customer ID, and country. This section outlines the data exploration and preparation conducted to prepare the dataset for analysis.

DATA PREPARATION

To conduct the required RFM model-based clustering analysis, the original dataset needs to be pre-processed. The key data wrangling steps, carried out using R, are summarised in Table 1**Error! Reference source not found.** below.

S/N	Observations	Action

1.	"Orderno", "prodcode", "custid", and "date" columns are in character data type.	Corrected "date" to datetime format and the rest to factor data type.
		`txn$Orderno <- as.factor(txn$Orderno)`
		`txn$prodcode <- as.factor(txn$prodcode)`
		`txn$custid <- as.factor(txn$custid)`
		`txn$date <- lubridate::dmy_hm(txn$date)`

2.	There are 1,170 and 108,237 missing values in the "proddesc" and "custid" columns	Removed 108,237 rows with missing customer ID since we will be performing customer segmentation and would be hard to interpret if we cannot identify these customers. Amongst these removed rows with missing customer ID were also rows with missing product description. `txn <- txn %>% filter(!if_all(c(custid), is.na))`
3.	There are 6,629 duplicates in the dataset.	Removed 6,629 rows since it is unlikely that there would be multiple transactions for the same product at the same time by the same customer. For the same items in an order, products would be in the same line item with increased quantities.

<table>
<tr><td></td><td></td><td><pre><code>txn <- txn %>%
  group_by_all() %>%
  filter(n()==1) %>%
  ungroup()</code></pre></td></tr>
<tr><td>4.</td><td>There are 8 customer IDs who have transactions for 2 different countries.</td><td>No further action as wholesalers could operate in various countries.</td></tr>
<tr><td>5.</td><td>Invoice IDs with prefix 'C' appears to be cancelled transactions as the quantities and unit prices are negative.</td><td>Filtered out order ID with the prefix ‘C’ from the dataset.<pre><code>txn <- txn %>%
  filter(!str_starts(Orderno, "C") & !str_starts(Orderno, "c"))</code></pre></td></tr>
<tr><td>6.</td><td>Transactional data included postage fees and discounts in addition to product purchases.</td><td>Postage fees were excluded from the list of transactions to ensure the analysis focuses solely on product-related data. Discount details were retained to ensure the analysis reflects the actual total values after discounts.<pre><code>txn <- txn %>%
  filter(!prodcode == "POST")</code></pre></td></tr>
<tr><td>8.</td><td>Dataset is in transactional level.</td><td>Grouped transactional data into customer level using customer ID.<pre><code>txn_bycust <- txn %>%
  mutate(salesamount = qty * uprice) %>%
  group_by(custid) %>%
  summarise(total_orderamt = sum(salesamount),
            latest_orderdate = max(date),
            earliest_orderdate = min(date),
            activedays = n_distinct(date(date)),
            total_numorders = n_distinct(Orderno),
            total_types = n_distinct(prodcode),
            total_qty = sum(qty),
            total_orderamt = round(sum(total_orderamt),0)
) %>%</code></pre></td></tr>
</table>

Table 1: Data Preparation Steps

DERIVING CLUSTERING VARIABLES

The study focuses on customer behavioural segmentation. The selected variables must therefore give insight into patterns of behaviour based on their actions. Inspired by both the RFM model and Yang and Balaji (2003)’s pattern-based approach, Table 2 details the features derived for each customer:

<table>
<tr><th>Category</th><th>Metric</th><th>What it measures</th><th>How it is derived</th></tr>
<tr><td>Time-related</td><td>Customer Lifespan</td><td>How long (in days) has the customer been purchasing from the retailer?</td><td>Difference between latest and earliest order date + 1 day.
1 day is added to the result to ensure that customers with only 1 order (where the latest and earliest order dates are the same) have a customer length of at least 1 day. This prevents the metric from showing a length of 0 days for customers and ensures all customers are accounted for.</td></tr>
</table>

	Recency	*How recent did the customer purchased from the retailer?*	(Days lapsed since last order + 1 day) * -1 • *1 day is added to avoid any 0 values.* • *Multiplying by -1 reverses the values so that more recent transactions have larger values. This way, a higher recency value represents a more recent transaction, making interpretation more intuitive.*
	Active days	*How engaged is the customer?*	Total number of days the customer placed at least 1 order
Quantity-related	Frequency (Total number of orders)	*How many orders has the customer made?*	Total number of orders made during the entire period
	Total distinct types of items	*How many unique products has the customer purchased?*	Number of unique items purchased across all orders
	Total quantity of products	*How many products did the customer purchase?*	Total quantity of products ordered across all transactions
	Average quantity per order	*On average, how many products did the customer purchase per transaction?*	Average quantity of all products per order
	Average quantity per item	*On average, how many of the same product has the customer purchased?*	Average quantity of each distinct item across all orders
Order-related	Monetary (Total order amount)	*How much has the customer spent in total?*	Sum of (Quantity of item * Unit Price) for all orders
	Average order amount	*On average, how much did the customer spend per transaction?*	Total order amount divided by the number of orders
	Average days between orders	*What is the average time interval between orders?*	(Max(Order Date) - Min(Order Date)) / (Number of Orders)

Table 2: Derived Variables

Cluster Analysis

UNIVARIATE DATA ANALYSIS

Our resulting data, after data preparation, consists of 4,324 customers and 11 variables which we will be using for our customer segmentation. Using *SAS Studio > Task > Summary Statistics* in SAS Studio, we observed in Figure 1 that most variables exhibited skewed distributions, and with some outliers present.

Variable	Mean	Std Dev	Minimum	Maximum	Median	N	95th Pctl
cust_lifespan	130.7092969	131.7799458	1.0000000	374.0000000	92.0000000	4324	357.0000000
recency	-92.7858464	100.1908284	-374.0000000	-1.0000000	-51.0000000	4324	-3.0000000
activedays	3.8283996	5.8935870	1.0000000	132.0000000	2.0000000	4324	12.0000000
total_numorders	4.1991212	7.5168029	1.0000000	205.0000000	2.0000000	4324	13.0000000
total_types	51.4220629	74.8551021	1.0000000	1631.00	29.0000000	4324	171.0000000
total_qty	944.5460222	4084.52	1.0000000	162840.00	301.5000000	4324	2789.00
avg_qty_perorder	195.8117484	689.1556287	1.0000000	40498.00	129.0000000	4324	486.0000000
avg_qty_peritem	41.9225254	670.7764389	1.0000000	40498.00	11.0000000	4324	66.0000000
total_orderamt	1614.84	7384.63	0	231262.00	529.5000000	4324	4517.00
avg_orderamt	323.4842738	1340.00	0	84236.00	229.0000000	4324	742.0000000
avg_daysbetweenorders	30.2504625	34.7669341	0	182.0000000	21.0000000	4324	98.0000000

Figure 1: Summary Statistics

Key Insights:

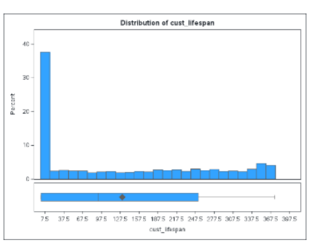
- Most customers are relatively new, with low customer lifespan and high recency, but there is also a smaller segment of customers who have been inactive for a longer period.
- Majority of customers have made <20 orders, with an average quantity of 196 per order.
- Customers purchase an average of 51 different product types.
- There is a large range in total order amount, with a mean of $1.6k per customer and a standard deviation of $7.4k.

While outliers were identified in several variables, they may represent distinct customer segments in this context and should not be removed. For example, the variable "avg_orderamt" is significantly skewed to the right, with 1 customer having an average order amount of $84k, as shown in Table 3. Such extreme values could reflect important high-value customer segments and are critical for a comprehensive analysis.

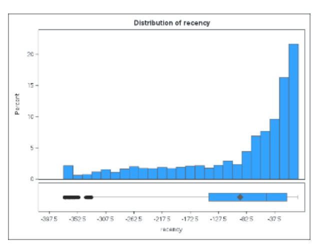
Custid	cust_lifespan	recency	activedays	total_numorders	total_types	total_qty	avg_qty_perorder	avg_qty_peritem	total_orderamt	avg_orderamt	avg_daysbetweenorders
16446	205	-1	2	2	2	80,996	40,498	40,498	168,471	84,236	102

Table 3: Outlier Point

Long-tailed distributions and large gaps in data points renders k-means clustering less effective and makes data visualisation challenging as high-ranged values would dominate the graphs. To address the skewed distributions, we used **log and cube-root transformations** to help increase the variability of data. For variables containing zero values, we opted for cube-root transformation, as log transformation would result in errors due to undefined values for zero. The histograms in Table 4Table 4 below show the distributions of variables before and after transformation.



No transformation performed



No transformation performed

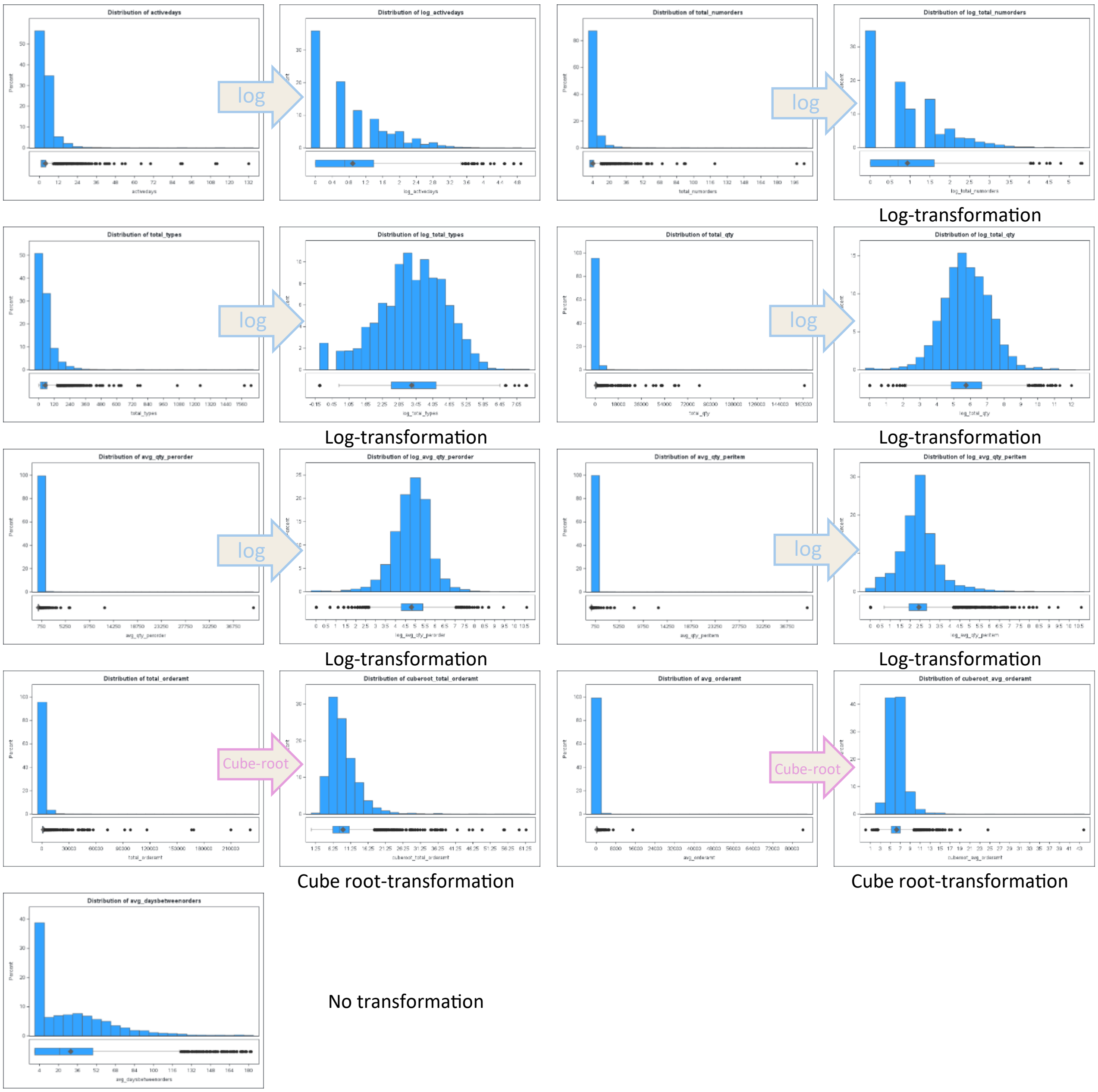


Table 4: Variable Transformations

BIVARIATE ANALYSIS

Next, we performed a correlation analysis to identify variables with high collinearity, as strong correlations suggest that some variables may not contribute uniquely to cluster identification.

The results revealed strong and statistically significant correlation between:

- "log_activedays" and "log_total_numorders" (0.99)
- "log_activedays" and "cust_lifespan" (0.85)
- "log_total_qty" and "cuberoot_total_orderamt" (0.84)
- "log_total_numorders" and "cust_lifespan" (0.83)

Given this high degree of correlation, we have decided to remove the variables “log_activedays”, “cust_lifespan”, and “log_total_qty” to reduce redundancy in our analysis. The correlation analysis was rerun using the remaining 8 variables, with all correlation coefficients falling below 0.8. Therefore, we will proceed with these variables for our clustering analysis.

Before	After removing highly-correlated variables

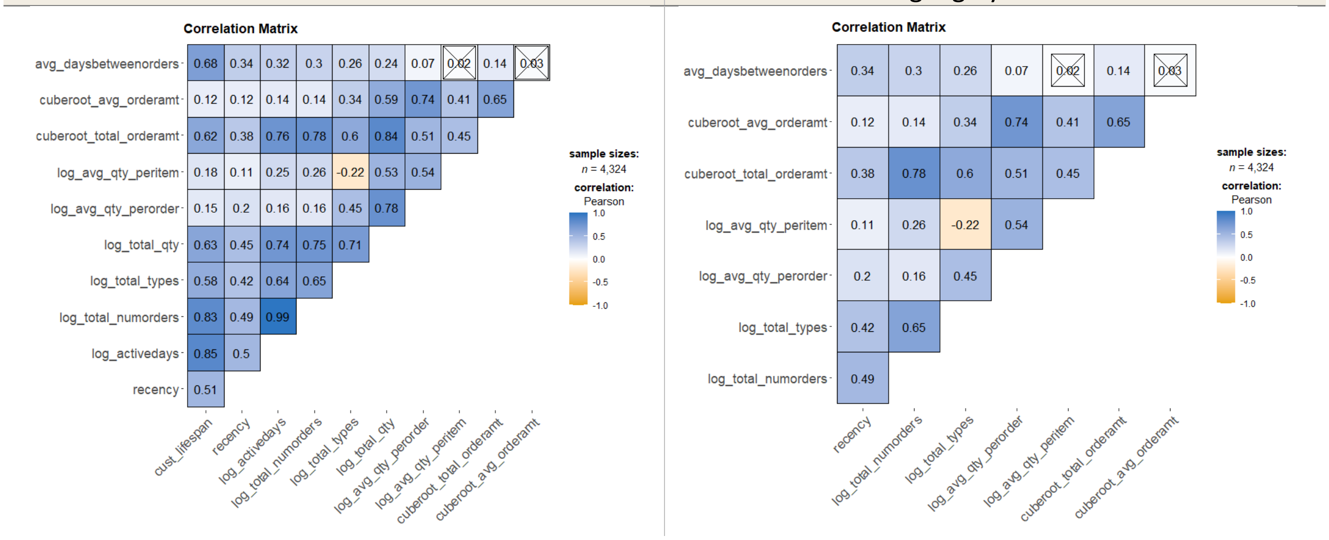


Figure 2: Correlation Matrix Before and After Removing Highly-Correlated Variables

K-MEANS CLUSTERING

K-means clustering was performed using the remaining 8 variables, and three popular methods were explored to determine the optimal number of clusters:

- **Elbow Method:** This graphical technique identifies the point where adding more clusters provides diminishing returns in reducing the sum of squared distances between points and their cluster centers (Syakur et al., 2018).
- **Gap Statistic:** This compares the total intracluster variation for different values of *k* to expected values under a null reference distribution (i.e., a distribution without evident clustering) (Tibshirani et al., 2001).
- **Average Silhouette:** This approach measures clustering quality by evaluating how well each point lies within its cluster. A high average silhouette width indicates a well-clustered dataset, with the optimal number of clusters being the one that maximizes the silhouette width (Rousseeuw, 1987).

Elbow Method	Gap Statistic	Average Silhouette

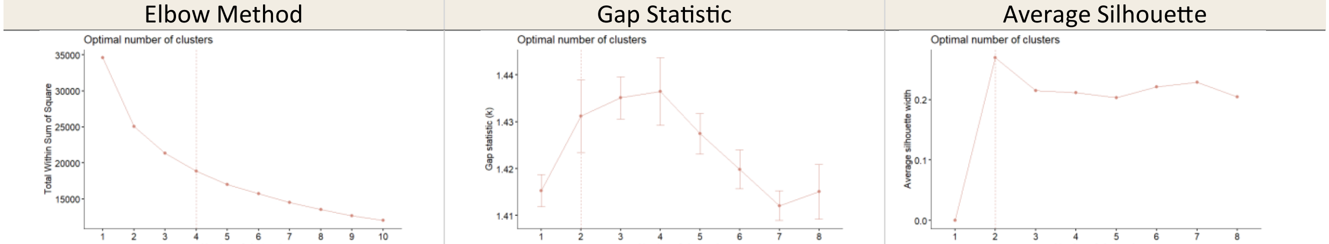


Figure 3: Comparison of Optimal Clusters by Various Methods

Key Insights from Figure 3:

- **Elbow Method:** The graph reveals an "elbow" at $k = 4$, indicating that 4 clusters is the optimal solution. After $k = 4$, the decrease in the within-cluster sum of squares becomes marginal.
- **Gap Statistic:** The results show that 2 clusters maximise the gap statistic, with 4 clusters as the second-best option.
- **Average Silhouette:** The analysis reveals that 2 clusters yield the highest average silhouette.

Although the gap statistic and average silhouette suggest 2 clusters as optimal, having only 2 clusters might oversimplify nuanced customer behaviour. Therefore, we opted for 4 clusters, as indicated by the elbow method and supported by the gap statistic's second-best option.

Since the variables have different measurement scales, we applied data standardisation before performing k-means clustering. This ensures that all variables contribute equally to the clustering process, as k-means is sensitive to the magnitude of the variables. By scaling the data to have a mean of 0 and a standard deviation of 1, we eliminate the risk of variables with larger scales disproportionately influencing the cluster formation.

C L U S T E R   S U M M A R Y

Among the 4 clusters, the number of customers ranges from 606 to 1,478. Notably, there are no extremely small clusters (e.g., with only 1-2 customers), indicating a balanced distribution across the clusters. The parallel coordinate plots in Table 5 illustrate the differences in customer behaviour.

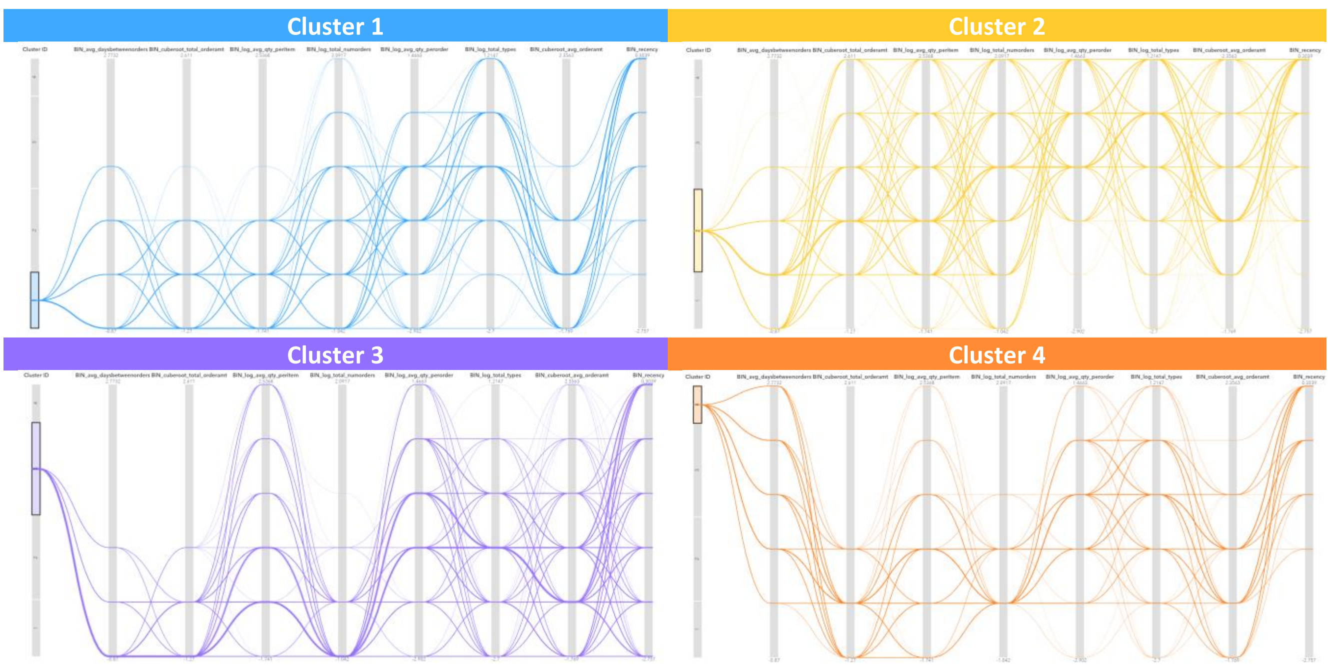


Table 5: Parallel Plots of 4 Clusters

Category	Metric	Variable Mean	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size			906 (21%)	1334 (31%)	1478 (34%)	606 (14%)
Time-related	Recency	-0.000741	0.3114	0.5357	-0.8514	0.4192
Quantity-related	Frequency (Total number of orders)	0.010446	0.0307	0.8844	-0.8867	0.0044
	Total distinct types of items	-0.014178	0.2644	0.7409	-0.8723	-0.0614
	Average quantity per order	-0.013890	-0.6492	0.6889	-0.2738	-0.1065
	Average quantity per item	-0.005575	-0.9010	0.5306	0.0143	-0.0228
Order-related	Monetary (Total order amount)	-0.011829	-0.3520	0.9823	-0.6665	-0.2409
	Average order amount	-0.017533	-0.5409	0.6786	-0.2830	-0.2484
	Average days between orders	0.013324	-0.0642	0.1037	-0.7550	1.9041

Table 6: Summary of Cluster Centroids

By analysing the parallel plots of the 4 clusters in Table 5: Parallel Plots of 4 Clusters, alongside the summary of the cluster centroid values in

Category	Metric	Variable Mean	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size			906 (21%)	1334 (31%)	1478 (34%)	606 (14%)
Time-related	Recency	-0.000741	0.3114	0.5357	-0.8514	0.4192
Quantity-related	Frequency (Total number of orders)	0.010446	0.0307	0.8844	-0.8867	0.0044
	Total distinct types of items	-0.014178	0.2644	0.7409	-0.8723	-0.0614
	Average quantity per order	-0.013890	-0.6492	0.6889	-0.2738	-0.1065
	Average quantity per item	-0.005575	-0.9010	0.5306	0.0143	-0.0228
Order-related	Monetary (Total order amount)	-0.011829	-0.3520	0.9823	-0.6665	-0.2409
	Average order amount	-0.017533	-0.5409	0.6786	-0.2830	-0.2484
	Average days between orders	0.013324	-0.0642	0.1037	-0.7550	1.9041

Table 6, we summarise of the key characteristics of each cluster below.

Category	Metric	Cluster 1 (Wavering)	Cluster 2 (Loyalists)	Cluster 3 (At Risk)	Cluster 4 (Promising)
Size		906 (21%)	1334 (31%)	1478 (34%)	606 (14%)
Time-related	Recency	Mid-High	High	Low-High	Mid-High
Quantity-related	Frequency (Total number of orders)	Low-Mid	Mid-High	Low	Mid
	Total distinct types of items	Mid-High	High	Low-Mid	Mid-High
	Average quantity per order	Low-Mid	High	Low-Mid	Mid-High
	Average quantity per item	Low	Mid-High	Mid-High	Low-Mid
Order-related	Monetary (Total order amount)	Low	Mid-High	Low	Low-Mid
	Average order amount	Low	High	Low-Mid	Low-Mid
	Average days between orders	Low-Mid	Low-Mid	Low	Mid-High

Table 7: Summary of Cluster Characteristics

CLUSTER INTERPRETATION

Customers could be further categorised into distinct groups, considering the significant inter-cluster differences (Saygili, 2021). This can be achieved by analysing the *Bin Details* of each variable, as well as *Profile by Segment and Overall* plots in SAS Viya's Model Sudio.

Output

Bin Details									
Variable	Bin ID	Lower Bound	Upper Bound	Bin Width	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
avg_daysbetweenorders	Missing				0				
	1	-Infty	-0.141		2329	-0.728	0.2286	-0.870	-0.151
	2	-0.141	0.5872	0.7287	961	0.2049	0.2032	-0.122	0.5681
	3	0.5872	1.3159	0.7287	590	0.8993	0.2077	0.5968	1.2871
	4	1.3159	2.0446	0.7287	249	1.6317	0.2097	1.3159	2.0350
	5	2.0446	2.7732	0.7287	110	2.3906	0.1953	2.0637	2.7540
	6	2.7732	Infty		85	3.2975	0.2574	2.7828	3.5019

Figure 4: Bin Details, and Profile By Segment and Overall for "avg_daysbetweenorders"

Table 8 below lists each cluster by risk of churn:

Churn Risk	Cluster Customer Profile	Characteristics
High	**3 (At Risk)**	
	Customers are primarily casual or one-time buyers. They exhibit the lowest frequency and monetary spending, and their variability in recency suggests that many of them could be at risk of churning. These customers typically make smaller, infrequent purchases and engage with a narrower range of products.	→ Variability in recency: 21% placed orders within the last 2 months, while ~16% made their last order nearly 1 year ago (356-366 days) → Lowest frequency: 84% made only 1 order → Days between orders: 94% have 0-25 days between orders → Lowest monetary: 72% spent <$321 in total → 69% spent $28-$861 per order → Low product diversity: 69% ordered 7-45 product types → 64% ordered 42-212 items per order → 75% purchased 5-26 quantity per item
Mid	**1 (Wavering)**	
	Wavering customers are not as frequent as loyal ones but show interest in trying a wide variety of products, although typically in smaller quantities.	→ High recency: 71% placed orders within the last 2 months → Medium frequency: 85% have made 1-5 orders → Days between orders: 49% have 0-25 days between orders, while 34% has 26-50 days → Medium monetary: 93% spent <$1122 in total → 60% spent $28-$658 per order → Medium product diversity: 77% ordered 18-118 product types → 80% ordered 42-212 items per order → 86% purchased 1-11 quantity per item
Low-Mid	**4 (Promising)**	
	The Promising customer has been with the retailer for a while and makes repeat purchases, but these tend to be sporadic and relatively small in scale. They generally buy fewer items per transaction and make moderate purchases.	→ Medium recency: 66% made orders within the last 2 months → Medium frequency: 81% have made 2-5 orders → Days between orders: 65% has 51-101 days between orders → Medium monetary: 72% spent $322-$1122 in total → 86% spent $28-$861 per order → Low product diversity: 93% with 7-118 types across all orders → 78% ordered 42-212 items per order → 77% purchased 5-26 quantity per items
Low	**2 (Loyalists)**	
	The Loyalist is a highly engaged customer who frequently makes large, high-value purchases. They	→ High recency: 80% made orders within the last 2 months

	tend to place large orders that include a wide variety of items, though the average quantity per item is not particularly high. Loyalists represent a key segment of the retailer's customer base due to their consistent and significant purchasing behaviour.	→ Medium frequency: Varying number of orders, with 76% who made >2 orders → Days between orders: 42% have 26-50 days between orders, while 37% have 0-25 days → High monetary: 70% spent $1125-$5322 in total → 71% spent $35-$2074 per order → High product diversity: 75% with 46-1631 product types → 80% ordered 94-483 items per order → 73% purchased 5-26 quantity per items

Table 8: Cluster Profiles

Recommendations

This section recommends the relevant customer relationship management strategies based on the identified segments and provides some suggestions for future studies.

Churn Risk	Cluster	Users (%)	Sales (%)	Recommendations
High	3 (At Risk)	34%	6%	For recent purchasers, the goal should be to strengthen brand loyalty and increase customer engagement. Strategies could include targeted (CRM campaigns, time-sensitive offers, or personalised promotions that encourage more frequent interactions with the retailer. For those who have not made a purchase in a while, efforts should focus on rekindling interest. Proactive outreach, such as "welcome back" campaigns, can help regain their attention.
Mid	1 (Wavering)	21%	7%	To encourage loyalty and convert these customers into regular buyers, focus on offering loyalty program benefits. Consider "Price Match" promotions or time-sensitive discounts to incentivise more frequent purchases. Personalised product recommendations based on their past purchases could help increase engagement and build a stronger connection to the brand.
Low-Mid	4 (Promising)	14%	5%	Though the churn risk is moderately low, CRM efforts should focus on driving higher engagement and increasing purchase frequency. Tailored campaigns featuring products they commonly buy can keep these customers interested. Offering time-sensitive promotions or exclusive deals could encourage them to make more frequent purchases and reduce the time lapsed between orders. A loyalty program with rewards for frequent or recurring orders might help solidify their commitment to the brand.
Low	2 (Loyalists)	31%	82%	Loyalists are key drivers of revenue, making up only a third of total users but contributing 82% of sales. Despite their low

				churn risk, it's important to prevent customer fatigue. Retention efforts should focus on making them feel valued through exclusive appreciation promotions or personalised campaigns. Implementing a loyalty program with benefits will reinforce their connection to the brand. These customers can be encouraged to become brand advocates by offering referral bonuses for bringing in new customers, further expanding the retailer's reach.

Table 9: Recommendations

Conclusion

This study aimed to create distinct subgroups of customers based on transactional data using k-means clustering. By deriving purchase-related behavioural variables and applying various methods to determine the optimal number of clusters, we identified 4 distinct customers groups and proposed tailored customer relationship management strategies for each segment of the online retailer's customer base. Adopting a more personalised approach to customer management could enhance customer satisfaction, loyalty, and advocacy, ultimately boosting profits.

While the findings provided invaluable insights, there are several ways to enhance customer segmentation:

- **Profit-Based Segmentation:** While this analysis focused on sales amount, segmenting customers based on profit would offer more targeted insights. By identifying and catering to the most profitable customers, rather than those with higher spending, marketing efforts could be optimised to focus on driving profitability.
- **Qualitative Insights:** Conduct qualitative research, such as customer satisfaction surveys, to deepen understanding of each segment. Customer sentiments towards products or experiences could provide clearer insights, allowing the retailer to optimise strategies for satisfaction and retention.
- **Predictive Analysis:** Combine clustering with predictive techniques, such as decision tree, to predict key customer metrics such as expenditure. Predicting revenue and identifying factors influencing spending can guide business decisions. Further analysis of product preferences within segments can also help predict future purchase rates, which would help in inventory planning and stock optimisation.
- **Alternative Clustering Methods:** Explore advanced clustering techniques, such as the Gaussian Mixture Model, which uses "soft" clustering technique to assign cluster membership weights to each cluster. This offers more flexibility compared to k-means' "hard" clustering. Additionally, Latent Class Analysis, a non-parametric method, can be useful in addressing the skewed distribution seen in this study, offering more robust segmentations.

References

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

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. doi: 10.1016/0377-0427(87)90125-7

Saygili, T. (2021). *Customer segmentation with RFM score.* Medium. https://medium.com/cits-tech/customer-segmentation-with-rfm-score-d3b2b1bf45de

Syakur, M. A., Khotimah, B. K., Rochman. E., M. S., & Satoto, B. D. (2018). Integration K-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering, 336*. doi: 10.1088/1757-899X/336/1/012017

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 63*(2), 411–423. doi: 10.1111/1467-9868.00293

Yang, Y., & Padmanabhan, B. (2003). Segmenting customer transactions using a pattern-based clustering approach. *Third IEEE International Conference on Data Mining*. doi:10.1109/icdm.2003.1250947