IE6400 Project 2:

TOPIC: Customer Segmentation using RFM Analysis

Course: Foundations of Data Analytics Engineering

Semester: Fall 2025

Group Members:

1) Rutuja Mahesh Dambir

2) Praniti Sunil Kale

3) Hiral Hitesh Rana

1. Introduction

Understanding customer behaviour is critical for designing targeted marketing strategies and improving customer retention. In this project, we apply RFM (Recency, Frequency, Monetary) analysis to a real eCommerce transaction dataset and build a customer segmentation model based on purchasing behaviour. The work follows the requirements specified in the IE6400 Project 2 description.

The main goals are:

- Compute RFM metrics for each customer.
- Score customers using RFM quartiles and derive behaviour-based segments.
- Apply K-Means clustering on RFM variables to automatically discover segments.
- Profile each segment and translate patterns into actionable marketing recommendations.
- Address the guideline questions on customer, product, time, geography, returns, and profitability, within the limits of the dataset.

All implementation was done in Python using libraries such as *pandas, NumPy, scikit-learn, matplotlib,* and *seaborn*.

---

2. Data and Methodology

2.1 Dataset Description

The analysis uses the Online Retail dataset originally hosted on Kaggle, containing transactional data from a UK-based online retailer. The dataset includes invoice-level records for one year of sales.

After loading the CSV file with appropriate encoding, the raw dataset contains 541,909 rows and 8 columns. The key fields are:

- InvoiceNo – unique identifier for each invoice (string).
- StockCode – product code.
- Description – textual product description.
- Quantity – number of units of the product in the invoice line (positive for purchases, negative for returns).
- InvoiceDate – timestamp of the invoice.
- UnitPrice – price per unit (in GBP).
- CustomerID – numeric customer identifier (float in raw file, later converted to integer).
- Country – customer's country of residence.

The time period covered is from 1 December 2010 to 9 December 2011.

2.2 Data Preprocessing

Several preprocessing steps were performed before RFM calculation:

1. Datetime conversion
   - InvoiceDate was converted from string to datetime. The converted date range confirmed the one-year period mentioned above.
2. CustomerID filtering
   - The notebook reported 0 rows with missing CustomerID after conversion, so no rows needed to be removed for this reason.
3. Handling returns
   - Negative Quantity values indicate product returns. For accurate RFM analysis, which focuses on positive purchase behavior, rows with Quantity < 0 were removed from the working dataset.
   - However, the original dataset was retained separately to quantify the overall return rate (Section 3.6).
4. Feature engineering
   - TotalPrice = Quantity × UnitPrice was created to represent line-level revenue.
   - Additional time features DayOfWeek (day name) and Hour (hour of day) were derived from InvoiceDate for time-based analysis.

After filtering out returned items, the cleaned dataset used for RFM analysis contains 397,924 rows and 11 columns(original 8 plus TotalPrice, DayOfWeek, and Hour).

2.3 RFM Metric Computation

RFM metrics were computed per customer using the cleaned transactional data:

- Recency (R):
  - Defined as the number of days between the customer’s most recent purchase and a fixed analysis date.
  - The analysis date was set to 10 December 2011, i.e., one day after the maximum InvoiceDate in the dataset.
  - For each CustomerID, recency = (analysis_date − max(InvoiceDate)) in days.
- Frequency (F):
  - Defined as the number of unique invoices associated with the customer.
  - This counts how many separate orders the customer placed, not how many line items.
- Monetary (M):
  - Defined as the sum of TotalPrice across all transactions of the customer.
  - This represents the total revenue generated by the customer during the observed period.

These metrics were aggregated via a group-by operation over CustomerID, resulting in an RFM table with 4,339 unique customers.

Descriptive statistics from this table:

Metric	Count	Mean	Median	Std Dev	Min	Max
Recency (days)	4,339	~92.5	51	~98.3	1	374
Frequency (orders)	4,339	~4.27	2	~6.65	1	210
Monetary (£)	4,339	~2,053.79	~674.45	~7,981	~0.01	~280,206

2.4 RFM Scoring

To transform continuous RFM metrics into categorical scores, we used quartile-based scoring:

- For Frequency and Monetary, higher values are better.
    - Quartiles were computed, and each customer was assigned a score from 1 to 4, where 4 corresponds to the top quartile and 1 to the bottom quartile.
- For Recency, lower values are better (more recent purchases).
    - The quartile scoring was inverted: customers with the smallest recency received a score of 4, and those with the largest recency received 1.

This produced three discrete variables: R_score, F_score, M_score. A concatenated RFM_Score, e.g., “444”, and a numeric RFM_Sum = R + F + M were created to help interpret overall customer quality.

2.5 Clustering Approach

Beyond manual RFM segmentation, we applied K-Means clustering on the continuous RFM metrics to discover behaviour-based segments:

1. Standardization
    - Recency, Frequency, and Monetary were scaled using StandardScaler to have zero mean and unit variance. This prevents Monetary from dominating the clustering due to its larger scale.
2. Model selection
    - K-Means was fitted for k = 2 to 6 clusters on the scaled RFM variables.
    - For each k, the silhouette score was computed to evaluate separation between clusters.
    - The best silhouette score was observed at k = 2 (≈ 0.896), substantially higher than scores for k = 3–6 (≈ 0.59–0.62).
3. Final model
    - A K-Means model with k = 2 clusters was selected.
    - Cluster labels (0 and 1) were assigned to each customer.
    - Cluster centroids were transformed back to the original RFM scale for interpretation.

---

3. Exploratory Data Analysis

3.1 Data Overview

- Dataset size (cleaned, positive-quantity transactions with known customers):
  - 397,924 records and 11 variables.
- Time coverage:
  - 1 December 2010 – 9 December 2011.
- Columns:
  - InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country, TotalPrice, DayOfWeek, Hour.

3.2 Customer Analysis

- Number of unique customers:
  - 4,339 customers.
- Orders per customer:
  - Distribution is right-skewed: most customers place between 1 and 5 orders, while a small group makes over 50 orders.
  - The top 5 customers by number of orders (unique invoices) are:
    - ID 12748 – 210 orders
    - ID 14911 – 201 orders
    - ID 17841 – 124 orders
    - ID 13089 – 97 orders
    - ID 14606 – 93 orders

These top customers likely correspond to wholesale or repeat B2B buyers, contributing significantly to revenue and justifying strong retention strategies.

- Customer active lifetime:
  - For each customer, we computed the time between first and last purchase.
  - The average active lifetime is approximately 130 days, indicating that on average customers remain active for about 4–5 months within the one-year period.

3.3 Product Analysis

- Top 10 most frequently purchased products by quantity:
  1. PAPER CRAFT , LITTLE BIRDIE (≈ 80,995 units)
  2. MEDIUM CERAMIC TOP STORAGE JAR (≈ 77,916 units)
  3. WORLD WAR 2 GLIDERS ASSTD DESIGNS (≈ 54,415 units)
  4. JUMBO BAG RED RETROSPOT (≈ 46,181 units)
  5. WHITE HANGING HEART T-LIGHT HOLDER (≈ 36,725 units)
  6. ASSORTED COLOUR BIRD ORNAMENT
  7. PACK OF 72 RETROSPOT CAKE CASES
  8. POPCORN HOLDER
  9. RABBIT NIGHT LIGHT
  10. MINI PAINT SET VINTAGE

  Many of these items are small gift or decorative products, consistent with the retailer’s focus.

- Average product price:
    - The mean UnitPrice is approximately £3.12, with many low-cost items and a long tail of higher-priced products.
- Top revenue-generating products:
    - When aggregated by TotalPrice, the highest revenue items include:
        - PAPER CRAFT , LITTLE BIRDIE (≈ £168,470)
        - REGENCY CAKESTAND 3 TIER (≈ £142,593)
        - WHITE HANGING HEART T-LIGHT HOLDER (≈ £100,448)
        - JUMBO BAG RED RETROSPOT (≈ £85,221)
        - MEDIUM CERAMIC TOP STORAGE JAR (≈ £81,417)
        - POSTAGE (≈ £77,804)

  This shows that some relatively low-priced items generate high revenue through volume, while others (like POSTAGE) reflect shipping-related charges.

3.4 Time-Based Analysis

- Orders by day of week:
    - The highest volume of transactions occurs on Thursdays, followed by Wednesdays and Tuesdays.
    - Fridays show the lowest number of orders among weekdays, while Sundays still have a substantial number of orders.
- Orders by hour of day:
    - Most activity is concentrated between 9:00 and 16:00, with a clear peak around midday (12:00). Very few orders occur in early morning or late evening hours.
- Monthly revenue trend:
    - Figure 1 (Monthly Revenue Time Series)
        - Data used: Monthly sum of TotalPrice, obtained by resampling the transaction data by calendar month.
        - What it shows: Revenue varies substantially across the year, with some pronounced peaks (likely during the holiday season near November–December) and calmer periods in the middle of the year.
        - Possible reasons: Seasonality driven by holiday and gifting periods, promotional campaigns, or bulk corporate orders.

3.5 Geographical Analysis

- Top 5 countries by number of orders:
    1. United Kingdom – 354,345 orders
    2. Germany – 9,042
    3. France – 8,342
    4. EIRE (Ireland) – 7,238
    5. Spain – 2,485

  The business is clearly UK-dominated, with a long tail of international customers.

- Average order value by country:
    - When computing mean TotalPrice per country, several non-UK countries such as Netherlands, Australia, Japan, Singapore, Sweden, and Denmark have relatively high average order values (often above £80–£100).
    - This suggests that international customers tend to place fewer but larger orders, possibly due to higher shipping costs and import constraints, while UK customers place frequent smaller orders.

3.6 Returns and Refunds

Unlike the cleaned dataset used for RFM, the original dataset includes both purchases and returns:

- Total rows: 541,909.
- Number of rows with Quantity < 0 (returns): 10,624.
- Return rate by line items: approximately 1.96%.

An example categorical comparison (using a derived Boolean is toy flag from product descriptions) shows that some product types such as toys may have higher return counts than others, highlighting that return behaviour is not uniform across product families.

Because the dataset does not provide explicit return reasons or cost, we focus on describing patterns rather than calculating exact profitability impact.

---

4. RFM Analysis Results

4.1 Distribution of Recency, Frequency and Monetary Value

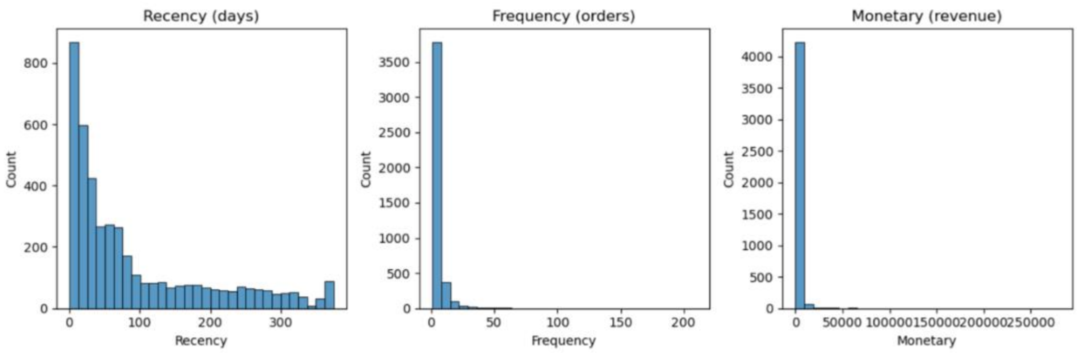

Figure 2 – *Histograms of Recency, Frequency, and Monetary*

- Data used: R, F, M values for all 4,339 customers in the RFM table.
- What it shows:

    - Recency: A large mass of customers has recency under ~100 days, with a long tail of customers who have not purchased for several hundred days.
    - Frequency: Most customers place 1–3 orders; only a small minority reaches >20 orders.
    - Monetary: Highly skewed distribution with many low-spend customers and a few extremely high spenders.
- Possible reasons:
    - The store likely acquires many one-time or occasional buyers, while only a small group develops long-term loyalty and very high spend.
    - The long tail in Monetary may reflect wholesale or corporate clients.

Text size in these plots should be kept sufficiently large (e.g., 10–12 pt labels) to remain readable when inserted in the report.

4.2 RFM-Based Customer Segments

The quartile scoring of R, F, and M enables us to construct RFM segments, such as:

- “444” customers: most recent purchasers, highest frequency, and highest monetary value – *best customers*.
- “144” or “244” customers: older purchases but still high spender – *potential win-back*.
- “111” customers: long ago, low frequency, low spend – *low-value or one-time buyers*.

Additionally, we created a feature Segment = RFM_Score or, alternately, used cluster labels (Section 5) as segment identifiers. The value counts show many customers clustered in lower combination scores, with relatively fewer customers in elite segments, reinforcing the Pareto principle (small % of customers drive a large % of revenue).

---

5. Clustering and Segment Profiling

5.1 K-Means Model Selection

Using standardized RFM metrics, we evaluated K-Means models with 2–6 clusters:

- Silhouette scores:
    - k=2: ~0.896
    - k=3: ~0.594
    - k=4: ~0.616
    - k=5: ~0.616
    - k=6: ~0.598

Although k=4 and k=5 showed modest improvements over k=3, k=2 produced by far the highest silhouette score, indicating very clear separation between two major behaviour modes. As a result, we finalized k = 2 clusters and summarized them by their size and revenue contribution, using the Monetary totals in Table B2.

Table B2. K-Means Cluster Profiles (k = 2)

Feature / Metric	Cluster 0 (Regular Customers)	Cluster 1 (VIP / Champions)
Number of Customers	4,313	26
Avg Recency (days)	~93	~6
Avg Frequency (orders)	~3.9	~66.5
Avg Monetary (£)	~1,548	~85,904
Total Monetary Contribution (£)	~6.68 million	~2.23 million
% of Total Revenue	~75%	~25%
% of Customer Base	99.4%	0.6%

Interpretation:

- Although Cluster 1 customers represent **less than 1%** of all customers, they contribute **a** quarter of total revenue—highlighting their strategic value.
- Cluster 0 customers are vital for volume, but individual spending is much lower.

5.2 Cluster Characteristics

The cluster centroids, mapped back to the original RFM scales, are summarized in Table 1.

Table 1 – Cluster Centroids in Original RFM Scale

Cluster	Recency (mean days)	Frequency (mean orders)	Monetary (mean revenue)
0	≈ 93.0	≈ 3.9	≈ £1,548
1	≈ 6.0	≈ 66.5	≈ £85,904

- Cluster 0 – “Regular / Low–Medium Value Customers”
    - Recency ≈ 93 days (less recent).
    - Median Frequency ≈ 2 orders; median Monetary ≈ £668.
    - Very large in size: 4,313 customers.
- Cluster 1 – “Champions / VIP Customers”
    - Extremely recent Recency ≈ 6 days.
    - Median Frequency ≈ 53 orders; median Monetary ≈ £59,797.
    - Very small in size: 26 customers, but extraordinarily high revenue per customer.

5.3 Revenue Contribution by Cluster

Using the aggregated Monetary sums:

- Cluster 1 contributes a disproportionately high share of total revenue despite being <1% of all customers.

- Cluster 0 contains the majority of the customer base but at a much lower average spend.

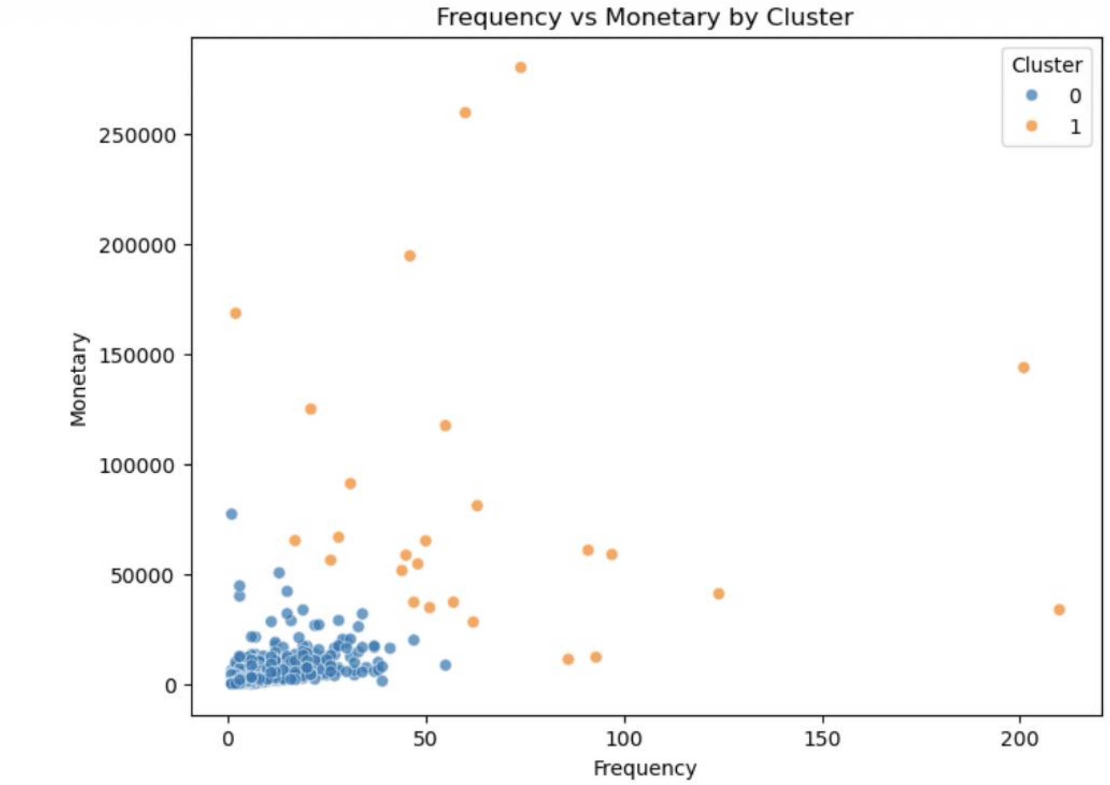

Figure 3 – Frequency vs. Monetary by Cluster

- Data used: Each point represents a customer, with x-axis = Frequency, y-axis = Monetary, coloured by cluster.
- What it shows:
  - Cluster 0 points are densely concentrated at low Frequency and low-to-moderate Monetary.
  - Cluster 1 points appear far to the right and high on the Monetary axis, clearly separated from the main mass.
- Possible reasons:
  - Cluster 1 likely captures key accounts or wholesale partners making many large purchases.
  - The separation supports the idea of designing distinct strategies for VIP vs. regular customers.

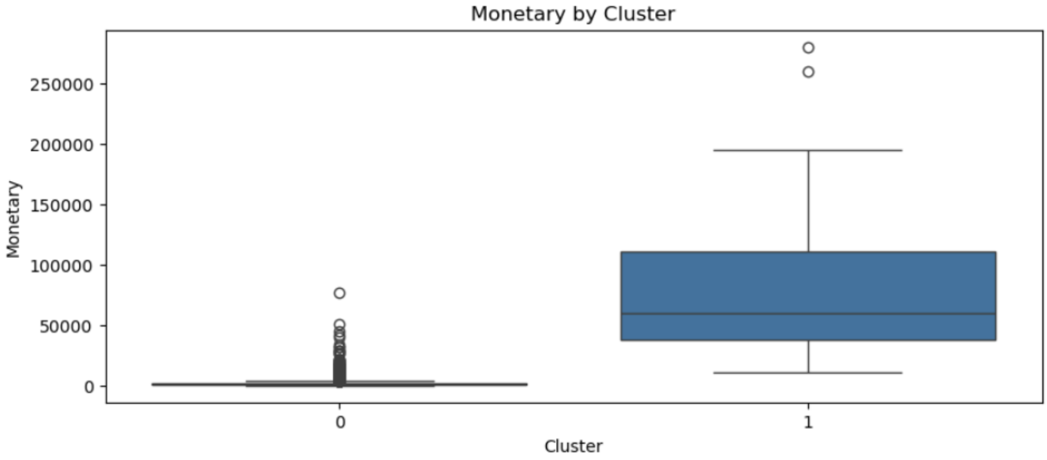

Figure 4 – Boxplot of Monetary by Cluster

- Data used: Monetary values for all customers in each cluster.
- What it shows:
  - The median Monetary for Cluster 1 is orders of magnitude higher than Cluster 0.
  - The box and whiskers are much taller for Cluster 1, reflecting wide variability among VIP customers.
- Possible reasons:
  - Different contract sizes or purchase cycles among key accounts; some may place extremely large orders.

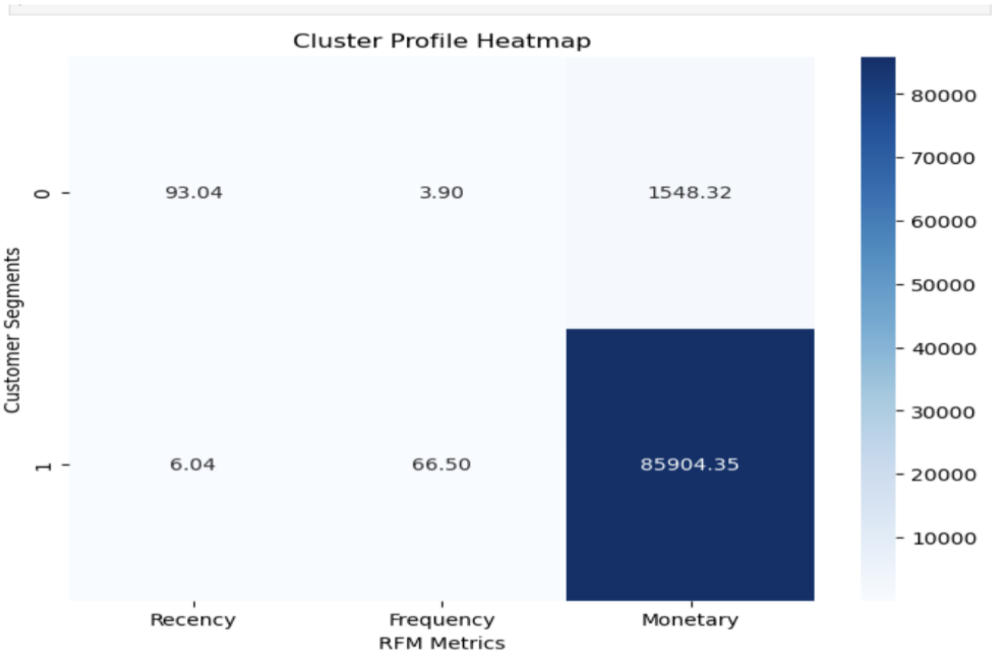

Figure 5 – Cluster Profile Heatmap

- Data used: For each cluster, average Recency, Frequency and Monetary values.

- What it shows:
    - Cluster 1 is dark in the Monetary column, indicating very high spending and high Frequency.
    - Cluster 0 is lighter and has higher Recency, indicating less recent activity.

These visualizations collectively confirm that two sharply separated customer segments exist in terms of engagement and value.

---

6. Marketing Recommendations

Based on RFM analysis and clustering, we derive the following segment-specific strategies.

6.1 Cluster 1 – “Champions / VIP Customers”

- Profile:
    - Extremely recent purchases, very high frequency, very high revenue.
    - Small in number (26 customers) but vital to total revenue.
- Recommended actions:
    - Implement a VIP loyalty program with exclusive discounts, early access to new collections, and dedicated account managers.
    - Offer personalized bundles and upsell premium/high-margin products.
    - Invite them to provide reviews or referrals; they may influence other customers.
    - Monitor this segment closely with churn alerts (e.g., if recency suddenly increases).

6.2 Cluster 0 – “Regular / Low–Medium Value Customers”

- Profile:
    - The majority of customers with moderate spend and limited order history.
    - Recency around 51–93 days, suggesting many are *at risk* of becoming inactive if not re-engaged.
- Recommended actions:
    - Use email campaigns and personalized product recommendations to increase order frequency.
    - Promote cross-sell opportunities (e.g., accessories, complementary items).
    - Offer small incentives such as free shipping thresholds or bundle discounts to raise average order value.

    - Introduce a points-based loyalty program to encourage repeat purchase.

6.3 Behaviour-Based RFM Segments

For more granular targeting, RFM scores can be translated into common marketing archetypes:

- “Champions” (e.g., R=4, F=4, M=4):
    - Reward heavily, invite to beta programs, treat as ambassadors.
- “Loyal but Low Spending” (e.g., high R and F, lower M):
    - Encourage trading up via bundles, premium product recommendations, and promotions.
- “At Risk” (e.g., low R score, mid or high F/M):
    - Run win-back campaigns with personalized discounts and reminders.
- “New / One-time Buyers” (e.g., high R but low F and M):
    - Focus on onboarding and education, reducing friction to second purchase.

---

7. Limitations and Future Work

Several limitations constrain the scope of this analysis:

1. Missing payment method information
    - The dataset does not contain payment type (e.g., credit card, PayPal). Therefore, we cannot analyze payment preferences or link payment method to order value.
2. No cost or margin data
    - We can compute revenue but not true profit or margin per product. Any profit analysis would require assumptions (e.g., a uniform 30% margin) and should be treated as approximate.
3. No explicit product categories
    - There is only a free-text Description field. Without structured categories, product-level insights rely on keyword heuristics and may miss higher-level patterns.
4. No customer satisfaction / ratings data
    - The dataset lacks reviews or rating fields, preventing sentiment analysis or satisfaction modelling.
5. Static one-year window
    - RFM metrics are calculated for a single period, not updated over time. A production system should recompute RFM periodically (e.g., monthly) to track customer lifecycle changes.

Future work could include:

- Enriching the dataset with product categories, cost, and payment data.
- Building predictive models for churn or CLV (Customer Lifetime Value).
- Integrating A/B testing to measure the impact of RFM-based campaigns.
- Extending clustering to more than 2 segments once additional behavioral features are available.

---

8. Conclusion

This project demonstrates how RFM analysis combined with K-Means clustering can be used to segment customers in an eCommerce setting:

- The dataset contains 4,339 customers and nearly 400k cleaned transaction lines over one year.
- RFM metrics reveal a heavily skewed distribution, with many low-engagement customers and a small but critical group of high-value buyers.
- K-Means clustering on Recency, Frequency, and Monetary identifies two clearly separated segments:
    - Cluster 1 (Champions/VIP) – extremely engaged customers with very high revenue.
    - Cluster 0 (Regular/Low–Medium value) – the majority, with moderate frequency and spend.
- These insights translate directly into differentiated marketing strategies for retention, cross-sell, and win-back.

Overall, the analysis illustrates how relatively simple metrics, when combined with clustering and visualization, can provide actionable and interpretable guidance for customer relationship management.