Customer Segmentation Report

Sinethemba Ngcongo - 26409852

Executive Summary

This report applies data-driven techniques to segment customers in an online retail dataset using RFM (Recency, Frequency, Monetary) analysis and clustering. The findings reveal that most revenue is driven by a small group of high-value customers, while many others purchase infrequently. Four key customer groups were identified: Champions, Loyal Customers, At Risk, and New/Low Spenders. Based on these insights, targeted strategies are recommended to retain valuable clients, re-engage and those at risk.

Introduction & Research Question

Understanding customer purchasing behaviour is critical for retail businesses aiming to increase profitability and improve customer retention. Transactional datasets hold rich signals about when customers buy, how often they purchase, and how much they spend. By segmenting customers into distinct groups, businesses can tailor engagement strategies to specific needs.

Research Question: *How can the retail platform identify and target groups of customers with different purchasing patterns in order to maximize value and improve customer retention?*

Data Preparation

The dataset contained transactions from a online retailer. Key cleaning and preparation steps included:

- **Excluding cancellations**: Transactions with invoice numbers starting with "C" were removed to focus on completed sales.
- **Filtering invalid values**: Negative or zero quantities and prices were excluded.

- **Removing missing customers**: Rows without a Customer ID were dropped.
- **Creating monetary value**: A new field, `amount = quantity × unit_price`, was introduced to calculate revenue per line item.
- **Snapshot definition**: To measure recency, the reference date was set to one day after the most recent transaction.

These steps ensured a reliable, business-ready dataset.

Exploratory Insights

Examining **recency, frequency, and monetary distributions** revealed highly skewed patterns:

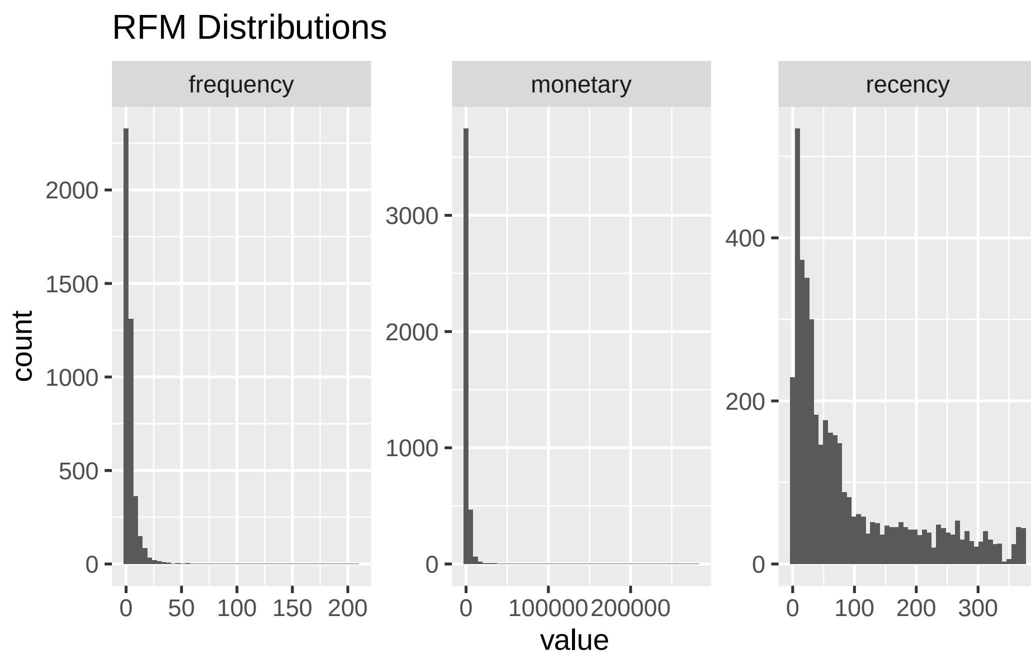

Figure 1: Figure 1: Distribution of Recency, Frequency, and Monetary Values

(Figure 1): The distributions are highly skewed, showing that most customers purchase rarely and spend little, while a small fraction contributes disproportionately to revenue. This confirms the importance of segmenting customers rather than applying a uniform marketing approach.

Customer Relationships

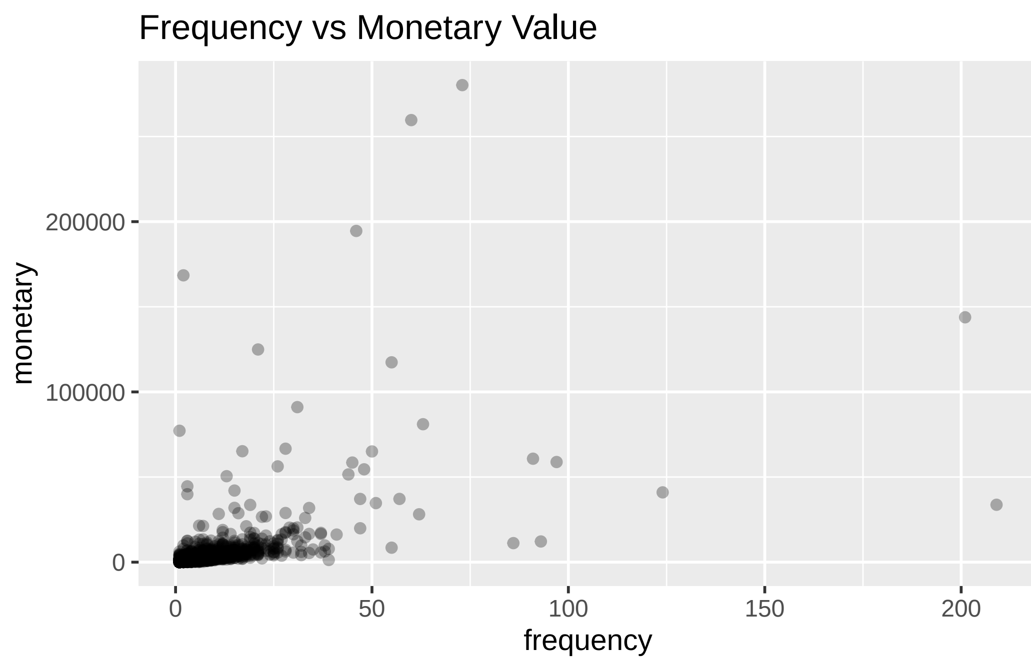

Figure 2: Figure 2: Relationship between Frequency and Monetary Value

(Figure 2): Customers who purchase more frequently also generate significantly higher revenue. Frequency is therefore a key driver of overall customer value and should be central in retention strategies.

Customer Segmentation (Clustering Results)

To group customers, log-transformed RFM features were standardised, and **k-means clustering** was applied. Both the elbow and silhouette methods suggested four optimal clusters.

(Figure 3): The elbow method shows a clear bend at four clusters, indicating this number provides the best balance between simplicity and explanatory power.

(Figure 4): The silhouette method supports four clusters, showing relatively strong internal cohesion and separation across groups.

```
# A tibble: 4 x 5
  cluster recency frequency monetary     n
  <fct>     <dbl>     <dbl>    <dbl> <int>
```

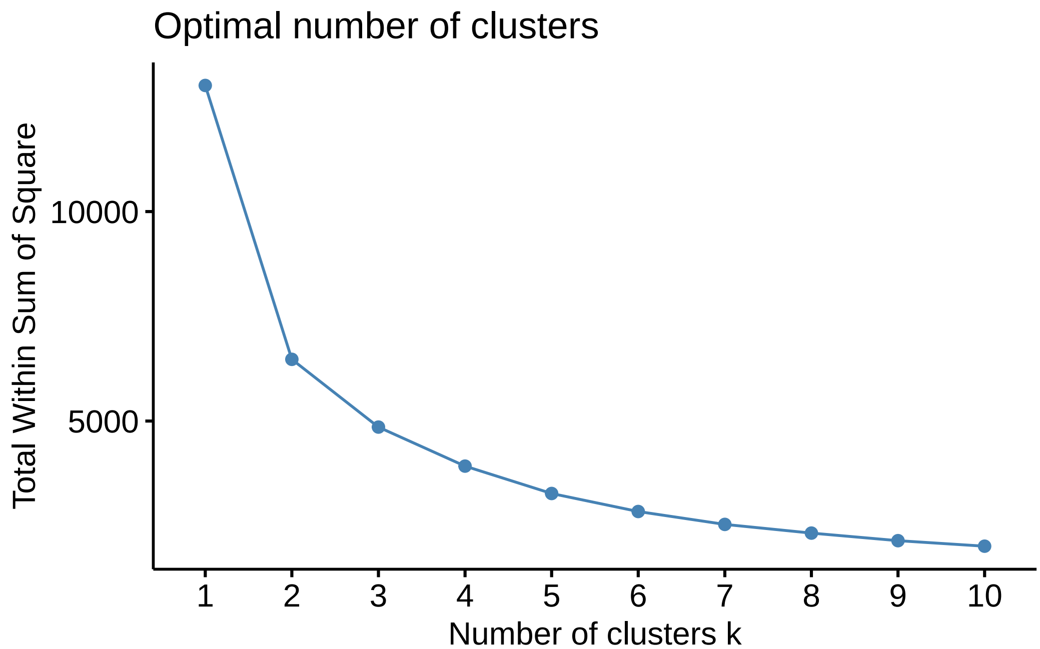

Figure 3: Figure 3: Elbow Method for Determining Optimal Number of Clusters

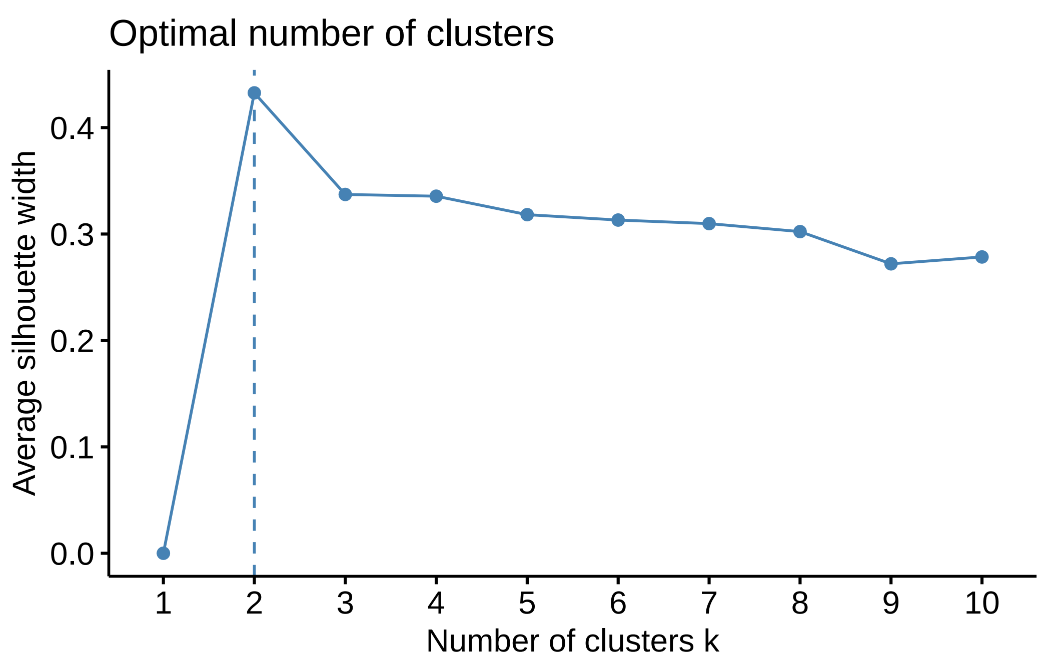

Figure 4: Figure 4: Silhouette Method for Cluster Validation

1	1	183.	1	300.	1579
2	2	19.0	2	452.	878
3	3	8.94	10	3722.	723
4	4	54.1	4	1353.	1158

(Figure 5): Each cluster shows distinct characteristics. **Cluster 1** ("Champions") includes recent and high-spending customers.**Cluster 2** ("Loyal Customers") purchase frequently but at moderate spend levels. **Cluster 3** ("At Risk") customers have high past spend but long gaps since their last purchase. **Cluster 4** ("New/Low Spenders") are new or low-value customers, representing growth opportunities.

Recommendations

- **Champions:** Reward with exclusive benefits and early access to products.
- **Loyal Customers:** Introduce loyalty programs and encourage upselling.
- **At Risk:** Run reactivation campaigns to bring them back.
- **New/Low Spenders:** Provide education and small incentives to build trust and loyalty.

Limitations & Future Work

- **Missing IDs:** Many transactions lacked Customer IDs, limiting full visibility.
- **Cancellations:** Excluding canceled invoices may mask dissatisfaction trends.
- **Seasonality:** No adjustment was made for holiday or seasonal demand patterns.
- **Scope:** Only RFM metrics were used. Incorporating demographics and product categories would enhance segmentation.

Conclusion

This segmentation confirms that a minority of customers generate most revenue, while many remain underdeveloped. By acting on these insights, the retailer can retain high-value customers, reduce churn, and nurture new segments for long-term growth. The clustering approach demonstrates the power of analytics to directly inform marketing and retention strategies.

Appendix - Full R Code

```
print(raw)
```

```
# A tibble: 541,909 x 8
   InvoiceNo StockCode Description        Quantity InvoiceDate         UnitPrice
   <chr>     <chr>     <chr>                 <dbl> <dttm>                  <dbl>
 1 536365    85123A    WHITE HANGING HEA~        6 2010-12-01 08:26:00      2.55
 2 536365    71053     WHITE METAL LANTE~        6 2010-12-01 08:26:00      3.39
 3 536365    84406B    CREAM CUPID HEART~        8 2010-12-01 08:26:00      2.75
 4 536365    84029G    KNITTED UNION FLA~        6 2010-12-01 08:26:00      3.39
 5 536365    84029E    RED WOOLLY HOTTIE~        6 2010-12-01 08:26:00      3.39
 6 536365    22752     SET 7 BABUSHKA NE~        2 2010-12-01 08:26:00      7.65
 7 536365    21730     GLASS STAR FROSTE~        6 2010-12-01 08:26:00      4.25
 8 536366    22633     HAND WARMER UNION~        6 2010-12-01 08:28:00      1.85
 9 536366    22632     HAND WARMER RED P~        6 2010-12-01 08:28:00      1.85
10 536367    84879     ASSORTED COLOUR B~       32 2010-12-01 08:34:00      1.69
# i 541,899 more rows
# i 2 more variables: CustomerID <dbl>, Country <chr>
```

```
print(df)
```

```
# A tibble: 397,884 x 9
   invoice_no stock_code description     quantity invoice_date        unit_price
   <chr>      <chr>      <chr>              <dbl> <dttm>                   <dbl>
 1 536365     85123A     WHITE HANGING ~        6 2010-12-01 08:26:00       2.55
 2 536365     71053      WHITE METAL LA~        6 2010-12-01 08:26:00       3.39
 3 536365     84406B     CREAM CUPID HE~        8 2010-12-01 08:26:00       2.75
 4 536365     84029G     KNITTED UNION ~        6 2010-12-01 08:26:00       3.39
 5 536365     84029E     RED WOOLLY HOT~        6 2010-12-01 08:26:00       3.39
 6 536365     22752      SET 7 BABUSHKA~        2 2010-12-01 08:26:00       7.65
 7 536365     21730      GLASS STAR FRO~        6 2010-12-01 08:26:00       4.25
 8 536366     22633      HAND WARMER UN~        6 2010-12-01 08:28:00       1.85
 9 536366     22632      HAND WARMER RE~        6 2010-12-01 08:28:00       1.85
10 536367     84879      ASSORTED COLOU~       32 2010-12-01 08:34:00       1.69
# i 397,874 more rows
# i 3 more variables: customer_id <dbl>, country <chr>, amount <dbl>
```

```r
print(rfm)
```

```
# A tibble: 4,338 x 5
   customer_id recency frequency monetary cluster
         <dbl>   <dbl>     <int>    <dbl> <fct>  
 1       12346  326.           1   77184. 4      
 2       12347    2.87         7    4310  3      
 3       12348   76.0          4    1797. 4      
 4       12349   19.1          1    1758. 2      
 5       12350  311.           1     334. 1      
 6       12352   36.9          8    2506. 4      
 7       12353  205.           1      89  1      
 8       12354  233.           1    1079. 1      
 9       12355  215.           1     459. 1      
10       12356   23.2          3    2811. 4      
# i 4,328 more rows
```

```r
print(cluster_profile)
```

```
# A tibble: 4 x 5
  cluster recency frequency monetary     n
  <fct>     <dbl>     <dbl>    <dbl> <int>
1 1        183.           1     300.  1579
2 2         19.0          2     452.   878
3 3          8.94        10    3722.   723
4 4         54.1          4    1353.  1158
```