

Customer Segmentation Project: A Full Documentation

BY TOBI MAUTIN

Table of Contents

| | |
|---|----|
| Customer Segmentation Project: Full Documentation | 1 |
| Executive Summary..... | 3 |
| 1. Introduction | 4 |
| 2. Methodology | 4 |
| 2.1 Expanded Methodology..... | 4 |
| 3. Clustering Analysis | 6 |
| 3.1 K-means Clustering | 6 |
| 3.2 DBSCAN Clustering | 7 |
| 3.3 Cluster Profiling | 8 |
| 3.3.1 Cluster 0 Profile..... | 8 |
| 3.3.2 Cluster 1 Profile..... | 9 |
| 4. Top Customers Analysis..... | 10 |
| 5. Customer Lifetime Value (CLV) Analysis | 11 |
| Top 10 customers by CLV: | 11 |
| 6. Cohort Analysis..... | 12 |
| 7. Conclusions and Recommendations | 13 |
| 7.1. Customer Segmentation Insights:..... | 13 |
| 7.2. Targeted Marketing Strategies: | 13 |
| 7.3. Customer Retention and Lifecycle Management: | 13 |
| 7.4. Personalization and Customer Experience:..... | 14 |
| 7.5. Data Quality and Customer Identification: | 14 |
| 7.6. Continuous Improvement and Monitoring: | 14 |
| 8. Limitations..... | 15 |
| 9. Implementation Plan | 15 |
| 10. Suggested Additional Analyses | 17 |
| 10.1. RFM Score Distribution Visualization | 17 |
| 10.2. Customer Lifecycle Stage Analysis: | 17 |
| 10.3. Predictive Churn Analysis:..... | 17 |
| 10.4. Monetary Value Pareto Analysis:..... | 17 |

Executive Summary

This customer segmentation analysis for an online retail business has successfully identified two distinct customer groups using RFM (Recency, Frequency, Monetary) modelling and advanced clustering techniques. Key findings include:

Cluster 0: A larger group of low-value, inactive customers (1,924 customers)

Cluster 1: A smaller group of high-value, active customers (2,267 customers)

Identification of exceptionally high-value customers, including an anomalous customer ID -1.0

Clear patterns in customer retention and lifecycle through cohort analysis

Recommendations focus on targeted marketing strategies, personalized customer experiences, and improved retention efforts for each segment. Implementation of these strategies is expected to significantly enhance customer lifetime value and overall business performance.

1. Introduction

This document presents a comprehensive analysis of customer segmentation for an online retail business. The project aims to identify distinct customer groups based on their purchasing behaviour, allowing for targeted marketing strategies and improved customer relationship management. By leveraging advanced analytics techniques, we seek to uncover actionable insights that can drive business growth and enhance customer satisfaction.

2. Methodology

The analysis employs the RFM (Recency, Frequency, Monetary) model as the foundation for customer segmentation. Two clustering algorithms, K-means and DBSCAN, were applied to identify customer segments. Additionally, a cohort analysis was performed to understand customer retention patterns over time. The project also incorporates Customer Lifetime Value (CLV) calculations to identify and analyse high-value customers.

2.1 Expanded Methodology

Data Preparation:

Source: Online Retail dataset

Cleaning: Handled missing values, removed duplicates, addressed outliers using IQR method

Feature Engineering: Created RFM features (Recency, Frequency, Monetary)

RFM Analysis:

Recency: Days since last purchase

Frequency: Number of purchases

Monetary: Total spend

Log transformation applied to handle skewness

Clustering Techniques:

K-means:

- Applied to RFM scores
- Optimal clusters (k=2) determined using silhouette score analysis

DBSCAN:

- Used to identify density-based clusters and potential outliers
- Parameters: eps=0.5, min_samples=5

Customer Lifetime Value (CLV) Calculation:

Simple CLV = (Frequency * Monetary) / Recency

Cohort Analysis:

Grouped customers by first purchase date

Analyzed retention rates over time

Tools Used:

Python with libraries: pandas, numpy, sklearn, matplotlib, seaborn.

3. Clustering Analysis

3.1 K-means Clustering

K-means clustering was applied to segment customers based on their RFM scores. The analysis resulted in two distinct clusters.

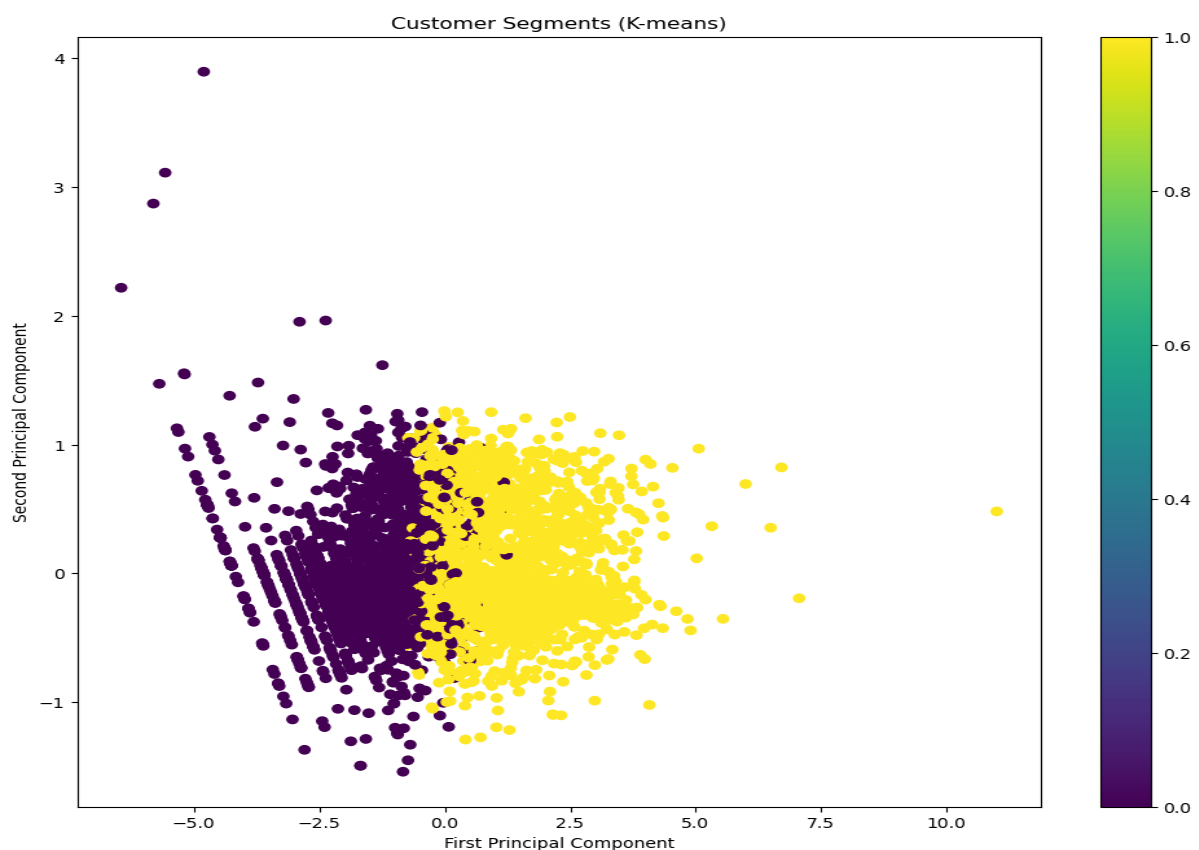


Figure 1. customer segments k-means

Key observations:

- The clusters are well-separated, indicating distinct customer segments.
- Cluster 0 (purple) appears to be larger and more spread out, suggesting a diverse group of customers.
- Cluster 1 (yellow) is more concentrated, potentially representing a more homogeneous group of customers.
- The optimal number of clusters was determined to be 2, based on silhouette score analysis.

3.2 DBSCAN Clustering

DBSCAN clustering was used as an alternative method to identify clusters based on density.

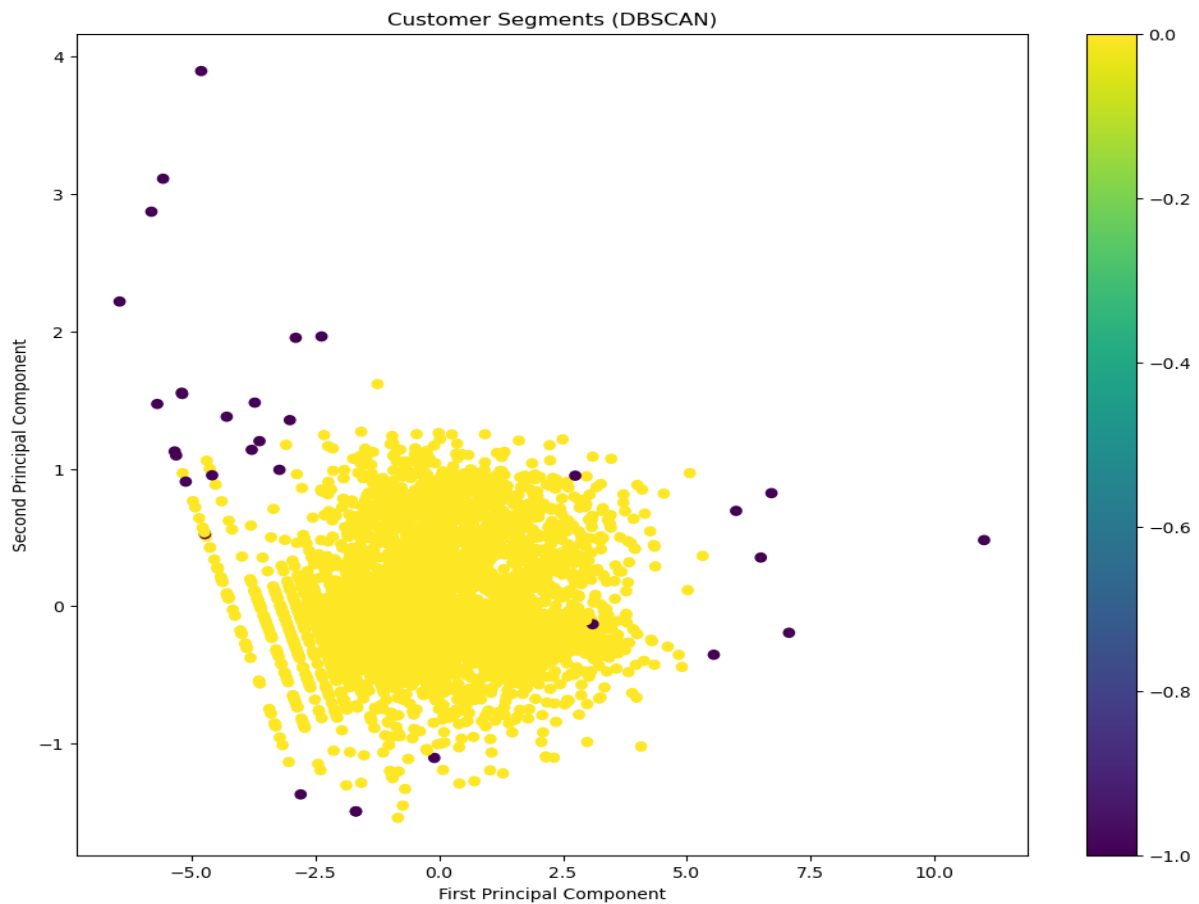


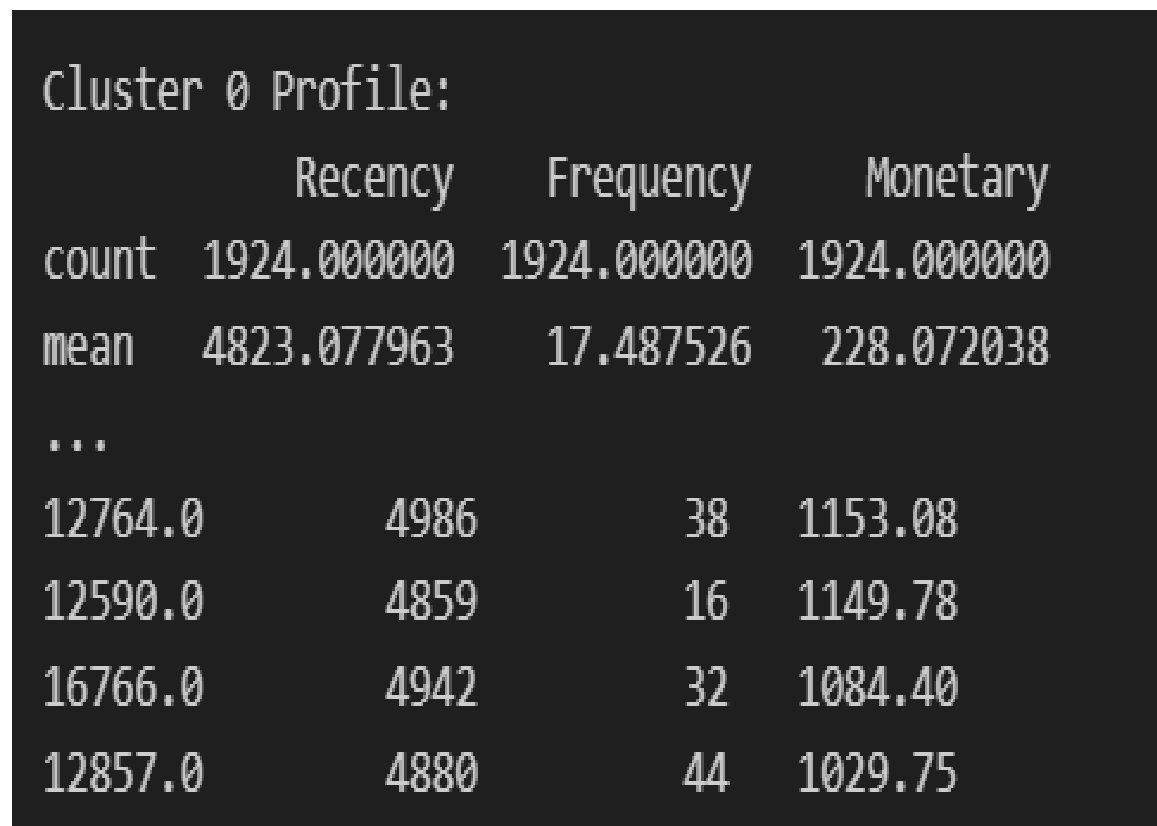
Figure 2. customer segments DBSCAN

Key observations:

- DBSCAN identified a large core cluster (yellow) and several smaller, outlier groups (purple).
- The outliers are spread across the plot, suggesting the presence of unique customer behaviours that don't fit into the main cluster.
- The core cluster (yellow) is dense and well-defined, representing the majority of customers with similar behaviours.

3.3 Cluster Profiling

3.3.1 Cluster 0 Profile



| | Recency | Frequency | Monetary |
|---------|-------------|-------------|-------------|
| count | 1924.000000 | 1924.000000 | 1924.000000 |
| mean | 4823.077963 | 17.487526 | 228.072038 |
| ... | | | |
| 12764.0 | 4986 | 38 | 1153.08 |
| 12590.0 | 4859 | 16 | 1149.78 |
| 16766.0 | 4942 | 32 | 1084.40 |
| 12857.0 | 4880 | 44 | 1029.75 |

Figure 3. cluster profile 0

- Size: 1,924 customers
- Recency: Average of 4,823 days since last purchase, indicating these are mostly inactive customers.
- Frequency: Average of 17.5 purchases, suggesting low engagement.
- Monetary: Average spend of £228.07, indicating lower-value customers.

This cluster likely represents inactive or churned customers who haven't made purchases in a long time and have low overall engagement.

3.3.2 Cluster 1 Profile

| Cluster 1 Profile: | | | |
|--------------------|-------------|---------------|---------------|
| | Recency | Frequency | Monetary |
| count | 2267.000000 | 2267.000000 | 2267.000000 |
| mean | 4707.194971 | 188.218350 | 2059.859824 |
| std | 42.494960 | 2492.523379 | 18351.507887 |
| min | 4669.000000 | 14.000000 | 133.250000 |
| 25% | 4678.000000 | 48.000000 | 602.810000 |
| 50% | 4693.000000 | 81.000000 | 1020.300000 |
| 75% | 4722.000000 | 144.500000 | 1946.160000 |
| max | 4984.000000 | 118120.000000 | 865571.280000 |

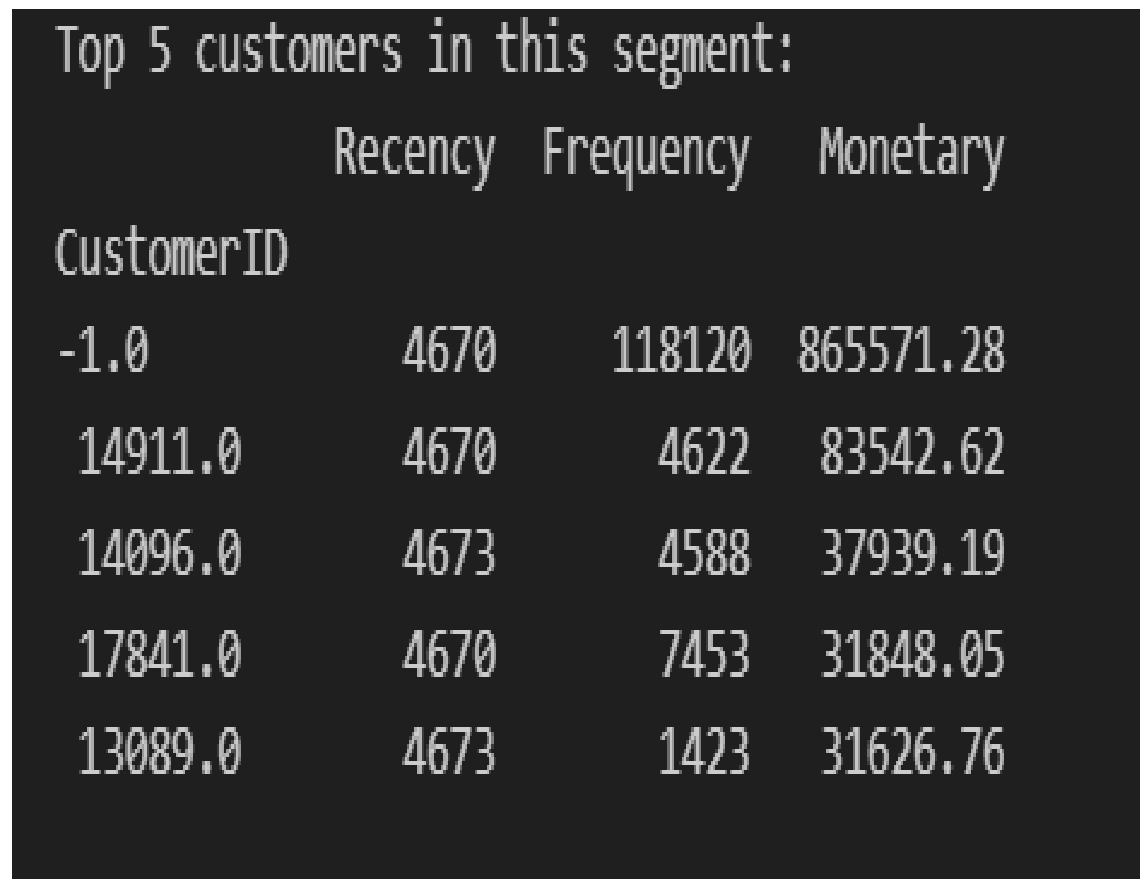
Figure 4. cluster profile 1

- Size: 2,267 customers
- Recency: Average of 4,707 days since last purchase, slightly better than Cluster 0 but still high.
- Frequency: Average of 188 purchases, significantly higher than Cluster 0.
- Monetary: Average spend of £2,059.86, much higher than Cluster 0.

This cluster represents more active and valuable customers. The high standard deviation in frequency (2,492) and monetary value (£18,351) suggests a wide range of behaviours within this cluster, possibly including some very high-value customers.

4. Top Customers Analysis

The top 5 customers identified in Cluster 1 represent the most valuable clients for the business. They show high engagement (frequency) and high monetary value, despite having high recency values. This suggests that while they haven't made recent purchases, their historical value to the company is significant.



| Top 5 customers in this segment: | | | |
|----------------------------------|---------|-----------|-----------|
| | Recency | Frequency | Monetary |
| CustomerID | | | |
| -1.0 | 4670 | 118120 | 865571.28 |
| 14911.0 | 4670 | 4622 | 83542.62 |
| 14096.0 | 4673 | 4588 | 37939.19 |
| 17841.0 | 4670 | 7453 | 31848.05 |
| 13089.0 | 4673 | 1423 | 31626.76 |

Figure 5. top 5 customer in the segment

Key observations:

- Customer ID -1.0 shows exceptionally high frequency and monetary values, warranting further investigation.
- There's significant variation in frequency and monetary values among top customers, indicating diverse high-value customer behaviours.
- The high recency values for all top customers suggest a need for re-engagement strategies even for the most valuable clients.

5. Customer Lifetime Value (CLV) Analysis

Top 10 customers by CLV:

```
Optimal number of clusters: 2

Top 10 Customers by CLV:
```

| CustomerID | Recency | Frequency | Monetary | CLV | KMeans_Cluster |
|------------|---------|-----------|-----------|--------------|----------------|
| -1.0 | 4670 | 118120 | 865571.28 | 2.189321e+07 | 1 |
| 14911.0 | 4670 | 4622 | 83542.62 | 8.268394e+04 | 1 |
| 17841.0 | 4670 | 7453 | 31848.05 | 5.082731e+04 | 1 |
| 14096.0 | 4673 | 4588 | 37939.19 | 3.724909e+04 | 1 |
| 12748.0 | 4669 | 4121 | 20938.31 | 1.848078e+04 | 1 |
| 13089.0 | 4673 | 1423 | 31626.76 | 9.630832e+03 | 1 |
| 15311.0 | 4669 | 2033 | 16282.59 | 7.089849e+03 | 1 |
| 14606.0 | 4670 | 2605 | 8865.82 | 4.945495e+03 | 1 |
| 15039.0 | 4679 | 1378 | 15638.83 | 4.605751e+03 | 1 |
| 14298.0 | 4677 | 849 | 21246.01 | 3.856716e+03 | 1 |

Figure 6. top 10 customers by CLV

Key observations:

1. All top 10 customers by CLV belong to Cluster 1, confirming it as the high-value cluster.
2. The customer with ID -1.0 has an exceptionally high CLV, over 260 times higher than the next highest customer.
3. There's a significant drop in CLV from the first to the second customer, indicating a highly skewed distribution of customer value.
4. Recency values are similar across all top customers, suggesting these high-value customers haven't made very recent purchases.

6. Cohort Analysis

The cohort analysis provides insights into customer retention over time:

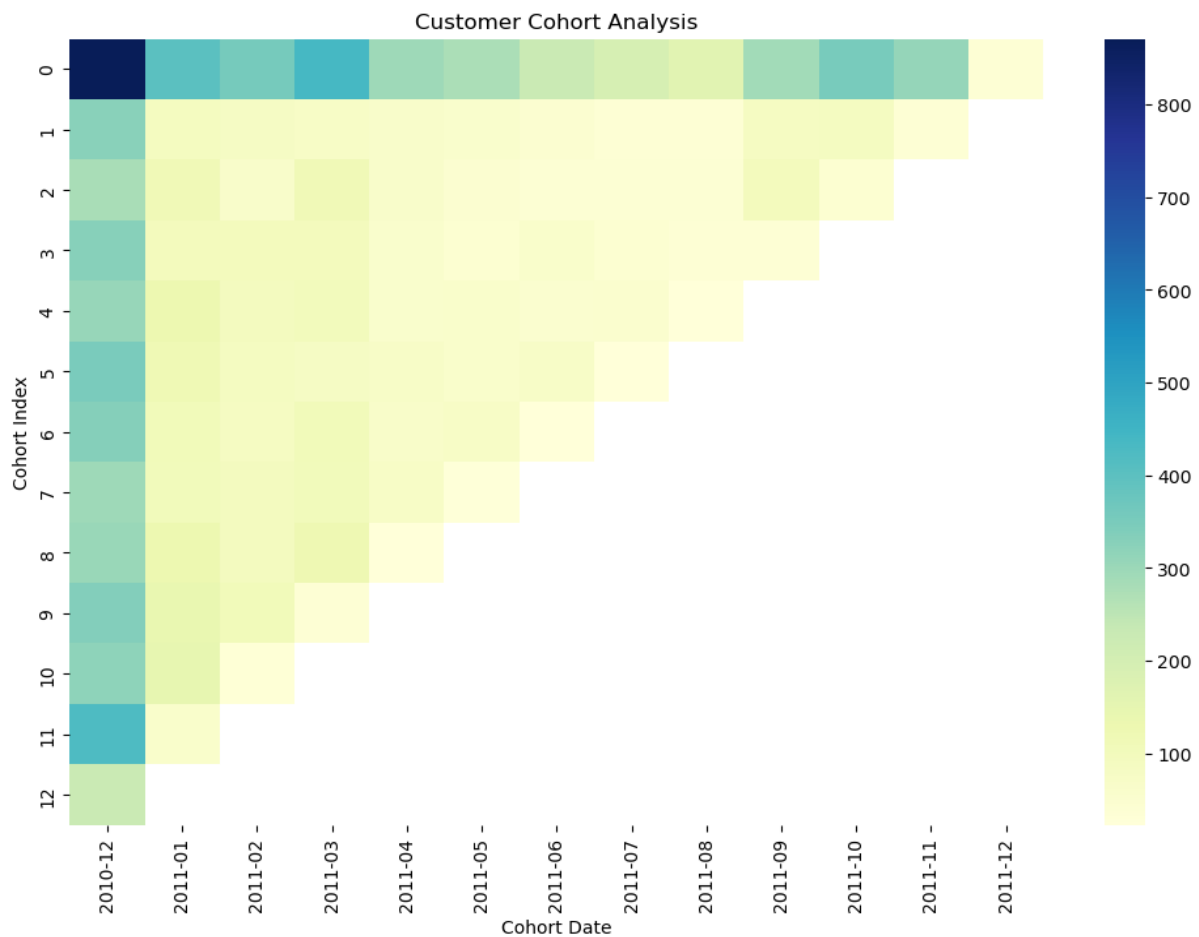


Figure 7. customer cohort analysis

- The x-axis represents cohort dates (when customers first made a purchase), while the y-axis shows the cohort index (months since first purchase).
- Darker colours indicate higher retention rates.
- There's a clear pattern of decreasing retention over time (moving down each column), which is typical in retail.
- Some cohorts (e.g., those starting in late 2011) show better long-term retention, indicated by darker colours persisting further down the column.
- The most recent cohorts (right side of the chart) show high initial retention but it's too early to assess their long-term behaviour.

7. Conclusions and Recommendations

The customer segmentation analysis has successfully identified distinct customer groups and provided valuable insights to inform targeted marketing strategies and improve customer relationship management. Here are the key conclusions and recommendations:

7.1. Customer Segmentation Insights:

- We identified two primary customer segments: a larger group of low-value, inactive customers (Cluster 0) and a smaller group of high-value, active customers (Cluster 1).
- DBSCAN analysis revealed a core group of customers with similar behaviours and several outlier groups, potentially representing niche market segments.

7.2. Targeted Marketing Strategies:

- For Cluster 0 (Low-value, Inactive Customers):
 - Implement targeted re-engagement campaigns using personalized email marketing with special "welcome back" offers.
 - Develop a "win-back" program featuring products similar to their past purchases.
- For Cluster 1 (High-value, Active Customers):
 - Create a tiered loyalty program with exclusive benefits for top spenders.
 - Implement cross-selling and upselling strategies based on purchase history and product affinities.

7.3. Customer Retention and Lifecycle Management:

- Develop cohort-specific retention strategies, focusing on critical periods identified in the cohort analysis.

- Implement a VIP program for the top 10% of customers by CLV, offering personalized services, early access to new products, and exclusive events.
- Create an "at-risk customer" intervention program based on the high recency values observed among top customers.

7.4. Personalization and Customer Experience:

- Utilize RFM profiles to tailor product recommendations, communication frequency, and channel preferences for each segment.
- Develop segment-specific customer journeys, ensuring touchpoints are optimized for each group's characteristics and preferences.

7.5. Data Quality and Customer Identification:

- Investigate and resolve the anomaly of customer ID -1.0, which shows exceptionally high values.
- Implement measures to reduce anonymous purchases and increase customer identification, possibly through incentivized account creation.

7.6. Continuous Improvement and Monitoring:

- Establish a quarterly review process to update the segmentation analysis and assess the effectiveness of targeted strategies.
- Develop KPIs for each customer segment to track improvements in engagement, retention, and CLV.

8. Limitations

Data Timeframe: The analysis is limited to the period covered in the dataset. Long-term trends may not be fully captured.

Anomalous Data: Customer ID -1.0 shows exceptionally high values, which may skew overall results if not properly addressed.

Limited Features: The analysis primarily relies on RFM metrics. Additional customer attributes could provide more nuanced segmentation.

Clustering Limitations:

K-means assumes spherical clusters and may not capture complex shapes

DBSCAN results are sensitive to parameter choices

CLV Calculation: The simple CLV formula used may not account for all factors influencing customer value.

External Factors: The analysis doesn't account for external factors (e.g., market conditions, seasonality) that might influence customer behaviour.

9. Implementation Plan

Immediate Actions (0-3 months):

Validate and clean customer ID -1.0 data

Implement basic email re-engagement campaign for Cluster 0

Develop tiered loyalty program for Cluster 1

Short-term (3-6 months):

Launch personalized product recommendation system

Implement "win-back" program for inactive customers

Develop and deploy customer lifecycle stage model

Medium-term (6-12 months):

Roll out comprehensive VIP program for top 10% CLV customers

Implement predictive churn model

Conduct Monetary Value Pareto Analysis and adjust resource allocation

Long-term (12+ months):

Develop automated, real-time customer segmentation system

Implement advanced customer journey optimization

Conduct full review of segmentation strategy effectiveness

Key Performance Indicators (KPIs):

Customer Retention Rate

Average Customer Lifetime Value

Conversion Rate of Re-engagement Campaigns

Revenue per Segment

Net Promoter Score (NPS)

Regular quarterly reviews will be conducted to assess progress and adjust strategies as needed.

10. Suggested Additional Analyses

To further refine our customer segmentation and enhance its business impact, we recommend the following key analyses:

10.1. RFM Score Distribution Visualization:

- Create a 3D scatter plot of Recency, Frequency, and Monetary values to identify micro-segments and optimize targeting strategies.

10.2. Customer Lifecycle Stage Analysis:

- Develop a model to classify customers into lifecycle stages, enabling more nuanced and stage-appropriate marketing interventions.

10.3. Predictive Churn Analysis:

- Build a churn prediction model to proactively identify at-risk customers and implement targeted retention efforts.

10.4. Monetary Value Pareto Analysis:

- Conduct a Pareto analysis to identify the top revenue-contributing customers and optimize resource allocation for maximum ROI.

By implementing these recommendations and conducting the suggested analyses, we can significantly enhance our customer segmentation strategy. This will lead to more effective targeted marketing, improved customer retention, and ultimately, increased customer lifetime value and overall business performance.