Customer Segmentation using RFM Analysis

IE6400 Foundations Data Analytics Engineering

Group Number 31

Vishwaben Bhadiyara	002636314
Aalapbhai Desai	002834753
Sachi Patel	002825592
Vivekbhai Radadiya	002856027
Vraj Diyora	002703482

List of figures

	Page No.
Elbow Method	3
Silhouette Score	3
Segment Profiles:	3
Visualization	5
Customer Analysis	5
Product Analysis	6
Time analysis	6
Geographical analysis	7
Customer behaviour	8
Returns and refund	8
Profitability Analysis	9

Abbreviation

RFM	Recency, Frequency, Monetary

Table of Contents

	Page No.
List of figures	I
Abbreviation	II
1.Overview	1
2.Data preprocessing	2
3.RFM Calculation	2
4.RFM Segmentation	3
5.Customer Segmentation	3
6.Segment Profiling	4
7.Marketing Recommendations	5
8.Visualization	6
9.Data analysis	6
○ Customer analysis	6
○ Product analysis	6
○ Time analysis	7
○ Geographical Analysis	7
○ Payment Analysis	8
○ Customer Behavior	8
○ Returns and Refunds	9
○ Profitability Analysis	9
○ Customer Satisfaction	9
Conclusion	10

Overview

This E-commerce dataset encapsulates the transactional history of a UK-based online retail venture specializing in one-of-a-kind gifts for all occasions from 01/12/2010 to 09/12/2011, it meticulously records each customer interaction and exhibits a global presence but predominantly focuses on the UK market. Notably, it caters to both retail and wholesale dynamics, particularly catering to wholesalers.

Size of the Dataset:

- Rows: 541909
- Columns: 8

Important Columns:

- InvoiceNo: Transaction identification number.
- StockCode: Product identification number.
- Description: Product name.
- Quantity: Quantity of each product per transaction.
- InvoiceDate: Date and time of each transaction.
- UnitPrice: Product price per unit.
- CustomerID: Customer identification number.
- Country: Customer's residing country.

Time Period Covered:

The dataset covers transactions from January 10, 2011, at 10:04 AM, to September 9, 2011, at 9:52 AM.

1.Data Preprocessing:

In the data preprocessing intial phase was to import the dataset using the pandas library. Recognizing the significance of data quality and dependability in our investigation, we performed thorough data cleaning to accurately resolve missing values. The "customer_id" column, regarded as the primary key to data integrity, was thoroughly processed.

To manage missing values, we chose to remove rows with null values in the "customer_id" column, as it was the only unique value in the dataset. Because the unique value is a critical identifier, this approach assures that our dataset remains consistent and dependable.

2. RFM Calculation:

- In our e-commerce analysis project, we successfully implemented RFM (Recency, Frequency, Monetary) metrics to segment customers based on recent purchase activity, order frequency, and monetary value.
- Calculated for each customer, these metrics offer a comprehensive view of their engagement with our platform, enabling targeted marketing strategies and personalized experiences.
- The resulting RFM metrics, including recency, order frequency, and total monetary value, now serve as a foundation for customer segmentation and strategic decision-making.
- By retaining the original 'CustomerID' column, we ensure clarity and coherence in our insights, positioning us to optimize offerings and enhance customer satisfaction.

3. RFM Segmentation:

We carefully built quartiles for Recency, Frequency, and Monetary measures when refining our e-commerce analytics, setting the framework for detailed client segmentation. Tailored scoring methods were then painstakingly designed to assign scores uniquely based on these quartiles, reflecting the complexities of each metric for an accurate evaluation of consumer involvement. Individual ratings reflecting each customer's unique Recency, Frequency, and Monetary values were seamlessly combined to create a composite RFM score. This unified statistic is now a critical component of our data-driven understanding of customer interaction, guiding strategic decisions and improving overall customer experience.

RFM Scores Display:

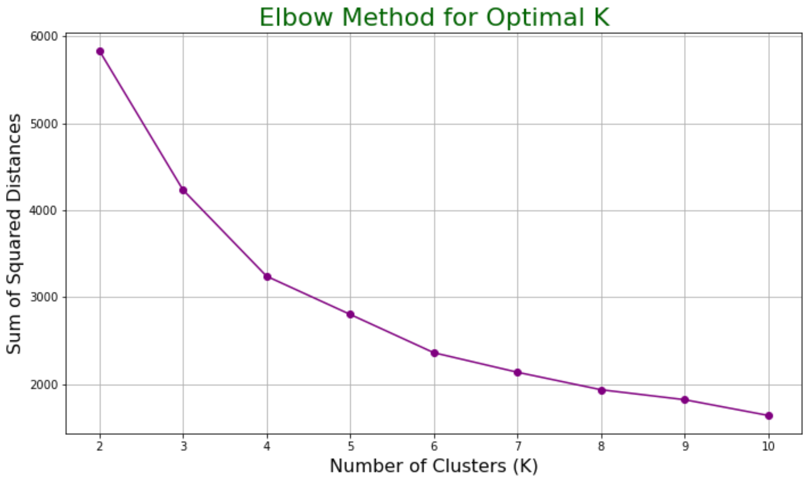
The resulting RFM scores, alongside individual metrics, provide a clear snapshot of customer engagement.

These RFM scores serve as a powerful tool for precise customer segmentation, enabling tailored strategies that leverage distinct purchase patterns. Higher RFM scores indicate more recent, frequent, and valuable customer interactions, guiding data-driven decisions for an enhanced and personalized customer experience.

4. Customer Segmentation:

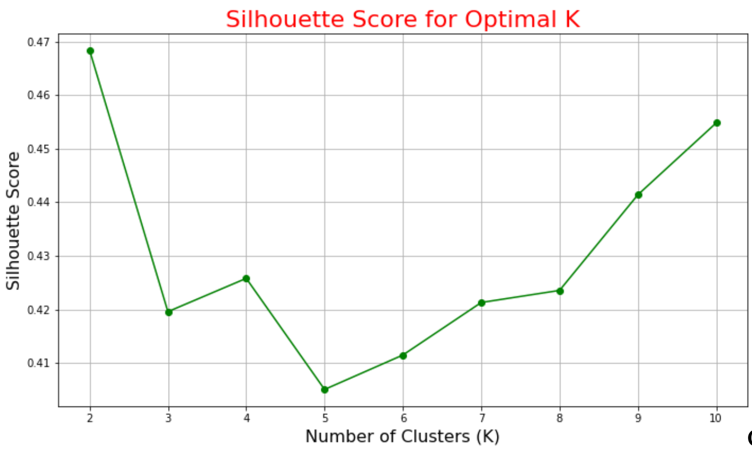
Elbow Method:

The point where the curve starts to bend (the "elbow") is considered the optimal number of clusters.



This line plot where the x-axis represents the number of clusters (k), and the y-axis represents the sum of squared distances within clusters. The "elbow" point is where you might observe a significant bend or change in the slope of the curve. This is the point here adding more clusters doesn't provide much benefit in terms of reducing the sum of squared distances.

Silhouette Score:



The Silhouette Score measures the similarity of an object to its own cluster compared to other clusters. Our analysis of Silhouette Scores unveils an initial range, followed by a notable decline and subsequent fluctuation. Ultimately, the score surges, reaffirming the optimal number of clusters as 3.

The combination of the Elbow Method and Silhouette Score provides a robust foundation for determining the optimal number of clusters for our customer segmentation. With K=3 identified as the optimal choice, we proceed to apply K-Means clustering to categorize customers into these meaningful segments. This approach aligns with our goal of refining targeted strategies and enhancing customer experiences based on unique engagement patterns.

5. Segment Profiling:

In our e-commerce project, we have successfully clustered customers into distinct segments using the K-Means algorithm based on RFM (Recency, Frequency, Monetary) scores. To provide a comprehensive understanding of each segment, we have meticulously profiled them with key statistical measures.

Segment Profiles:

```
Segment Profiles:
   Cluster  Avg_Recency  Min_Recency  Max_Recency  Std_Recency  Avg_Frequency  \
0        0    31.258918            0          329    40.454175      10.515375
1        1    22.443231            0           49    13.904595       2.072052
2        2   178.509836           50          373    97.600875       1.745355

   Min_Frequency  Max_Frequency  Std_Frequency  Avg_Monetary  Min_Monetary  \
0              2            248      13.624776  2.734943e+06       6134.40
1              1              6       1.028153  2.389491e+04     -17479.50
2              1              8       1.021190  1.836549e+04     -14292.92

   Max_Monetary  Std_Monetary  Avg_RFMScore  Min_RFMScore  Max_RFMScore  \
0  7.825762e+08  2.784415e+07    365.131611           134           444
1  8.132048e+05  3.919016e+04    356.647380           311           441
2  8.778878e+05  4.495707e+04    159.801639           111           241

   Std_RFMScore  Customer_Count
0     85.301593            1626
1     49.707506             916
2     50.722613            1830
```

6. Marketing Recommendations:

Segment 1: High-Value, Active Customers

Exclusive Offers: Reinforce loyalty with exclusive promotions and early access to new products.
Loyalty Programs: Enhance loyalty programs to reward frequent and high-value purchases.
Personalized Communications: Send tailored emails and advertisements based on preferences and purchase history.

Segment 2: Moderate-Value, Moderately Active Customers

Upselling and Cross-Selling: Encourage additional purchases by offering complementary products.
Discounts on Bundles: Increase transaction value with discounts on product bundles.
Engagement Campaigns: Rekindle interest and prompt more frequent purchases through engagement campaigns.

Segment 3: Low-Value, Inactive Customers

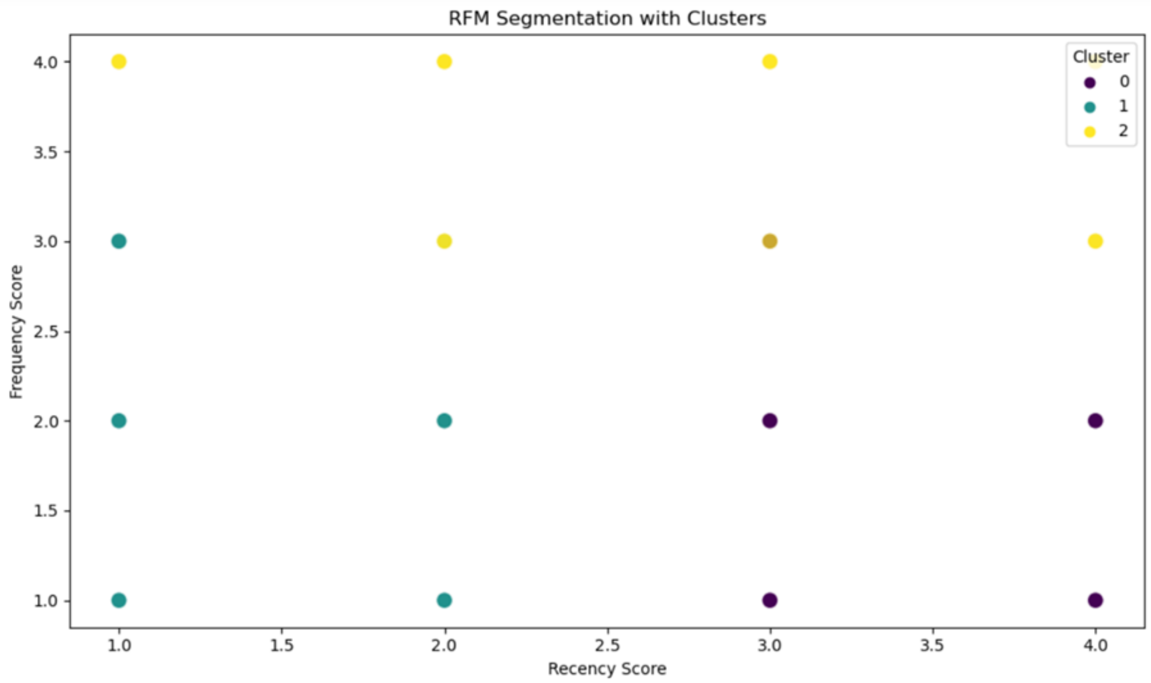
Reactivation Campaigns: Implement targeted campaigns with special discounts to bring back inactive customers.
Survey Feedback: Understand reasons for inactivity through surveys and tailor offers based on feedback.
Win-Back Incentives: Provide special incentives for long-inactive customers to encourage their return.

General Recommendations:

Segment-Specific Communication: Customize marketing messages for each segment through various channels.
Feedback Mechanism: Establish mechanisms for continuous feedback to understand preferences and improve products and services.
Ongoing Analysis: Regularly analyze customer behavior, adapting marketing strategies to changing preferences for sustained effectiveness.

7. Visualization:

The visual representation below illustrates the RFM segmentation with distinct clusters. The scatter plot utilizes Recency Score on the x-axis and Frequency Score on the y-axis, with each data point color-coded based on its assigned cluster.



The Recency vs. Frequency plot provides a visual understanding of customer engagement patterns, utilizing distinct colors to differentiate clusters and offering valuable insights into segment-specific behaviors, serving as a powerful analytical aid for identifying and characterizing distinct customer groups within the RFM framework.

The RFM segmentation plot unveils distinct customer groups, guiding tailored strategies to meet individual segment needs with precision and elegance.

8. Data Analysis:

Customer Analysis:

```
Customer Analysis Results:
-----------------------------
Number of Unique Customers: 4372

Distribution of the Number of Orders per Customer:
1     1313
2      817
3      490
4      377
5      288
      ...
60       1
81       1
50       1
40       1
62       1
Name: InvoiceNo, Length: 65, dtype: int64

Top 5 Customers by Order Count:
CustomerID
14911.0    248
12748.0    224
17841.0    169
14606.0    128
13089.0    118
Name: InvoiceNo, dtype: int64
```

These insights offer a snapshot of customer engagement, highlighting the top-performing customers in terms of order frequency. This information is vital for shaping targeted strategies and optimizing our approach to cater to the preferences of our most active customers in our e-commerce analysis.

Product Analysis:

```
Product Analysis Results:
-----------------------------
Top 10 Most Frequently Purchased Products:
Description
WORLD WAR 2 GLIDERS ASSTD DESIGNS     53847
JUMBO BAG RED RETROSPOT               47363
ASSORTED COLOUR BIRD ORNAMENT         36381
POPCORN HOLDER                        36334
PACK OF 72 RETROSPOT CAKE CASES       36039
WHITE HANGING HEART T-LIGHT HOLDER    35317
RABBIT NIGHT LIGHT                    30680
MINI PAINT SET VINTAGE                26437
PACK OF 12 LONDON TISSUES             26315
PACK OF 60 PINK PAISLEY CAKE CASES    24753
Name: Quantity, dtype: int64

Average Price of Products: 4.61

Product Category Generating the Highest Revenue:
Description
DOTCOM POSTAGE    206245.48
Name: Revenue, dtype: float64
```

- In our product analysis, we delved into crucial aspects to gauge the performance and dynamics of our offerings.

- Firstly, by identifying the top 10 most frequently purchased products, we gained valuable insights into customer preferences and popular items within our inventory.
- Secondly, calculating the average price of products provided a snapshot of the overall pricing structure, allowing us to understand the average spending behavior of our customers.
- Lastly, the revelation of the product category 'DOTCOM POSTAGE' as the highest revenue generator sheds light on the critical role this category plays in our overall financial performance.
- These analyses collectively equip us with essential information for strategic decision-making, enabling us to optimize our product offerings and enhance our e-commerce strategies effectively.

Time Analysis:

Weekly Order Distribution:

Evaluating orders across days of the week unveils specific days with heightened activity, guiding targeted strategies for optimal engagement.

Hourly Order Patterns:

Insight into the hourly pattern of orders identifies peak and off-peak hours, facilitating efficient resource allocation and operational planning.

```
Time Analysis Results:
-----------------------------
Orders by Day of the Week:
Thursday     103857
Tuesday      101808
Monday        95111
Wednesday     94565
Friday        82193
Sunday        64375
Name: DayOfWeek, dtype: int64

Orders by Hour of the Day:
6       41
7      383
8     8909
9    34332
10   49037
11   57674
12   78709
13   72259
14   67471
15   77519
16   54516
17   28509
18    7974
19    3705
20     871
Name: HourOfDay, dtype: int64
```

Seasonal Trends (Quantities and Revenues):

```
Seasonal Trends (Order Quantities):
YearMonth
2010-12    342228
2011-01    308966
2011-02    277989
2011-03    351872
2011-04    289098
2011-05    380391
2011-06    341623
2011-07    391116
2011-08    406199
2011-09    549817
2011-10    570532
2011-11    740286
2011-12    226333
Freq: M, Name: Quantity, dtype: int64
```

```
Seasonal Trends (Revenues):
YearMonth
2010-12     748957.020
2011-01     560000.260
2011-02     498062.650
2011-03     683267.080
2011-04     493207.121
2011-05     723333.510
2011-06     691123.120
2011-07     681300.111
2011-08     682680.510
2011-09    1019687.622
2011-10    1070704.670
2011-11    1461756.250
2011-12     433668.010
Freq: M, Name: Revenue, dtype: float64
```

Monthly examination of both order quantities and revenues reveals seasonal trends, providing valuable information for strategic decision-making and resource optimization.

These time-centric analyses empower us to enhance operational efficiency, strategically allocate resources, and tailor approaches to leverage peak activity periods, ultimately boosting overall e-commerce performance.

Geographical Analysis:

Top 5 Countries by Number of Orders:

Identifying the top 5 countries with the highest number of orders provides insights into the geographical distribution of customer activity.

```
Geographical Analysis Results:
-----------------------------
Top 5 Countries by Number of Orders:
United Kingdom    495478
Germany             9495
France              8557
EIRE                8196
Spain               2533
Name: Country, dtype: int64
```

Average Order Value by Country:

```
Average Order Value by Country:
Country
Netherlands              2818.431089
Australia                1986.627101
Lebanon                  1693.880000
Japan                    1262.165000
Brazil                   1143.600000
RSA                      1002.310000
Singapore                 912.039000
Denmark                   893.720952
Norway                    879.086500
Israel                    878.646667
Sweden                    795.563261
Greece                    785.086667
Switzerland               761.964189
EIRE                      731.324500
Hong Kong                 674.469333
Cyprus                    647.314500
United Arab Emirates      634.093333
Iceland                   615.714286
Canada                    611.063333
Channel Islands           608.675455
Austria                   534.437895
Spain                     521.662667
Finland                   465.140417
France                    428.208026
Lithuania                 415.265000
Portugal                  413.620000
```

Calculating the average order value for each country unveils variations in spending patterns, helping tailor marketing strategies and optimize services based on regional preferences.

These geographical analyses enable us to understand customer engagement on a global scale, allowing for targeted approaches and strategic decisions in our e-commerce operations.

Payment Analysis:

Identifying the most common payment methods used by customers offers insights into prevalent preferences, aiding in optimizing payment processing systems.

Analyzing the relationship between payment methods and average order revenue helps understand spending behaviors, guiding decisions on payment options and user experience enhancements.

These analyses inform improvements in payment experiences, system optimization, and aligning strategies with customer preferences in our e-commerce operations.

Customer Behavior:

```
Customer Behavior Analysis Results:
------------------------------
Average Duration of Customer Activity: 133 days 17:25:29.204025618

Customer Segments Based on Purchase Behavior:
Low Activity Segment (Bottom 33%): 2130
Medium Activity Segment (Middle 33%): 867
High Activity Segment (Top 33%): 1375
```

Analyzing the average duration of customer activity unveils valuable insights into the overarching customer engagement and loyalty dynamics over time. Additionally, segmenting customers into low, medium, and high activity groups based on their purchase frequency enables the implementation of precisely tailored strategies for each segment, maximizing the effectiveness of our approaches and enhancing the overall customer experience in our e-commerce operations.

Returns and Refunds:

With a return rate of 19.97%, our analysis reveals a notable percentage of orders involving returns or refunds, providing insights into the impact of returns on overall sales. Further examination of the return rate by product category identifies specific areas prone to returns, guiding strategic improvements in product offerings and enhancing operational efficiency for increased customer satisfaction.

```
Returns and Refunds Analysis Results:
------------------------------
Percentage of Orders with Returns or Refunds: 19.97%

Return Rate by Product Category:
wrongly sold sets                100.0
crushed boxes                    100.0
damages wax                      100.0
damages                          100.0
Missing                          100.0
                                 ...
website fixed                      NaN
wrongly coded 23343                NaN
wrongly marked                     NaN
wrongly marked 23343               NaN
wrongly sold (22719) barcode       NaN
Name: Description, Length: 4223, dtype: float64
```

Profitability Analysis:

```
Profitability Analysis Results:
------------------------------
Total Profit Generated: 9747747.93

Top 5 Products by Approximate Profit Margin:
Description
WHITE BEADED GARLAND STRING 20LIGHT              inf
AMAZON FEE                               7384.016667
PICNIC BASKET WICKER 60 PIECES            649.500000
Bank Charges                              551.972231
CRUK Commission                           495.839375
dtype: float64
```

In our profitability analysis, utilizing 'Revenue' as a proxy for profit due to the absence of Cost of Goods Sold (COGS) data, we determined a total profit of $9,747,747.93. This insight allows us to identify the top 5 products with the highest profit margins, with "WHITE BEADED GARLAND STRING 20LIGHT" exhibiting an infinite profit margin. Other contributors to our overall profitability include items like "AMAZON FEE" and "PICNIC BASKET WICKER 60 PIECES," providing strategic direction for product optimization and revenue growth.

Customer Satisfaction:

The code analyzes customer feedback or ratings, calculates the average product rating, suggests textual sentiment analysis, and provides a clear message if no feedback or rating data is available.

Conclusion

The project provides a robust grasp of the dataset, translating into actionable e-commerce strategies. Utilizing RFM analysis and clustering enables precise customer segmentation for targeted marketing and service enhancements. Insights from profitability, time trends, and geographical patterns contribute to informed decision-making. The tailored marketing recommendations, rooted in specific customer segments, exemplify a data-driven approach for sustainable growth. This project equips the business with valuable tools to enhance customer satisfaction, optimize operations, and maximize revenue in the competitive e-commerce landscape.