CUSTOMER SEGMENTATION

ABSTRACT

This project employs RFM analysis on an online retail dataset to uncover customer segments, providing insights for tailored marketing strategies. The findings offer a roadmap for enhancing customer retention and maximizing revenue in the ecommerce domain.

Akshit Dubey
Jaque Sarnoff
Rashmi Daga
Rishabh Madhani
Shuhan Li

GROUP 33

Introduction

The project leverages the rich dataset provided by the UCI Machine Learning Repository, titled "Online Retail," which captures real transactions spanning from December 1, 2010, to December 9, 2011. This dataset encapsulates the intricacies of a UKbased nonstore online retail business that is both registered and established. The primary focus of this enterprise is the sale of distinctive alloccasion gifts, catering to a diverse customer base that notably includes wholesalers.

Dataset Source: The dataset originates from the UCI Machine Learning Repository, a reputable source for machine learning datasets. The repository hosts the dataset under the title "Online Retail."

Temporal Scope: The dataset encompasses a period from December 1, 2010, to December 9, 2011, providing a comprehensive view of transactional activities over the span of almost a year.

Business Nature: The UKbased online retail business operates as a nonstore entity, emphasizing its digital presence. The company specializes in the sale of unique alloccasion gifts, indicative of a diverse product catalog.

Customer Composition: The dataset sheds light on a customer base that includes wholesalers, suggesting a business model catering to both individual consumers and bulk purchasers.

The use of this dataset in our project allows us to delve into the intricacies of customer behavior and purchasing patterns, utilizing RFM (Recency, Frequency, Monetary) analysis. By applying this powerful segmentation technique, we aim to unravel valuable insights for crafting targeted marketing strategies and enhancing customer retention initiatives in the dynamic landscape of online retail.

Objective

RFM (Recency, Frequency, Monetary) segmentation is a robust methodology employed by businesses to categorize customers according to their recent purchasing behaviour, purchase frequency, and monetary value. This strategic approach enables organizations to develop precise and tailored marketing campaigns, ultimately fostering more effective customer engagement.

RFM Segmentation Significance: The objective of this project is rooted in the recognition of the power and efficacy of RFM segmentation. By dissecting customer behaviour into Recency, Frequency, and Monetary metrics, businesses gain a nuanced understanding of their clientele.

Purpose of RFM Analysis: The primary goal is to perform RFM analysis on the provided dataset derived from actual transactions in the UKbased online retail sector. This analysis aims to segment customers into distinct groups based on their RFM scores, emphasizing the following dimensions:

Recency (R): How recently a customer made a purchase.

Frequency (F): How often a customer makes a purchase.

Monetary Value (M): The total monetary value of a customer's purchases.

Insights for Strategies: The ultimate objective is to derive valuable insights from these customer segments. By understanding the diverse behaviors within the customer base, businesses can formulate targeted marketing strategies and implement effective customer retention initiatives. These insights are pivotal for navigating the competitive landscape of online retail and establishing a sustainable and customercentric business approach.

TASKS

1. Importing the Dataset and Initial Overview:

The analysis commenced with the importation of the dataset from the 'data.csv' file using the Pandas library. This initial step was followed by an exploration of the dataset's structure and information, providing crucial insights into column names, data types, and potential missing values.

Date Conversion and Cleaning:

To facilitate temporal analysis, the 'InvoiceDate' column, containing transaction timestamps, underwent conversion into a DateTime object. Additionally, a new 'Date' column was created, focusing solely on the date component to streamline subsequent recency calculations. A meticulous check for missing values ensured the dataset's completeness and readiness for analysis.

Handling Missing Values and Duplicates:

Ensuring data integrity, records lacking a 'CustomerID' were excluded from the dataset. Further, duplicate records were identified and removed, mitigating redundancy and ensuring a clean dataset for indepth analysis.

2. Returned Transactions and RFM Table Construction:

Transactions marked by negative quantities, indicative of returns, were isolated for a nuanced understanding of customer interactions. The RFM (Recency, Frequency, Monetary) table was then crafted, introducing the 'total_amount' variable to represent the total monetary value of transactions. Frequency calculations were adjusted to account for returned transactions, providing a more accurate depiction of customer engagement.

With a cleaner dataset, RFM metrics were systematically computed. The `pd.to_datetime` function converted `InvoiceDate` into a DateTime object, facilitating temporal analyses. Recency, representing the days since a customer's last purchase, was calculated by finding the difference between the current date and the maximum invoice date. Frequency was derived from the count of positive quantity transactions, and Monetary value, a key metric, was determined by multiplying the `Quantity` and `UnitPrice` columns. These calculations laid the foundation for our subsequent RFM segmentation.

3. RFM Segmentation:

The RFM scores, fundamental to our analysis, were assigned to each customer based on quartiles. This process involved creating bins for each RFM metric and assigning scores, enabling us to categorize customers effectively. The `merge` function consolidated these metrics into a comprehensive RFM table. The resulting dataset underwent outlier removal and normalization for robust analysis. This segmentation approach sets the stage for detailed customer profiling and strategic decisionmaking.

These meticulous steps in data preprocessing and RFM calculation are pivotal, shaping the course of our analysis and paving the way for deeper insights into customer behavior and segmentation.

4. Customer Segmentation:

Cluster Analysis:

Using the KMeans clustering algorithm, we segmented customers into different clusters based on Recency, Frequency, and Monetary (RFM) scores.

Cluster 0 (New Customers):

Recency: Customers in this cluster have likely made purchases recently.

Frequency: They have a lower frequency since they are new and have had less time to make repeat purchases.

Monetary: Their monetary value might be lower as they have only made a few purchases so far.

Profiling: This group represents new customers. Welcome campaigns and initial engagement efforts could help move them to a more loyal status.

Cluster 1 (Loyal Customers):

Recency: These customers have made purchases very recently.

Frequency: They purchase frequently, showing a pattern of regular engagement.

Monetary: They have a high monetary value, contributing significantly to revenue.

Profiling: This segment is the most valuable and should be the focus of retention strategies. They are likely to respond well to loyalty programs and exclusive offers.

Cluster 2 (Champions/HighValue Customers):

Recency: Similar to Cluster 2, these customers have also made recent purchases.

Frequency: The frequency of purchases is high, indicating consistent patronage.

Monetary: They have the highest monetary scores, indicating that they spend a lot.

Profiling: They are the 'Champions' of your customer base and could also be leveraged for advocacy and referral programs.

Cluster 3 (Potential Loyalists/At Risk):

Recency: This cluster might have a mix of recent and less recent purchases.

Frequency: Customers in this cluster could be those who used to purchase frequently but have seen a drop in their purchasing frequency.

Monetary: They might have a moderate to high monetary value, indicating that they were once valuable customers.

Profiling: These could be past loyal customers who are at risk of churning. Targeted reengagement campaigns might be effective.

5. Segment Profiling:

Cluster 0 (New Customers):

Recency: Customers in this cluster have likely made purchases recently.

Frequency: They have a lower frequency since they are new and have had less time to make repeat purchases.

Monetary: Their monetary value might be lower as they have only made a few purchases so far.

Profiling: This group represents new customers. Welcome campaigns and initial engagement efforts could help move them to a more loyal status.

Cluster 1 (Loyal Customers):

Recency: These customers have made purchases very recently.

Frequency: They purchase frequently, showing a pattern of regular engagement.

Monetary: They have a high monetary value, contributing significantly to revenue.

Profiling: This segment is the most valuable and should be the focus of retention strategies. They are likely to respond well to loyalty programs and exclusive offers.

Cluster 2 (Champions/HighValue Customers):

Recency: Similar to Cluster 2, these customers have also made recent purchases.

Frequency: The frequency of purchases is high, indicating consistent patronage.

Monetary: They have the highest monetary scores, indicating that they spend a lot.

Profiling: They are the 'Champions' of your customer base and could also be leveraged for advocacy and referral programs.

Cluster 3 (Potential Loyalists/At Risk):

Recency: This cluster might have a mix of recent and less recent purchases.

Frequency: Customers in this cluster could be those who used to purchase frequently but have seen a drop in their purchasing frequency.

Monetary: They might have a moderate to high monetary value, indicating that they were once valuable customers.

Profiling: These could be past loyal customers who are at risk of churning. Targeted reengagement campaigns might be effective.

Segment profiling provides a deeper understanding of each cluster's characteristics, enabling businesses to tailor their marketing strategies effectively for different customer segments.

6.Marketing Recommendations

1. New Customers

Characteristics: Recently made their first purchases.

Strategies:

1. Welcome Campaigns: Utilize targeted welcome campaigns for new customers, emphasizing the ease of the purchasing process.

2. FirstTime Buyer Offers: Introduce special offers or discounts on subsequent purchases to encourage repeat transactions.

3. Educational Content: Provide resources and tutorials on product usage and benefits to enhance their overall experience.

4. Feedback Solicitation: Actively seek feedback on their initial purchase experience, demonstrating a commitment to continuous improvement.

2. Loyal Customers

Characteristics: Regularly purchase over a long period.

Strategies:

1. Loyalty Programs: Implement a loyalty program with tiered rewards to recognize and appreciate frequent buyers.

2. Exclusive Offers: Provide exclusive promotions or early access to new products, reinforcing their valued customer status.

3. Personalized Communication: Leverage their purchase history for personalized recommendations, enhancing engagement.

4. Customer Appreciation Events: Host exclusive events or offer premium services to express gratitude and reinforce loyalty.

3. Champions/HighValue Customers

Characteristics: Frequent buyers, high spenders, and brand advocates.

Strategies:

1. VIP Treatment: Extend VIP treatment with premium services, prioritized customer support, and exclusive benefits.

2. Referral Programs: Encourage these highvalue customers to become brand advocates through referral programs.

3. Exclusive Access: Provide early access to new products, limited editions, or exclusive merchandise.

4. Community Building: Foster a sense of community by involving them in events, focus groups, or cocreation opportunities.

4. Potential Loyalists/At Risk

Characteristics: Customers who have shopped more than once but are not frequent buyers.

Strategies:

1. Reengagement Campaigns: Deploy targeted campaigns to reengage customers, emphasizing the value proposition and benefits.

2. Personalized Offers: Tailor promotions based on past purchase behavior or items of interest, reigniting their interest.

3. Surveys and Feedback Requests: Seek feedback to understand reasons for infrequent purchases, with a commitment to address concerns.

4. Educate About Products/Services: Provide comprehensive information about the full product range, showcasing additional offerings that align with their preferences.

These tailored marketing strategies, informed by the RFM analysis and customer segmentation, aim to optimize customer retention and revenue generation by addressing the specific needs and behaviors of each segment.

INSIGHTS

1. Data Overview:

1.1Size of the Dataset:

The dataset encompasses a substantial volume, comprising 541,909 rows and 8 columns, offering a detailed and expansive representation of the company's transactions.

1.2 Column Descriptions:

a. Invoice Number: This unique identifier plays a crucial role in accurately documenting income for tax and accounting purposes, tracking items bought in each transaction.

b. Stock Code: Employed for the identification and tracking of inventory items, the Stock Code column is instrumental in maintaining a comprehensive record of products.

c. Description: Providing a concise description of each item, this column enhances the interpretability of the dataset, aiding in the understanding of the nature of each transaction.

d. Quantity: Reflecting the volume of items involved in each transaction, the Quantity column quantifies the extent of customer purchases.

e. Invoice Date: Serving as a timestamp for each transaction, the Invoice Date column captures the temporal aspect of customer interactions, enabling timebased analyses.

f. Unit Price: This column denotes the price per unit of the purchased item, contributing to the overall monetary value of transactions.

g. Customer ID: A unique identifier assigned to each customer, facilitating customerspecific analyses and insights.

h. Country:Reflecting the geographical location of customers, the Country column provides valuable information regarding the market distribution of the company's products.

1.3 Time Period Covered:

The dataset encapsulates a significant temporal span, ranging from January 12th, 2010, to September 12th, 2011. This timeframe provides an extensive window into the company's transactions, allowing for indepth analyses of purchasing patterns and trends over the course of nearly two years.

2. Customer Analysis (Visualization Insights):

2.1 Unique Customers:

The visualization reinforces our understanding of the dataset's customer base, indicating a total of 4,372 unique customers.

2.2 Distribution of Orders per Customer:

The histogram illustrates the distribution of orders per customer, showcasing a central tendency around an average of 93.05 orders. This average provides a benchmark for understanding the typical engagement level of customers with the company.

2.3 Top 5 Customers by Order Count:

The identified top 5 customers, characterized by their high order count, are as follows:

a. Customer ID: 17841

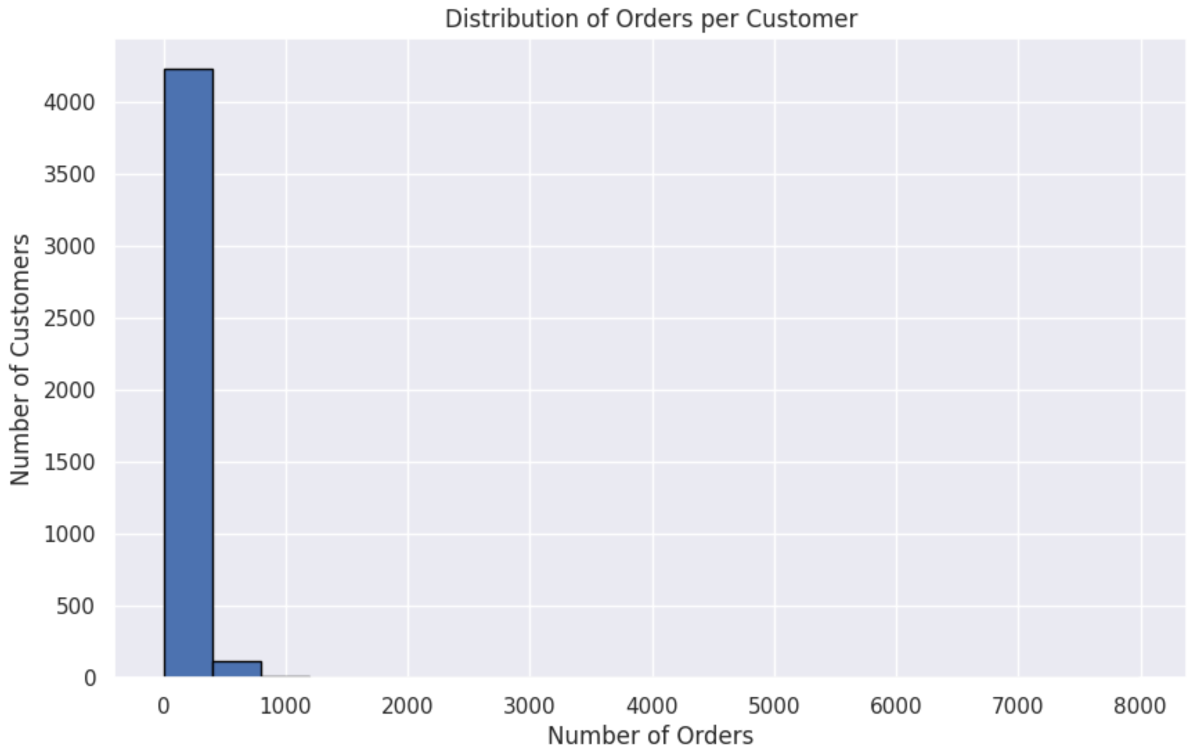
b. Customer ID: 14911

c. Customer ID: 14096

d. Customer ID: 12748

e. Customer ID: 14606

These top customers play a pivotal role in shaping the company's revenue stream and warrant strategic attention for personalized engagement and retention efforts.



These visualization insights complement the quantitative findings, providing a more holistic view of customer dynamics within the dataset. Moving forward, we'll delve into Product Analysis to uncover patterns in product preferences and sales.

3. Product Analysis (Insights):

3.1 Top 10 Most Frequently Purchased Products:

a. The top 10 most frequently purchased products and their respective purchase counts are as follows:

i. 85123A 2313 purchases

ii. 22423 2203 purchases

iii. 85099B 2159 purchases

iv. 47566 1727 purchases

v. 20725 1639 purchases

vi. 84879 1502 purchases

vii. 22720 1477 purchases

viii. 22197 1476 purchases

ix. 21212 1385 purchases

x. 20727 1350 purchases

3.2 Average Price of Products:

The average price of products across the dataset is 4.61 Pounds. This metric provides a baseline understanding of the general pricing structure and customer spending patterns.

3.3 Product Category with Highest Revenue:

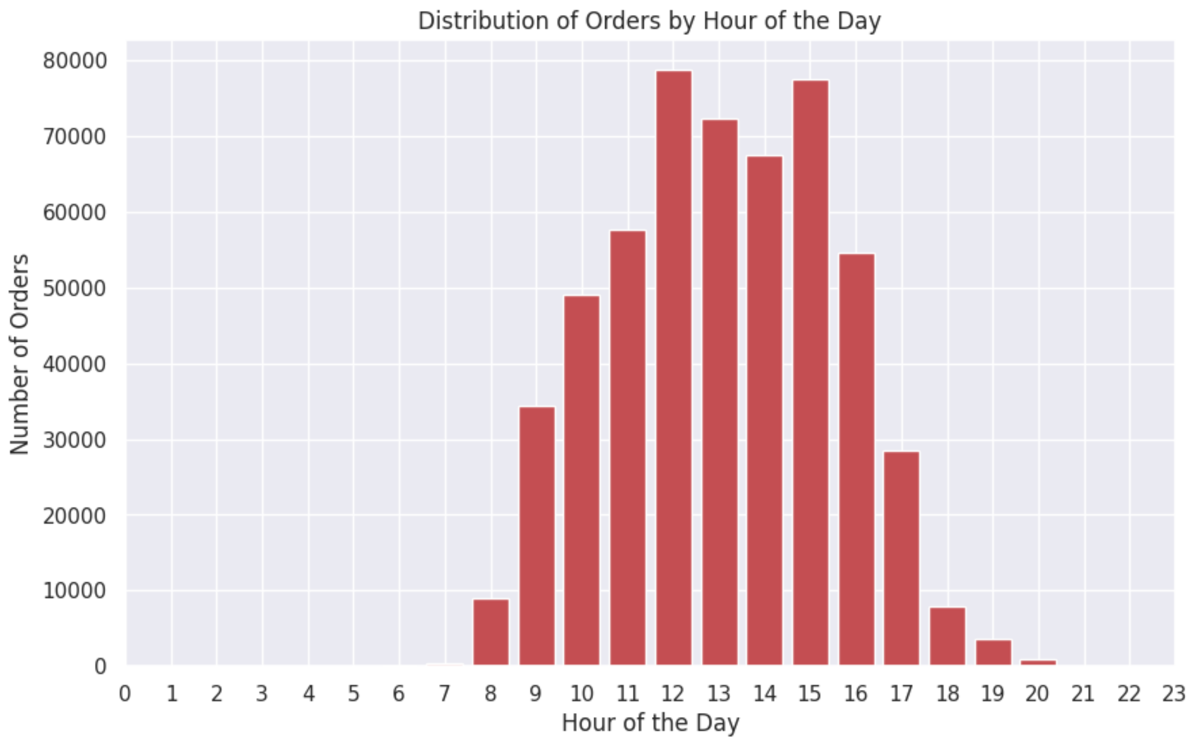
The product category or stock code that generates the highest revenue is the Amazon Fee, contributing a total revenue of £249,042.68 Pounds. This insight is crucial for strategic focus on highimpact product categories that significantly contribute to the company's revenue stream.

These product centric insights offer a deeper understanding of customer preferences, pricing dynamics, and revenue generating product categories, laying the groundwork for targeted marketing and sales strategies. Next, we'll explore Time Analysis to uncover temporal trends in the dataset.

4. Time Analysis :

4.1 Specific Day or Time of Day for Most Orders:

Based on the analysis of the dataset, it is evident that the peak time for order placements is at 12 PM. This peak is closely followed by 3 PM, collectively indicating that the most active period for order placements occurs between noon and 3 PM. This finding is crucial for optimizing resource allocation and ensuring efficient order processing during these peak hours.



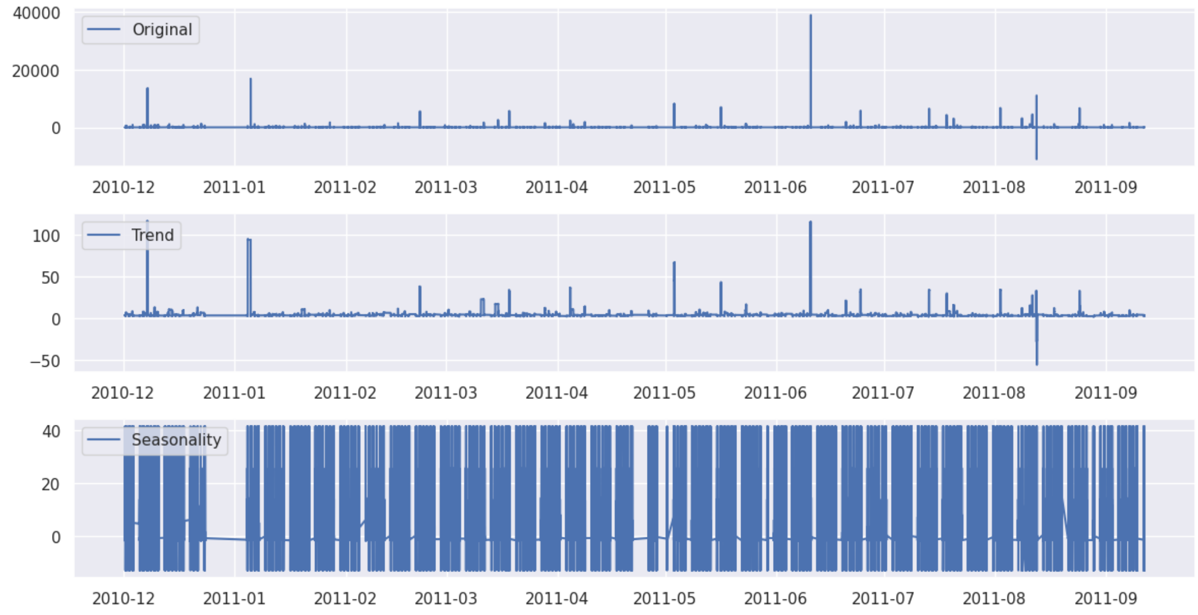
4.2 Average Processing Time:

The dataset's 'OrderProcessingTime' column captured the time duration between the earliest and latest transactions within each invoice. The resulting average order processing time is approximately 23 seconds. This duration is considered a holistic measure, encapsulating the entire order processing cycle. Given that only invoice dates are available, it is reasonable to interpret this time as the total processing time for each order.

This swift processing time aligns with the streamlined nature of an online retail platform, where orders are efficiently managed with a simple click of a button. The recorded 23second average processing time underscores the platform's efficiency in swiftly handling customer orders. This insight provides a comprehensive understanding of the dataset, highlighting the company's commitment to prompt and reliable order fulfillment services.

4.3 Seasonal Trends:

Analyzing the seasonal trends in the dataset reveals interesting patterns. According to the graph/data, the months that exhibit an upward trend in sales are December, January, and June. This suggests that the company experiences increased sales activity during the holiday season, as well as potentially during the midyear period. The overall trend indicates a positive trajectory throughout the year, emphasizing the absence of distinct seasonal downturns.



These insights into the timing of orders, particularly the peak hours, and seasonal trends provide valuable information for optimizing operational processes and tailoring marketing strategies to capitalize on peak sales periods. Next, we'll move on to Geographical Analysis to explore the distribution of orders across different countries.

5. Geographical Analysis (Insights):

5.1 Top 5 Countries with the Highest Number of Orders:

The analysis reveals the top 5 countries with the highest number of orders, showcasing the extent of customer engagement in each country:

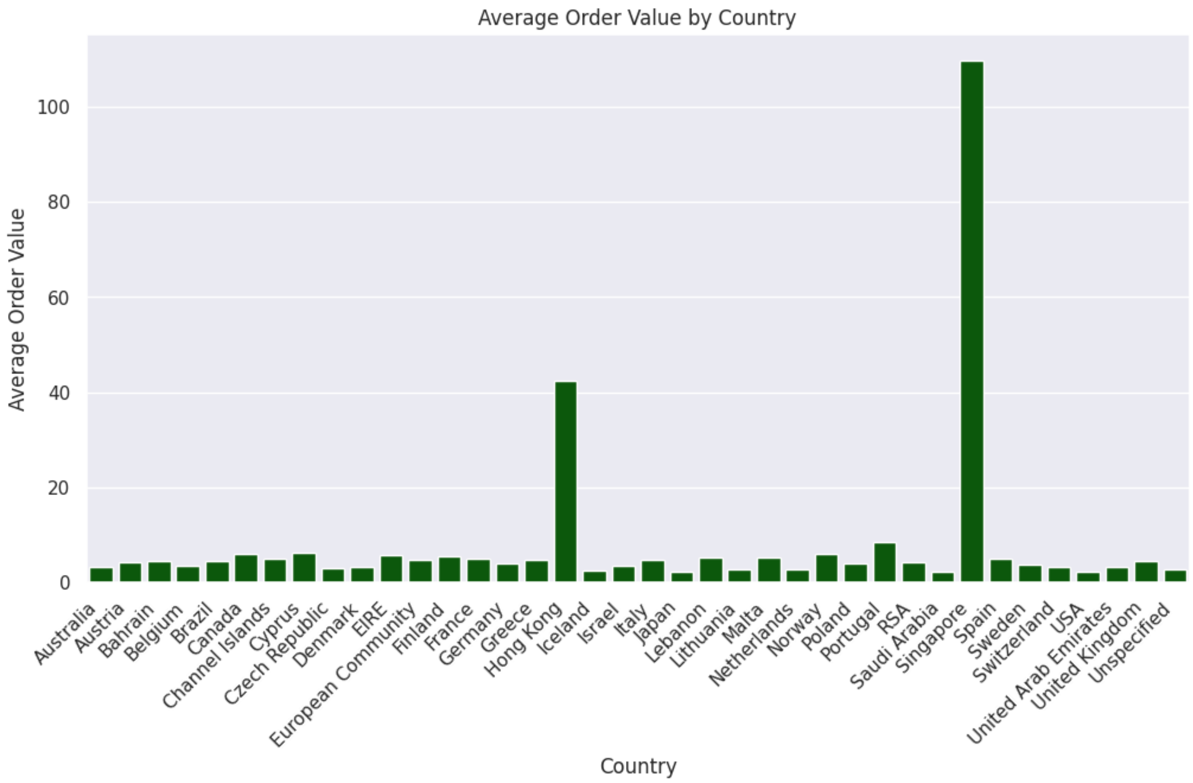
i. United Kingdom 495,478 orders

ii. Germany 9,495 orders

iii. France 8,557 orders

iv. EIRE (Ireland) 8,196 orders

v. Spain 2,533 orders



5.2 Correlation between the Country of the Customer and Average Order Value:

The calculated correlation coefficient is 0.9, indicating a strong positive association. This suggests a robust correlation between the country of the customer and the average order value. In other words, there is a tendency for higher average order values in certain countries, providing valuable insights for targeted marketing strategies and pricing optimization.

These geographical insights enhance our understanding of customer behavior across different regions, facilitating more informed decision making in terms of marketing, sales, and customer engagement strategies.

6. Payment Analysis :

In the given dataset, there is a limitation as it does not provide information regarding the payment method. Unfortunately, this absence of data makes it challenging to identify the most common mode of payment used by customers. Additionally, it hinders our ability to establish any relationship between the payment method and order amounts.

This insight underscores the importance of comprehensive data collection, as the availability of payment related information could offer valuable insights into transaction patterns and contribute to a more nuanced understanding of customer behavior. Moving forward, we'll explore Customer Behavior to gain insights into the duration of customer activity and potential segmentation based on purchase behavior.

Below is a concise algorithm outlining the steps for payment analysis, which we would have followed had the data been provided:

1. Explore Unique Payment Methods:

Identify and list the unique payment methods available in the dataset.

2. Calculate Frequency of Each Payment Method:

Count the frequency of each payment method to determine their prevalence.

Formula: {Frequency of Payment Method} = {Count(Payment Method)}

3. Visualize Payment Method Distribution:

Create a visualization, such as a bar chart, to represent the distribution of payment methods.

4. Analyze Average Order Amount by Payment Method:

Calculate the average order amount for each payment method.

Formula: {Average Order Amount} = {Sum(Order Amount)}/{Count(Orders)}

5. Correlation Analysis:

Conduct correlation analysis between payment methods and other variables

6. Segmentation Based on Payment Behavior:

If applicable, segment customers based on their payment behavior.

7. Customer Behavior (Comprehensive Insight):

7.1Average Customer Activity Duration:

The average duration of customer activity, calculated from the dataset, is approximately 133 days, 17 hours, 25 minutes, and 29 seconds. This metric represents the typical span of engagement from a customer's initial to final transaction.

The calculated average duration of customer activity provides a nuanced understanding of how long, on average, customers actively engage with the platform.

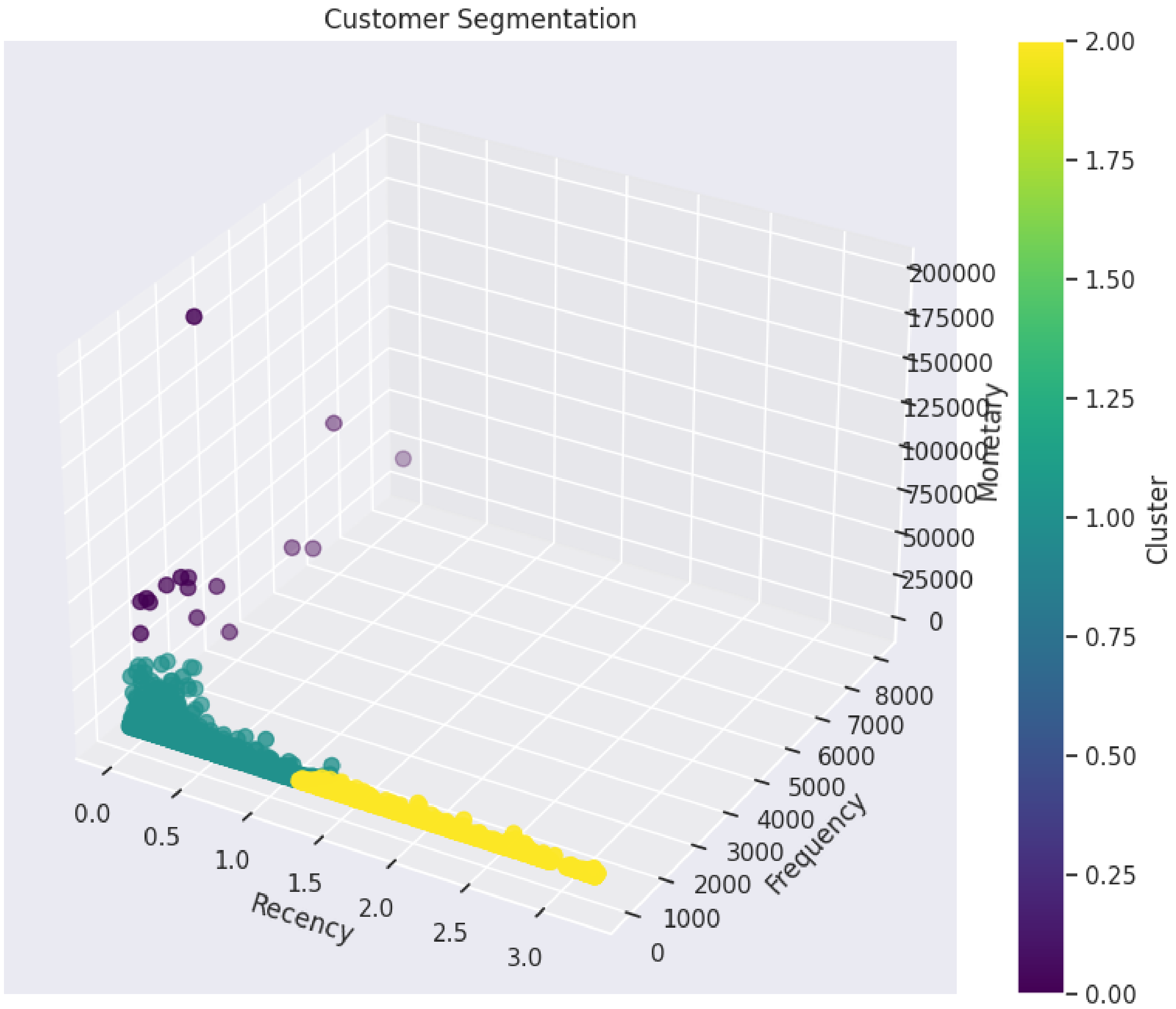
7.2 Customer Segmentation Based on Purchase Behavior

Cluster Characteristics:

1. Cluster 0: Exhibits the highest recency metrics, indicating recent purchases. Focused on current products or offerings.

2. Cluster 1: Customers in this segment made recent purchases, but they exhibit an average monetary contribution. The most significant monetary contributions from Cluster 1 come from less recent purchases, with the most being £50,000 but averaging between £25,000 to £50,000.

3. Cluster 2: Displays the highest monetary metrics, signifying substantial spending. The most significant contributions come from customers in this cluster, with the most being £175,000 but averaging between £50,000 to £75,000 from less recent purchases.



Strategic Implications:

1. Recency Focus (Cluster 0): Targeted marketing for recent transactions—promotions, product recommendations, or loyalty programs.

2. Balanced Spending with Recent and HighValue Contributions (Cluster 1): Despite recent purchases, customers in Cluster 1 demonstrate an average monetary contribution. The most substantial contributions come from less recent purchases, indicating potential highvalue customers. Tailored strategies, such as enticing offers or exclusive promotions, can encourage increased spending and foster loyalty.

3. HighValue and Less Recent (Cluster 2): Customers in Cluster 2, particularly those with the most significant monetary contributions from less recent purchases, represent highvalue segments. Tailored strategies, such as exclusive rewards or personalized engagement, can enhance their overall experience.

8. Returns and Refunds Analysis:

8.1 Percentage of Orders with Returns or Refunds:

The dataset does not explicitly indicate returned or refunded orders. However, orders that were canceled can be considered as a proxy for returns or refunds.

There are 9,288 canceled orders out of a total of 541,909 orders.

Percentage of orders canceled: 1.71%

8.2 Correlation between Product Category and Likelihood of Returns:

Unfortunately, the dataset does not provide information on product categories, hindering the establishment of any relationship between product categories and the likelihood of returns.

Understanding the percentage of orders with returns or refunds helps in assessing the impact on overall business operations and adjusting strategies accordingly.

The absence of product category information limits the ability to identify specific categories prone to returns, preventing targeted interventions in those areas.

9. Profitability Analysis:

9.1 Total Profit Generated:

Determining the total profit necessitates the availability of cost price data for each product, which is currently unavailable in the dataset. Consequently, the calculation of the total profit generated by the company during the dataset's time period is not feasible without the cost price information.

9.2 Top 5 Products with Highest Profit Margins:

In the absence of cost price data, identifying the top 5 products with the highest profit margins remains unattainable. The lack of cost price information limits the assessment of profitability metrics, hindering the determination of products yielding the highest profit margins.

These insights underscore the necessity of cost price data to perform a comprehensive profitability analysis, enabling the assessment of profit margins and identification of topperforming products based on profitability.

10. Customer Satisfaction Analysis:

10.1 Data on Customer Feedback or Ratings:

Regrettably, the dataset does not provide explicit information on customer feedback or ratings for products or services. The absence of such data limits the ability to gauge customer satisfaction through direct feedback.

10.2 Handling the Absence of Data:

If customer feedback or rating data were available, sentiment analysis tools or natural language processing (NLP) techniques could be employed to analyze sentiment or feedback trends.

Sentiment analysis algorithms could assess customer sentiments, categorizing feedback as positive, negative, or neutral.

Trends in sentiment over time or in response to specific events or promotions could be identified, providing valuable insights into customer satisfaction dynamics.

Conclusion

In conclusion, the RFM analysis conducted on the eCommerce dataset yielded valuable insights into customer behavior, offering a foundation for targeted marketing and customer retention strategies. Key findings include the identification of distinct customer segments based on Recency, Frequency, and Monetary metrics, although certain analyses, such as profitability and customer satisfaction, were hindered by data limitations. The dataset's time period spans from January 12th, 2010, to September 12th, 2011, encompassing 541,909 transactions. The top five countries with the highest number of orders are the United Kingdom, Germany, France, EIRE, and Spain.

While the absence of certain data, such as customer feedback or cost price information, limited the depth of the analysis, the RFM segmentation provided actionable insights for targeted marketing and customer engagement. The analysis highlighted the significance of understanding customer purchasing behavior, allowing businesses to tailor their strategies to specific customer segments.

Reiterating the importance of customer segmentation for businesses, RFM analysis enables personalized approaches, leading to more effective marketing campaigns and enhanced customer retention. By categorizing customers based on their recency, frequency, and monetary value, businesses can optimize resource allocation, tailor promotional efforts, and build stronger relationships with their diverse customer base.

Future Work

For future work, several avenues can be explored to enhance the depth and scope of customer segmentation analysis:

1. Refining Segmentation Criteria:

Investigate additional segmentation criteria beyond RFM, such as demographic or behavioral factors, to create more nuanced customer segments.

Explore dynamic segmentation that adapts to changing customer behavior patterns over time.

2. Advanced Machine Learning Techniques:

Implement advanced machine learning algorithms, such as clustering algorithms beyond kmeans, to uncover more complex patterns in customer behavior.

Incorporate predictive modeling techniques to forecast future customer behavior, aiding in proactive customer engagement strategies.

3. Enriching Dataset:

Acquire and integrate additional data sources, including customer feedback, product categories, and cost price information, to enable more comprehensive analyses.

Extend the time frame of the dataset to capture evolving trends and seasonality over a more extended period.

4. Sentiment Analysis:

If customer feedback data becomes available, implement sentiment analysis to gain insights into customer satisfaction trends and sentiments.

Utilize natural language processing (NLP) techniques to extract valuable information from textual feedback.

These future directions aim to refine and expand the current RFM analysis, providing businesses with a more detailed understanding of customer behavior and facilitating targeted strategies for sustained growth and customer satisfaction.