Customer Segmentation using RFM Analysis Report

IE 6400 Foundations for Data Analytics Engineering Project 2

2023/12/01

Group 1

Aliya

Qiuyi Chen

Siyuan Gao

Xinyue Niu

Yiqian Ning

Table of Contents

Executive Summary

This RFM analysis project aimed to stratify eCommerce customers into segments based on their recency, frequency, and monetary values of purchases. The goal was precise targeting for customized marketing. The analysis uncovered four primary clusters:

1. Champions (Cluster 0) - Recent, frequent, high-value buyers
2. Loyal Customers (Cluster 1) - Regular, less frequent, moderate-value buyers
3. Potential Loyalists (Cluster 2) - Sporadic but promising newer customers
4. At-Risk Customers (Cluster 3)- Infrequent, low-spending customers

Key Findings:

- Champions Customers delivers substantial revenue concentration risk despite small size
- Loyal and Potential Loyalists Customers present growth potential through engagement strategies
- At-Risk Customers urgently needs reactivation to prevent further attrition

Recommendations:

- Reward Champions via loyalty programs to maintain advocacy
- Convert Potential Loyalists through personalized recommendations
- Reengage At-Risk with tailored win-back offers
- Continuously track segments using RFM to evolve retention tactics

In summary, this RFM analysis enabled granular customer segmentation to derive actionable targeting strategies per cluster, with significant potential to boost engagement, loyalty and revenues. Continued monitoring will allow for optimization over time.

Introduction

In an increasingly competitive eCommerce landscape, precise customer segmentation is critical for companies seeking to optimize marketing efforts and customer engagement. RFM (recency, frequency, monetary) analysis has emerged as a key segmentation technique based on when, how often, and how much revenue customers generate from purchases.

This project applies RFM analysis to an online retail dataset from the UCI Machine Learning Repository (E-Commerce Data, n.d.) containing actual transactions from 2010-2011 for a UK-based online gift retailer. The dataset offers a unique opportunity to examine customer purchase behaviors as the transactions span a full year.

Core Objectives

1. Stratify customers using recency, purchase frequency, and spending patterns.
2. Derive actionable insights on targeted marketing opportunities per segment.
3. Inform customer engagement strategies to drive loyalty and revenues.

By examining transaction history through an RFM lens, companies can identify their most valuable customers as well as at-risk segments requiring reactivation. These granular insights catalyze the development of tailored marketing campaigns, personalized product recommendations, and retention initiatives across the customer lifecycle.

In summary, this project demonstrates how RFM analysis powers precision segmentation at scale, enabling businesses to optimize resource allocation, nurture customer relationships, and promote growth. The ability to continually refresh segments provides an adaptive framework for long-term success.

Data Preprocessing

Data Overview

The dataset under review encompasses a record of online retail transactions, delineating a timeline from December 1, 2010, to December 9, 2011. It is comprised of 541,909 entries, each detailed across eight distinct columns. These columns chronicle a range of information from invoice specifics (such as the invoice number and date) to product attributes (including stock code, description, quantity, and unit price), as well as the customer's unique identifier and geographical data.

Data Cleaning

The following preprocessing steps were taken to ensure data quality:

- Missing product descriptions were uniformly labeled as 'Unknown'.
- Absent customer IDs were substituted with a placeholder value of -1.
- The 'InvoiceDate' field was divided into separate columns to distinctly represent time and date.
- We transformed the 'CustomerID' field to an integer data type to standardize its format.

Data Filtering

During data filtering, we identified and excised canceled orders, which were discernible by invoice numbers prefixed with 'C', and purged records where the quantity or unit price were negative, as they do not constitute valid transactions. Data cleaning reduced dataset from

541,909 to 531,283 rows. Notice that the canceled orders and returned orders (quantity as negative) are used in future additional analysis for customer behavior.

The processed dataset now provides clean information on valid transactions to enable robust analysis on customer purchase behaviors.

RFM Analysis

RMF Calculation

To derive meaningful customer segments from the eCommerce data, we employed a methodical approach in calculating Recency, Frequency, and Monetary (RFM) values:

- **Recency (R)**: Determined by the number of days since a customer's last purchase, with the calculation anchored to the day following the latest transaction date in the dataset. This metric is indicative of customer engagement, underpinning the notion that more recent interactions are suggestive of ongoing customer relationships.
- **Frequency (F)**: Ascertained by counting unique invoice numbers per customer, proxying purchase frequency and customer loyalty.
- **Monetary (M)**: Represented by the aggregate spend of a customer, computed by multiplying the quantity of products purchased by the unit price, thereby encapsulating the customer's value contribution.

These computations relied on the following assumptions:

- Anchoring recency to the day after the last transaction provides appropriate context for recent activity.
- Unique invoice numbers represent individual transactions without duplication.
- Total spend adequately captures customer monetary value despite its simplicity.

Merging normalized RFM metrics enabled nuanced segmentation, with segmentation refined further by standardizing the values with '***StandardScaler***' to equalize influence before clustering. This methodical RFM computation and standardization facilitates precise, behavioral-based customer segmentation.

Customer Segmentation and Profiling

We employed K-Means clustering, renowned for its simplicity and precision, to uncover customer behavior nuances and effectively segment the base. Using the Elbow method on the within-cluster sum of squares (WCSS), diminishing returns emerged beyond four clusters, indicating it as the optimal number. Upon a comprehensive review of the cluster summaries and these scores, the choice of four clusters emerged as the most balanced approach.

Number of Clusters	Silhouette Score	Davies-Bouldin Score
3	0.593	0.710
4	**0.616**	**0.752**
5	0.617	0.717
6	0.598	0.627

- **Silhouette Score**: This metric showed a gradual increase from a score of 0.59 in a 3-cluster solution to 0.616 for 4 clusters, before plateauing at 0.617 with 5 clusters and then dipping to 0.598 with 6 clusters. The peak at 4 clusters signaled well-defined and cohesive groups.
- **Davies-Bouldin Score**: The score rose from 0.71 for 3 clusters to 0.752 with 4 clusters, slightly decreased to 0.717 for 5 clusters, and then to 0.627 for 6 clusters. While not the lowest, the score for the 4-cluster solution was indicative of a reasonable separation between clusters.

The 4-cluster solution groups customers into distinct segments with varying recency, frequency, and monetary properties. Each cluster contains a reasonable number of customers for segment-level analysis. Splitting into 5 clusters risks creating groups that are too small for actionability. With 4 segments, there is sufficient granularity for tailored initiatives while maintaining interpretability for key insights. This balance of detail and comprehension provides a meaningful clustering that powers targeted marketing strategies across the customer base.

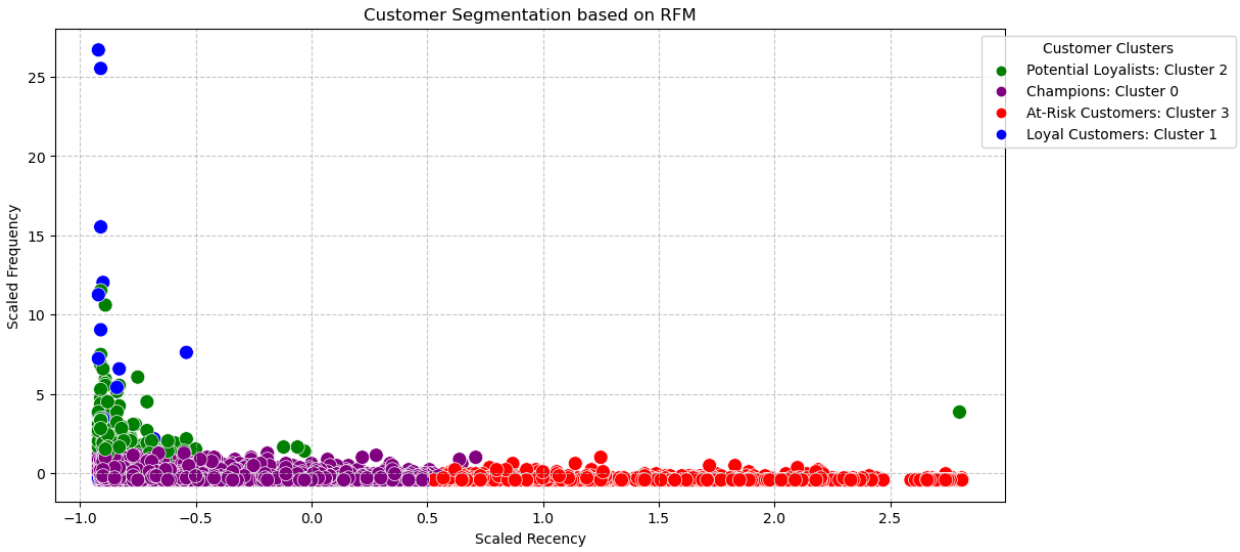
Customer Type	Cluster	Recency	Frequency	Monetary	Count
Champions	0	44.48	3.66	1349.38	3054
Loyal Customers	1	7.62	82.69	127338.31	13
Potential Loyalists	2	16.12	22.05	12453.23	211
At-Risk Customers	3	249.17	1.55	478.19	1061

- **Champions (Cluster 0)**: Moderate recency and frequency with significant monetary value, suggesting engaged, valuable customers.
- **Loyal Customers (Cluster 1)**: Highly recent and frequent with an exceptional monetary contribution, indicating a small but extremely valuable customer group.
- **Potential Loyalists (Cluster 2)**: Fair recency and good frequency with a substantial monetary contribution, representing an engaged, profitable segment.
- **At-Risk Customers (Cluster 3)**: The least recent and frequent, with modest monetary value, pointing to a segment with potential for re-engagement.

Visualization and Marketing Recommendations

RFM Visualization

The provided scatter plot is a visual representation of the customer segmentation based on RFM analysis:



Overall, the scatter plot clearly segments customers into distinct groups based on their shopping behavior, allowing for tailored marketing strategies.

Marketing Recommendations

Based on the characteristics of each customer cluster, here are some key insights and marketing recommendations for each cluster:

Champions (Group 0)
Marketing
- Implement loyalty programs to retain these valuable customers.
- Offer exclusive deals, early access, or personalized promotions to encourage repeat purchases.
- Gather feedback from this group to understand what keeps them engaged and satisfied.

Communication
- Regularly communicate new product launches, promotions, or updates to keep them engaged.
- Consider personalized newsletters with product recommendations based on their purchase history.

Retention
- Offer tiered loyalty rewards for reaching specific spending or frequency milestones.
- Provide excellent customer service to enhance their overall experience.

Loyal Customers (Group 1)
Marketing

- Focus on maintaining brand loyalty through targeted marketing campaigns.
- Offer subscription-based services or exclusive memberships with added benefits.
- Upsell premium or complementary products to increase average transaction value.

Communication

- Send personalized recommendations based on their purchase history.
- Seek feedback on their experience and use it to improve products or services.

Retention

- Provide early access to sales, events, or new product launches.
- Create a referral program to encourage them to bring in new customers.

Potential Loyalists (Group 2)

Marketing

- Implement targeted marketing campaigns to increase engagement and encourage repeat purchases.
- Offer special promotions or discounts to incentivize more frequent transactions.
- Highlight the unique value propositions of your products to encourage brand loyalty.

Communication

- Use targeted email campaigns introducing them to new products or features.
- Leverage social media to showcase customer testimonials and success stories.

Retention

- Provide a welcome discount to encourage the second purchase.
- Consider bundling products to increase the average order value.

At-Risk (Group 3)

Marketing

- Re-engage this group with win-back campaigns offering special discounts or incentives.
- Identify pain points and address them through targeted marketing efforts.
- Consider creating special offers to bring them back into the purchasing cycle.

Communication

- Send personalized emails addressing their inactivity and offering exclusive deals.
- Implement targeted social media ads to remind them of your brand and products.

Retention

- Conduct surveys to understand the reasons for their decreased engagement.
- Provide a limited time offer to encourage them to make a new purchase.

In summary, the segmentation visual enables tailored strategies based on engagement levels - from VIP rewards to re-activation outreach. Continuous refresh of clusters will ensure adaptive approaches as customers evolve across their lifecycles with the brand.

Additional Analysis

In the intricate tapestry of our dataset, numerous avenues await exploration to unravel richer insights and actionable intelligence. As we embark on this additional analysis journey, our aim is to delve into key aspects that extend beyond the initial clustering exercise, shedding light on

various facets of customer behavior, product dynamics, temporal patterns, geographical trends, and operational efficiency.

Beyond the clusters that have provided valuable segmentation, there exists a wealth of information waiting to be unearthed. Each query opens a door to a realm of understanding, offering strategic perspectives that can drive targeted decision-making and enhance our overall business strategy.

Customer Analysis

The dataset comprises 4,372 unique customers. The distribution of orders per customer reveals that while the average customer has 90 orders, there is significant variability (standard deviation of 218 orders) indicating a wide range of behaviors. The maximum order count is 7,983, demonstrating extremely high engagement among top customers.

Specifically, the top 5 customers by order count are:

1. Customer 17841 with 7,983 orders
2. Customer 14911 with 5,903 orders
3. Customer 14096 with 5,128 orders
4. Customer 12748 with 4,596 orders
5. Customer 14606 with 2,700 orders

This skewed distribution shows a small subset of buyers contributing a very high number of transactions that significantly exceeds the average. The company's top customers demonstrate tremendous loyalty that warrants tailored retention initiatives. In contrast, the long tail of buyers with fewer purchases represents an opportunity to nurture higher engagement through targeted marketing outreach.

Product Analysis

The top 10 most frequently purchased products by customers are:

- White hanging heart t-light holder (2,369 purchases)
- Regency cakestand 3 tier (2,200 purchases)
- Jumbo bag red retrospot (2,159 purchases)
- Party bunting (1,727 purchases)
- Lunch bag red retrospot (1,638 purchases)
- Assorted colour bird ornament (1,501 purchases)
- Set of 3 cake tins pantry design (1,473 purchases)
- Pack of 72 retrospot cake cases (1,385 purchases)
- Lunch bag black skull (1,350 purchases)
- Natural slate heart chalkboard (1,280 purchases)

This list reveals the company's most popular products, indicating strong customer demand likely driven by product quality, style, and effectiveness in gift-giving use cases.

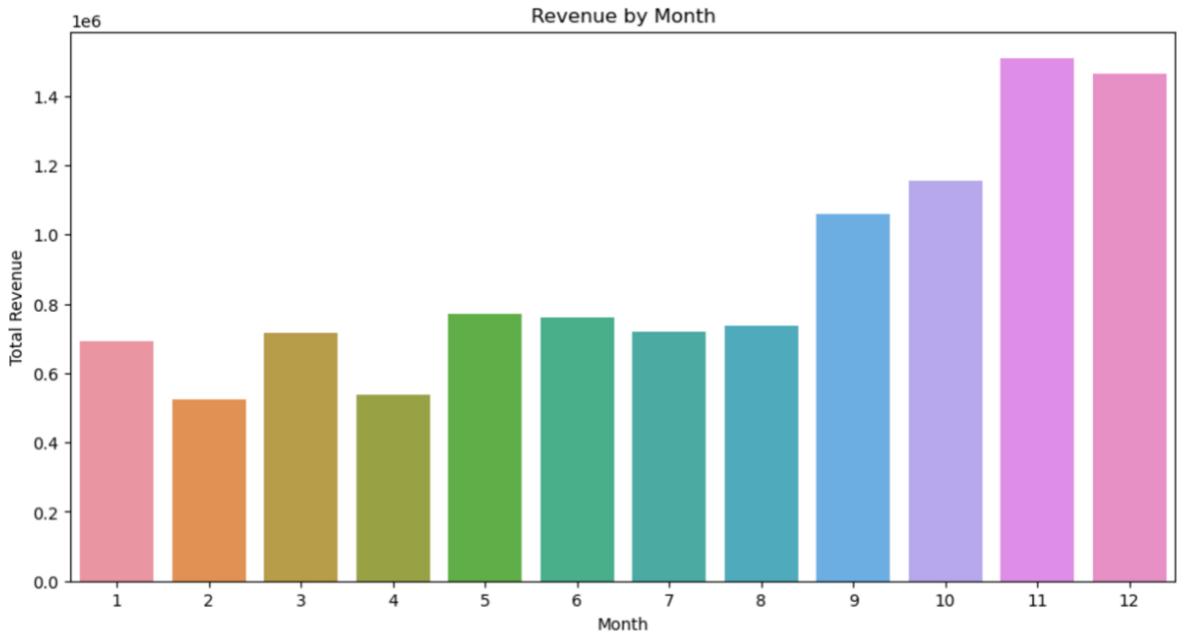
The average price of all products in the dataset is approximately £4.61, representing affordable gifts and home goods. However, the 'DOTCOM POSTAGE' product category generated the highest total revenue of £206,245.48, underlining the sales volume from the company's online retail channel.

Product Analysis

Granular analysis of order timing reveals Thursday at 12 PM as the consistent peak ordering period. Over the full dataset, Thursdays accounted for 19% of total purchases compared to 12-18% on other weekdays. 12 PM also stands out, representing 14.54% of daily orders. It is noticeable that starting from 11 AM to 4 PM, the orders volume is high.

The prominence of the Thursday lunch hour for transactions suggests office workers may be placing online orders from their workplace. The company could explore B2B gift purchase behaviors incorporates to further understand this trend. Inventory availability and prompt order fulfillment should be prioritized for Thursday middays.

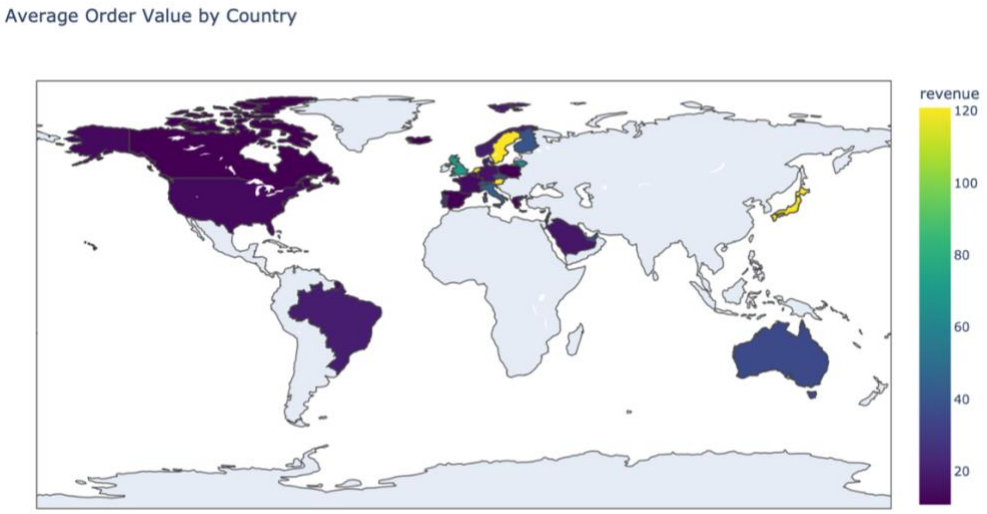
Examining seasonal patterns shows November as the top month for both order volume (83,498 order) and revenue, followed by December. September-December sustains higher order rates and financial performance versus other months, pointing to holiday seasonal impacts.



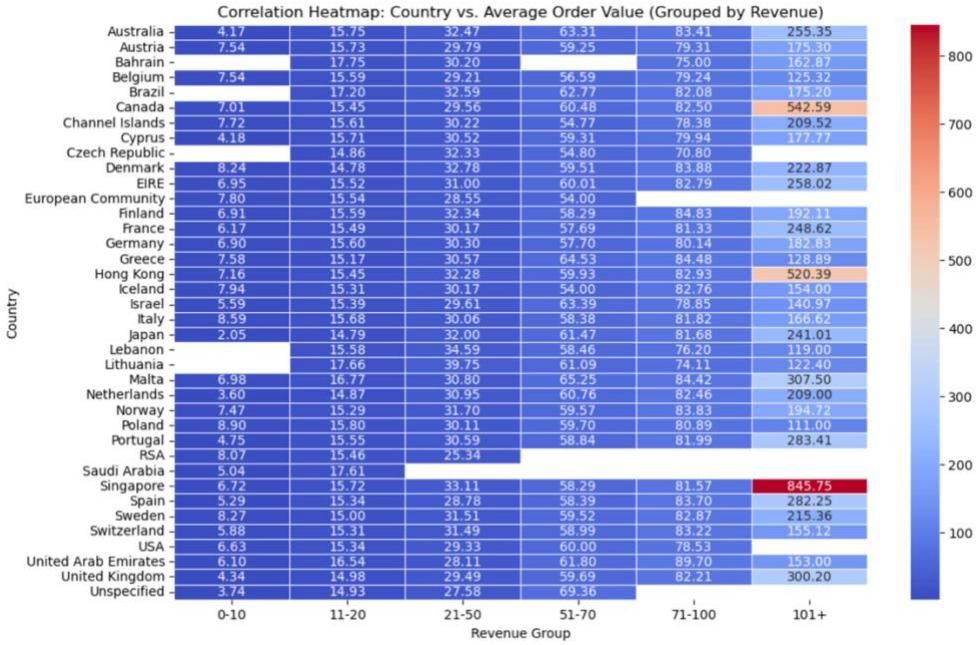
In summary, the Thursday midday spikes complement a broader late-year holiday buying trend. Tactical focus on stocking and promotions during these periods of heightened customer activity can further boost orders and revenue. Granular order timing analysis enables optimization tied to predicted purchasing cycles.

Geographical Analysis

The top 5 countries by order volume are the UK, the Netherlands, Germany, France, and Ireland. Among these, the Netherlands has the highest average order value, representing an attractive high-spending customer profile (Customer Segmentation, n.d.).



Examining correlations between country and average purchase size reveals additional patterns. Singapore demonstrates the strongest association with order values above £100, pointing to a propensity for higher spending. Overall, there is a minor negative correlation of -0.06 between the numerical country designation and order amount, indicating a weak relationship between location and average ticket size in this dataset.



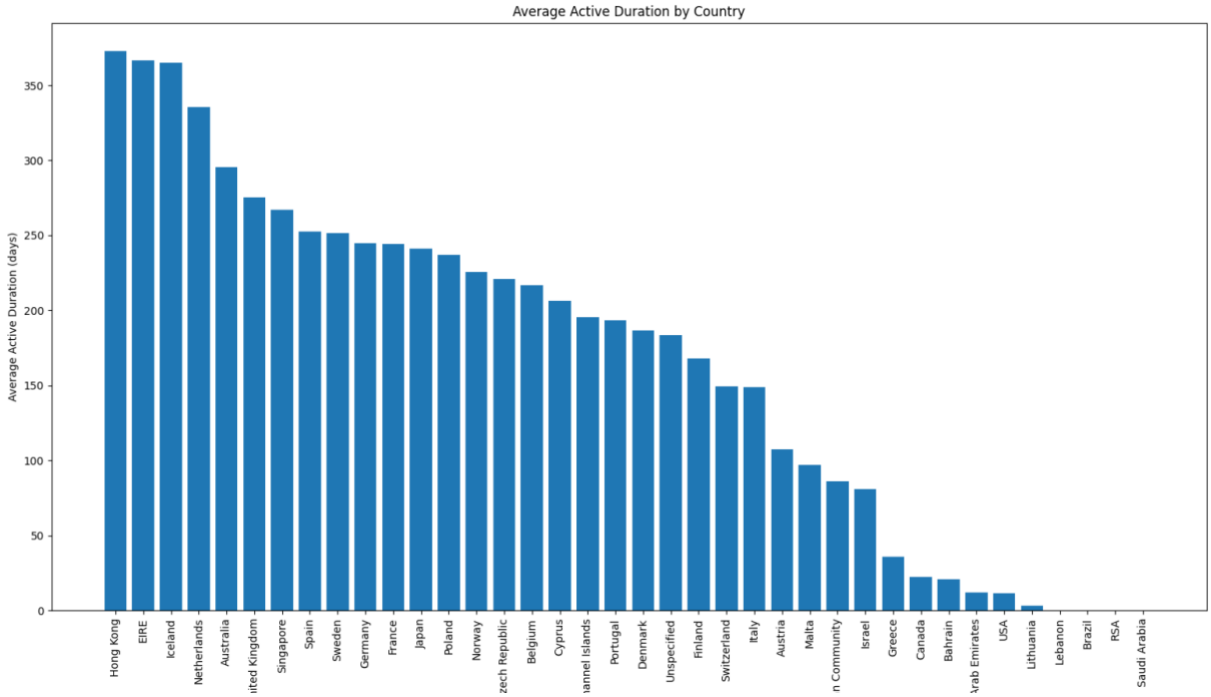
In conclusion, while geographical location has minimal linkage to transaction value, key buyer markets have emerged based on order and revenue contribution. The Netherlands leads with the highest average purchase amounts while Saudi Arabia over-indexes on large orders. Targeting higher spending segments in these potential growth geographies can yield disproportional returns.

Payment Analysis

The absence of complete payment information restrains us from conducting a thorough analysis of payment methods utilized by customers. Vital questions related to the prevalence of specific payment modes and their potential correlation with order amounts remain suspended, awaiting a more comprehensive dataset for a conclusive exploration.

Customer Behavior

The customer lifespan between first and most recent purchase averages 130.8 days across the full dataset. However, significant variation emerges at the country level. Hong Kong, Ireland, and Iceland exhibit the longest engagement cycles while the US, Saudi Arabia, and Lithuania fall well below average. Tailoring retention initiatives based on expected active durations within key target countries can optimize results.



Additionally, buyers can be segmented by their purchasing frequency over time. Established "Loyal" customers order consistently in a sustained cadence, "Potential Loyalists" are still building lifetime engagement, and "At Risk" shoppers show early signs of dormancy. Aligning customer use cases and offerings to refresh and reactivate various segments promotes organic growth.

Returns and Refunds

1.71% of orders are cancelled outright based on designated invoice numbers. Of the remaining purchases, 1,336 orders (0.25%) contain negative quantity values indicating item returns. In total, nearly 2% of transactions incur returns or cancellations - a sizable downstream cost.

Further analyzing return rates by product category uncovers additional trends. Apparel gifts sourced from the USA correlate to higher return volumes of 38.49%, potentially caused by improper sizing or product unsuitability. By contrast, categories like small decor items or utensils see negligible returns due to standardized fits. The top 10 Return Rate Country is as below.

Country	Return Rate
USA	38.48%
Czech Republic	16.67%
Malta	11.81%
Japan	10.34%
Saudi Arabia	10.00%
Australia	5.88%
Italy	5.60%
Bahrain	5.26%
Germany	4.77%
EIRE	3.68%

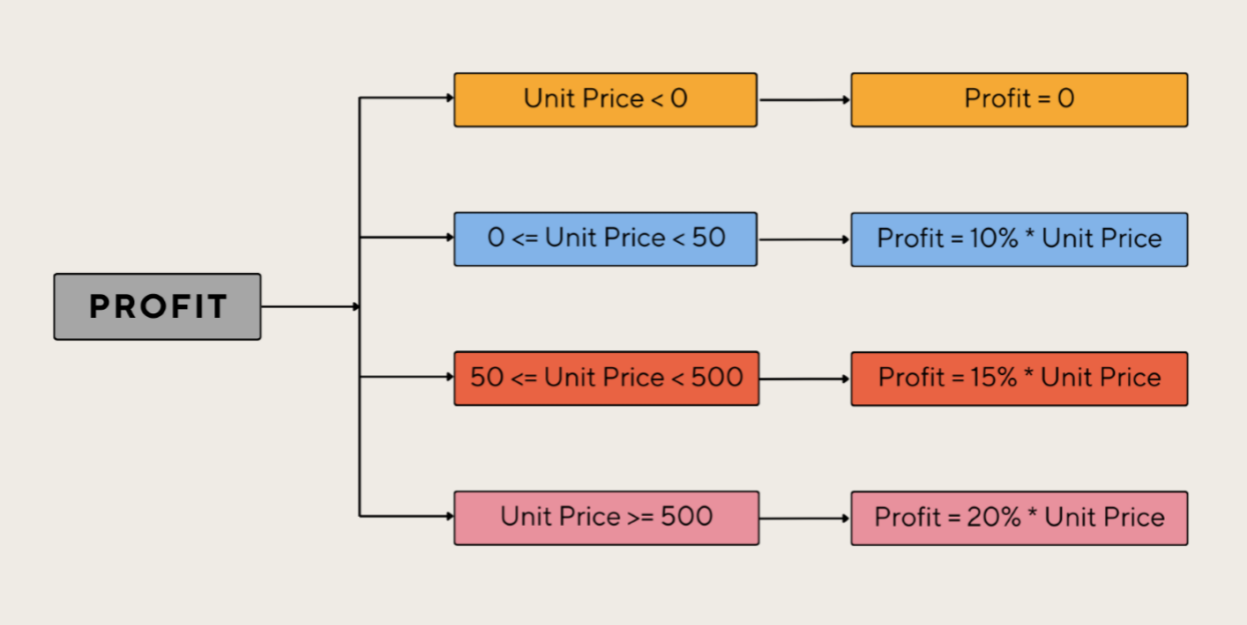
In summary, 2% of purchases result in refund or fulfillment expenses. Variations by product category present an opportunity to limit financial risk if amplified in specific merchandise from high return geographies. Diversifying domestic item selection and enhancing sizing guidance are potential levers to mitigate excessive returns.

In exploring the profitability analysis of our dataset, we encountered a challenge due to the absence of direct profit information for products. To address this, we devised a rule for profit calculation based on unit prices, creating a new 'Profit' column. This rule considers varying profit percentages based on different price ranges, enabling us to estimate the profitability of each product. Subsequently, we introduced a 'TotalProfit' column by multiplying the calculated profit with the quantity of items sold.

Profitability Analysis

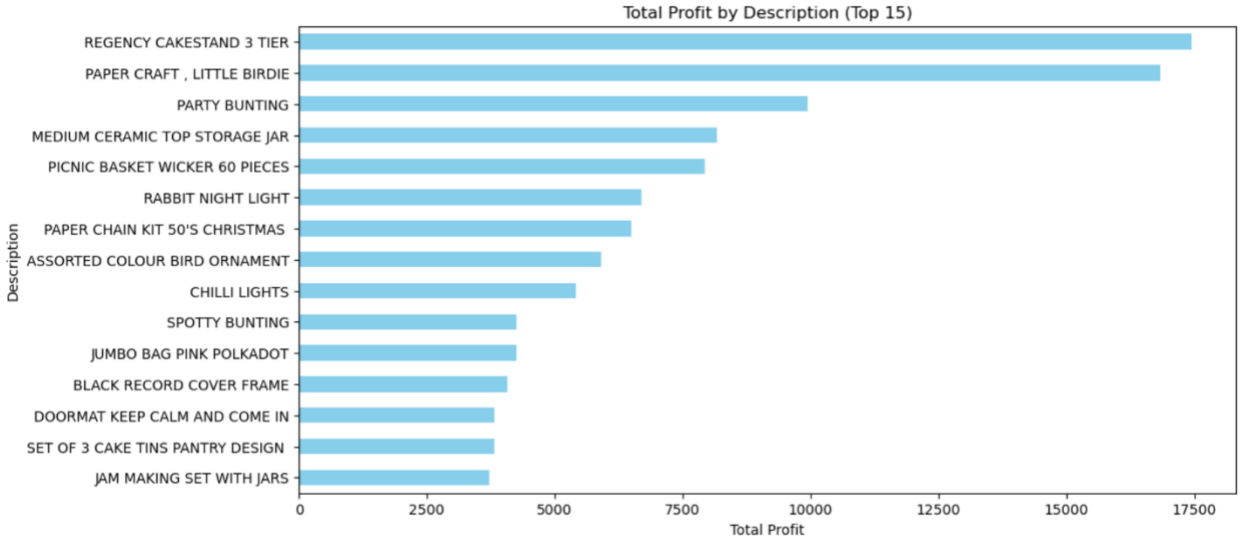
In exploring the profitability analysis of our dataset, we encountered a challenge due to the absence of direct profit information for products. To address this, we devised a rule for profit calculation based on unit prices, creating a new 'Profit' column. This rule considers varying profit

percentages based on different price ranges, enabling us to estimate the profitability of each product. Subsequently, we introduced a 'TotalProfit' column by multiplying the calculated profit with the quantity of items sold.

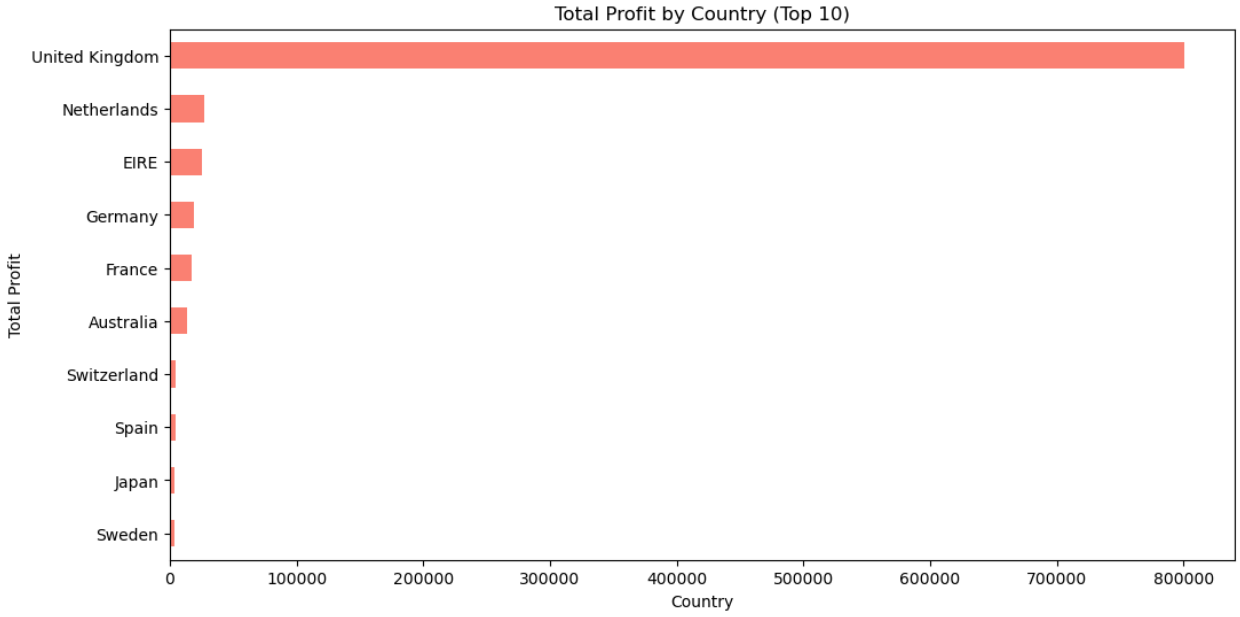


However, during our investigation, we identified peculiar transactions indicated by non-numeric stock codes. To ensure a focused analysis on product profits, we selectively excluded these transactions by creating a new data frame, *'df_profit*,' filtered based on specific criteria, including numeric stock codes and stock code length.

The chart ranks products according to their total profit contribution, with 'REGENCY CAKESTAND 3 TIER' generating the most profit of the top 15 items listed. There is a substantial variance in profitability among the top 15 products. The most profitable item generates significantly more profit than the others. The products appear to be a mix of household items, festive or seasonal goods, and general wares, indicating a diverse product offering. Products like 'PARTY BUNTING' and 'PAPER CHAIN KIT 50'S CHRISTMAS' suggest seasonal sales trends, which may require strategic stock management around specific periods.

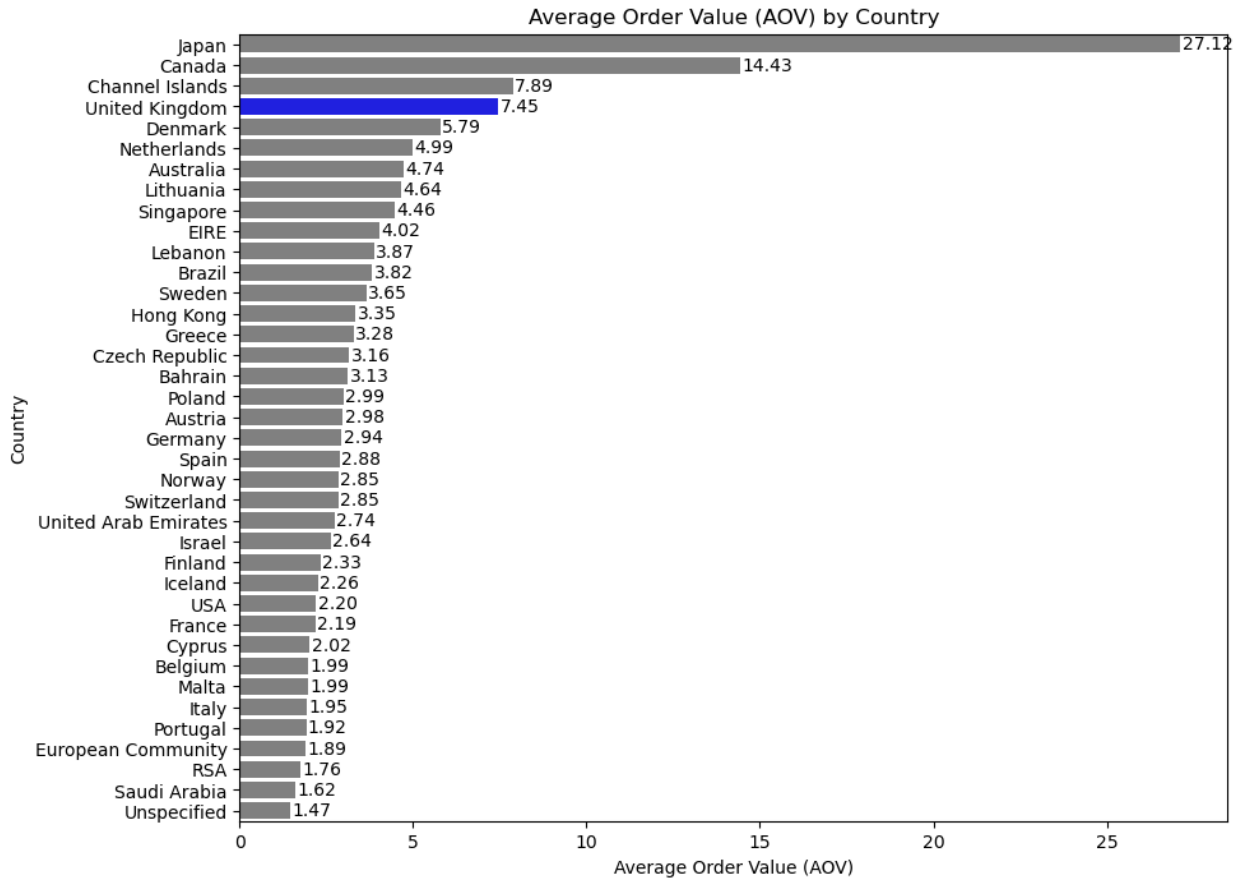


Transitioning to a broader perspective, we visualized the total profit by country through a bar chart. The United Kingdom stood out as the primary market, contributing significantly more profit than other countries. The Netherlands, EIRE (Ireland), and Germany followed, suggesting a global footprint for the business. The data hinted at potential opportunities for market expansion, particularly in countries where profit contributions were substantial but not as dominant as in the UK.



Considering the UK's prominence, we turned our attention to the average order value (AOV) in this market. The calculated AOV for the UK, at 1.70, indicated a relatively lower average profit per order compared to some international markets. A comparison revealed higher AOVs in countries such as Australia and the Netherlands, suggesting potential areas for strategic focus and market optimization. Furthermore, the examination of AOVs in Japan and Sweden hinted at

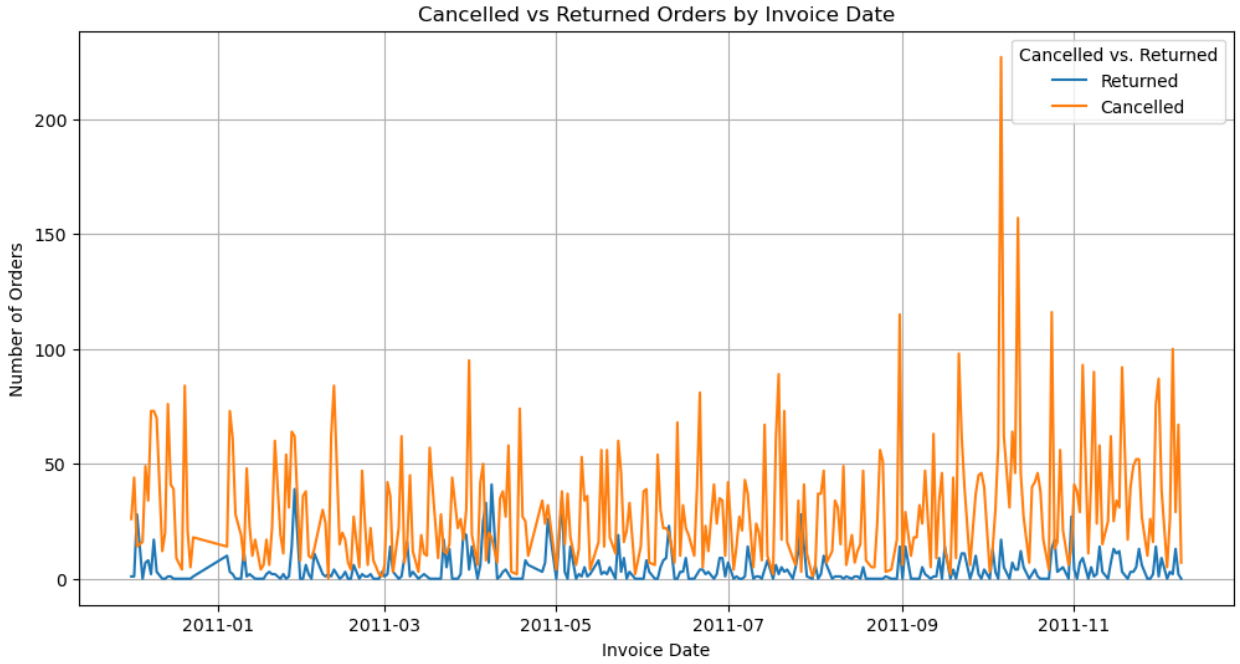
unique customer behaviors, possibly characterized by higher-value transactions, opening avenues for targeted marketing or product offerings in these regions.



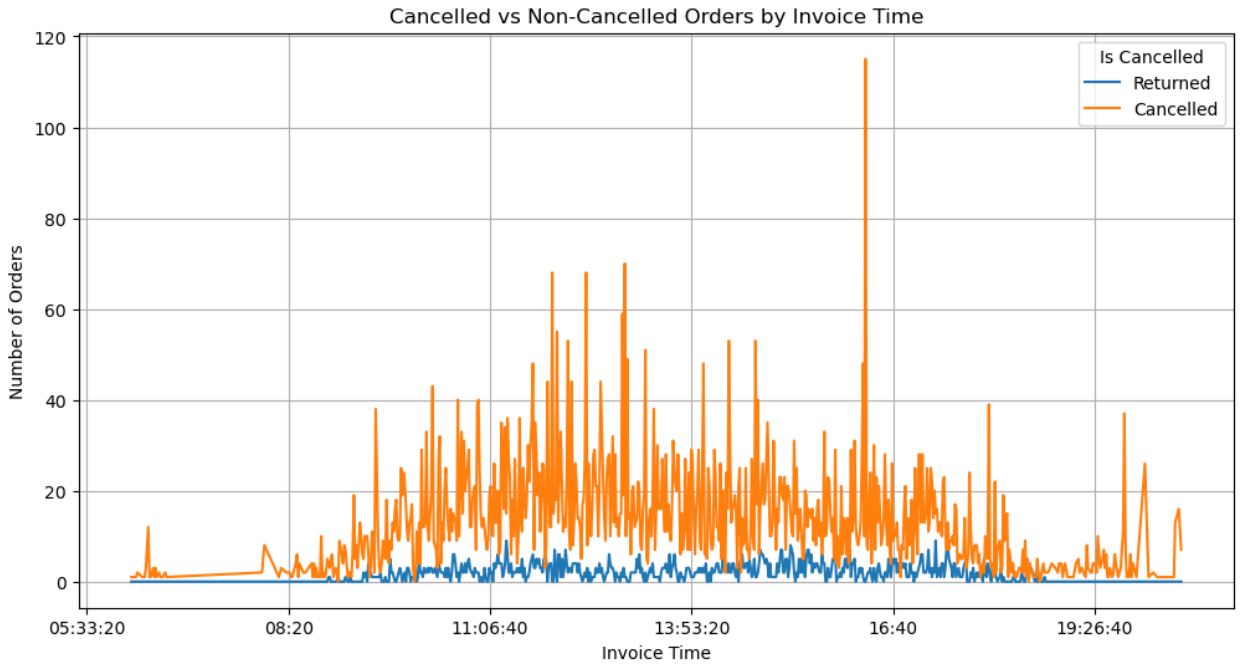
Customer Satisfaction

Despite the absence of direct customer feedback or ratings in our dataset, we leveraged returned items as a proxy to analyze customer satisfaction. Within this subset, 87.42% of orders were marked as canceled (E-Commerce Sales Forecast , n.d.), indicating a prevalent trend of customers choosing to cancel post-purchase. Non-canceled returned orders, while representing only 12.57% of all returns, constitute a mere 0.25% of total orders, suggesting infrequent instances of customers returning items without prior cancellation. This minimal percentage implies that customers tend to exhibit higher satisfaction upon receiving their orders, reducing the likelihood of dissatisfaction leading to subsequent returns.

The analysis of Cancelled vs. Returned Orders by Invoice Date revealed a seasonal pattern with peaks occurring at regular intervals. Outliers, particularly for cancellations, suggest potential data entry errors or unusually high sales volumes on specific dates, warranting further investigation.



Additionally, examining Cancelled vs. Non-Cancelled Orders by Invoice Time unveiled fluctuations in cancellations throughout the day, potentially indicating operational stress during peak hours of working time between 10:00AM – 5:00PM. We can also see here Cancelled orders have another smaller peak around 7:30 PM, indicate people normally cancel the orders after working hours. These insights provide valuable operational considerations, highlighting the importance of managing system load during peak hours to minimize the likelihood of cancellations.



While direct customer feedback remains unavailable, these data-driven insights provide a valuable perspective on customer satisfaction trends and operational considerations. The observed patterns can inform strategies to minimize cancellations, enhance customer experiences, and optimize operational efficiency during peak periods.

Conclusions

The project's RFM analysis has carved out distinct customer segments within the eCommerce dataset, leading to critical insights that are pivotal for crafting tailored marketing strategies. The segmentation reveals a spectrum of customer engagement, from the highly valuable Champions and Loyal Customers to the Potential Loyalists and At-Risk groups, each requiring a unique approach to maximize their business potential.

Key findings from the analysis underscore the substantial revenue generated from a small group of top-tier customers, suggesting a concentrated risk but also a significant opportunity for targeted engagement. The identification of Potential Loyalists and At-Risk Customers provides a strategic vantage point to enhance growth and re-engage dwindling customer relations, respectively.

While the project delivers actionable strategies to drive customer retention and augment revenue streams, it also recognizes the limitations inherent within the dataset, including the absence of direct customer feedback and comprehensive payment data. These constraints slightly curtail the depth of customer satisfaction and payment preference analysis that could be conducted.

Reflecting on the project, the inferences drawn from the RFM analysis demonstrate the robust potential of data-driven customer segmentation. However, a more enriched dataset encompassing a broader range of customer interaction metrics could further refine the insights and recommendations.

Despite these limitations, the project represents a significant stride in understanding customer behaviors and provides a scalable approach to customer segmentation. By marrying analytical rigor with strategic marketing, it lays the groundwork for enhanced customer experiences and a stronger market position in the competitive landscape of eCommerce.

Bibliography

Customer Segmentation. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/code/fabiendaniel/customer-segmentation

E-Commerce Data. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/datasets/carrie1/ecommerce-data

E-Commerce Sales Forecast . (n.d.). Retrieved from Kaggle: https://www.kaggle.com/code/allunia/e-commerce-sales-forecast