E-Commerce Data Analysis Report

NUID	Name
002242934	Anjali Ingle
002772623	Avani Kala
002291838	Ameya Deshmukh
002734199	Wenzhe Zhang
002922799	Yangwenyu Peng

Table of Contents

1. Executive Summary

This report provides a succinct overview of an extensive e-commerce data analysis, covering the period from December 2010 to December 2011. Key stages of data processing were meticulously executed to curate a robust dataset for in-depth analysis. The study focused on customer purchasing patterns, revealing distinct behavioral segments, and pinpointing areas for targeted marketing engagement.

RFM segmentation identified key customer groups, including regulars and high-value, infrequent shoppers, informing tailored strategies to increase loyalty and spending. The UK emerged as the dominant market in the geographical analysis, guiding a region-specific approach to marketing and sales. Product analysis highlighted top performers and categories with high return rates, signaling opportunities for product optimization.

While insights into revenue trends were gained, limitations in cost data hindered profitability analysis, and a lack of payment data restricted payment behavior insights. The report sets forth foundational knowledge for strategic initiatives aimed at enhancing customer engagement, optimizing product offerings, and refining sales strategies to drive revenue growth.

2. Data Preprocessing

The data analysis report herein pertains to the E-Commerce Dataset procured from the UCI Machine Learning Repository. This dataset is characterized as a transnational compilation of transactions from a non-store-based online retail entity in the UK, spanning from December 1, 2010, to December 9, 2011.

Initial analysis revealed the presence of missing entries in the 'Description' and 'CustomerID' columns. To maintain the dataset's completeness and utility, these missing entries were substituted with 'No Description' and 'Unknown Customer' respectively. Furthermore, essential data type transformations were executed. The 'InvoiceDate' was altered to a datetime format, incorporating error coercion to address any discrepancies, thereby facilitating more nuanced temporal analysis. Additionally, the conversion of 'InvoiceNo' and 'CustomerID' to string formats was instrumental in enabling the following effective string manipulations and categorical evaluations.

A distinctive cleaning process was employed for invoice numbers commencing with 'C', indicative of cancellations as per the guidelines from the UCI Machine Learning Repository's Variable Table. Recognizing these cancellations as reversals of previous transactions, they were treated as non-contributory to the count of actual transactions. To operationalize this, transactions were categorized based on their 'InvoiceNo' prefix, segregating cancellations from regular transactions. A comparative method based on tuples was deployed to align and subsequently exclude these cancellations from the dataset. However, the original dataset, which did not receive this specific processing, is replicated for use in later analysis to determine the return rate.

Further refinement of the dataset involved the removal of records exhibiting negative values in the 'Quantity' and 'UnitPrice' fields, a crucial step to avert potential skewing of analytical outcomes. Moreover, an Interquartile Range (IQR) technique was adopted to methodically eliminate outliers from each numerical column, thus bolstering the dataset's integrity.

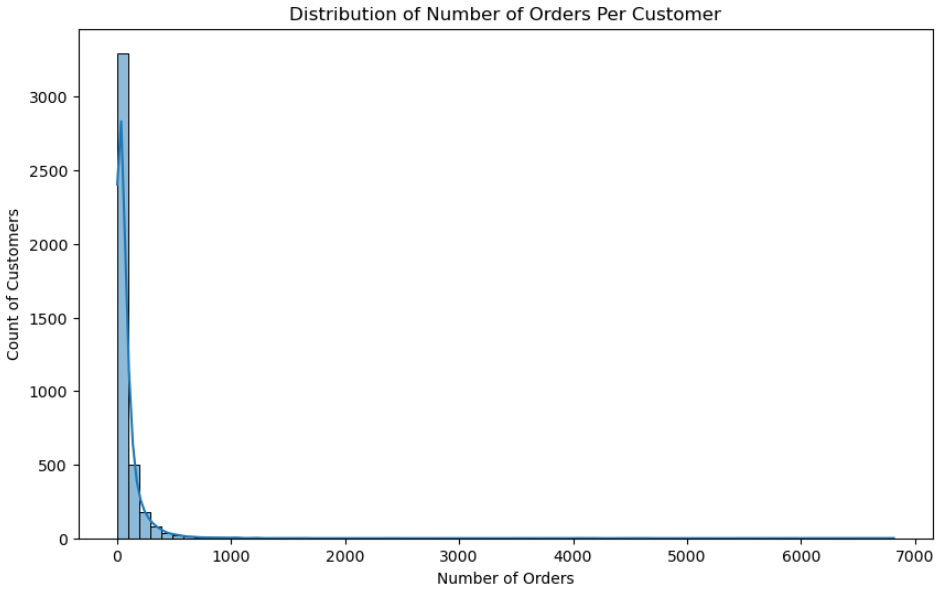
Following the completion of these procedures, the resultant dataset was meticulously refined, comprising 438,396 rows and 8 columns. This

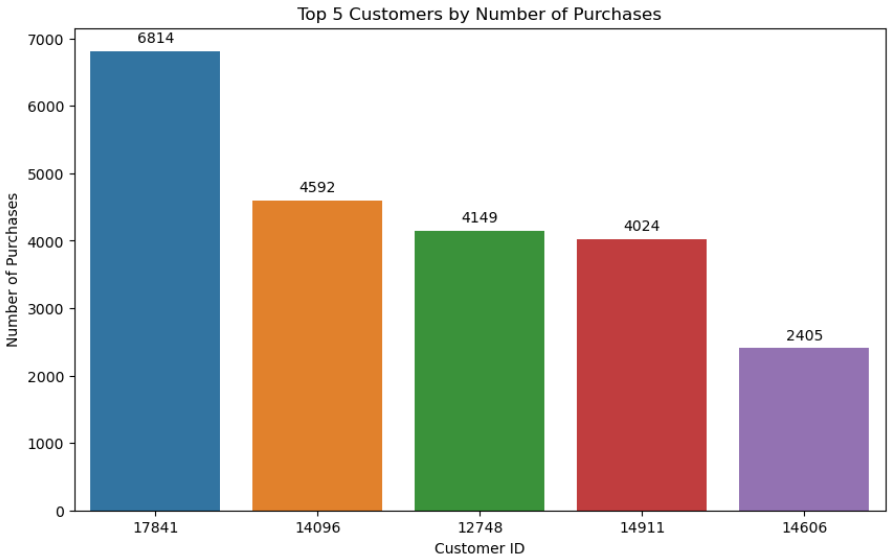
dataset encapsulates transaction records spanning from December 1, 2010, at 08:26:00 to December 9, 2011, at 12:50:00.

3. Customer Analysis

During the data cleaning phase, entries labeled as 'Unknown Customer' act as stand-ins for absent information. While they may offer some understanding of product trends, they fail to yield useful insights about specific customer habits or inclinations. Since these entries often represent collective data from guest checkouts or non-registered users, they are not suitable for segmentation or profiling, potentially distorting the outcomes of segmentation studies. Consequently, eliminating these entries before conducting customer analysis allows for a concentration on comprehensive and significant data.

The dataset reveals a total of 4,156 distinct customers. The following histogram shows a right-skewed distribution of the number of orders per customer. This indicates that a large number of customers have a relatively small number of orders, and fewer customers have a large number of orders.





From the bar chart above, the customer with ID 17841 is observed to vastly outnumber the purchases of other identified customers, who have purchase counts ranging from approximately 4,000 to 4,500. There is a noticeable disparity between the number of purchases made by the top customer and the fifth customer—nearly three times as many. This indicates that the top customer may be a very high-value customer. The variability in the purchase volumes among the top customers could be indicative of different purchasing behaviors or customer engagement levels.

1) RFM Calculations and Customer Segmentation

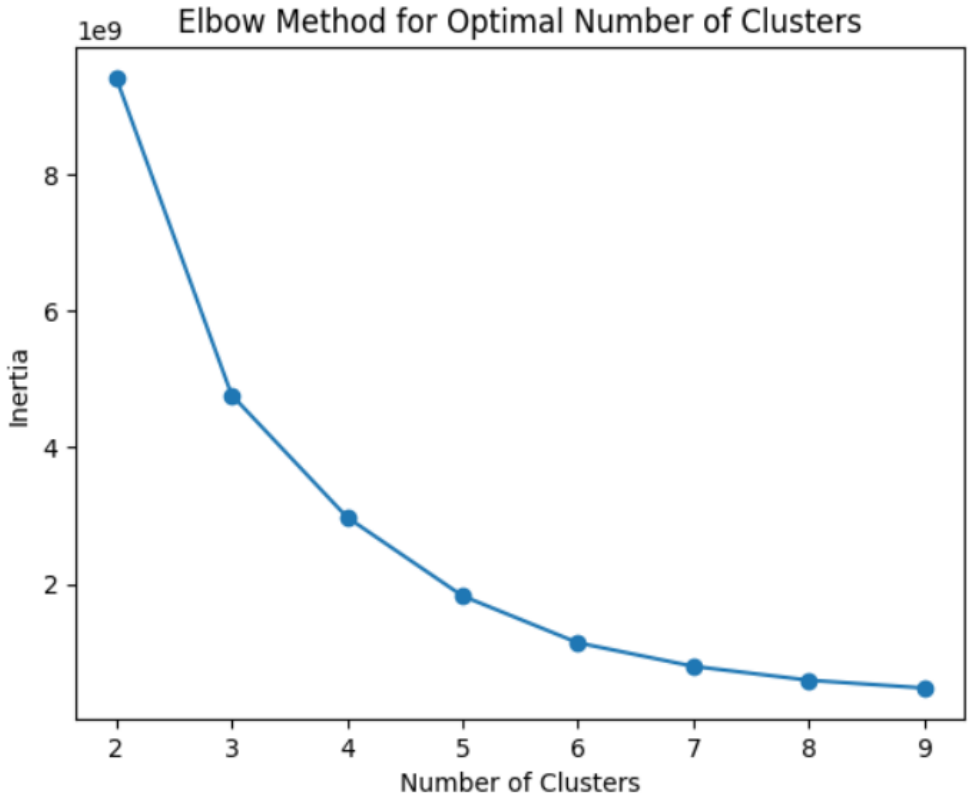
RFM analysis is conducted to categorize customers into distinct segments according to their RFM scores. The recency score is determined by identifying the most recent invoice date for each customer and calculating the days elapsed since that transaction. The data is then aggregated by 'CustomerID' to assess frequency, indicated by the number of invoice dates, and the monetary value, calculated as the sum of 'Total', which is derived from the product of 'UnitPrice' and 'Quantity'.

Each RFM component is assigned a specific weight, leading to the computation of an overall RFM score based on these weighted values, which can be adjusted according to the specific requirements of the business. In this instance, the weights are set as: Recency at 0.4, Frequency at 0.3, and Monetary at 0.3.

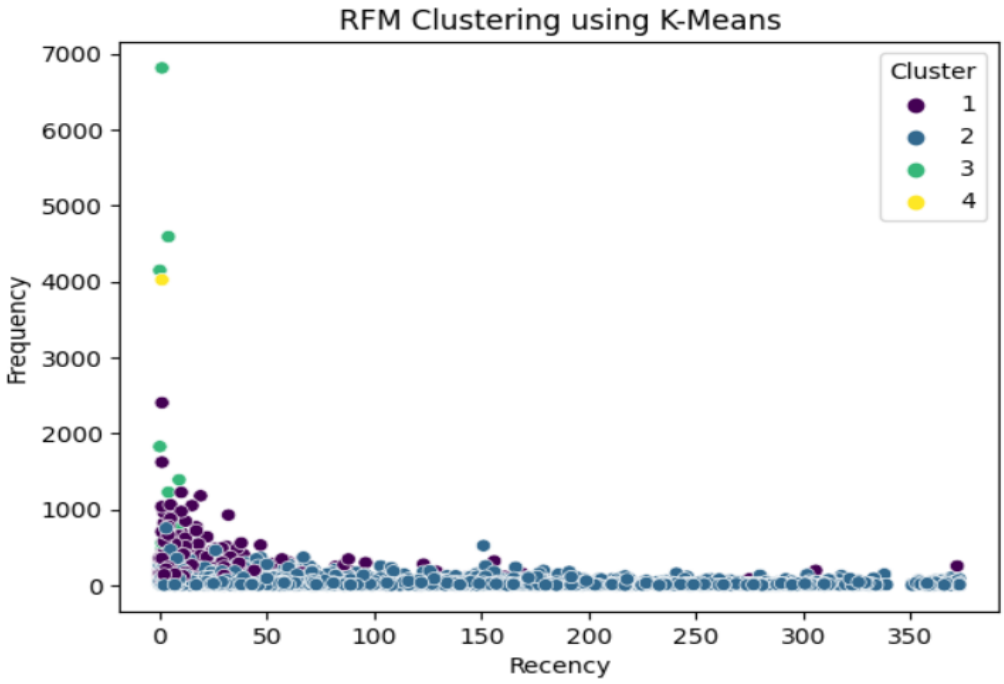
The analysis employs quartiles to classify the RFM scores, assigning each customer a corresponding quartile ranking. The truncated result is demonstrated as below：

	CustomerID	LastPurchaseDate	Recency	Monetary	Frequency	RFM_Combined	RFM_Quartile
0	12347.0	2011-12-07	2	2866.77	141	903.131	4
1	12348.0	2011-04-05	248	17.00	1	104.600	1
2	12349.0	2011-11-21	18	1155.75	61	372.225	4
3	12350.0	2011-02-02	310	274.00	15	210.700	3
4	12352.0	2011-11-03	36	971.98	56	322.794	3

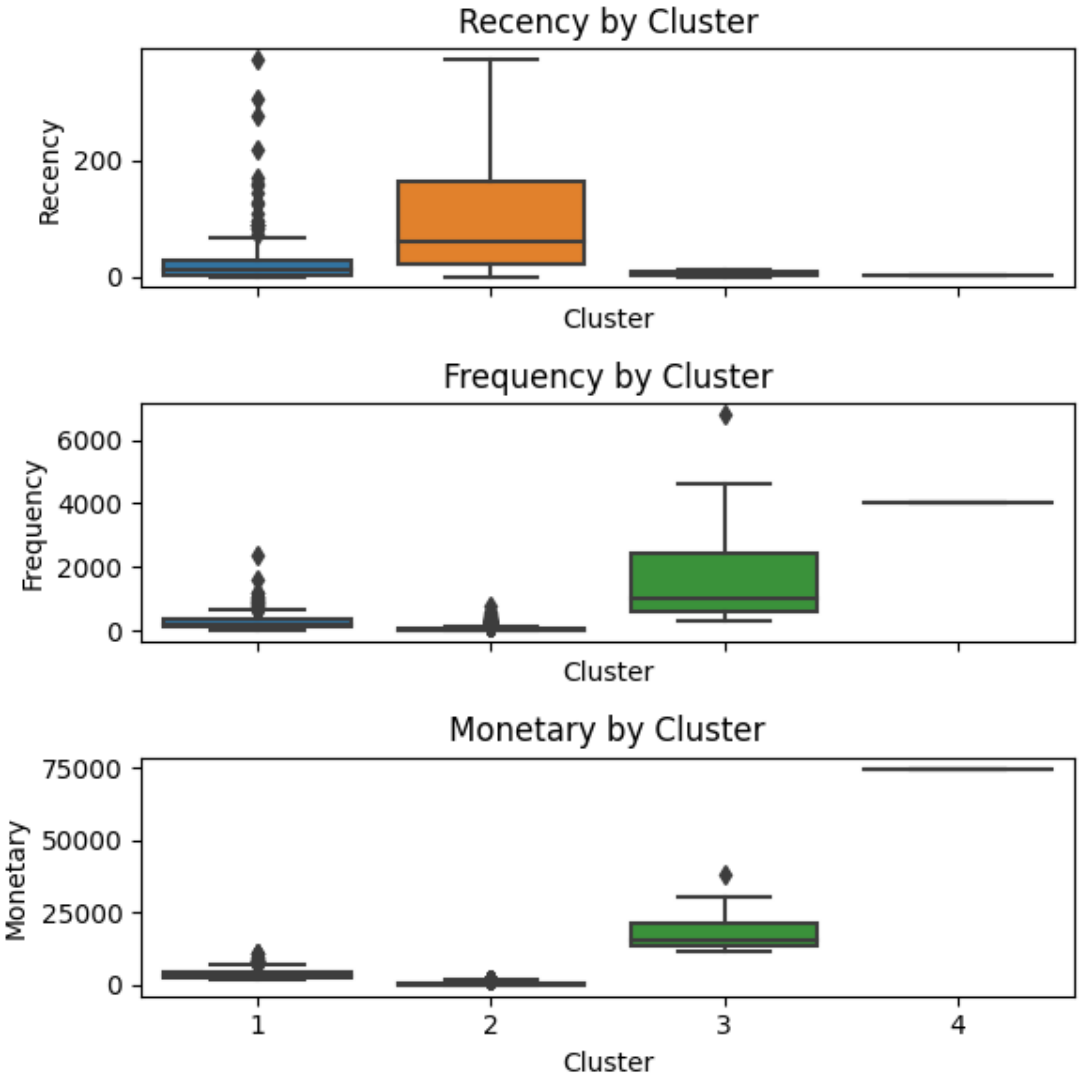
We then employed K-means clustering technique to segment customers based on their RFM scores. The optimal number of clusters in a K-means clustering algorithm is determined through the elbow method, which involves plotting the inertia values against the number of clusters (ranging from 2 to 9) and identifying the point where the rate of decrease in inertia, or explained variance, notably diminishes. This point, known as the "elbow," indicates the most suitable number of clusters for the data.



In our analysis, after generating and examining the elbow curve, it became evident that the optimal number of clusters is 4. This conclusion was reached as the elbow formation was observed between the 3rd and 4th cluster, suggesting that increasing the number of clusters beyond 4 yields diminishing returns in terms of explained variance.



The distribution of RFM values and the resulting clusters are depicted in the boxplot as follows:



Recency by Cluster: This boxplot shows the distribution of days since the last purchase for each cluster. Cluster 1 exhibits the most compact distribution, indicating consistency in recent purchases. Cluster 2 has a wider range, suggesting more variation in the time since the last purchase. Clusters 3 and 4 have very tight distributions, with Cluster 4 showing an extremely low median recency, close to zero, which indicates very recent interactions with the business.

Frequency by Cluster: The frequency distribution shows the number of purchases per customer in each cluster. Cluster 1 again shows a relatively tight distribution, suggesting similar purchase behavior among its customers. Cluster 2 has a wider spread, but less so than Cluster 3, which has a very wide distribution, indicating significant variability in purchase frequency among its customers. Cluster 4 is not as widely spread as Cluster 3, indicating more uniformity in purchase frequency within this

group.

Monetary by Cluster: The monetary boxplot displays the distribution of total spend per customer in each cluster. Clusters 1 and 2 show tight distributions with relatively low median monetary values, suggesting these customers spend less per transaction. Cluster 3 has the widest distribution, indicating a significant variance in spending, with some customers spending much more than others. Cluster 4, similar to its recency, has a very low spread and a high median value, suggesting customers in this cluster tend to spend large amounts.

2) Segmentation Profiling and Marketing Recommendations

The K-Means clustering resulted in four distinct customer segments. Here is the profiling of each segment based on the statistical summary provided in the analysis:

Cluster 1: The Engaged Regulars

- Recency (R): These customers have a low average recency (22.77 days), showing regular interaction. The wide range in recency (0-372 days) suggests a combination of highly active customers and those whose engagement is waning.
- Frequency (F): With an average frequency of 279.8 transactions, this group includes both consistent and occasional shoppers, indicated by the broad range (35-2405 transactions). This variability points to a segment with diverse engagement levels, potentially reflecting different customer needs or life cycles.
- Monetary (M): Their spending is moderate but shows significant variability ($3,786.84 on average, with a range from $2,155.02 to $11,120.14). This suggests a mix of middle-tier spenders and some premium customers, possibly reflecting a broad demographic spread.
- Other Attributes: Standard deviation in monetary value and frequency indicates that this segment may contain sub-segments with distinct shopping behaviors.

Cluster 2: The Infrequent Shoppers

- Recency (R): A higher average recency of 101.11 days characterizes this segment as less engaged, with some customers potentially at risk of churning.
- Frequency (F): They have a relatively low average frequency of 48.11 transactions, and the low standard deviation (58.75) suggests that this segment's shopping behavior is more uniform and infrequent.
- Monetary (M): This cluster has the lowest average spend ($553.87), aligning with their infrequent purchases. The narrow range of spending (from $0 to $2,164.25) could indicate a segment with consistently low spend per transaction.
- Other Attributes: The narrow range in both frequency and monetary values suggests that this group may consist of occasional buyers or deal-seekers.

Cluster 3: The Loyal High Spenders

- Recency (R): The very low average recency (4.83 days) suggests that this segment is highly engaged and has interacted with the business very recently.
- Frequency (F): With a high average frequency of 1948.67 transactions, this group stands out as the most frequent shoppers. The high standard deviation (2088.69) reflects a wide variety in shopping habits, potentially indicating a mix of highly loyal customers and those making bulk purchases.
- Monetary (M): The high average spend of $19,257.38 and the large range in spending (from $11,870.26 to $38,290.66) confirm that this cluster contributes significantly to revenue, marking them as valuable customers.
- Other Attributes: The segment's high spending and frequency could reflect a group with a preference for premium products or services, potentially including business clients or collectors.

Cluster 4: The One-Time High Rollers

- Recency (R): An average recency of 1 day suggests a very recent interaction, likely a significant purchase event.
- Frequency (F): The frequency is reported singularly at 4024, which is exceptionally high and likely indicative of an outlier, suggesting a data anomaly or a significant one-time bulk purchase.

- Monetary (M): An exceptionally high monetary value of $74,816.69, with no reported variability, further suggests this could be a one-time event or outlier.

Based on RFM segments, we formulated actionable marketing recommendations for each customer segment.

Cluster 1: The Engaged Regulars

For the Engaged Regulars, a nuanced approach is essential. Personalized communication, particularly through email marketing, can showcase new products or services that align with their previous purchases, sparking renewed interest. Loyalty programs should be emphasized, with rewards that escalate with the frequency and value of purchases to encourage higher spending. Additionally, considering the variation within this group, targeted promotions should be strategically deployed, offering special discounts to reactivate those whose engagement levels may have dipped.

Cluster 2: The Infrequent Shoppers

To draw back Infrequent Shoppers, the brand could initiate re-engagement campaigns, perhaps with an emotional appeal like a "we miss you" message paired with a compelling discount. Establishing a feedback loop could provide insights into their sporadic engagement, allowing for more tailored offerings. Data-driven recommendations could leverage their previous purchases to suggest new products that resonate with their known preferences, thereby increasing transaction frequency.

Cluster 3: The Loyal High Spenders

For the Loyal High Spenders, a strategy of exclusivity and recognition could be effective. VIP treatments, such as access to exclusive products or services and invitations to brand events, can make these customers feel valued. A tiered rewards system, where high-frequency and high-value transactions are acknowledged with increasingly attractive benefits, can reinforce their loyalty. Experiential engagement, such as exclusive events or personalized services, can deepen their connection with the brand.

Cluster 4: The One-Time High Rollers

With One-Time High Rollers, the brand should focus on understanding the nature of their significant purchases and ensuring satisfaction. A follow-

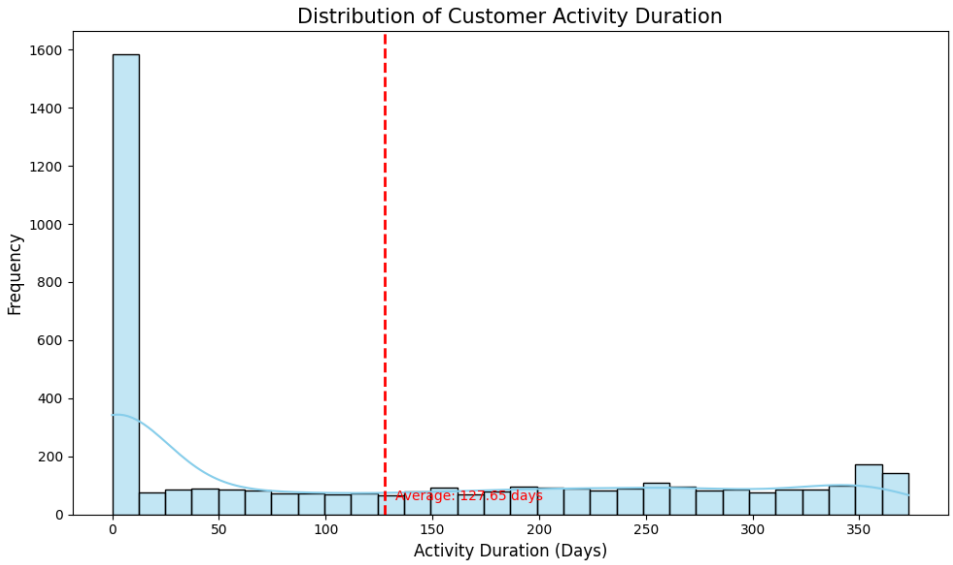
up after their purchase to check satisfaction levels and to suggest complementary products or services can promote repeat business. Offering a loyalty point bonus for a subsequent purchase within a specified timeframe may incentivize these customers to engage with the brand again.

Overarching Strategies

To enhance customer retention and revenue across all segments, optimizing the customer experience is paramount. This includes a seamless shopping journey from browsing to post-purchase service. Personalization, driven by deep data analytics, should inform all customer communications. Collecting and integrating feedback can help refine the brand's offerings. Engaging customers across multiple channels, maintaining consistent messaging, and utilizing dynamic content that reflects individual customer behaviors can make marketing efforts more effective. Lastly, fostering a sense of community can create an emotional bond with the brand, enhancing customer loyalty.

In conclusion, by tailoring marketing strategies to the distinct behaviors and preferences of each customer segment identified through RFM analysis, the business can improve customer satisfaction, increase the lifetime value of customers, and achieve sustainable revenue growth.

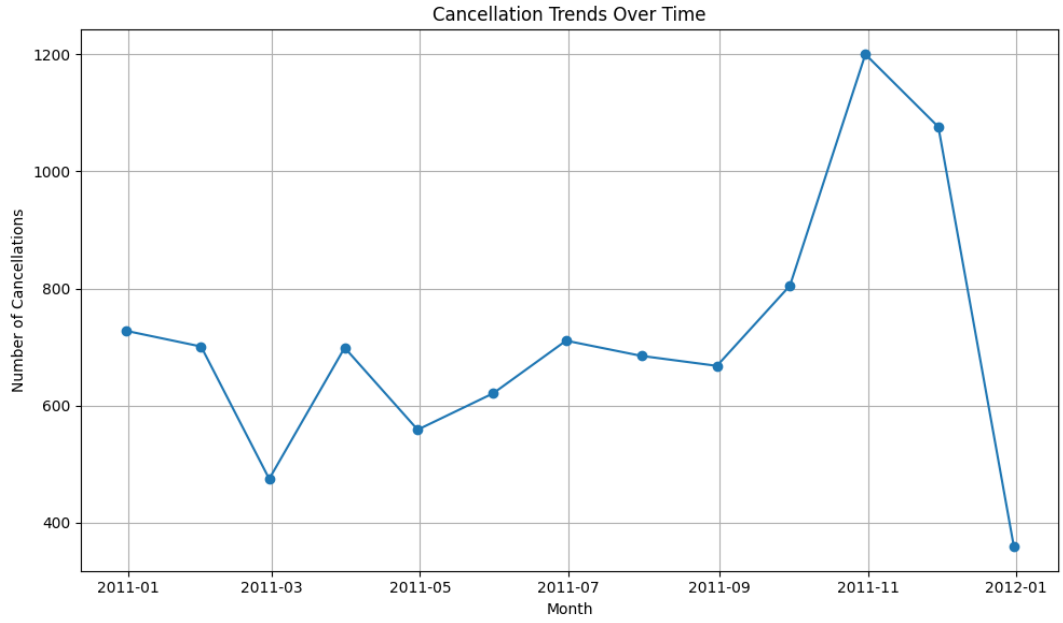
3) Customer Behavior Analysis

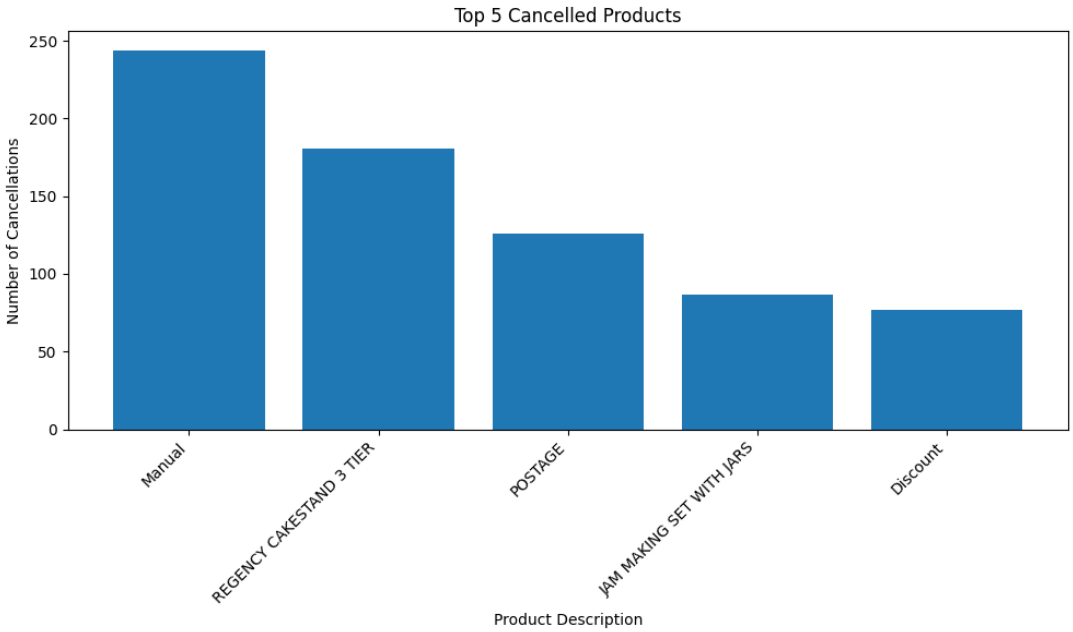


The distribution of customer activity duration is heavily right skewed. There is a very high frequency of customers with an activity duration close to 0 days. This could imply that many customers make their first and last purchase almost immediately. The average activity duration is approximately 127 days (about 4 months). Given the high number of customers with a short activity duration, the company should invest in customer retention strategies. This could include loyalty programs, personalized marketing, or special offers to encourage repeat purchases.

4) Customer Satisfaction Analysis

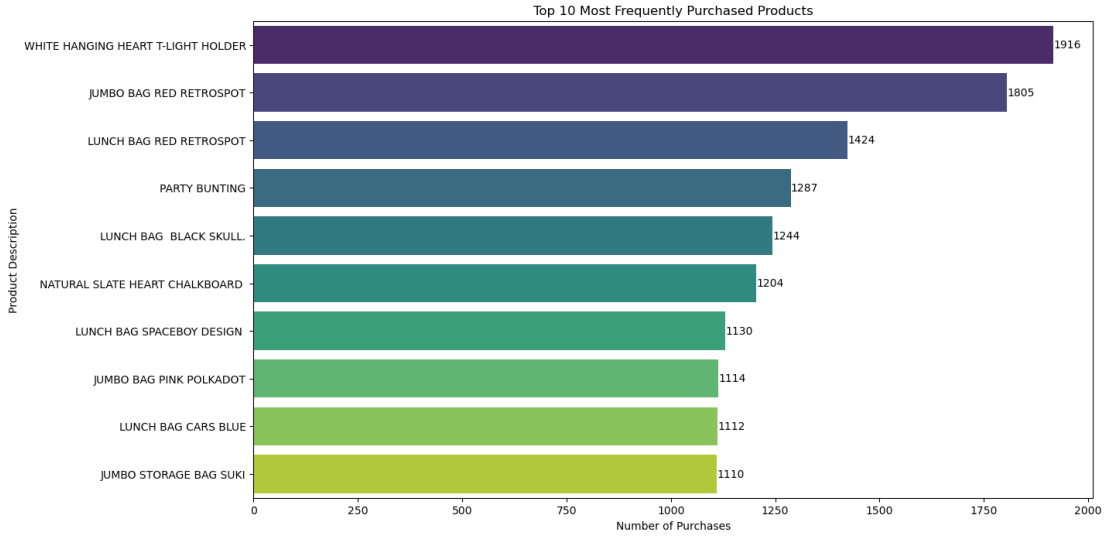
Although there is a lack of data in customer feedback, customer satisfaction can somehow be captured by examining the 'InvoiceNo' column. An InvoiceNo starting with 'C' signifies a canceled order, which, to a degree, implies customer dissatisfaction with the product.





The 'Manual' category has the highest number of cancellations, suggesting that there may be issues with this product or how it is sold. The 'REGENCY CAKESTAND 3 TIER' and 'POSTAGE' also have a high number of cancellations. This indicates potential issues with these products or services as well. The company should implement a feedback system to gather reasons for cancellations directly from customers. This will provide valuable insights into the root causes and help the company to address specific issues.

4. Product Analysis



The horizontal bar chart titled "Top 10 Most Frequently Purchased Products" displays the most popular products based on the number of times they were purchased. All products on the list have been purchased over 1,100 times, which highlights their significance to the retailer's sales.

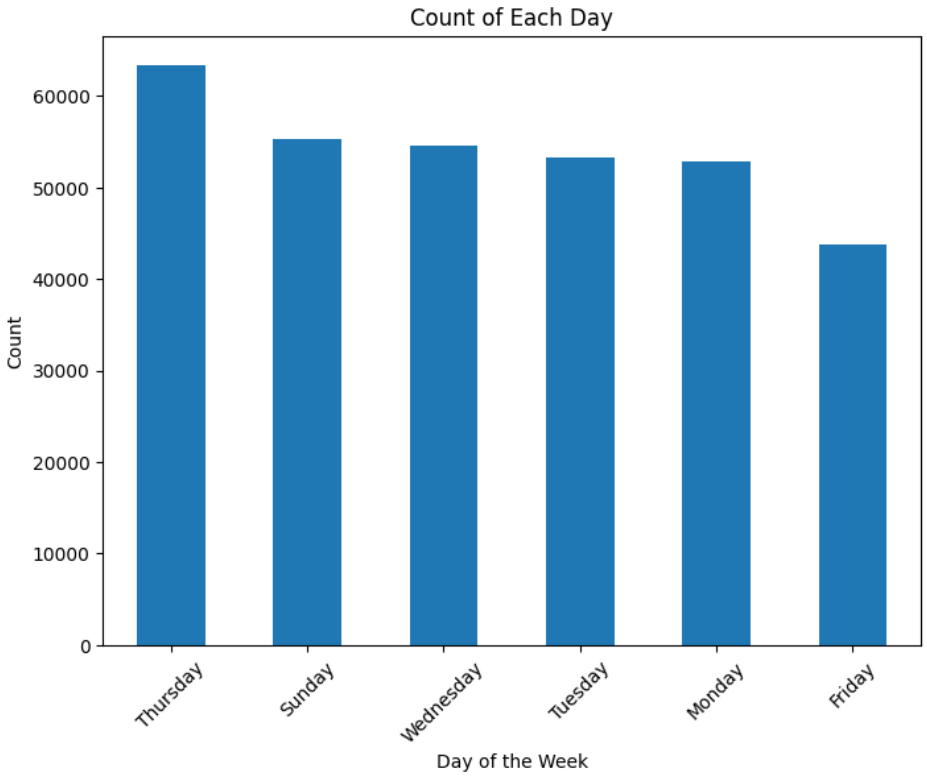
Leading the chart, the 'white hanging heart t-light holder' stands out with 1,916 purchases, indicating its popularity among customers. Following closely is the 'jumbo bag red retrospot' with 1,805 purchases. Other items in the top ten include various lunch bags with distinct designs (Red Retrospot, Black Skull, Spaceboy, and Cars Blue), each ranging from 1,112 to 1,424 purchases, demonstrating a strong customer preference for these designs.

Home decor items, such as 'party bunting' and 'natural slate heart chalkboard,' also make a strong showing, with 1,287 and 1,204 sales respectively, indicating consumer interest in these categories as well.

To compute the average product price, the total revenue from all transactions is aggregated and then divided by the cumulative quantity of products sold. This average price at 2.22, therefore, is weighted according to the sales volume of each product, yielding a nuanced measure of the dataset’s pricing landscape.

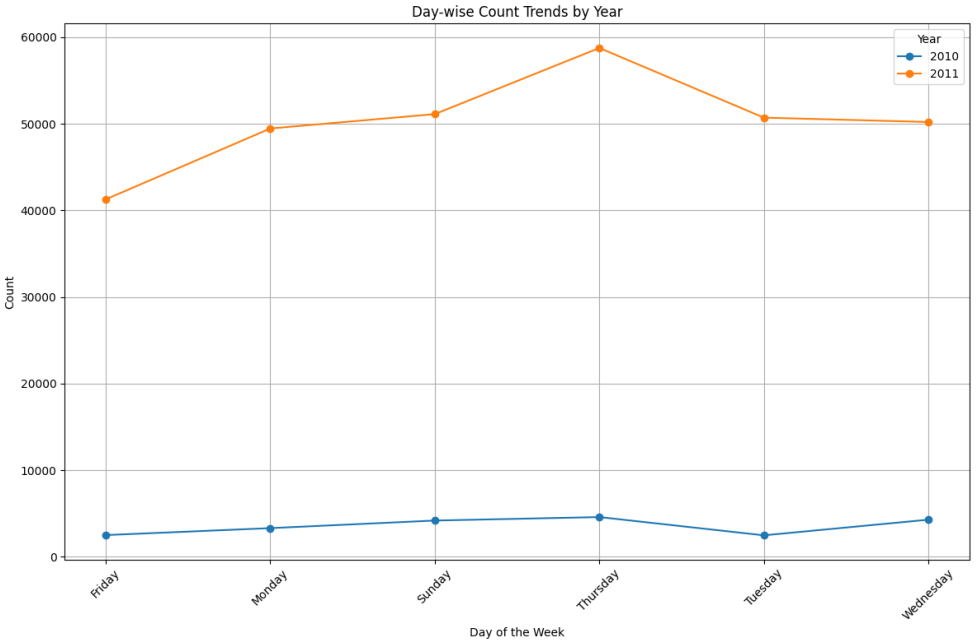
It is observed that 'Party Bunting' emerges as the product category contributing the most to revenue, amassing 33,819.75 British pounds in sales.

5. Time Analysis

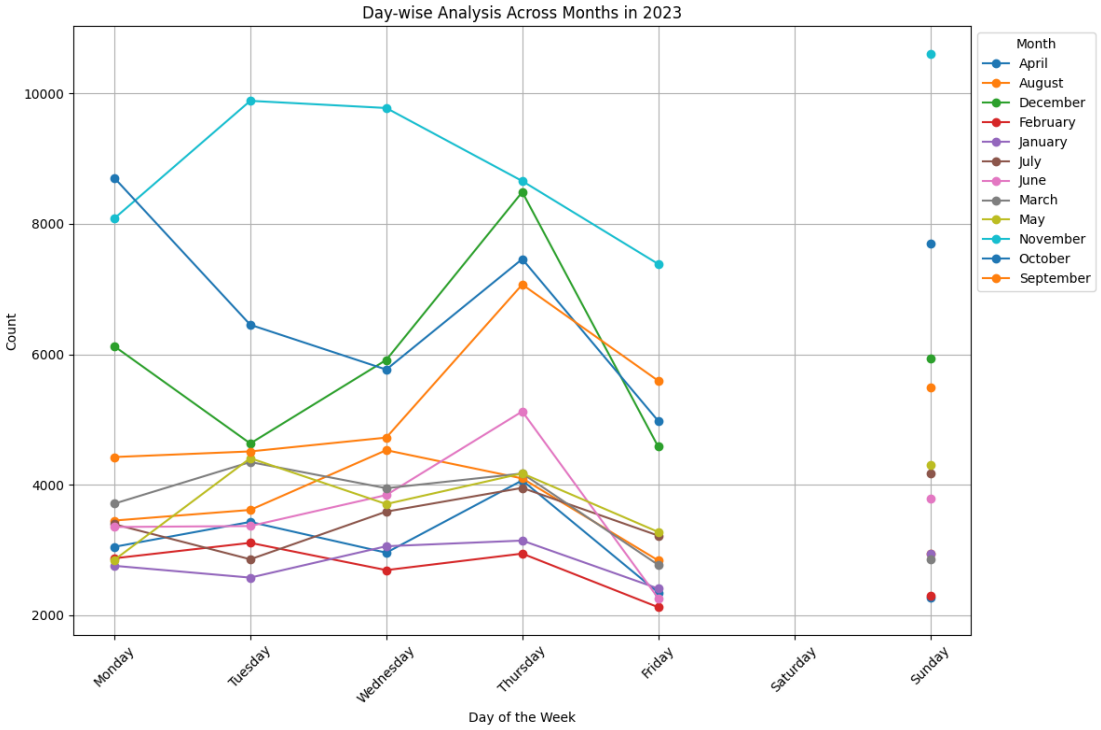


It's clear that Thursdays are the busiest days for orders. People seem to be ordering a lot more on Thursdays, showing a regular pattern of increased demand.

On the flip side, Fridays consistently have the fewest orders throughout the whole period we're looking at. This means that not many people are placing orders on Fridays. It might be interesting to figure out why there's this consistent drop in customer activity on Fridays.

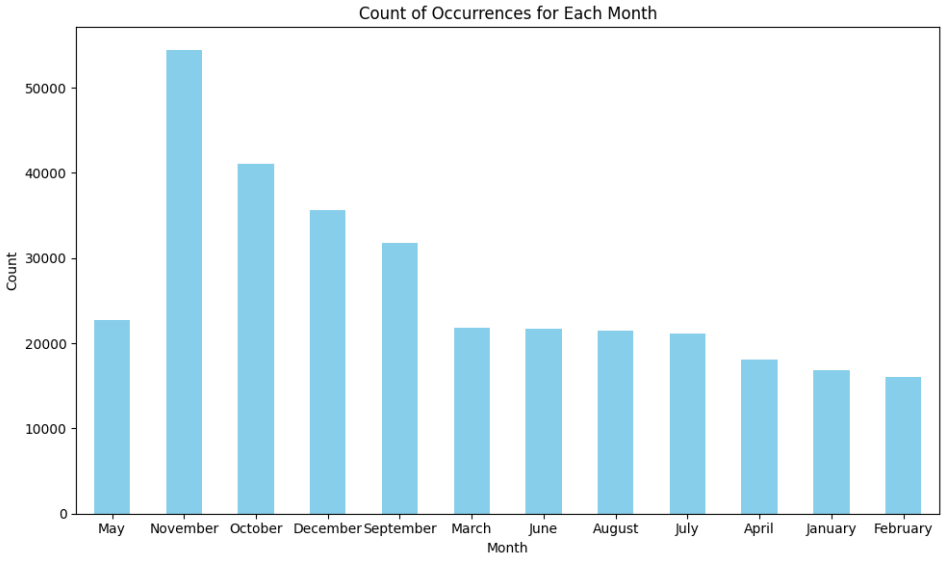


Next, we scrutinized the day-wise order counts across individual years within the dataset. In 2010, the total order count appears evenly distributed across all weekdays. In contrast, 2011 exhibits fluctuations in the order count per day. Specifically, Thursdays have the highest order count, while Sundays experience a notable decline in the total number of orders during 2011.



The day-wise trends is analyzed across various months:

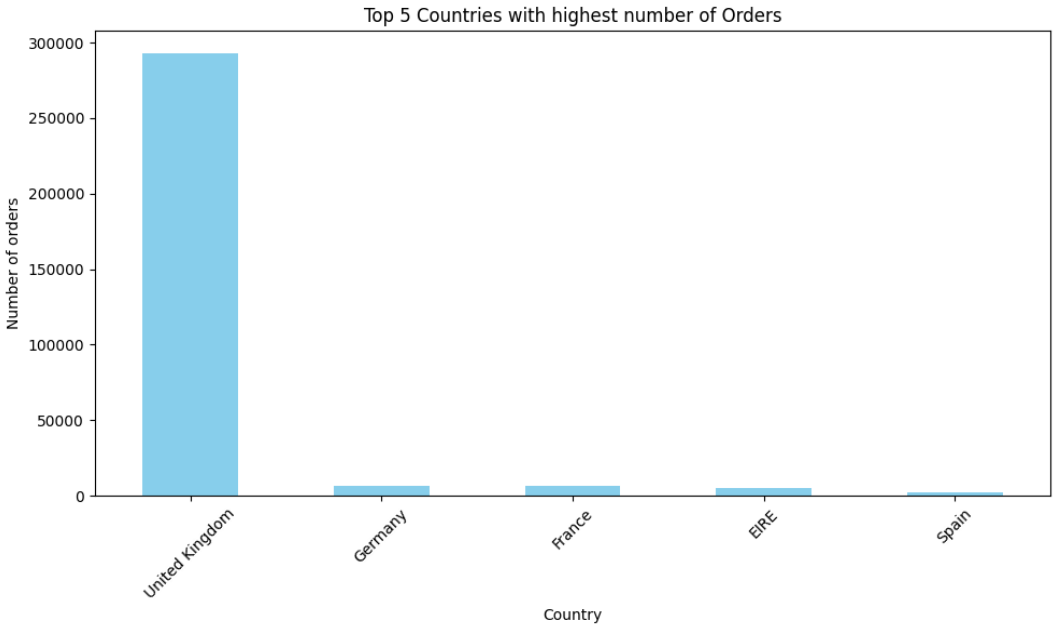
- In January and February, Mondays witness the highest order counts, while Fridays consistently show the lowest.
- March sees the highest orders on Tuesdays, with Fridays marking the lowest counts.
- April indicates Thursday as the peak day for orders, while Fridays persist as the lowest.
- May portrays Tuesdays with the highest orders and Fridays with the lowest.
- June records Thursdays with the highest order counts and Fridays with the lowest.
- July showcases Mondays as the peak day and Tuesdays as the lowest.
- August displays Thursdays as the peak and Mondays as the lowest.
- September depicts Tuesdays with the highest counts and Fridays with the lowest.
- October shows Mondays as the peak and Fridays as the lowest.
- November highlights Tuesdays as the highest and Thursdays as the lowest.
- December illustrates Thursdays with the highest counts and Tuesdays with the lowest.

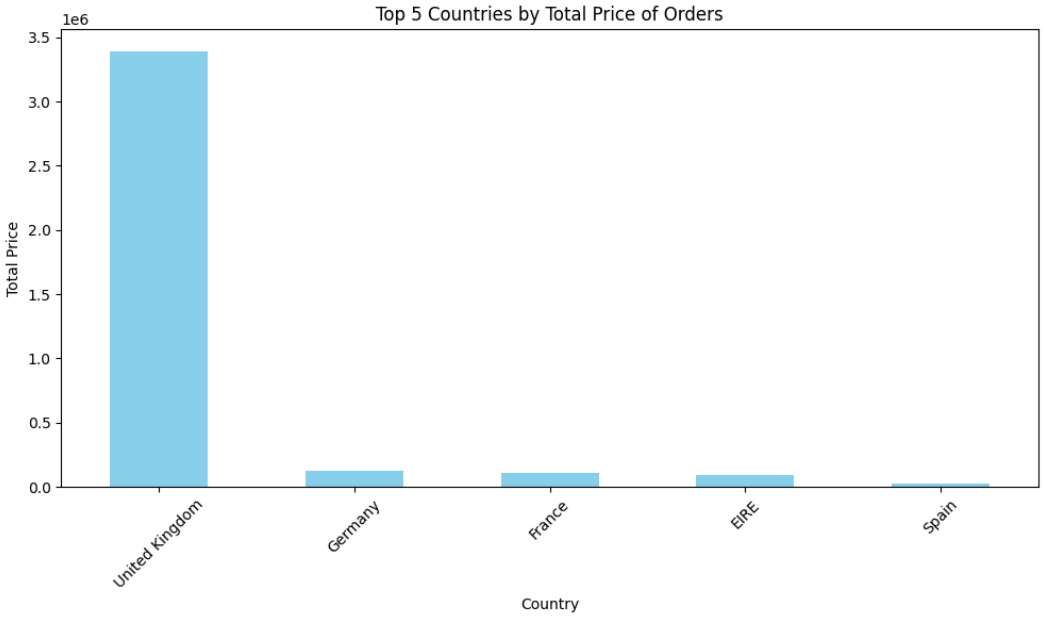


Subsequently, it's evident that the total volume of orders peaks in November and reaches its lowest point in February. One plausible explanation for this pattern could be attributed to increased order volumes during the Thanksgiving and Christmas seasons, possibly prompting higher ordering activities in November.

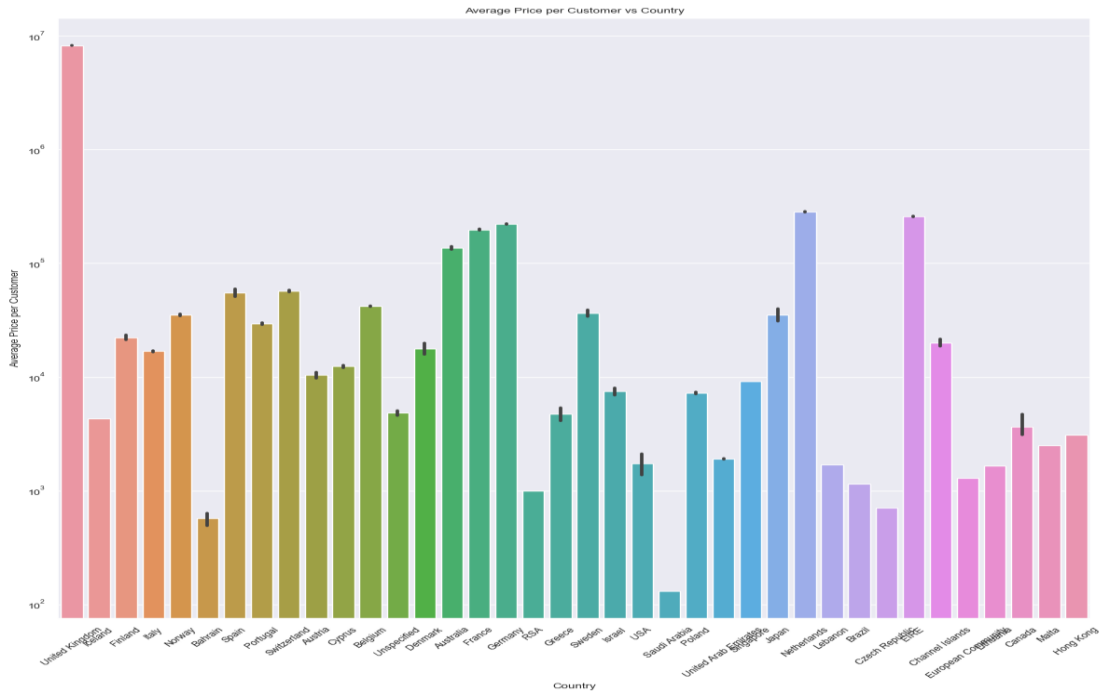
6. Geographical Analysis

In our study of different geographic regions, it's evident that the United Kingdom consistently stands out by having the highest number of orders. This means that compared to other locations, the UK is where we see the most significant quantity of orders being placed. The data consistently points to a notable trend of increased order activity in the United Kingdom when compared to other regions in our analysis. Exploring the reasons behind this heightened order volume in the UK could provide valuable insights into the factors driving customer engagement and demand in that specific geographic area.





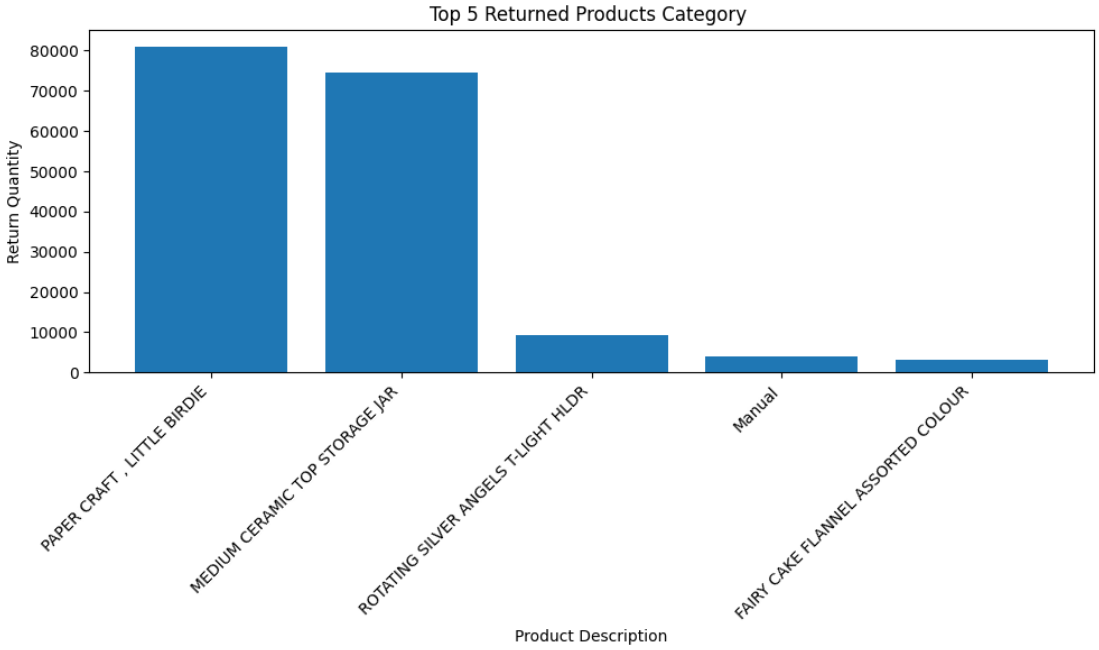
The United Kingdom clearly leads the pack with the highest order volume. Following closely behind are the top five countries: UK, Netherlands, EIRE, Germany, and France. This indicates that these nations, especially the UK, consistently have a substantial number of orders, making them key players in terms of customer engagement and demand.



Upon examination, it's discernible that customers from the United Kingdom tend to showcase a significantly higher average spending per transaction compared to those from Saudi Arabia. Conversely, countries like Australia and Germany display a more moderate average order value, suggesting a consistent mid-range expenditure pattern. The United States also aligns with this mid-level trend in average order values. This analysis underscores the diversity in spending behaviors across distinct countries, illustrating the variance in average spending per customer, with the UK exhibiting notably higher averages and Saudi Arabia reflecting comparatively lower averages.

7. Returns and Refunds

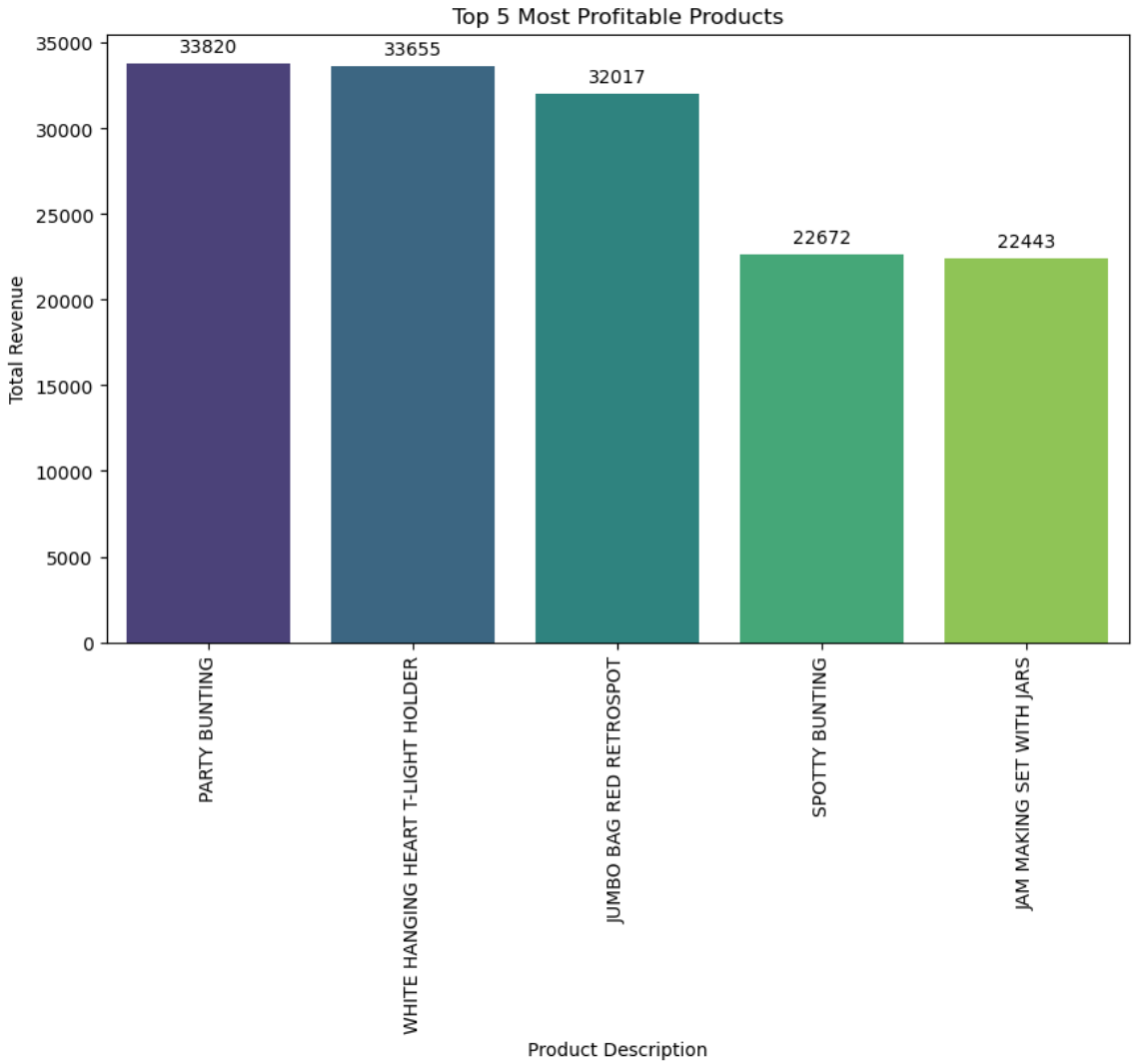
Products marked as returns have a 'Quantity' below 0 and 'InvoiceNo' starting with 'C'. We noted two types of negative quantity entries: those with 'C' invoice numbers and transactions with a unit price of 0, no CustomerID, and descriptions like "damages", "showroom", "thrown away", etc. These appear to be write-offs, not actual returns or refunds. Thus, we excluded them from our analysis. We focused only on transactions with 'C' starting invoice numbers. By comparing these to the total orders, we determined that the return or refund rate is 1.71%.



The 'PAPER CRAFT - LITTLE BIRDIE' and 'MEDIUM CERAMIC TOP STORAGE JAR' categories have significantly higher return quantities compared to other categories. This could suggest a higher likelihood of returns in these categories.

8. Revenue Analysis (Profitability)

From December 1, 2010, to December 9, 2011, the company's total revenue reached £4,788,851.55. The five leading products in terms of revenue contribution were: party bunting, generating £33,819.75; white hanging heart t-light holder, with £33,655.32; jumbo bag red retrospot, at £32,016.72; spotty bunting, which brought in £22,671.55; and the jam making set with jars, contributing £22,443.06.



Nevertheless, due to the lack of cost data, it was not possible to perform profitability analysis.

9. Payment Analysis

During the Data Preprocessing stage, we could not find any particular data for payments. Because of this, we were unable to produce any meaningful analysis of payments.

10. Conclusion

The in-depth analysis conducted on the e-commerce data revealed several actionable insights. Firstly, customer engagement levels vary significantly, indicating the potential for more personalized marketing strategies to enhance retention and increase transaction frequency. The identification of specific customer segments through RFM analysis suggests differentiated marketing approaches, from loyalty rewards for frequent buyers to targeted re-engagement for those less active.

The pronounced skew in customer activity duration points to a need for strategies that foster longer-term customer relationships. Implementing programs that incentivize repeat purchases and enhance customer experience could help mitigate the high attrition rates indicated by the data.

Product popularity and revenue analysis underscore the importance of stock optimization and product placement. Products that contribute significantly to revenue, like 'Party Bunting', warrant strategic promotion and inventory planning to maintain sales momentum. Moreover, a high return rate in specific categories calls for a review of product quality, return policies, and post-purchase customer support.

Geographically, the focus should remain on the UK market, but with an eye on expanding reach and adapting strategies to tap into the purchasing power evident in other regions. Investment in a feedback system to capture customer satisfaction and reasons for returns would enable more nuanced product and service adjustments.

While the dataset provided substantial revenue figures, the absence of detailed cost and payment data represents an opportunity for future analysis. To build on these findings, the company would benefit from integrating cost data to assess profitability and from capturing payment trends for a more comprehensive understanding of customer purchasing behavior.

In conclusion, the insights from this analysis offer a roadmap for strategic enhancement across the business. By leveraging data-driven customer insights, refining product offerings, and optimizing operations for peak

periods and key markets, the business can aim to not only sustain but also expand its market share and revenue streams.