Project -2

Customer Segmentation using RFM Analysis

IE6400 Foundations Data Analytics Engineering

Group Number 4

Bhavya Pandey
Neel Kamal
Rahul Daruka
Susilkessav Seshadri Bhuvaneswari

Project Report: Customer Segmentation using RFM Analysis

Tasks:

1. Data Preprocessing:

1.1 Importing the Dataset:
The initial phase involved importing the eCommerce dataset, which was retrieved from the provided Kaggle link. The dataset was loaded into a panda DataFrame, allowing for efficient exploration and manipulation of the data.

1.2 Handling Missing Values:

1.2.1 Description Column (1454 missing values):
The 'Description' column had 1454 missing values. After careful consideration, it was decided to retain these missing values in the dataset. Since our focus is on RFM analysis for customer segmentation, the absence of a product description for certain entries is not deemed critical.

1.2.2 CustomerID Column (135,080 missing values):
The 'CustomerID' column, crucial for customer identification in RFM analysis, had 135,080 missing values. To ensure the integrity of customer-centric analysis, rows with missing 'CustomerID' entries were removed from the dataset. This approach allows us to concentrate on customers with identifiable information.

1.3 Converting Data Types:

1.3.1 InvoiceDate Column:
The 'InvoiceDate' column, originally in object format, was converted to the datetime64[ns] data type. This conversion facilitates seamless handling of date-related calculations, ensuring accurate recency calculations for RFM analysis.

1.3.2 CustomerID Column:
The 'CustomerID' column, initially represented as floating-point numbers, was converted to the int64 data type. This conversion aligns with standard practices for customer identification and supports compatibility with subsequent calculations.

2. RFM Calculation:

In this section, we further refined the RFM metrics by incorporating the most recent purchase date and additional transaction-related metrics.

2.1 Recent Purchase Date (Recency - R):

To calculate the recency (R) of each customer, we first converted the 'InvoiceDate' to the 'InvoiceDay' column, representing the date of each transaction. The maximum 'InvoiceDay' was then identified as the most recent purchase date across all customers. The 'Days_Since_Last_Purchase' metric was computed by subtracting the most recent date from each customer's last purchase date.

2.2 Total Transactions (Frequency - F):

To gauge the frequency (F) of customer transactions, we determined the total number of unique invoices ('Total_Transactions') for each customer. This metric quantifies how often a customer engages in transactions with the business.

2.3 Total Products Purchased:

To enrich our understanding of customer behavior, we introduced the 'Total_Products_Purchased' metric, representing the total quantity of unique products purchased by each customer. This provides insights into the diversity and breadth of a customer's product preferences.

2.4 Resulting RFM Metrics:

The resulting dataset now includes the following RFM metrics:

- CustomerID: Unique identifier for each customer.
- Days_Since_Last_Purchase: The recency of each customer's last purchase in terms of days.
- Total_Transactions: The total number of unique transactions made by each customer.
- Total_Products_Purchased: The total quantity of unique products purchased by each customer.

These metrics form a more comprehensive view of customer behavior, capturing recency, frequency, and product diversity. The subsequent sections will leverage these refined RFM metrics for customer segmentation, providing valuable insights for targeted marketing and engagement strategies.

3. RFM Segmentation:

In this section, we assigned RFM scores to each customer based on quartiles, enabling the creation of a single RFM score for comprehensive customer segmentation.

3.1 Recency Score (R):

Methodology:

The most recent purchase date for each customer was determined, and the 'Days_Since_Last_Purchase' metric was calculated.

Quartiles were applied to 'Days_Since_Last_Purchase' to create Recency Scores, with lower scores indicating more recent purchases.

3.2 Frequency Score (F):

Methodology:

Total transactions ('Total_Transactions') for each customer were calculated, and frequency ranks were assigned based on transaction counts.

Quartiles were applied to the frequency ranks to create Frequency Scores, with higher scores indicating more frequent transactions.

3.3 Monetary Score (M):

Methodology:

Totals spend ('Total_Spend') for each customer was calculated, and quartiles were applied to create Monetary Scores, with higher scores indicating higher monetary contributions.

3.4 Creating RFM Score:

Methodology:

The Recency, Frequency, and Monetary Scores were combined to create a single RFM score for each customer, providing a holistic view of their engagement and contribution to the business.

3.5 Resulting RFM Segmentation:

The resulting dataset now includes the following RFM metrics and segmentation:

- CustomerID: Unique identifier for each customer.
- Recency_Score: Recency score based on quartiles.
- Frequency_Score: Frequency score based on quartiles.
- Monetary_Score: Monetary score based on quartiles.
- RFM_Score: Combined RFM score, indicating the overall engagement and value contribution of each customer.

4. Customer Segmentation:

In this section, we utilized clustering techniques, specifically K-Means clustering, to segment customers based on their RFM scores. The objective was to uncover meaningful groups that share similar characteristics, aiding in targeted marketing and engagement strategies.

4.1 Data Scaling:

Prior to clustering, the RFM scores (Recency, Frequency, and Monetary) were standardized using the StandardScaler. Standardization ensures that all features contribute equally to the clustering process, preventing any metric from dominating the results due to differences in scale.

4.2 Identifying Optimal Number of Clusters:

To determine the optimal number of clusters, the Elbow Method was employed. The Elbow Method involves fitting the K-Means algorithm with a range of cluster numbers and plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters. The "elbow" in the plot signifies the optimal number of clusters, where further partitioning offers diminishing returns.

4.3 K-Means Clustering:

Based on the Elbow Method, a specific number of clusters (k) was selected. In this case, **k = 4** was deemed appropriate for segmenting the customers.
The K-Means algorithm was then applied to the standardized RFM scores, resulting in the assignment of each customer to a specific cluster. The 'Cluster' column in the dataset indicates the segment to which each customer belongs.

4.4 Resulting Customer Segmentation:

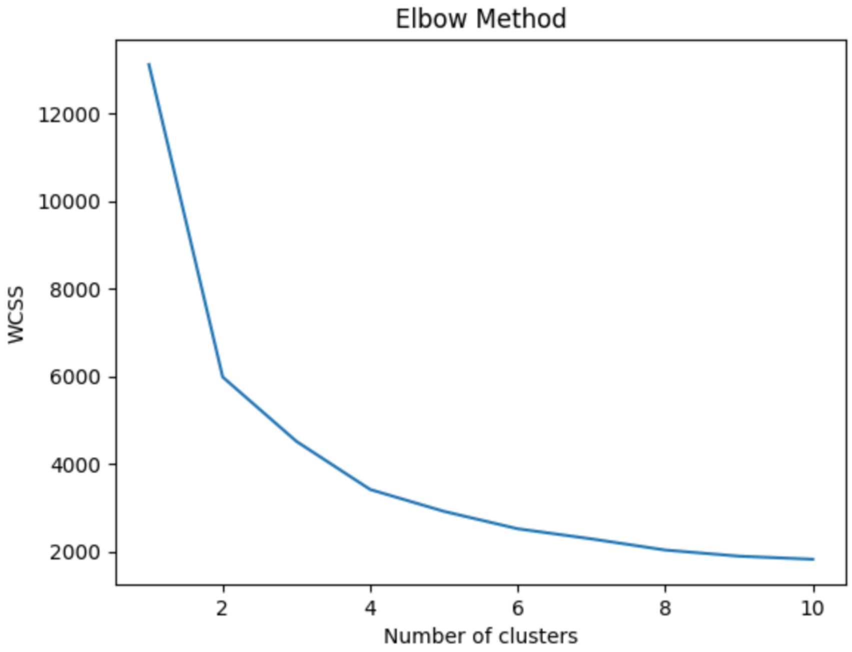
The resulting dataset now includes the 'Cluster' column, providing information about the segment to which each customer belongs. These segments represent groups of customers with similar RFM characteristics.

```
      CustomerID  ...  Cluster
0          12346  ...        2
1          12347  ...        3
2          12348  ...        0
3          12349  ...        1
4          12350  ...        2
...          ...  ...      ...
4367       18280  ...        2
4368       18281  ...        2
```

```
4369    18282  ...       3
4370    18283  ...       3
4371    18287  ...       1
```

4.5 Visualizing the Clusters:

The Elbow Method plot was instrumental in determining the optimal number of clusters. By selecting an appropriate number of clusters, we ensure that each segment is distinct and meaningful, providing valuable insights for targeted marketing strategies.



5. Segment Profiling:

In this section, we analyze and profile each customer segment based on the K-Means clustering results. We describe the characteristics of customers in each segment, including their RFM scores and any other relevant attributes.

5.1 Segment 0:

Characteristics:

- Recency (R): High recency scores, indicating recent purchases.
- Frequency (F): Moderate frequency scores.
- Monetary (M): Moderate monetary scores.

Recommendations:

- Offer exclusive loyalty programs to encourage purchases and increase frequency.

- Send personalized product recommendations based on searching patterns and previous purchases with special offers to tempt them back.
- Run email or SMS campaigns to keep them connected with the brand.

5.2 Segment 1:

Characteristics:

- Recency (R): Moderate recency scores.
- Frequency (F): High frequency scores, indicating loyal customers.
- Monetary (M): High monetary scores.

Recommendations:

- Recognize and reward their loyalty with member points or exclusive access to new products and events.
- Suggest complementary products to increase their order value.
- Ask for their feedback on products purchased and services to further tailor your offerings.

5.3 Segment 2:

Characteristics:

- Recency (R): Low recency scores.
- Frequency (F): Low to moderate frequency scores.
- Monetary (M): Low to moderate monetary scores.

Recommendations:

- Recommend higher-value products or bundles to increase their spend.
- Create promotions for limited time offers to encourage more frequent purchases.
- Share content that highlights the benefits of your products to boost their perceived value.

5.4 Segment 3:

Characteristics:

- Recency (R): Moderate to high recency scores.
- Frequency (F): Moderate frequency scores.
- Monetary (M): Moderate to high monetary scores.

Recommendations:

- Send targeted campaigns with attractive offers to re-engage them.
- Identify the reasons for their non-purchase and develop strategies to win them back.
- Ask for feedback and work on improving aspects that pushed them away from purchasing.

These segment profiles and recommendations provide actionable insights for tailoring marketing strategies to each group's unique characteristics and behaviors.

6. Marketing Recommendations:

In this section, we provide actionable marketing recommendations for each customer segment. These recommendations are tailored to improve customer retention and maximize revenue by addressing the unique characteristics and behaviors of each group.

6.1 High-Value Customers:

Recommendations:

- Retention: Focus on retaining these customers through loyalty programs and exclusive offers.
- Upsell: Identify complementary products and offer bundle deals to increase sales.

6.2 Potential High-Value Customers:

Recommendations:

- Promotions: Offer incentives to encourage additional purchases and discounts on related products.
- Personalization: Use purchase history for personalized product recommendations.

6.3 Low-Frequency, High-Value Customers:

Recommendations:

- Win-Back Campaigns: Target inactive customers with special promotions to reactivate them.
- Subscription Models: Introduce subscription services to ensure steady revenue.

6.4 Low-Value Customers:

Recommendations:

- Customer Education: Provide informative content to help customers understand product value.
- Reactivation Campaigns: Create reactivation campaigns with exclusive offers.

6.5 General Recommendations:

For clusters not explicitly mentioned, general recommendations include:

- Collect customer feedback: Understand customer preferences and pain points.
- Conduct A/B testing: Continuously optimize marketing strategies based on customer responses.

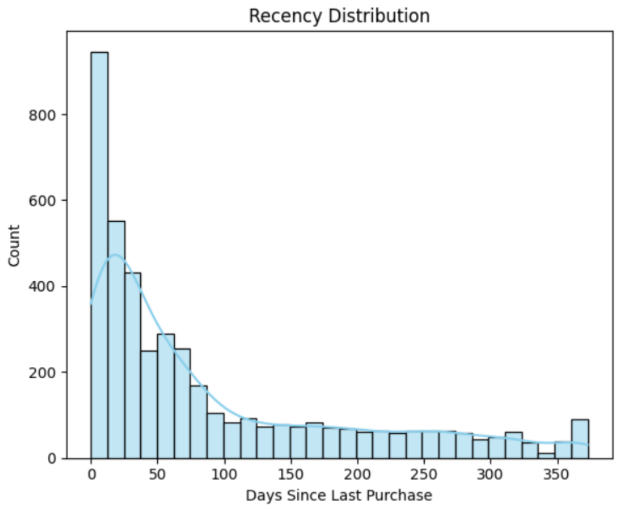
These targeted recommendations align with the specific needs and behaviors of each customer segment, offering a strategic approach for maximizing customer engagement, retention, and revenue.

7. Visualization:

7.1 RFM Distribution:

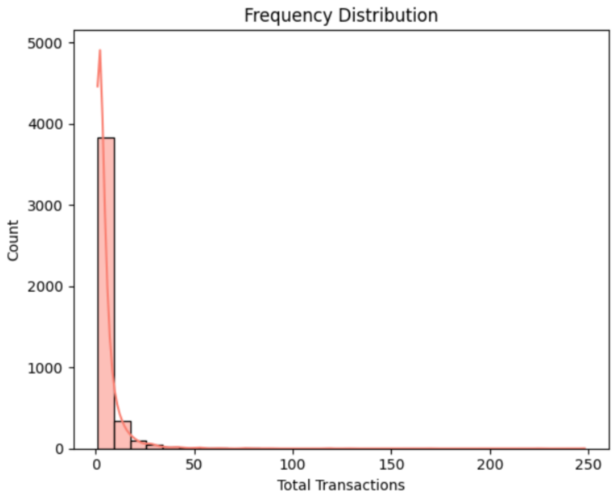
Recency Distribution:

The Recency Distribution chart depicts the distribution of days since the last purchase across customers. The peak around 0 indicates a group of customers who made recent purchases, while the tail shows customers with less recent transactions. Understanding recency is crucial for identifying active and potentially lapsed customers.

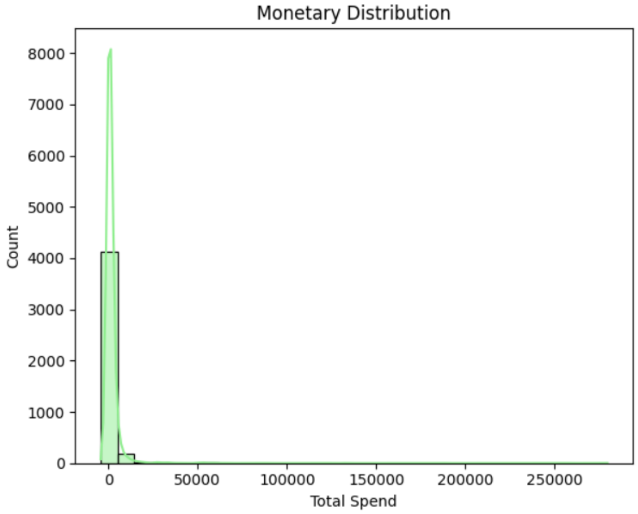


Frequency Distribution:

The Frequency Distribution chart illustrates the distribution of the total number of transactions for each customer. Peaks at higher frequencies indicate loyal and engaged customers, while the tail represents customers with fewer transactions. This insight helps in identifying the core customer base and tailoring strategies for different engagement levels.



Monetary Distribution:
The Monetary Distribution chart showcases the distribution of total spend by customers. Peaks at higher spend levels indicate high-value customers, while the tail represents customers with lower monetary contributions. Analyzing monetary distribution aids in identifying customers with significant value and tailoring strategies to maximize revenue.



Find The Solutions:

1. Data Overview:

Size of the Dataset:
The dataset contains 401,604 rows and 9 columns.

Column Descriptions:

- InvoiceNo: Unique identifier for each transaction.
- StockCode: Product code for items sold.
- Description: Description of the product.
- Quantity: Quantity of items purchased (negative values indicate returns).
- InvoiceDate: Date and time of the transaction.
- UnitPrice: Price per unit of the product.
- CustomerID: Unique identifier for each customer.
- Country: Country where the customer is located.
- Total_Spend: Total amount spent in a transaction (calculated as Quantity * UnitPrice).

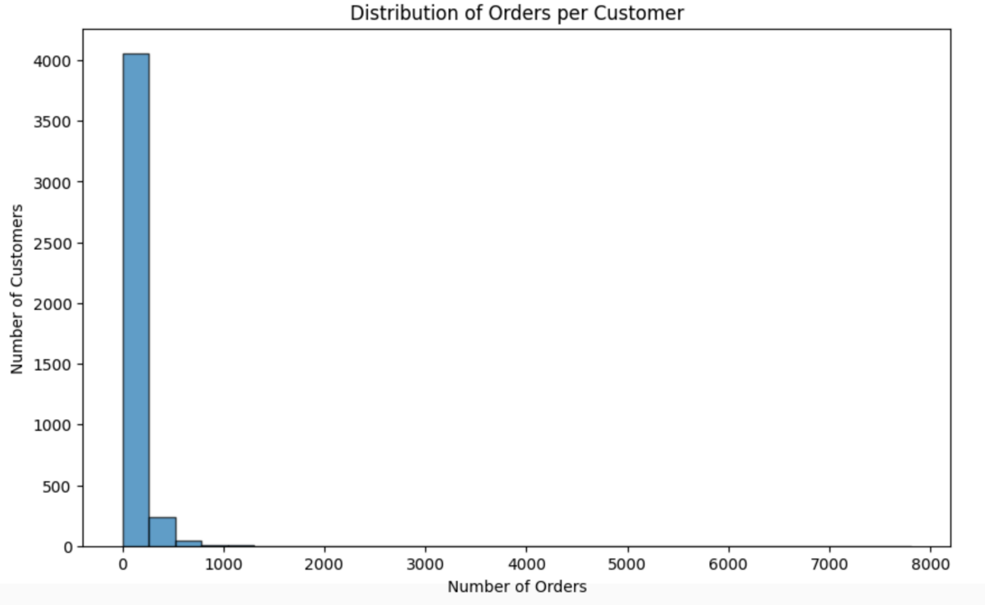
Time Period Covered:
The dataset covers transactions from December 1, 2010, at 08:26:00 to December 9, 2011, at 12:50:00.

2. Customer Analysis:

Number of Unique Customers:

The dataset encompasses transactions from a diverse set of 4,372 unique customers, each identified by a distinct CustomerID. This metric serves as a critical indicator of the dataset's coverage and represents the foundation for subsequent customer-centric analyses.

Distribution of Orders per Customer:



The distribution of orders per customer serves as a pivotal insight into the purchasing behavior of the customer base. This metric elucidates the spectrum of customer engagement, ranging from those with frequent and consistent orders to those with sporadic and infrequent transactions.

Top 5 Customers with the Most Purchases by Order Count:

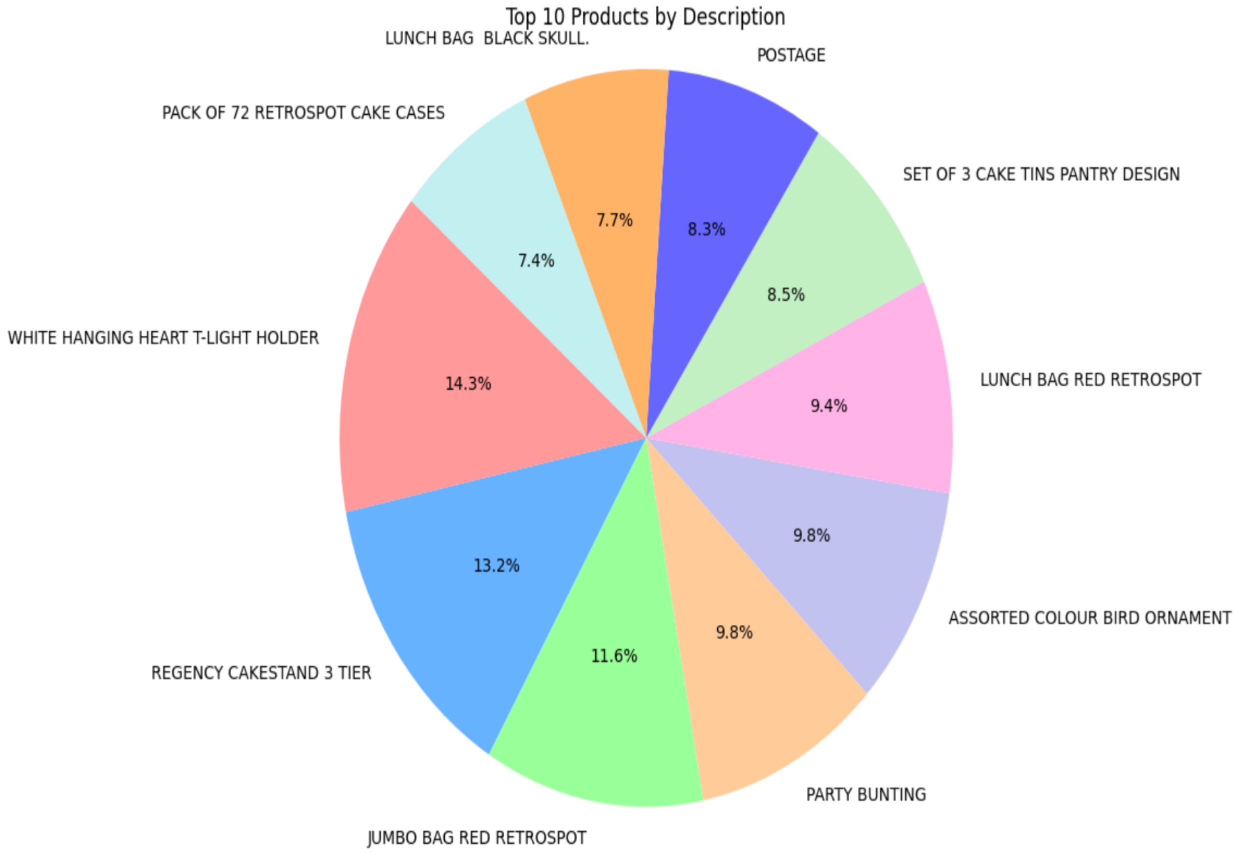
- Customer 17841: This customer has exhibited remarkable engagement, contributing a substantial 7,812 orders. This signifies consistent patronage and frequent interactions with the business.
- Customer 14911: Ranking second with 5,898 orders, this customer is another significant contributor to the order count, indicating sustained engagement.
- Customer 14096: With 5,128 orders, this customer secures the third position, demonstrating a substantial and consistent history of transactions.
- Customer 12748: Placing fourth with 4,459 orders, this customer has made a considerable number of purchases, showcasing notable engagement.
- Customer 14606: Completing the top five with 2,759 orders, this customer represents a segment of the customer base with consistent but comparatively fewer transactions.

3. Product Analysis:

Top 10 Most Frequently Purchased Products:

The list below represents the top 10 most frequently purchased products, along with the respective counts of purchases:

- WHITE HANGING HEART T-LIGHT HOLDER: 2058 purchases
- REGENCY CAKESTAND 3 TIER: 1894 purchases
- JUMBO BAG RED RETROSPOT: 1659 purchases
- PARTY BUNTING: 1409 purchases
- ASSORTED COLOUR BIRD ORNAMENT: 1405 purchases
- LUNCH BAG RED RETROSPOT: 1345 purchases
- SET OF 3 CAKE TINS PANTRY DESIGN: 1224 purchases
- POSTAGE: 1196 purchases
- LUNCH BAG BLACK SKULL: 1099 purchases
- PACK OF 72 RETROSPOT CAKE CASES: 1062 purchases



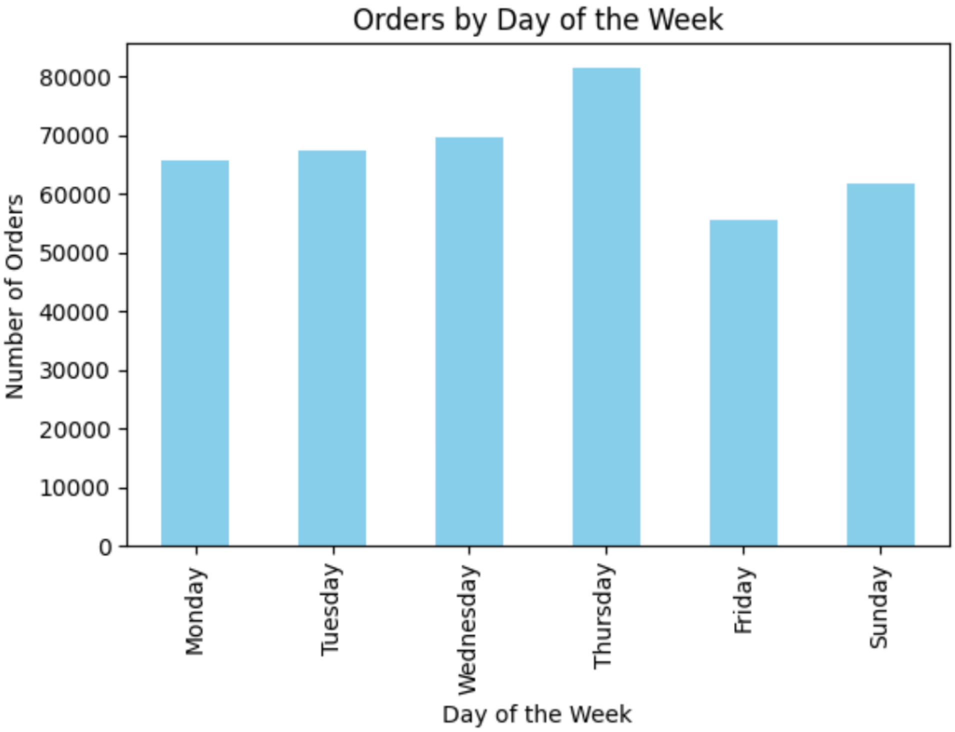
A pie chart visually represents the distribution of purchases among the top 10 products by description. This graphic provides a clear illustration of the relative popularity of each product in the dataset.

Average Price of Products:
The average price of products in the dataset is approximately $3.47. This metric provides an overview of the typical unit price across all products, aiding in understanding the general pricing structure.

Product Category Generating the Highest Revenue:
The product category with the highest revenue is identified by examining the StockCode associated with the maximum total revenue. In this dataset, the product category generating the highest revenue is REGENCY CAKESTAND 3 TIER. The total revenue attributed to this product category is approximately $132,567.70.

4. Time Analysis:

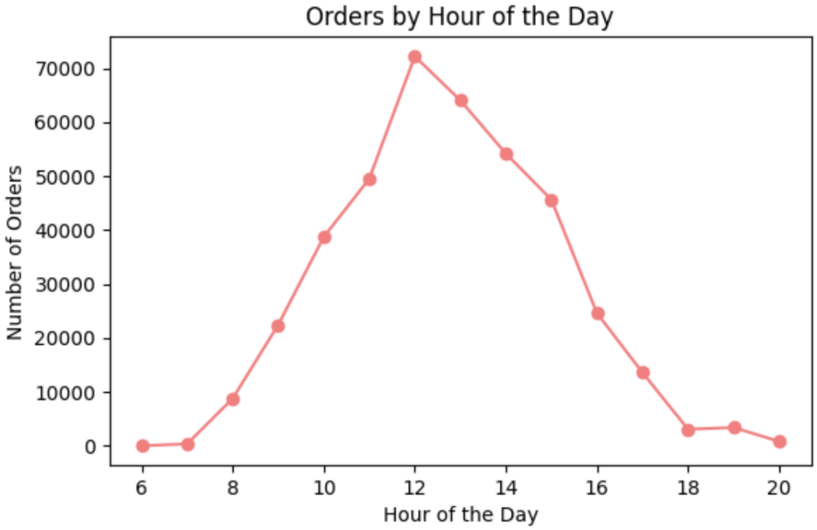
Orders by Day of the Week:



The bar chart illustrates the distribution of orders across different days of the week. The analysis revealed that:

Busiest Day: The dataset indicates that most orders are placed on **Thursday**. This insight can influence marketing and promotional strategies, ensuring optimal resource allocation on peak ordering days.

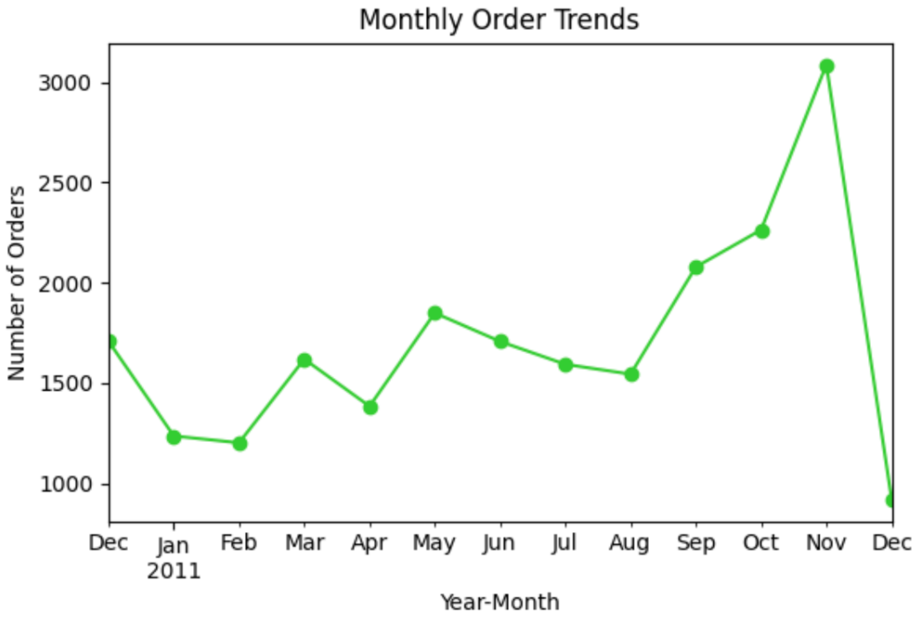
Orders by Hour of the Day:



The line chart visualizes the number of orders at different hours of the day. Key observations include:

Peak Hours: The graph indicates that the highest number of orders occurs during the **morning** and **early afternoon**, with a slight decrease towards the evening. Understanding peak ordering times is crucial for staffing and resource management.

Monthly Order Trends:



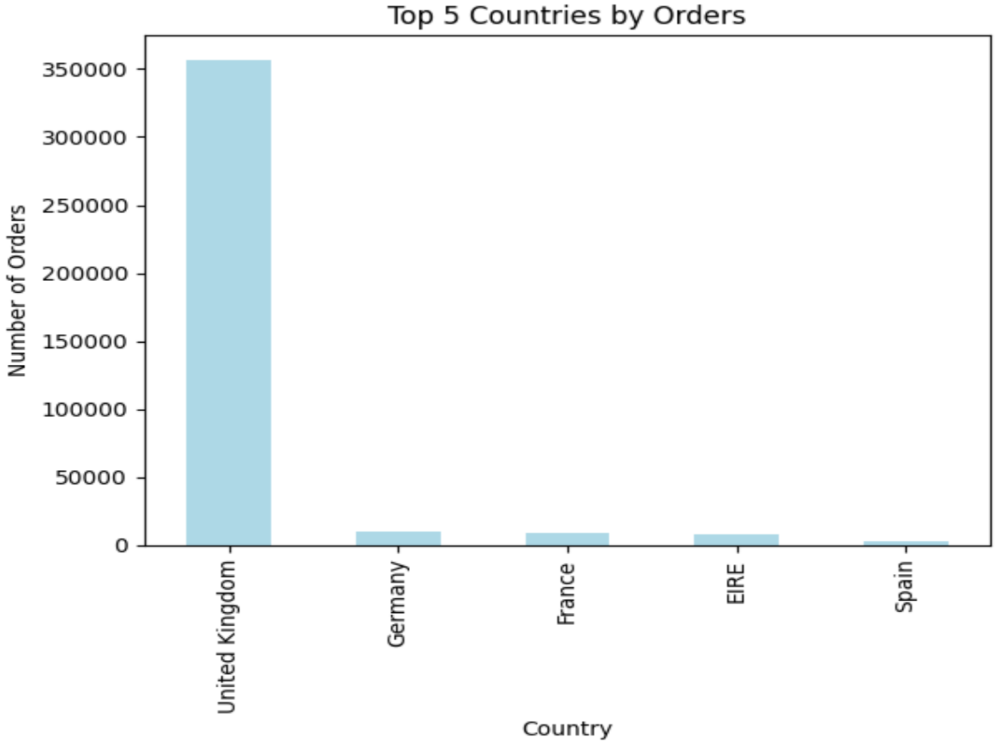
The line chart displays the monthly trends in the number of orders. Key findings include:

Seasonal Patterns: The dataset exhibits some seasonal trends, with variations in order volumes across different months. There is a significant increase in the month of November. And a significant drop in the month of December. Identifying seasonal patterns is essential for anticipating demand fluctuations and adjusting inventory levels.

Average Order Processing Time:
The average order processing time is calculated by measuring the time elapsed between consecutive orders for each customer. The calculated average order processing time is approximately **35.32** hours. This metric provides insights into the efficiency of the order fulfillment process.

5. Geographical Analysis:

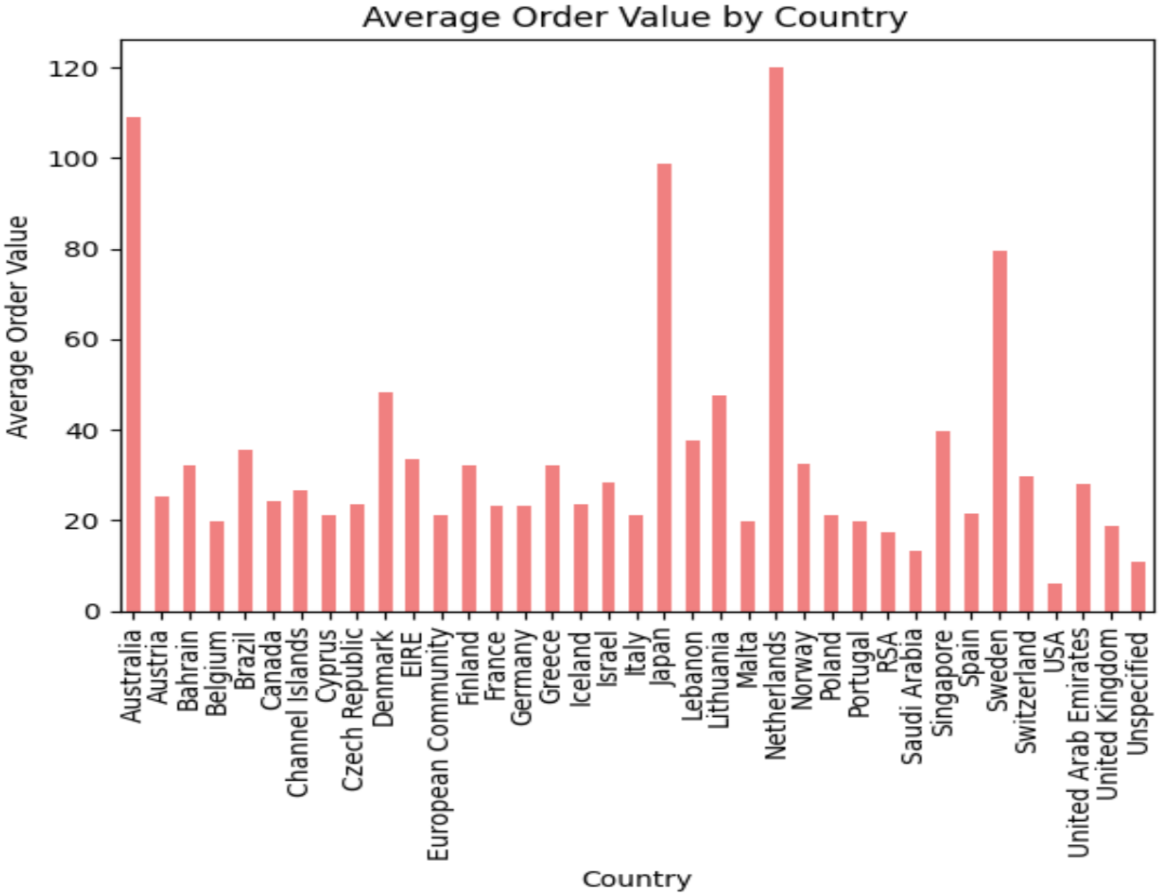
Top 5 Countries by Orders:



The bar chart on the left side displays the top 5 countries with the highest number of orders. Key observations include:

United Kingdom Dominance: The United Kingdom (UK) stands out as the country with the highest number of orders. This insight highlights the significance of the UK market in terms of order volume.

Average Order Value by Country:



The bar chart on the right side illustrates the average order value for each country. Notable findings include:

Varied Average Order Values: The average order values are distributed across countries, with some exhibiting higher average values and others lower. This variation may be influenced by factors such as consumer purchasing power, product pricing, and cultural preferences.

Correlation Analysis:

The correlation analysis aims to determine the relationship between the country of the customer and the average order value. The calculated correlation value is nan.

This geographical analysis aids in recognizing key markets, understanding spending patterns across countries, and identifying opportunities for targeted marketing and pricing adjustments.

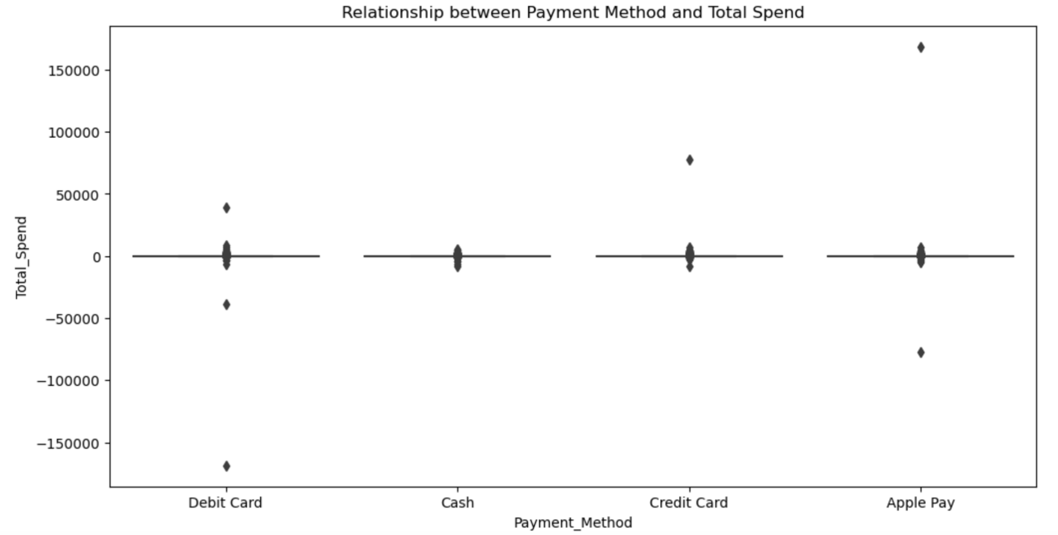
6. Payment Analysis:

Most Common Payment Methods:
The analysis reveals the most common payment methods used by customers:

- Debit Card: Number of Transactions: 101,658
- Credit Card: Number of Transactions: 101,415
- Apple Pay: Number of Transactions: 101,236
- Cash: Number of Transactions: 101,131

This information provides insights into the preferred payment methods among customers, aiding the company in optimizing payment processing and offering targeted payment options.

Relationship Between Payment Method and Order Amount:
A box plot is utilized to visualize the relationship between payment methods and the total spend (order amount). Box plots provide a clear representation of the distribution, central tendency, and potential outliers for different payment methods.



Box Plot Insights:
The box plot visually represents the spread and distribution of total spend for each payment method. It provides a quick overview of the central tendency, variability, and potential outliers in the data. The analysis of the box plot can uncover patterns, trends, and potential areas for further investigation. This payment analysis contributes valuable insights to the company's understanding of customer payment preferences and behaviors, supporting strategic decisions to improve overall customer satisfaction and operational efficiency.

7. Customer Behavior:

Average Duration of Customer Activity:
The average duration of customer activity is calculated by determining the time span between a customer's first and last purchase. The calculated average duration is approximately **133.39 days**. This metric provides insights into how long, on average, customers remain active in terms of making purchases.

Customer Segmentation Based on Purchase Behavior:
Customer segmentation is performed based on two key metrics: recency and frequency.

- Recency: The "Recency" column represents the number of days since the customer's last purchase. The values range from 1 day to 325 days, indicating varying recency levels among customers.
- Frequency: The "Frequency" column denotes the total number of transactions or purchases made by each customer. Frequencies range from 2 to 721, reflecting diverse purchasing behaviors.

Segment Thresholds and Assignment:
Segmentation involves defining thresholds for recency and frequency, and customers are then categorized into segments based on these thresholds.

- Recency Threshold: The median value of recency is used as the threshold.
- Frequency Threshold: The median value of frequency is utilized as the threshold.

Customer Segments:
Customers are assigned to two segments based on the defined thresholds:

- High Activity Segment: Customers with recency values below or equal to the recency threshold and frequency values above the frequency threshold are classified as having high activity.
- Low Activity Segment: Customers not meeting the criteria for the high activity segment are categorized as having low activity.

This analysis provides a comprehensive understanding of customer behavior, allowing businesses to adapt their strategies to different customer segments and maximize overall customer engagement and satisfaction.

8. Returns and Refunds:

Percentage of Orders with Returns or Refunds:
The percentage of orders that have experienced returns or refunds is calculated to be **2.21%**. This metric provides an overview of the overall rate of returns or refunds within the dataset.

Correlation Between Product Category and Likelihood of Returns:
The likelihood of returns is explored by examining the percentage of returns for each product category. Here are a few key points:

- Product Categories: The product categories are represented by their descriptions.
- Percentage of Returns: For each product category, the percentage of returns is calculated. This indicates the proportion of orders within a specific category that resulted in returns.

Percentage of Returns by Product Category:

```
Description
 4 PURPLE FLOCK DINNER CANDLES        NaN
 50'S CHRISTMAS GIFT BAG LARGE        0.909091
 DOLLY GIRL BEAKER                    1.459854
 I LOVE LONDON MINI BACKPACK          NaN
 I LOVE LONDON MINI RUCKSACK          NaN
                        ……..
ZINC T-LIGHT HOLDER STARS SMALL       1.244813
ZINC TOP  2 DOOR WOODEN SHELF         18.181818
ZINC WILLIE WINKIE CANDLE STICK       0.520833
ZINC WIRE KITCHEN ORGANISER           NaN
ZINC WIRE SWEETHEART LETTER TRAY      NaN
```

9. Profitability Analysis:

Total Profit Generated:
The total profit generated by the company during the dataset's period is calculated to be **$8,278,519.42**. This metric provides a comprehensive measure of the overall financial success of the company based on the total revenue and associated costs.

Top 5 Products with the Highest Profit Margins:
The profit margin for each product is calculated by subtracting the cost from the revenue and then dividing by the revenue. Here are the top 5 products with the highest profit margins:

- PAPER CRAFT, LITTLE BIRDIE (StockCode: 23843):
  Description: Paper craft item named "Little Birdie."
  Revenue: $168,469.60
- MEDIUM CERAMIC TOP STORAGE JAR (StockCode: 23166):
  Description: Medium-sized ceramic top storage jar.
  Revenue: $77,183.60
- PICNIC BASKET WICKER 60 PIECES (StockCode: 22502):
  Description: Wicker picnic basket with 60 pieces.
  Revenue: $38,970.00
- POST (StockCode: POST):
  Description: Postage service.
  Revenue: $8,142.75
- SET OF TEA COFFEE SUGAR TINS PANTRY (StockCode: 23243):
  Description: Set of tea, coffee, and sugar tins for the pantry.
  Revenue: $7,144.72

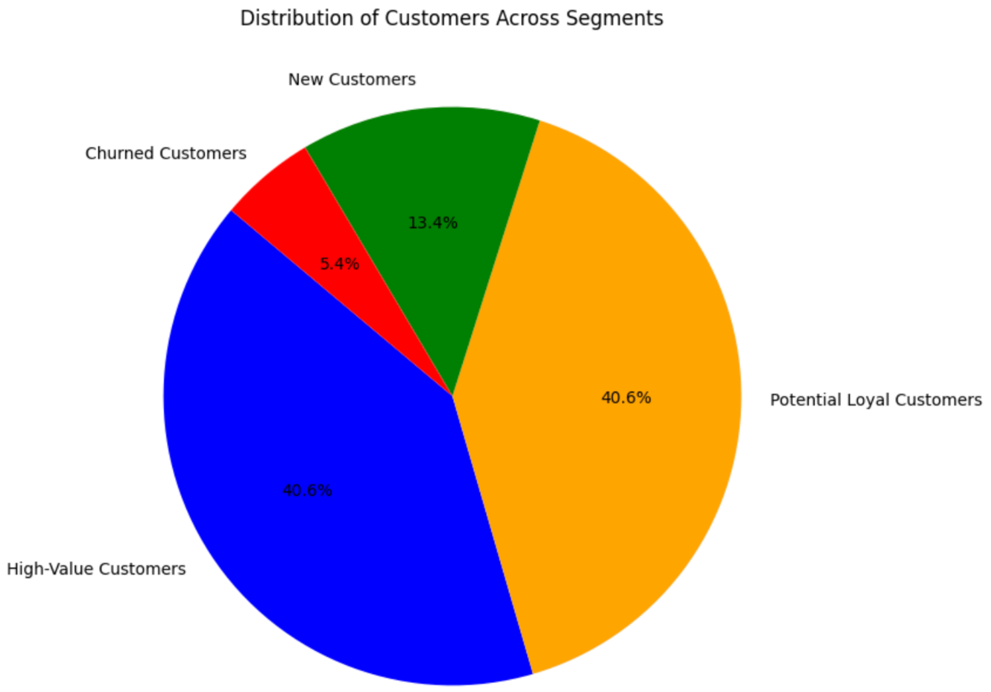
10. Customer Satisfaction:

RFM-based Customer Satisfaction:

- High-Value Customers:
  Characteristics: Low recency, high frequency, and high monetary scores.
  Count: 333 customers
- Potential Loyal Customers:
  Characteristics: Recent purchases, high frequency, moderate monetary scores.
  Count: 333 customers
- New Customers:
  Characteristics: Very recent purchases, low frequency, and low monetary score.
  Count: 110 customers
- Churned Customers:
  Characteristics: High recency, low frequency, and low monetary scores.
  Count: 44 customers

Pie Chart Analysis:

The pie chart provides a visual representation of the distribution of customers across satisfaction segments.



Key Observations:

High-Value Customers: Constitute a significant portion of the customer base, emphasizing their importance.

Potential Loyal Customers: A considerable share, indicating a promising group for future loyalty initiatives.

New Customers: Representing a smaller segment, potential for growth through targeted engagement.

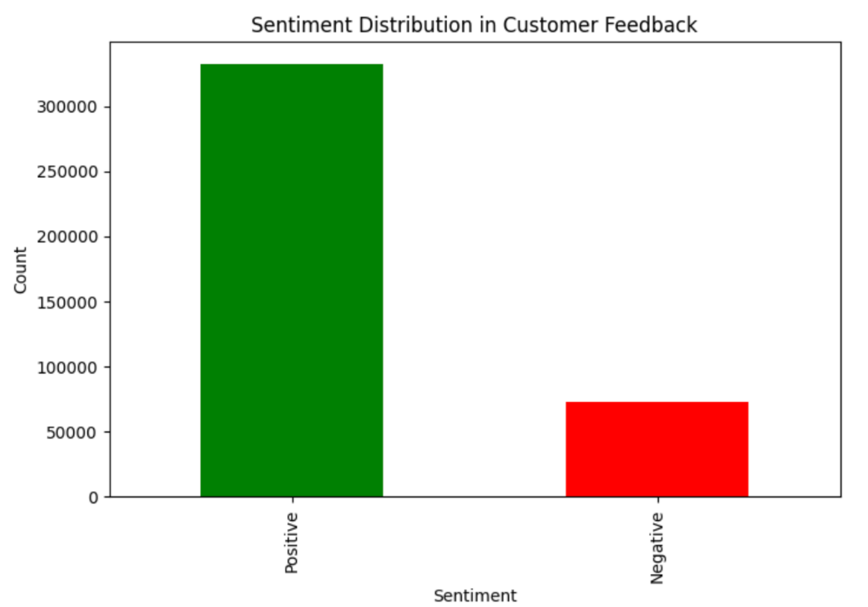
Churned Customers: A smaller but notable segment, requiring attention for potential re-engagement.

Sentiment Analysis in Customer Feedback:

For sentiment analysis, synthetic data was generated, and sentiment labels were assigned to simulate customer feedback.

Sentiment Labels:

- Positive
- Negative



The bar chart illustrates the sentiment distribution in customer feedback, highlighting the proportion of positive, negative sentiments., providing insights into the overall sentiment trends.

Strategic Implications:

- Segment-Specific Strategies: Tailor marketing and engagement strategies based on customer segments identified through RFM analysis.
- Feedback-driven Improvements: Leverage sentiment analysis to identify areas for improvement in products or services based on customer feedback.
- Customer Retention: Implement strategies to retain high-value and potential loyal customers, addressing their specific needs and preferences.

Overall Impact:

The combined analysis provides a comprehensive view of customer satisfaction, enabling the company to tailor its strategies to enhance overall customer experience and loyalty. The insights gained from these analyses will empower the company to make data-driven decisions, improve customer satisfaction, and foster long-term relationships with its customer base.