**IE 6400: Foundations of**

**Data Analytics**

**Project 2**

**Customer Segmentation**

**using RFM Analysis**

**Group 18**

Member Details:

Manogna Devalla : [devalla.m@northeastern.edu](mailto:devalla.m@northeastern.edu)

Amogha Gadde : [gadde.am@northeastern.edu](mailto:gadde.am@northeastern.edu)

Divya Babulal Shah : [shah.divyab@northeastern.edu](mailto:shah.divyab@northeastern.edu)

Kumari Simran : [simran.k@northeastern.edu](mailto:simran.k@northeastern.edu)

Abhishek Kumar Sinha : [sinha.ab@northeastern.edu](mailto:sinha.ab@northeastern.edu)

Contributions:

Percentage of Effort Contributed by Manogna: 20%

Percentage of Effort Contributed by Amogha: 20%

Percentage of Effort Contributed by Divya: 20%

Percentage of Effort Contributed by K.Simran: 20%

Percentage of Effort Contributed by Abhishek: 20%

Submission Date: Dec-01-2023

# Acknowledgements

Foremost, we would like to express our sincere gratitude to our Professor Sivarit Sultornsanee for giving us the opportunity to make this project. We would also like to thank our TAs for the continuous support of our study and research while making this project, for their patience, motivation and enthusiasm and immense knowledge.

We are thankful that we got an amazing team and for all the stimulating discussions and for all the experience and knowledge that we have gained by working together. We would also like to thank our parents and all our friends that have motivated and for never doubting our ability to complete this project.

# Contents

|  |  |
| --- | --- |
| Acknowledgement | Page 2 |
| Introduction | Page 4 |
| 1. Objective | Page 4 |
| 1. Data Set | Page 4 |
| 1. Data Inspection and Cleaning | Page 4 |
| 1. RFM Analysis | Page 4-5 |
| 1. Customer Segmentation | Page 6-8 |
| 1. Customer Profiles and Marketing Recommendations | Page 8 |
| Solutions to given questions | Page 9 - 23 |
| 1. Data Overview | Page 9 - 11 |
| 1. Customer Analysis | Page 11 - 12 |
| 1. Product Analysis | Page 13 |
| 1. Time Analysis | Page 14 -16 |
| 1. Geographical Analysis | Page 16 - 18 |
| 1. Payment Analysis | Page 18 - 20 |
| 1. Customer Behavior | Page 20 - 22 |
| 1. Returns and Refunds | Page 22 - 23 |
| 1. Profitability Analysis | Page 24 - 25 |
| 1. Customer Satisfaction | Page 25 - 26 |

# Introduction

In this project assignment, the focus is on leveraging the eCommerce dataset. The primary goal is to employ the RFM (Recency, Frequency, Monetary) analysis method, a robust technique widely used in the business domain for customer segmentation. RFM analysis enables the grouping of customers based on their recent purchasing behavior, purchase frequency, and monetary value. This segmentation empowers businesses with the ability to tailor marketing strategies and customer engagement approaches more effectively.

## Objective

In this project, our main objective of this project is to conduct RFM analysis on the provided eCommerce dataset and subsequently segment the customers into distinct groups based on their RFM scores. By performing this segmentation, we aim to uncover meaningful patterns within the customer base, allowing businesses to gain valuable insights into customer behavior. By categorizing customers into specific groups, businesses can develop targeted marketing campaigns and customer retention strategies. Ultimately, the objective is to enhance the overall understanding of customer dynamics and optimize strategies for improved customer satisfaction and loyalty.

# Dataset

We used a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

# Data Inspection and Cleaning

1. There is a significant percentage of CustomerID values missing, we replaced them with 0.
2. We also replaced the missing Description values with 'Unknown'.
3. The dataset contained 5268 duplicate rows that needed to be removed.
4. There was an issue with the one-to-one mapping between StockCode and Description.
5. There were negative values in Quantity and UnitPrice.

# RFM Analysis

For RFM Analysis, we dropped the negative values of quantity and unit price. We have also dropped the rows with empty Customer ID. The columns that we need for RFM Analysis are Customer ID, Unit Price, Quantity, and Invoice Date. We get the recency value by subtracting the last available date with the date of the row. The frequency is obtained by getting the count of invoice per customer id and monetary values is obtained by getting the sum of revenue per customer ID.

Once we calculate these values, we create a separate table as follows:



We then give an RFM Score to each customer as follows:

A screenshot of a table

Description automatically generated

Segmentation using the above scores:

We have defined the customer segments as follows:

'Best Customers': ['RFM\_Segment', '111'],

'Loyal Customers': ['RFM\_Segment', '211'],

'Almost Lost': ['RFM\_Segment', '134'],

'Passerby Customers': ['RFM\_Segment', '444']}

A pie chart with numbers and text

Description automatically generated

Fig 1: Pie chart showing customer segments using the RFM Scores

# Customer Segmentation

RFM segmentation readily shows customer for any business-like Best Customers, Loyal Customer, Customers on the verge of losing, Highest revenue-generating customers etc.

For doing the customer segmentation, we will use K-means clustering.

K-means Clustering gives the best result under the following conditions:

- Data’s distribution is not skewed

- Data is standardized (i.e. mean of 0 and standard deviation of 1).

First, we checked the distribution of the recency, frequency, and monetary values.

A graph with numbers and lines

Description automatically generated

Fig 2: Showing the distribution of Recency, Frequency and Monetary values

From the above plots, we found out that the data is skewed. We treated the data to remove the skewness.

A graph of a normal distribution

Description automatically generated

Fig 3: Showing the distribution of Recency, Frequency and Monetary values after treating data

We then worked on standardizing the data. After Scaling, the data looked like following:

A white background with black and white clouds

Description automatically generated

We used the Elbow method to find the optimal number of clusters for clustering.

A graph with a line

Description automatically generated

Fig 4: Showing the Line plot formed from Elbow method

From the above, we decided to test with 3, 4 and 5 clusters. We got the following graphs after clustering the customers into 3, 4 and 5 cluster. From the figure below, we can see that cluster 4 gives the best results. So we decided to use that as the segmentation values.

A screenshot of a graph

Description automatically generated

Fig 5: Showing the Clusters formed when n=3 and n=4

A screenshot of a graph

Description automatically generated

Fig 6: Showing the Clusters formed when n=5

From the above, we see that we get the best segmentation with n-4. Therefore, we get the following customer clusters:

A white background with black dots

Description automatically generated

# Customer Profiles and the Marketing Recommendations for them.

#### Cluster 0 - Loyal Customers

Customers in this cluster exhibit a remarkable level of loyalty and substantial spending. They have a relatively recent purchase history, suggesting ongoing engagement with the business. This group can be identified as "Loyal Customers," and fostering their loyalty through personalized offers and exclusive perks may contribute to long-term customer retention.

#### Cluster 1 - Passerby Customers

Customers in this cluster exhibit characteristics of passerby behavior. Their last purchase was relatively distant (R=4), and they have made few purchases with minimal spending (F=4, M=4). To convert them into permanent members, the business needs innovative strategies. Implementing enticing promotions, improving product visibility, and enhancing the overall customer experience can be pivotal in turning these passersby into loyal customers.

#### Cluster 2 - Best Customers

This segment comprises customers who demonstrate exemplary purchasing behavior. They are characterized by recent purchases (R=1), high purchase frequency (F=1), and substantial monetary spending (M=1). These customers can be classified as the "Best Customers" segment, representing a valuable group for the business. Tailoring marketing strategies to enhance their loyalty and satisfaction could yield significant benefits.

#### Cluster 3 - Almost Lost Customers

This cluster represents customers who, despite a history of frequent purchases and significant spending, have lapsed in recent activity (R=3). Termed as "Almost Lost" customers, they present an opportunity for re-engagement. Implementing targeted win-back strategies, such as special promotions or personalized reactivation campaigns, could be effective in revitalizing their connection with the brand.

# Find the solutions to these questions:

# 1. Data Overview

**What is the size of the dataset in terms of the number of rows and columns?**

The dataset has 541909 rows and 8 columns

**Can you provide a brief description of each column in the dataset?**

count of not null and data type of the columns:

A screenshot of a computer code

Description automatically generated

Quantity:

* The average quantity of products in a transaction is approximately 9.55.
* The quantity has a wide range, with a minimum value of -80995 and a maximum value of 80995. The negative values indicate returned or cancelled orders.
* The standard deviation is quite large, indicating a significant spread in the data. The presence of outliers is indicated by a large difference between the maximum and the 75th percentile values.

UnitPrice:

* The average unit price of the products is approximately 4.61.
* The unit price also shows a wide range, from -11062.06 to 38970, which suggests the presence of errors or noise in the data, as negative prices don't make sense.
* Similar to the Quantity column, the presence of outliers is indicated by a large difference between the maximum and the 75th percentile values.

CustomerID:

* There are 406829 non-null entries, indicating missing values in the dataset which need to be addressed.

InvoiceNo:

* There are 25900 unique invoice numbers, indicating 25900 separate transactions.
* The most frequent invoice number is 573585, appearing 1114 times, possibly representing a large transaction or an order with multiple items.

StockCode:

* There are 4070 unique stock codes representing different products.
* The most frequent stock code is 85123A, appearing 2313 times in the dataset.

Description:

* There are 4223 unique product descriptions.
* The most frequent product description is "WHITE HANGING HEART T-LIGHT HOLDER", appearing 2369 times.
* There are some missing values in this column which need to be treated.

Country:

* The transactions come from 38 different countries, with a dominant majority of the transactions (approximately 91.4%) originating from the United Kingdom.

**What is the time-period covered by this dataset?**

Time period covered in the dataset is from 1/10/2011 10:04 to 9/9/2011 9:52

# 2. Customer Analysis

**How many unique customers are there in the dataset?**

There are 4373 unique customers in the dataset

**What is the distribution of the number of orders per customer?**

Distribution of the numbers of order per customer:

**A graph of distribution of orders

Description automatically generated**

Fig 7: Showing distribution of the numbers of order per customer

**Can you identify the top 5 customers who have made the most purchases by order count?**

Top 5 customers who have made the most purchase by order count:

**A graph of a number of customers

Description automatically generated**

Fig 8: Showing Top 5 customers who have made the most purchase by order count

# 3. Product Analysis

**What are the top 10 most frequently purchased products?**

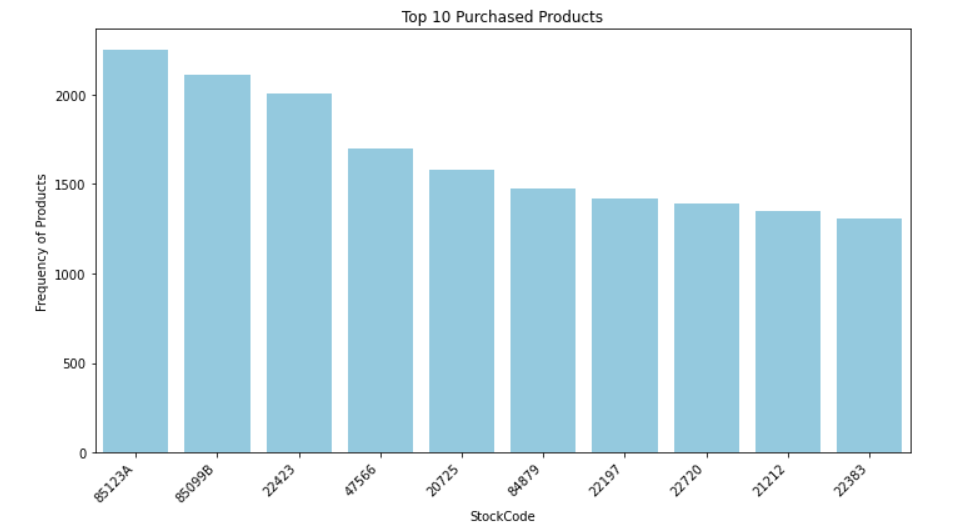
****

Fig 9: Showing Top 10 Frequently purchased products

The highest-purchased product, identified by the Stock Code 85123A, exhibits a frequency of 2250. Subsequently, Stock Code 85099B, Stock Code 22423, Stock Code 47566, and Stock Code 20725 follow in descending order. While a gradual decline is observed in the plot, Stock Codes 22197, 22720, 21212, and 22383 show a comparatively less significant decrease in the frequency of purchased products.

**What is the average price of products in the dataset?**

The average price of products in the dataset is 4.632655674831347

**Can you find out which product category generates the highest revenue?**

The StockCode DOT emerges as the product category contributing the highest revenue, totaling 206,245.480. This is succeeded by StockCode 22423, StockCode 47566, and StockCode 85123A.

Potential explanations for StockCode DOT attaining the highest revenue could include:

* It commands a premium price
* It enjoys widespread popularity, attracting a large customer base

# 4. Time Analysis

**Is there a specific day of the week or time of day when most orders are placed?**

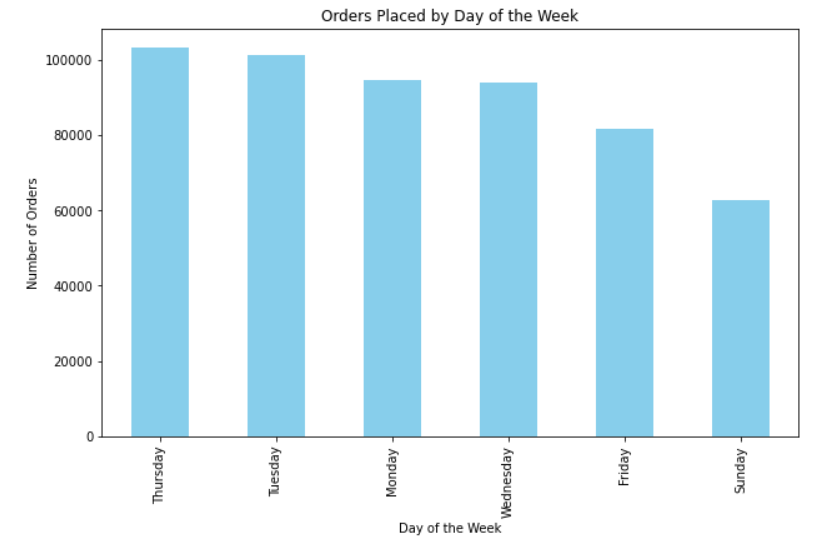
****

Fig 10: Showing Orders Placed by day of the week

The graph shows that the number of orders placed peaks on Thursdays and is at its lowest on Sundays. This is likely due to the fact that people are more likely to run errands on weekends when they are not working.

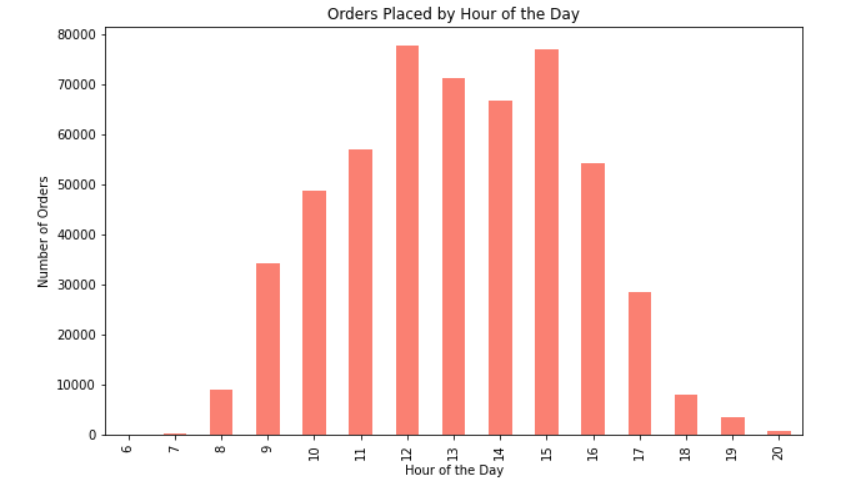
****

Fig 11: Showing Orders Placed by hour of the day

The graph shows the day timeline observed over one year, depicting the hours during which most orders are placed. The most orders are placed in the afternoon, between 12 pm and 6 pm. The number of orders placed gradually increases from 8 am to 12 pm, and then decreases from 6 pm to 10 pm. The fewest orders are placed in the early morning and late evening.

There are a few possible explanations for why most orders are placed in the afternoon. First, people are more likely to be at work or school in the morning, so they have less time to place orders. Second, people may be more likely to place orders after lunch, when they have had a chance to eat and think about what they need. Third, people may be more likely to place orders in the afternoon if they are shopping for items that they need for the next day.

The graph also shows that there is a spike in orders placed between 3 pm and 4 pm. This is likely because people are placing orders for items that they need for dinner or other evening activities. Fewest orders are placed in the early morning and late evening. This is likely because people are sleeping or otherwise engaged during these times.

**What is the average order processing time?**

Regarding this query, insufficient data was available. Ordinarily, order processing time is determined as the time elapsed between the order date and the order fulfillment date. To conduct this analysis, we introduced a new column with random dates, incorporating a buffer period of one to ten days beyond the order dates. The computed average order processing time stands at 5 days and 11 hours.

**Are there any seasonal trends in the dataset?**

The graphs show the seasonal decomposition of customer data from 2011. The top graph shows the observed customer data and the second graph shows the trend component.

The trend component shows the overall long-term trend in the data. In this case, the trend shows that the number of customers has been increasing over time.

The number of customers is increasing over time, but there is a clear seasonal pattern to the data, with more customers visiting the store during the winter months and fewer customers visiting the store during the summer months.

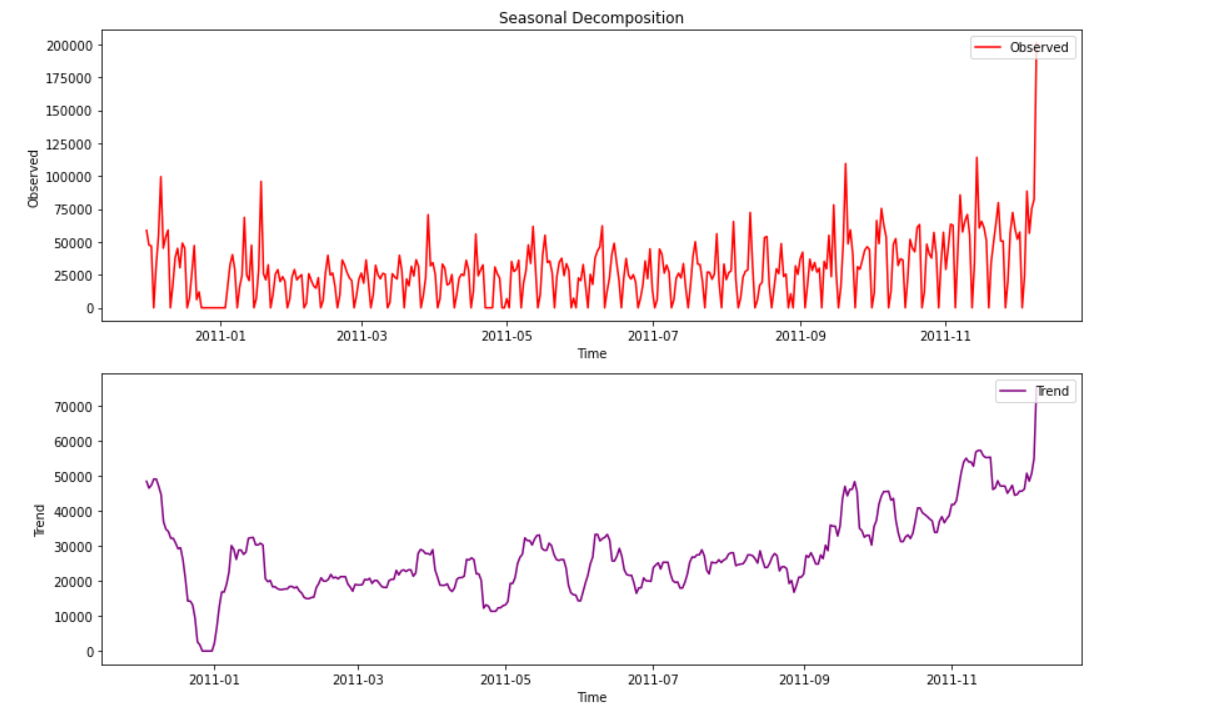
****

Fig 12: Showing Decomposition

# 5. Geographical Analysis

**Can you determine the top 5 countries with the highest number of orders?**

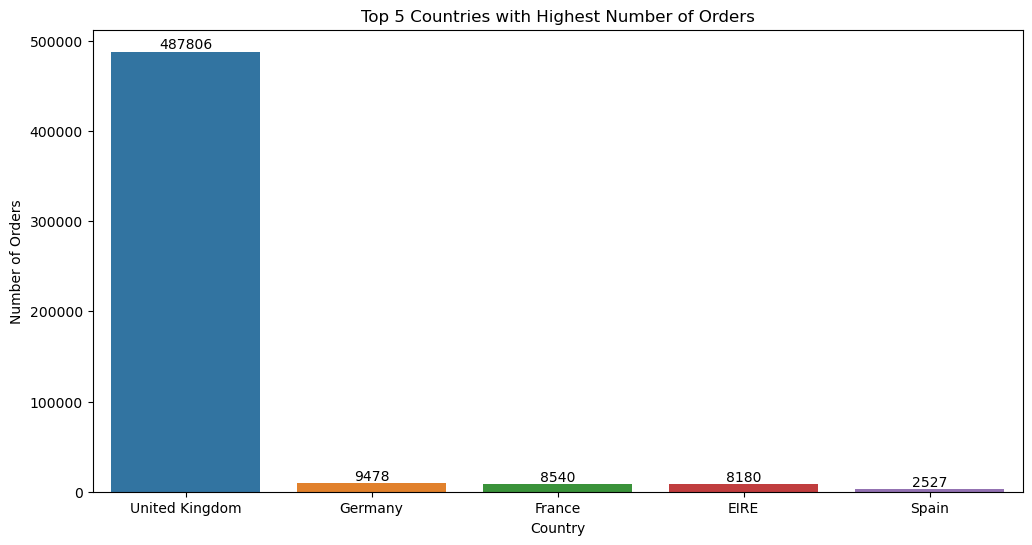
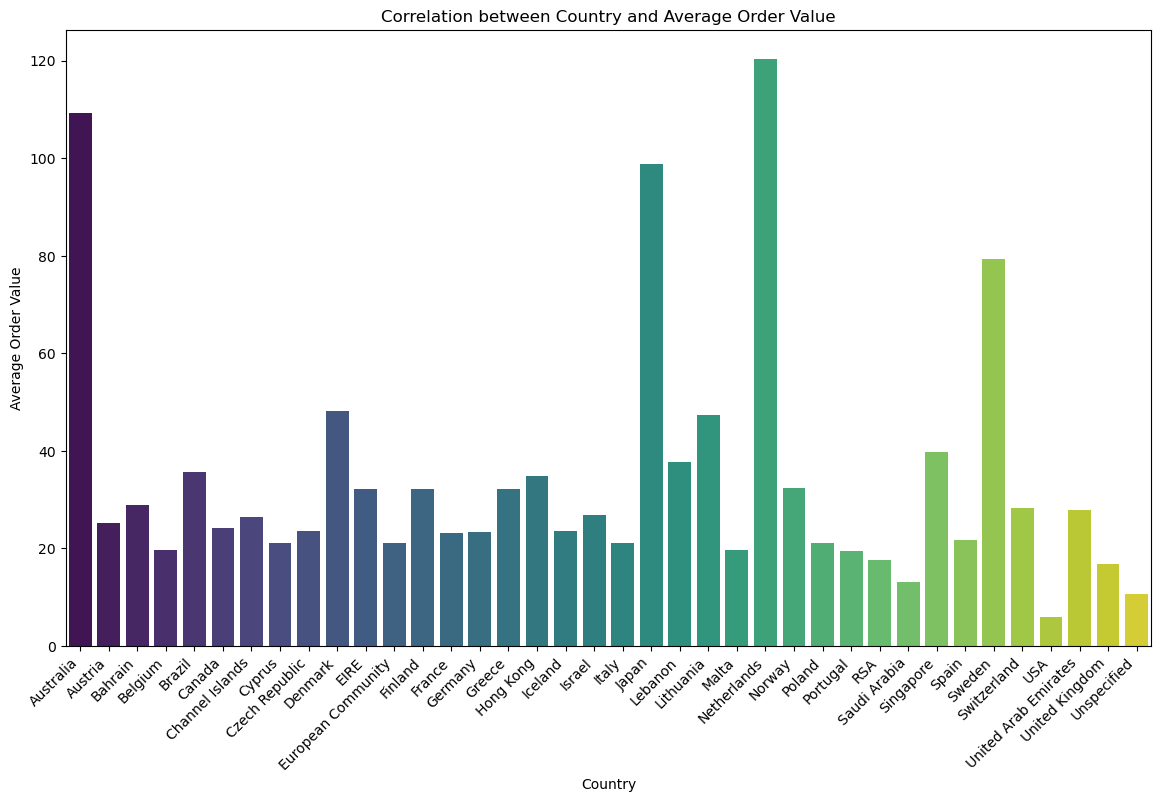


Fig 13: Showing top 5 countries with highest number of orders

Top 5 countries with the highest number of orders:

1. United Kingdom
2. Germany
3. France
4. EIRE
5. Spain

**Is there a correlation between the country of the customer and the average order value?**

Fig 14: Showing the correlation between Country and Average Order Value

Correlation between country and average order value: -0.47053032390485033

1. Negative Correlation:

The negative sign indicates a negative correlation. As the values of the 'Country' variable increases, the average order value tends to decrease, and vice versa.

1. Strength of Correlation:

The magnitude of the correlation coefficient is -0.4705. This value is moderate in strength. It suggests a discernible but not extremely strong relationship between the 'Country' variable and the average order value.

1. Interpretation:

The negative correlation suggests that there is a tendency for certain countries to have lower average order values.

In other words, customers from certain countries (Eg. USA, Saudi Arabia, United Kingdom etc.) might, on average, make smaller purchases compared to customers from other countries (Eg. Australia, Netherlands etc.).

1. Practical Implications:

This information could be valuable for marketing and sales strategies. It may prompt further investigation into the factors influencing purchasing behavior in different   
 countries. Tailoring marketing efforts or promotions based on the characteristics of countries with lower average order values could be considered.

# 6. Payment Analysis

The original dataset, as provided, did not include information about the payment methods used by customers. To address this absence, a simulated 'PaymentMode' column was introduced using predefined payment methods and associated probabilities. This column reflects a hypothetical distribution of payment methods, allowing for subsequent analyses related to payment preferences and potential associations with order amounts.

**What are the most common payment methods used by customers?**

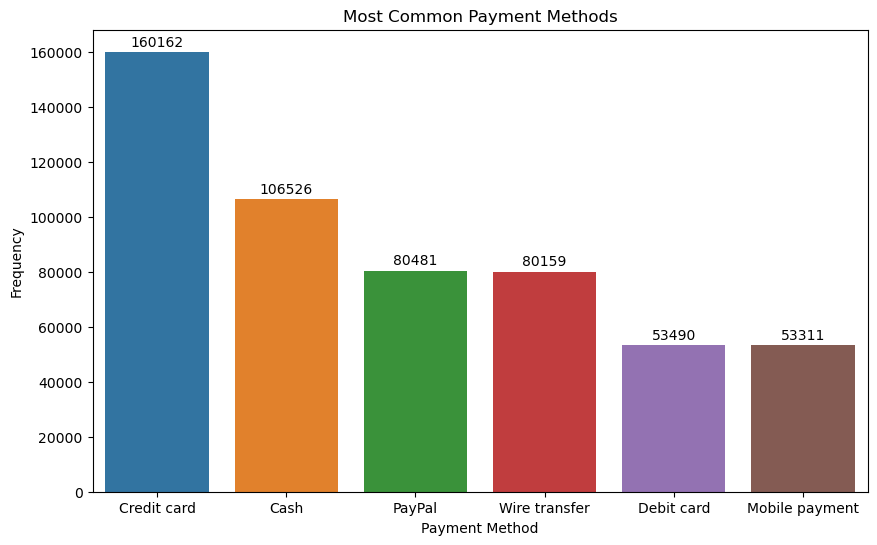


Fig 15: Showing the most common payment methods

From the above graph we can see that Credit card is the most common payment method, followed by cash, PayPal, wire transfer, debit card, and mobile payment.

**Is there a relationship between the payment method and the order amount?**

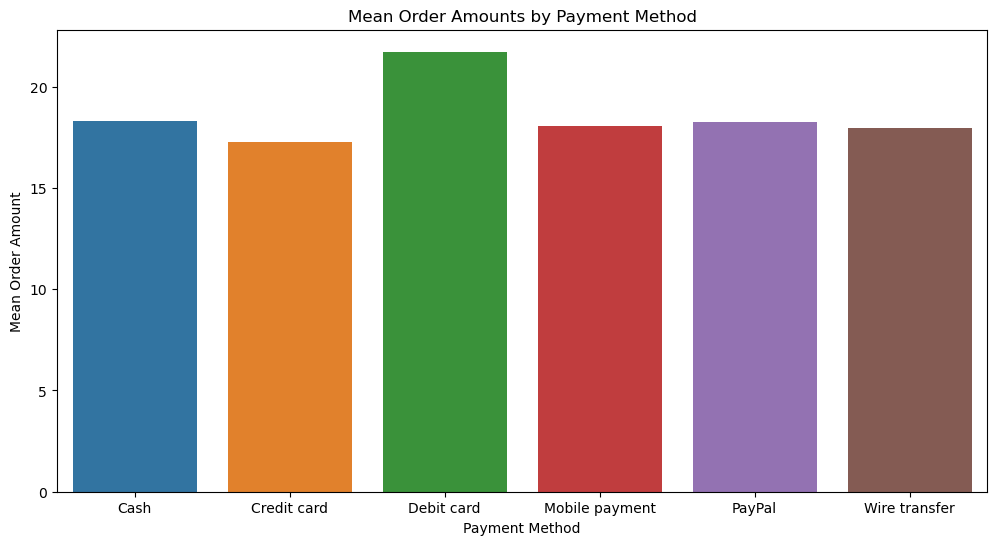
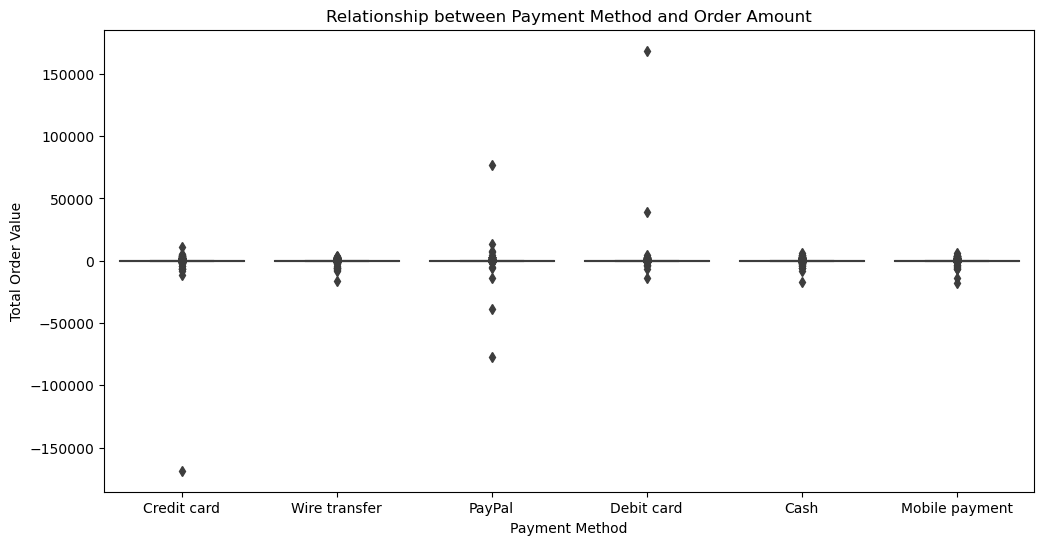


Fig 16: Showing the mean order amounts by Payment Method

Payment method and order amount relationship:

There is a weak positive correlation between payment method and order amount. This means that customers who use debit cards tend to have higher order amounts than customers who use other payment methods.

Fig 17: Showing the relationship between Payment Method and Order Amount

From the above graph, we can observe the following patterns:

* Average Order Values:

Users of debit cards tend to exhibit higher spending habits, while credit card and PayPal users show moderate spending patterns.

* Standard Deviation and Variability:

Debit card transactions showcase diverse spending behaviors, while credit card transactions display varied spending patterns. Other payment modes exhibit lower variability.

* Minimum and Maximum Order Values:

Debit card transactions show a wide range of spending, including both low and high values. Credit card transactions have a relatively narrow range of total order values.

* Negative Total Order Values:

Some transactions have negative total order values, especially in the Cash and Mobile payment categories, possibly due to refunds or adjustments.

# 7. Customer Behavior

**How long, on average, do customers remain active (between their first and last purchase)?**

A graph of a number of people

Description automatically generated

Fig 18: Showing the Distribution of Customer Activity Duration

From the above graph, we can see that the majority of customers tend to make purchases relatively frequently, with the average duration between two purchases typically falling within the 40 days. This suggests that a significant portion of customers engages with the business on a regular basis, possibly indicating consistent or periodic purchasing behavior. Understanding this pattern can be valuable for tailoring marketing strategies and promotions to maintain and enhance customer engagement, encouraging repeat purchases within a relatively short timeframe.

**Are there any customer segments based on their purchase behavior?**

#### Cluster 0 - Loyal Customers

Customers in this cluster exhibit a remarkable level of loyalty and substantial spending. They have a relatively recent purchase history, suggesting ongoing engagement with the business. This group can be identified as "Loyal Customers," and fostering their loyalty through personalized offers and exclusive perks may contribute to long-term customer retention.

#### Cluster 1 - Passerby Customers

Customers in this cluster exhibit characteristics of passerby behavior. Their last purchase was relatively distant (R=4), and they have made few purchases with minimal spending (F=4, M=4). To convert them into permanent members, the business needs innovative strategies. Implementing enticing promotions, improving product visibility, and enhancing the overall customer experience can be pivotal in turning these passersby into loyal customers.

#### Cluster 2 - Best Customers

This segment comprises customers who demonstrate exemplary purchasing behavior. They are characterized by recent purchases (R=1), high purchase frequency (F=1), and substantial monetary spending (M=1). These customers can be classified as the "Best Customers" segment, representing a valuable group for the business. Tailoring marketing strategies to enhance their loyalty and satisfaction could yield significant benefits.

#### Cluster 3 - Almost Lost Customers

This cluster represents customers who, despite a history of frequent purchases and significant spending, have lapsed in recent activity (R=3). Termed as "Almost Lost" customers, they present an opportunity for re-engagement. Implementing targeted win-back strategies, such as special promotions or personalized reactivation campaigns, could be effective in revitalizing their connection with the brand.

A graph of a distribution of customers

Description automatically generated

Fig 19: Showing Customer Segmentation based on their purchase behaviors

# 8. Returns and Refunds

**What is the percentage of orders that have experienced returns or refunds?**

From the analysis of the data, we found that 19.97% of all the orders are returned or refunded.

A pie chart with a triangle in the middle

Description automatically generated

Fig 20: Showing Percentage of Orders with Refunds or Returns

**Is there a correlation between the product category and the likelihood of returns?**

The Category of the products were never given.

For the purpose of demonstration, we have created a column with the category of the product. We have identified a few key phrases in the description to identify the category. For the cases we were unable to identify the category, we categorized them as other. From the graph below, we see that home décor items and Christmas decor items are more likely to be returned from the identifiable categories.

A graph of a number of bars

Description automatically generated with medium confidence

Fig 21: Showing Barplot of returns percentage by category

A graph with numbers and squares

Description automatically generated

Fig 22: Showing Heatmap of the return count vs the category of product.

# 9. Profitability Analysis:

**Can you calculate the total profit generated by the company during the dataset's time period?**

* Transaction details include InvoiceNo, StockCode, Quantity, InvoiceDate, UnitPrice, CustomerID and Country.
* Inorder to calculate the total profit, the cost price is needed since Total Profit = Selling Price – Cost Price but the dataset doesn’t include it.
* Therefore, the cost price column was generated randomly so as to perform the profit computation.
* Dataset spans from December 1, 2010, 08:26:00, to December 9, 2011, 12:50:00.
* Total profit during the dataset's time period is $7,624,301.55.
* Substantial profit indicates effective cost management and revenue generation.
* The line plot for the analysis over months is as follows:

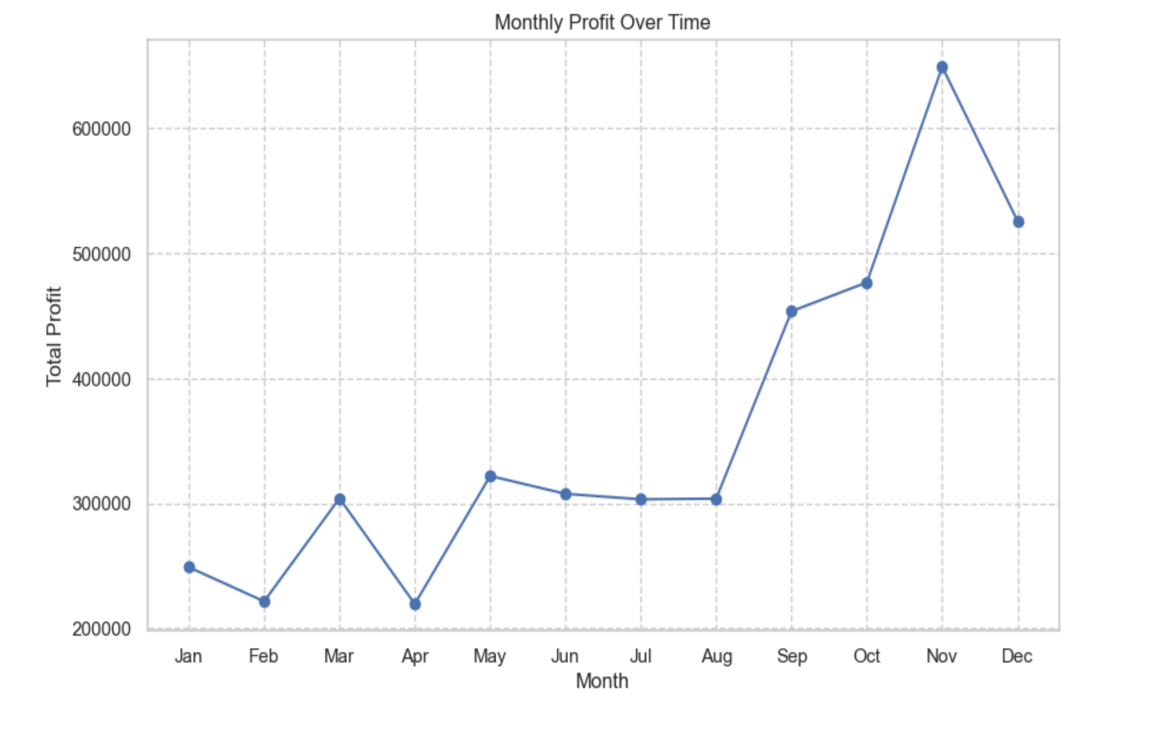


Fig 23: Showing the line graph for the monthly profit over time

**What are the top 5 products with the highest profit margins?**

* The profit margin is calculated by the formula, product\_margin = (product\_profit / product\_revenue) \* 100.
* Hence, the top 5 highest profit margins are:

1. Product: TEATIME TEA TOWELS

* + - Profit Margin: $101,676.84

2. Product: SET/5 RED SPOTTY LID GLASS BOWLS

* + - Profit Margin: $22,525.33

3. Product: MISELTOE HEART WREATH CREAM

* + - Profit Margin: $18,771.11

4. Product: MINI HIGHLIGHTER PENS

* + - Profit Margin: $16,424.72

5. Product: LUNCH BAG RED SPOTTY

* + - Profit Margin: $15,642.59
* A bar chart for the above has been plotted:

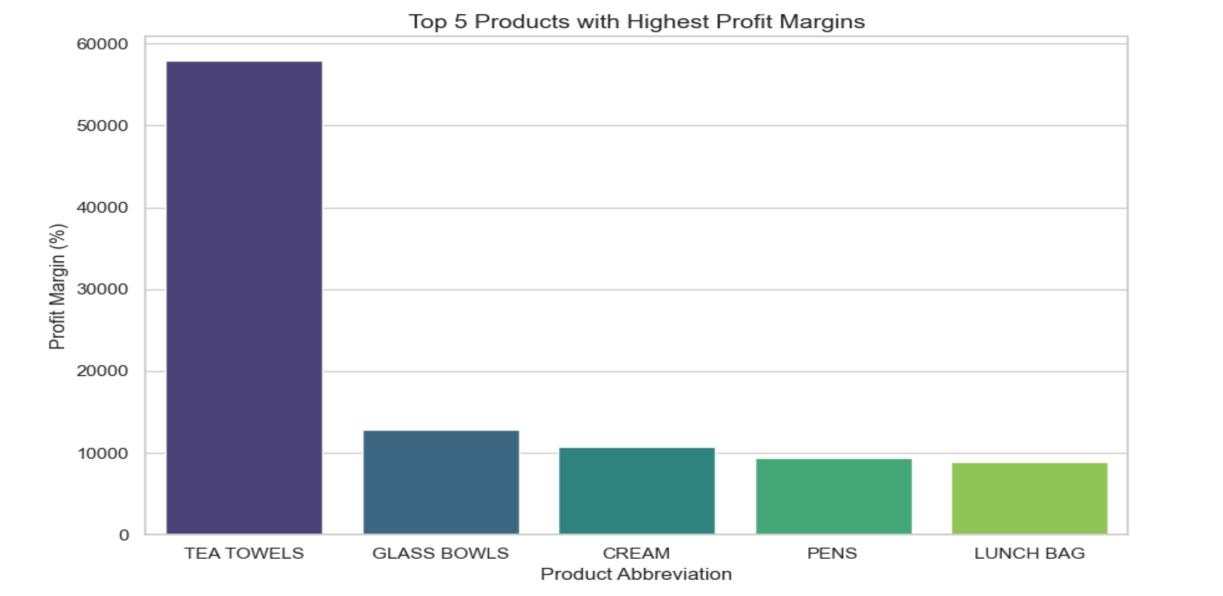


Fig 24: Showing the bar chart for the top 5 products with highest profit margins

# 10. Customer Satisfaction

**Is there any data available on customer feedback or ratings for products or services?**

* The dataset does not initially include a customer feedback or ratings column for products or services.
* To facilitate sentiment analysis, a synthetic 'Customer Feedback' column was generated, containing random feedback values for each of the 534,129 entries in the dataset.
* This additional column will aid in the analysis of customer sentiments.

**Can you analyze the sentiment or feedback trends, if available**

* The sentiment analysis involved calculating the polarity of each customer comment, indicating the positivity or negativity of the feedback.
* Subsequently, a 'SentimentCategory' column was created to categorize sentiments as Positive, Negative, or Neutral based on the calculated polarity values.
* The resulting Data Frame displays the original customer feedback, the sentiment polarity score, and the categorized sentiment for each entry.
* This was done using TextBlob, a natural language processing library in Python.
* A few observations made from the result are:
  + **Positive Shipping Experience:** The feedback "Fast shipping." has a positive sentiment score of 0.20, indicating satisfaction with the shipping speed.
  + **Consistent Praise for Customer Service:** Feedback like "Excellent customer service." consistently receives a high positive sentiment score of 1.00, reflecting ongoing positive customer service experiences.
  + **Negative Quality Perception:** The comment "Poor quality." has a negative sentiment score of -0.40, indicating dissatisfaction with the product's quality.
  + **Concerns about Value for Price:** The feedback "Not worth the price." carries a negative sentiment score of -0.15, suggesting a perception that the product does not justify its cost.
* Based on the analysis, we could recommend that:
* **Positive Sentiments:** Recognize and leverage positive sentiments, especially those praising fast shipping and excellent customer service. Consider highlighting positive feedback in marketing materials or on the company website to build a positive brand image.
* **Negative Sentiments:** Address concerns related to poor quality and perceived value for the price. Investigate specific instances of negative feedback to identify potential areas for improvement in product quality or pricing.
* **Customer Engagement:** Engage with customers who provide feedback to show responsiveness and a commitment to improvement. Encourage customers to share more detailed feedback to gain deeper insights into specific aspects of their experiences.
* The pie chart was plotted which shows the distribution of the sentiment category based on the customer feedback and it shows that the maximum share of customers have a positive impression on the products or services.

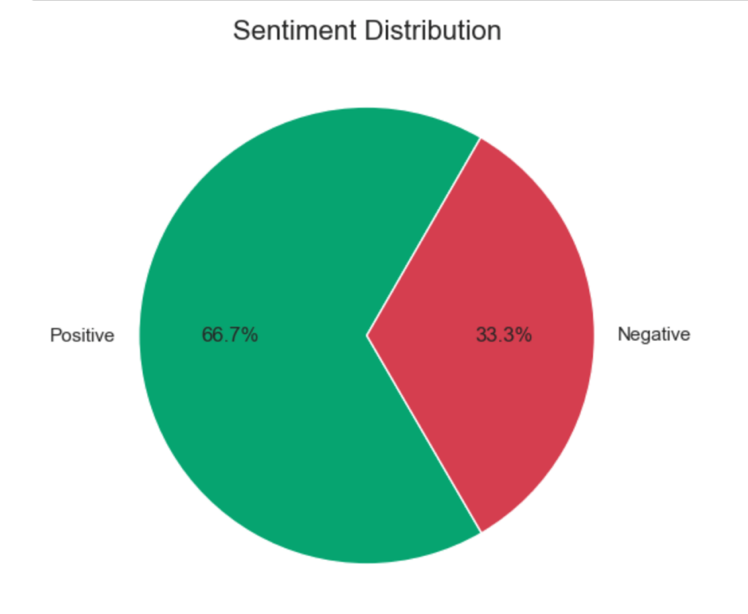


Fig 25: Showing a pie chart which indicates the sentiment distribution.