CUSTOMER SEGMENTATION USING RFM ANALYSIS

PROJECT REPORT

of

IE 6400: FOUNDATIONS FOR DATA ANALYTICS ENGINEERING

BY

GROUP NUMBER 15:

ARCHIT SINGH (002813253)

ANIRUDH HEGDE (002813268)

RAHUL ODEDRA (002835990)

SHUBHI SINHA (002201029)

SANCIA SEROPHENE SALDANHA (002851577)

DEPARTMENT OF COLLEGE OF ENGINEERING
NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS – 022115

NOVEMBER 2023

ABSTRACT

In this research project, we delve into the realm of eCommerce analytics, employing the powerful RFM (Recency, Frequency, Monetary) analysis technique to uncover valuable insights into customer behavior. The dataset under scrutiny, sourced from Kaggle, serves as the foundation for constructing a robust customer segmentation model. The overarching goal is to enhance marketing strategies and customer retention by grouping customers based on their recent purchasing patterns, buying frequency, and monetary contributions.

The project encompasses an array of tasks beginning with data preprocessing, involving importation, cleaning, and handling missing values to ensure data integrity. Subsequently, the calculation of RFM metrics, including recency, frequency, and monetary aspects, is executed to provide a comprehensive understanding of individual customer profiles. These metrics lay the groundwork for RFM segmentation, where customers are assigned scores based on quartiles or custom-defined bins. Utilizing clustering techniques, particularly K-Means clustering, enables the formation of distinct customer segments, each characterized by unique RFM scores.

The research extends further into segment profiling, as each customer group is scrutinized to uncover distinct characteristics. Marketing recommendations are then formulated, tailoring strategies for each segment to improve customer retention and maximize revenue. The project is fortified with visualizations, ranging from bar charts to scatter plots, offering a visual representation of the RFM distribution and the resultant customer clusters. A comprehensive report wraps up the research, summarizing findings, methodologies, and recommendations, coupled with clear code documentation for transparency and replicability.

In conclusion, this research project not only explores the intricacies of eCommerce customer data but also demonstrates the practicality and effectiveness of RFM analysis in guiding targeted marketing efforts, ultimately fostering business growth and customer satisfaction.

ACKNOWLEDGEMTS

We wish to express our deep gratitude to those individuals who played essential roles in the successful completion of this data analysis report. Our collaborative efforts, commitment, and teamwork were the driving forces behind this project. We are thankful for the guidance and support provided by the following people:

Professor Sivarit (Tony) Sultornsanee
Associate Teaching Professor of Mechanical and Industrial Engineering

Professor Sivarit Sultornsanee's expertise and mentorship were instrumental in shaping the direction of our analysis. We are appreciative of the valuable insights and guidance he provided during the project.

Teacher Assistant - Venkat Navneeth Burla

Venkat Navneeth Burla, our dedicated teacher assistant, played a significant role in facilitating our progress. His timely assistance and responsiveness to our inquiries were greatly beneficial.

Team Members
Archit Singh, Anirudh Hegde , Rahul Odedra, Shubhi Sinha, Sancia Saldanha

Our exceptional team members deserve our profound thanks for their unwavering commitment and collaboration. Together, we addressed various aspects of this project, including data sourcing, data cleaning, data analysis, and reporting. The project's quality and success would not have been achievable without their hard work and dedication.

This report stands as evidence of the exceptional teamwork and camaraderie that characterized our project. We take pride in working with such talented and cohesive team members. Our heartfelt appreciation goes out to everyone involved in this endeavour for their invaluable contributions.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.0 INTRODUCTION

In the fast-paced landscape of eCommerce, understanding and engaging with customers on a personalized level is pivotal for business success. The era of one-size-fits-all strategies has given way to more nuanced approaches, and RFM (Recency, Frequency, Monetary) analysis stands out as a potent tool for unravelling the intricacies of customer behaviour. This research project delves into the depths of customer segmentation, using RFM analysis to categorize customers based on their recent purchasing activities, buying frequency, and monetary contributions. By doing so, businesses can gain invaluable insights into their customer base, enabling them to tailor marketing efforts with precision and enhance overall customer satisfaction.

The dataset under examination, curated from Kaggle's diverse eCommerce data, forms the foundation of this research endeavour. The project embarks on a meticulous journey of data pre-processing, ensuring the integrity and reliability of the dataset. This involves tasks such as importing data, cleaning, and addressing missing values. The subsequent calculation of RFM metrics, including Recency, Frequency, and Monetary aspects, provides a comprehensive snapshot of each customer's engagement with the platform. These metrics set the stage for RFM segmentation, a process wherein customers are categorized into distinct groups based on quartiles or custom-defined bins.

Beyond the technical intricacies, this research project incorporates a practical dimension by delving into the marketing implications of RFM analysis. Clustering techniques, notably K-Means clustering, are deployed to form customer segments, each with its unique RFM scores. These segments are then profiled, allowing businesses to tailor marketing strategies that resonate with the distinct characteristics of each group. The visualizations, ranging from bar charts to scatter plots, not only provide a visual representation of the RFM distribution but also serve as a powerful communication tool for stakeholders, conveying complex data insights in a digestible format. The aim is to empower businesses with actionable recommendations for improving customer retention and maximizing revenue, fostering a symbiotic relationship between customers and commerce.

CHAPTER 2: DATA SOURCING AND CLEANING

2.1 Source of Data:

The dataset, primarily composed of actual e-commerce transaction data from 2010 and 2011, is a rare find in the public domain due to the proprietary nature of such data. It is hosted by the UCI Machine Learning Repository and can be located under the title "Online Retail."

2.2 Content Overview:

This international dataset details transactions from December 1, 2010, to December 9, 2011, for a UK-based, online-only retail store specializing in unique gifts for various occasions, with a customer base that includes a significant number of wholesalers.

Data Preparation Steps for the E-commerce Dataset

Preliminary Assessment:

1. Initial examination to identify missing, incomplete, or duplicate data entries.

2. Analysis of data types and structures for consistency.

Data Cleaning and Transformation:

1. Removal of irrelevant columns such as internal reference codes or extraneous identifiers.

2. Identification and removal of duplicate records to ensure data integrity.

3. Standardization of customer ID formats for consistency.

Addressing Missing and Inaccurate Data:

1. Handling missing values in key fields like 'Quantity' and 'Price'.

2. Correction of anomalies and standardization of transaction dates to a uniform datetime format.

3. Validation of transaction records against known inventory levels.

Enhancing Dataset Usability:

1. Categorization and labelling of product descriptions for clarity and analysis.

2. Creation of additional attributes, such as customer segmentation based on purchase history.

Error Correction and Data Integrity:

1. Removal of transactions with negative quantities, likely representing returns or errors.

2. Ensuring all price values are positive and logically consistent.

Post-Cleaning Review and Restructuring:

1. Comprehensive review to ensure the effectiveness of cleaning steps.

2. Statistical analysis to confirm data integrity.

3. Restructuring the dataset to facilitate analysis, such as customer behaviour modelling or sales forecasting.

Final Steps and Quality Assurance:

1. Running test analyses to confirm dataset readiness for advanced analytical tasks.

3. Documentation of the cleaning process, and assumptions, made during the data preparation process.

Final Dataset Structuring:

1. Consolidation of transaction data to reflect total sales per product and customer.

2. Implementation of a schema that aligns with the objectives of the intended analysis, such as customer behaviour modelling or sales forecasting.

Validation and Quality Assurance:

1. Running test queries and analyses to confirm the dataset's readiness for in-depth analytical tasks.

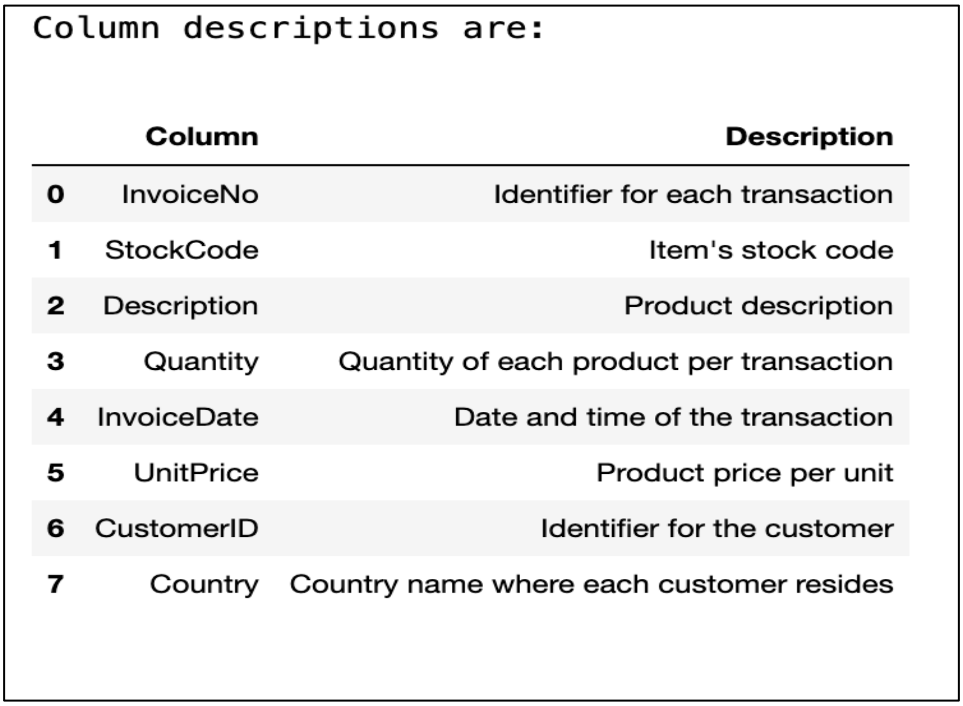
CHAPTER 3: DATA VISUALIZATION AND ANALYSIS

3.1 Data Overview:

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

The dataset comprises a substantial total of 525,460 records, each characterized by 8 distinct columns. This extensive collection of data encompasses transactions spanning an entire year, beginning from December 1, 2010, at 08:26 AM, and concluding on December 9, 2011, at 12:50 PM. This extensive timeframe provides a comprehensive view of the transactional activities over a year, allowing for in-depth analysis of trends, customer behaviours, and product performance across different periods.

3.1.2 Brief descriptions of the columns :



Column descriptions are:

	Column	Description
0	InvoiceNo	Identifier for each transaction
1	StockCode	Item's stock code
2	Description	Product description
3	Quantity	Quantity of each product per transaction
4	InvoiceDate	Date and time of the transaction
5	UnitPrice	Product price per unit
6	CustomerID	Identifier for the customer
7	Country	Country name where each customer resides

Figure 3.1.2.1 Column Descriptions

The data set contains 8 columns with descriptions given as below:

The screenshot provided contains a table that describes the columns of a dataset. Each column in the dataset serves a specific purpose in representing transactional data:

1. InvoiceNo: This column acts as a unique identifier for each transaction. It can be used to track individual sales or orders.

2. StockCode: This represents the item's stock code, which is a unique identifier for each product. This code is used to manage inventory and identify products across transactions.

3. Description: This contains the product description. It gives details about the item sold, which can include the name, size, colour, or other attributes.

4. Quantity: This column records the quantity of each product per transaction. It shows how many units of a product were bought in each sale.

5. InvoiceDate: This field captures the date and time of the transaction. It is crucial for analysing sales trends over time and understanding customer purchasing patterns.

6. UnitPrice: This represents the product price per unit. It is used to calculate the total sales amount and can be used for pricing analysis.

7. CustomerID: This is an identifier for the customer who made the transaction. It can be used to track customer purchase history and for customer relationship management.

8. Country: This column indicates the country name where each customer resides. It is important for market segmentation and geographical sales analysis.

These columns together provide a comprehensive view of the sales transactions, offering insights into what products are being sold, in what quantity, at what price, when, and to whom. Such information is vital for business analysis, inventory management, and strategic planning.

3.1.3 Time Period provided by the data set :

The dataset encompasses a period starting from the 1st of December, 2010, and extends to the 9th of December, 2011. This time range represents a full year and a few additional days, allowing for a thorough analysis of the dataset's annual transactional trends, seasonal variations, and customer behaviours within the given timeframe. The inclusion of both the start and end dates in the dataset provides an opportunity to study the patterns from the beginning of December, which is typically a significant time for retail due to the holiday season, through the same period in the following year, capturing the entirety of the holiday sales cycle.

3.2 Customer Analysis :

3.2.1 Number of unique customers in the dataset:

The dataset contains transactions from a diverse customer base, as evidenced by the unique customer count. A total of 4,339 individual customers are represented, indicating the breadth of the dataset's customer-related information. This figure highlights the dataset's potential for customer segmentation and behaviour analysis, allowing for a detailed understanding of the purchasing patterns across a wide range of customers. Such a considerable number of unique customers also suggests the possibility of rich insights into market penetration and customer loyalty.

3.2.2 Distribution of a number of orders per customer:

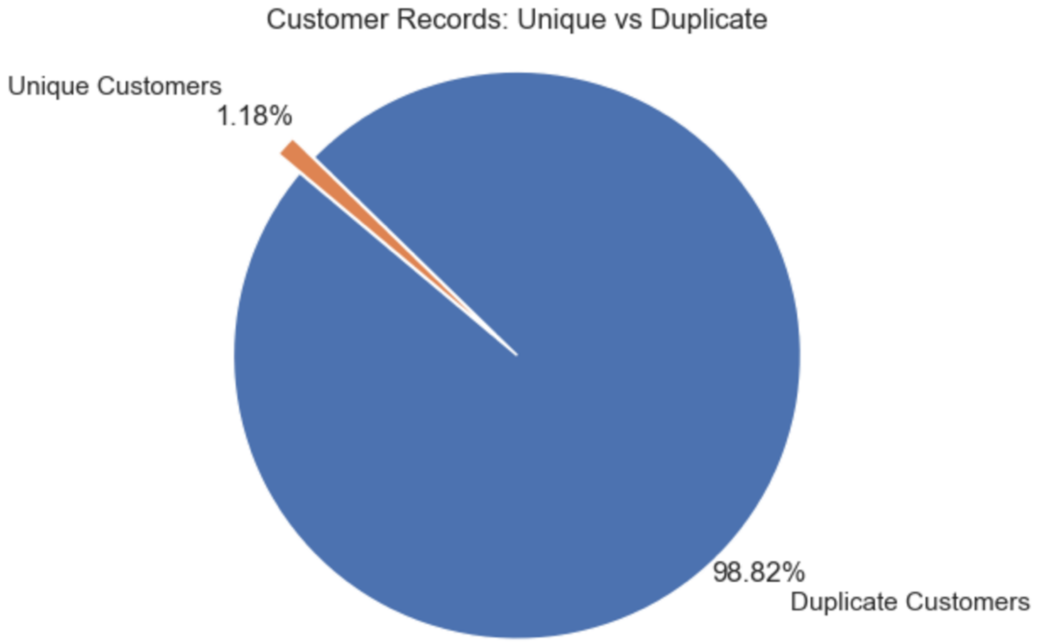


Figure 3.2.2.1 Unique Customers

The pie chart visualizes the proportion of unique and duplicate customer records within the dataset. It indicates that a vast majority, precisely 98.82%, of the customer records are duplicates, which suggests repeated transactions or entries for these customers. Conversely, only a small fraction, 1.18%, represents unique customer records, implying these customers have been recorded just once in the dataset.

Given this context, the dataset seems to have a high level of repeated entries for a majority of customers, which could be normal in a transactional dataset where customers make multiple purchases over time. This information could be pivotal for businesses focusing on customer retention strategies and understanding customer buying patterns. However, if these duplicates are not intentional (i.e., they are not separate transactions), this could suggest a need for data cleaning to remove unintended duplicates and ensure the accuracy of the data analysis.

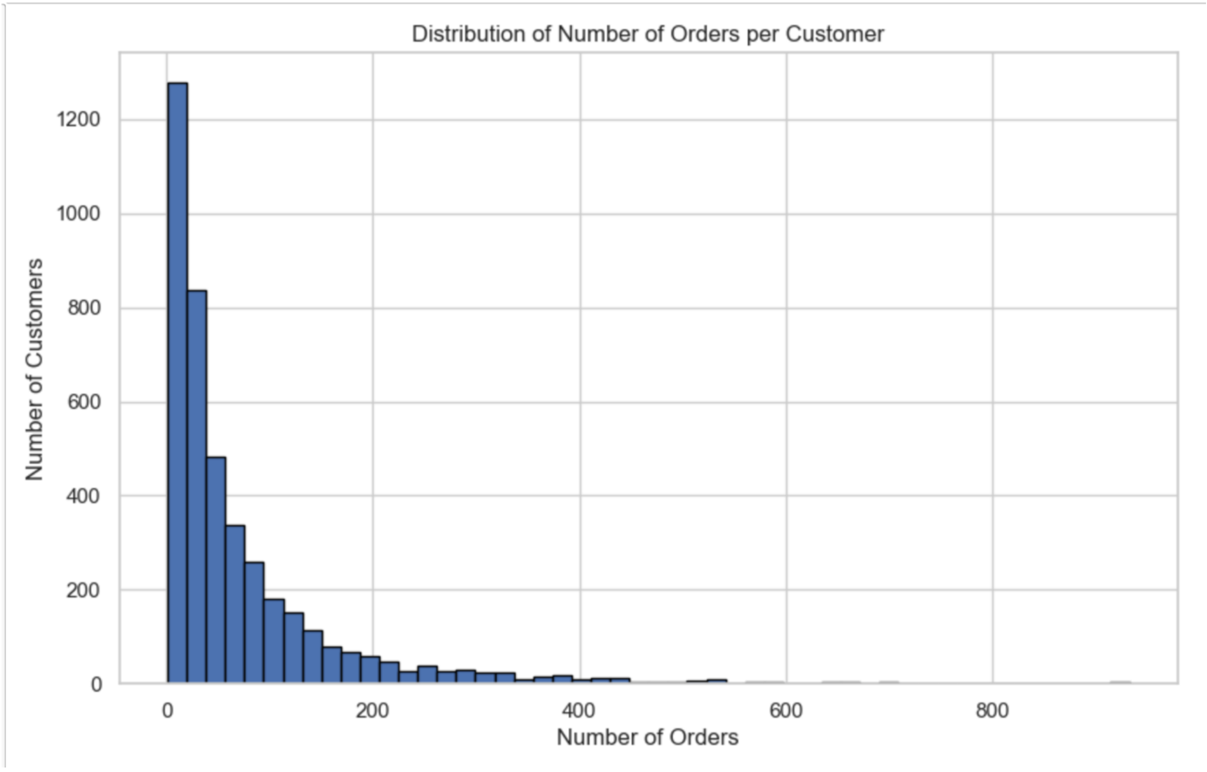


Figure 3.2.2.2 Number of orders per customer distribution

The histogram provided depicts the distribution of the number of orders per customer within the dataset. It shows that a large number of customers have placed a relatively small number of orders, as indicated by the tall bar at the leftmost side of the histogram. As we move to the right, representing a greater number of orders, the number of customers who have placed that many orders decreases significantly.

The distribution is right-skewed, meaning that there are a few customers with a very high number of orders, but these are exceptions. The majority of customers appear to have a low to moderate number of orders. This pattern is typical in consumer behaviour, where a small segment of customers often contributes to a large portion of orders, sometimes referred to as the Pareto principle or the 80/20 rule.

This graph is useful for businesses to understand their customer base and the frequency of their purchases. It might also inform strategies for customer engagement, retention, and loyalty programs, focusing efforts on either the broad base of infrequent buyers or the smaller group of highly engaged customers.

3.2.3 The top 5 customers who have made the most purchases by orders :

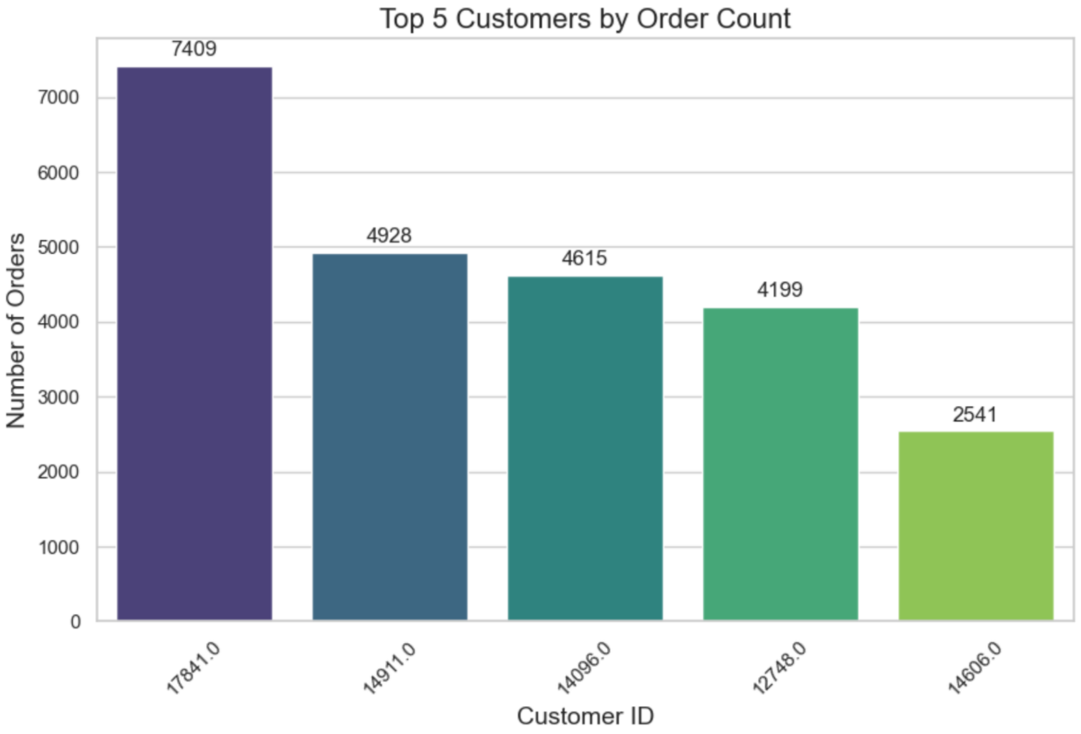


Figure 3.2.3.1 Top 5 Customers Distribution

The bar chart presents the top 5 customers from the dataset based on their order count. Each bar represents a unique customer, identified by their Customer ID, and the height of the bar indicates the total number of orders placed by that customer.

The tallest bar corresponds to the customer with ID '17841.0', who has the highest number of orders at 7,409. The next customer, '14911.0', has placed 4,928 orders, followed by '14096.0' with 4,615 orders, '12748.0' with 4,199 orders, and '14606.0' with 2,541 orders.

The graph visually emphasizes the variation in order count among the top customers, showing a significant difference between the customer with the highest number of orders and the others. This kind of analysis is crucial for identifying key customers and understanding their buying patterns, which can inform targeted marketing strategies and customer relationship management.

3.3 Product Analysis

3.3.1 RFM Analysis

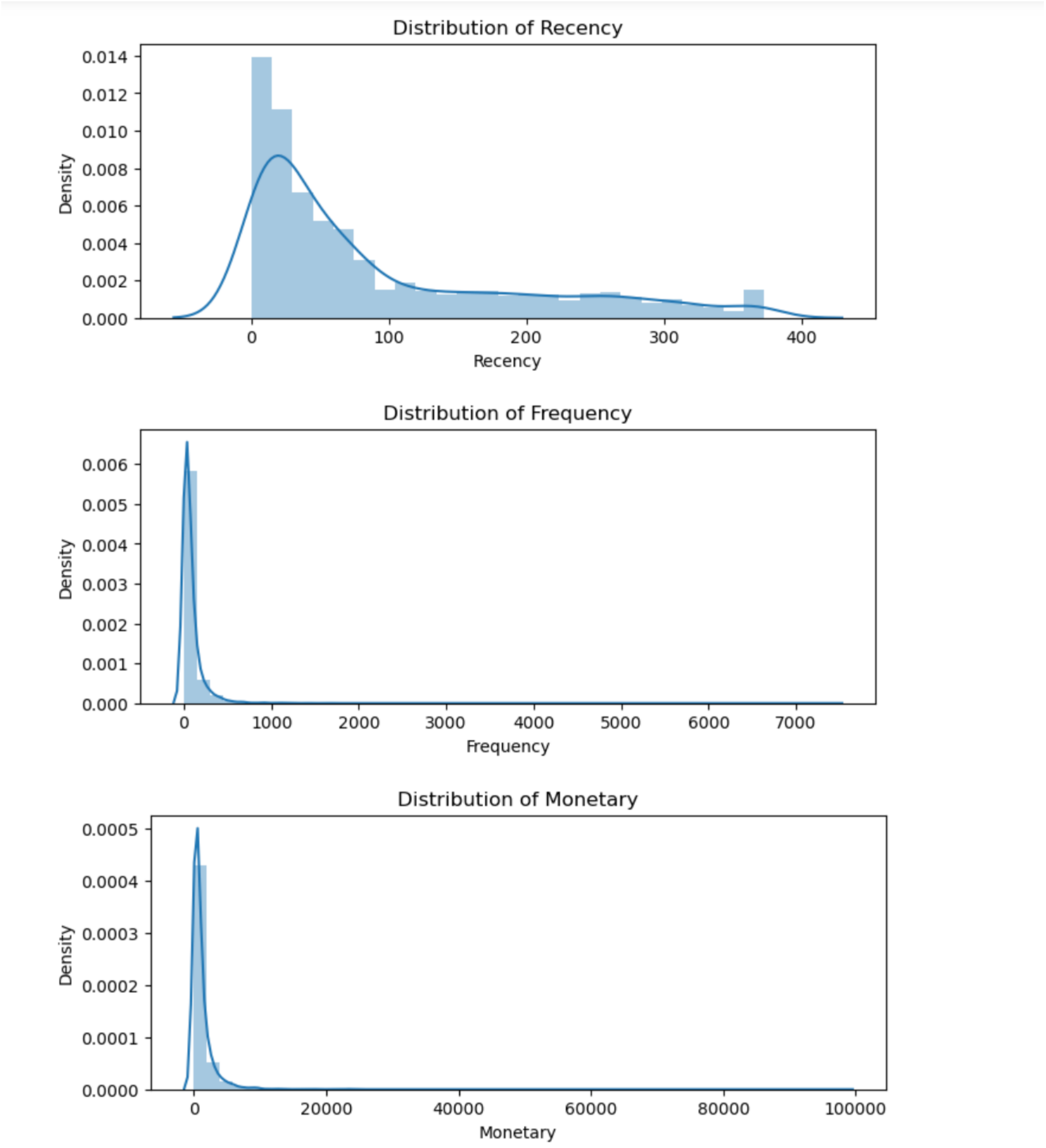


Figure 3.3.1.1 RFM analysis graphical distribution

The visualizations above represent the distribution of RFM (Recency, Frequency, Monetary) metrics for each customer:

Recency Distribution:

This histogram shows how recently customers have made a purchase. A lower recency value indicates more recent activity.

Frequency Distribution:

This histogram displays the total number of orders for each customer. Most customers have a relatively low frequency of orders, as indicated by the concentration on the left side of the histogram.

Monetary Distribution:

This histogram shows the total spend for each customer. Similar to frequency, most customers are concentrated at the lower end of the spending spectrum. In all three histograms, a logarithmic scale is used on the y-axis for 'Frequency' and 'Monetary' to better visualize the wide range of values. These metrics are crucial for understanding customer behaviour and segmenting customers based on their purchasing patterns

3.3.2 Top 10 most frequently purchased products :

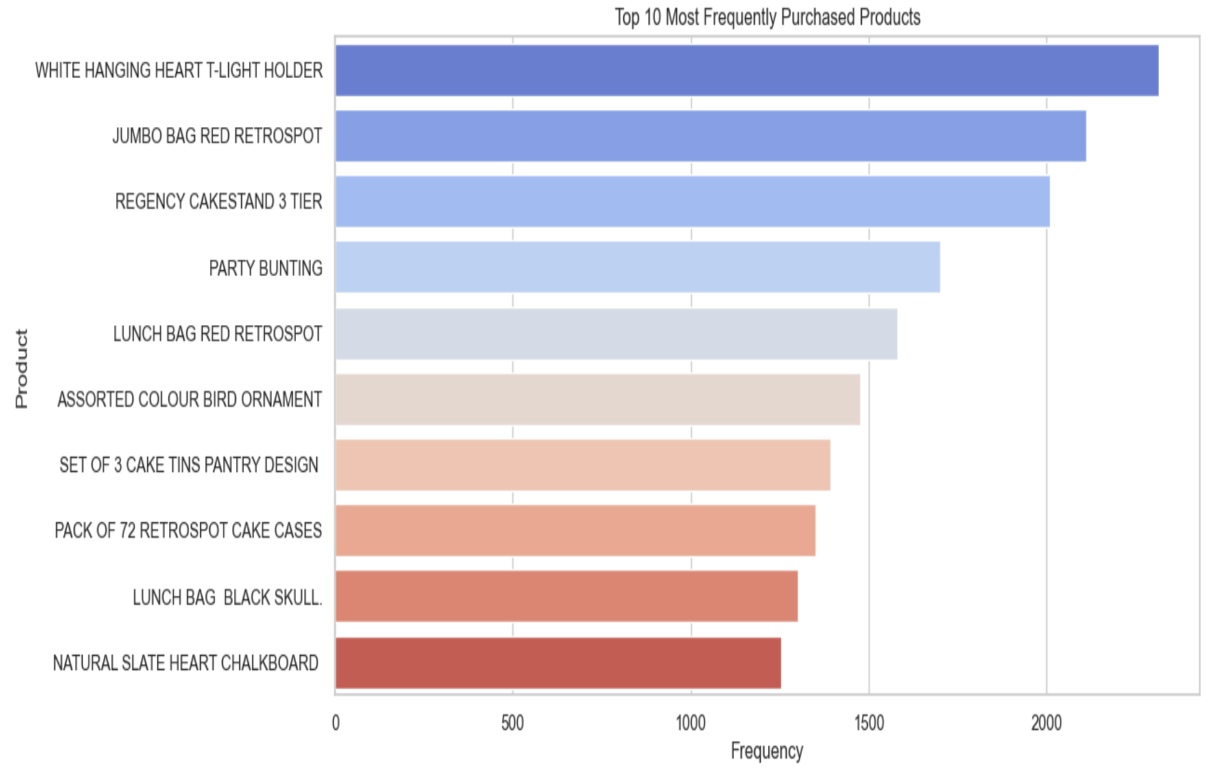
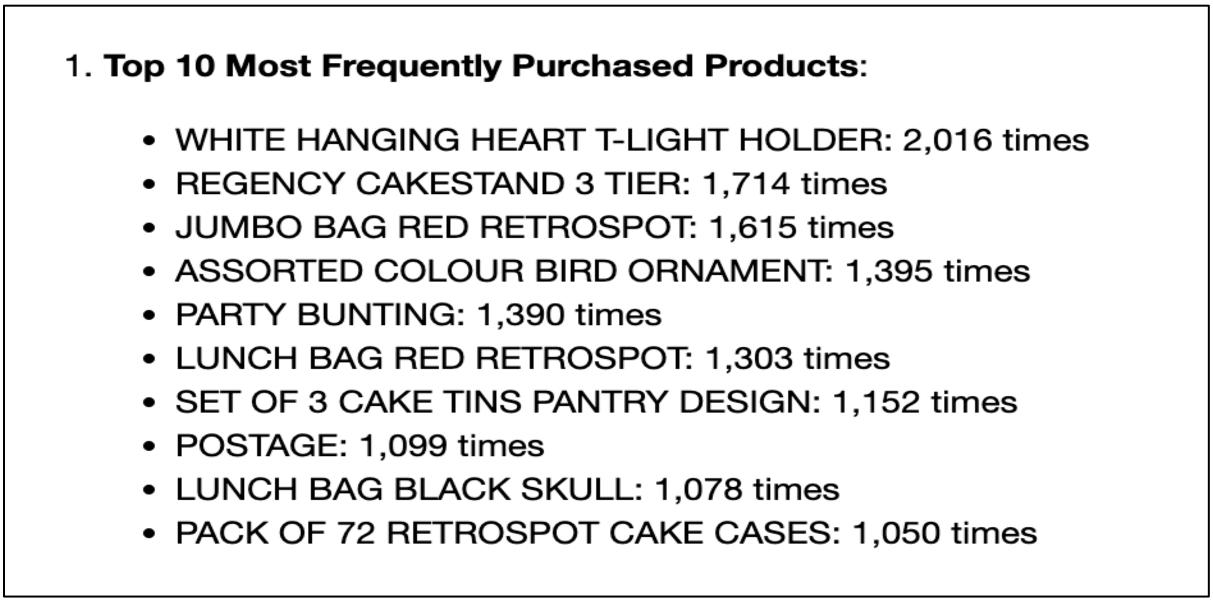


Figure 3.3.2.1 Top 10 most frequently purchased products

The horizontal bar chart illustrates the top 10 most frequently purchased products from the dataset. Each bar represents a different product, with the length of the bar corresponding to the frequency of purchase, which is indicated on the horizontal axis.

The product with the highest purchase frequency is "WHITE HANGING HEART T-LIGHT HOLDER," followed by "JUMBO BAG RED RETROSPOT," "REGENCY CAKESTAND 3 TIER," and others. The colour coding may represent different categories or simply be a visual aid to differentiate between products. The frequency values start at 0 and extend past 2000 for the most purchased product.

This chart helps identify the most popular products, which can be critical for inventory management, marketing strategies, and understanding consumer preferences. It appears that decorative items and practical goods like bags and cake stands are among the most purchased, indicating a potential trend or preference within the customer base.



1. **Top 10 Most Frequently Purchased Products**:
    - WHITE HANGING HEART T-LIGHT HOLDER: 2,016 times
    - REGENCY CAKESTAND 3 TIER: 1,714 times
    - JUMBO BAG RED RETROSPOT: 1,615 times
    - ASSORTED COLOUR BIRD ORNAMENT: 1,395 times
    - PARTY BUNTING: 1,390 times
    - LUNCH BAG RED RETROSPOT: 1,303 times
    - SET OF 3 CAKE TINS PANTRY DESIGN: 1,152 times
    - POSTAGE: 1,099 times
    - LUNCH BAG BLACK SKULL: 1,078 times
    - PACK OF 72 RETROSPOT CAKE CASES: 1,050 times

Figure 3.3.2.2 Top 10 products with quantities

3.3.3 The average price of products in the dataset :

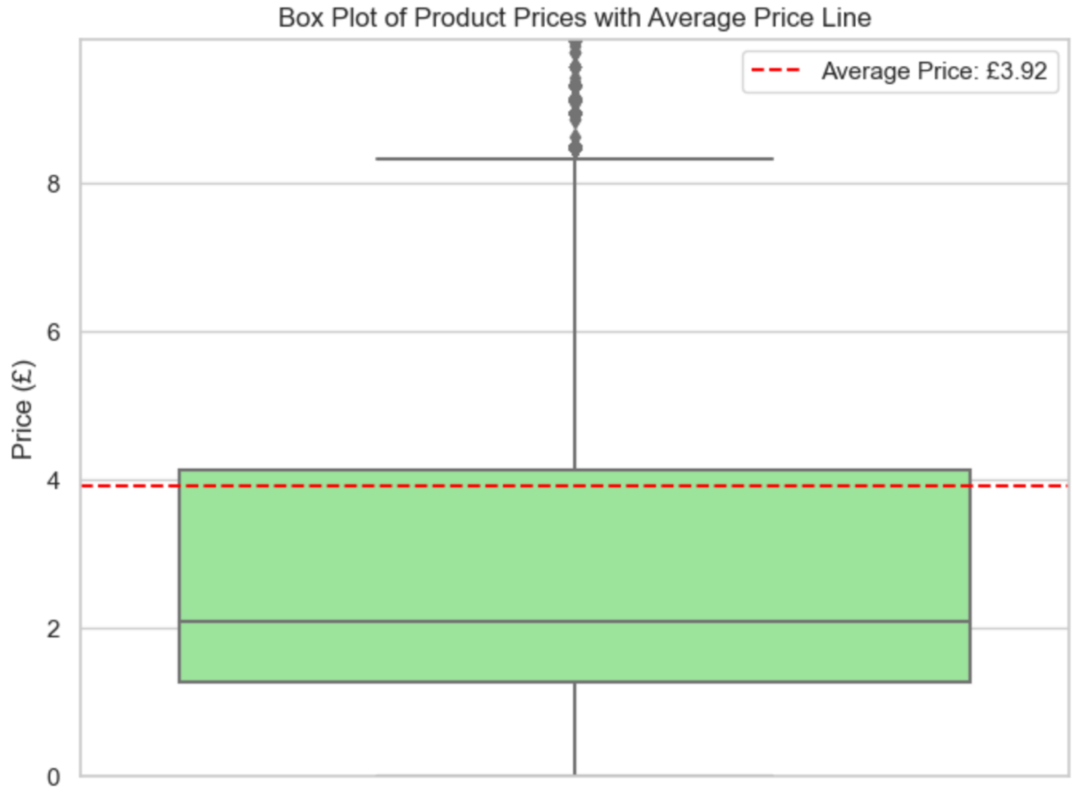


Figure 3.3.3.1 Average price box plot

The graph shown is a box plot that represents the distribution of product prices within the dataset. Here’s a breakdown of the chart:

- Central Rectangle (The Box): The main body of the box plot represents the interquartile range (IQR) of the product prices, with the bottom and top edges of the box indicating the first quartile (Q1) and third quartile (Q3), respectively. The IQR contains the middle 50% of the data.

- Horizontal Line in the Box (The Median): The line within the box marks the median price, which is the middle value when the data is ordered from lowest to highest.

- Dashed Line (The Mean): The red dashed line indicates the average price of the products, which is £3.92. This line provides a point of reference to compare the median and shows if the data is skewed.

- Whiskers: The lines extending vertically from the box (the "whiskers") indicate the range of the data, excluding outliers. Typically, they extend to the smallest and largest values within 1.5 times the IQR from the first and third quartiles.

- Outliers: The individual points above the upper whisker represent outliers, which are prices that are notably higher than the rest of the data (usually considered to be more than 1.5 times the IQR above the third quartile).

This box plot provides a visual summary of the central tendency and spread of the product prices. It also highlights the presence of outliers, which could be particularly expensive products. The average price line being close to the top of the box suggests that the distribution is slightly skewed, with a bulk of the products priced below the average.

3.3.4 Which product generates the highest revenue :

The product that generates the highest revenue is:

DOTCOM POSTAGE, 206248.77

3.4 Time Analysis :

3.4.1 Specific Day of the week or time of the day when most orders are placed:

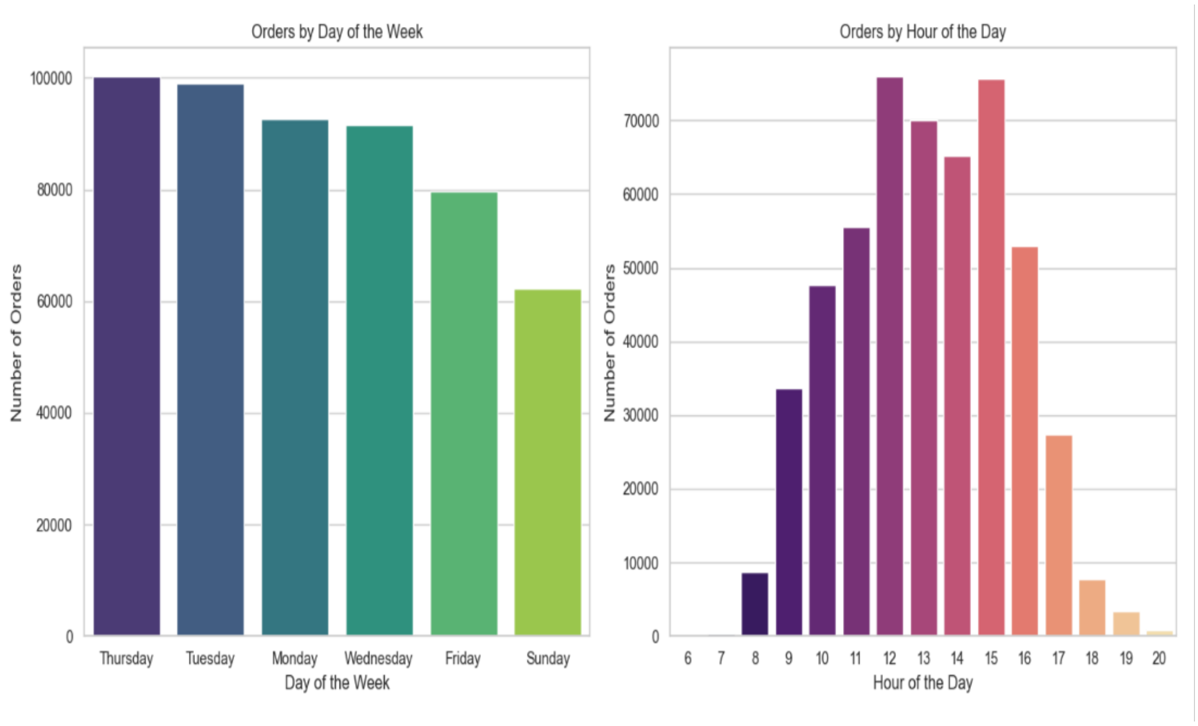


Figure 3.4.1.1 Orders by day of week and hours of the day

The provided contains two separate bar charts, each depicting different aspects of order patterns from the dataset.

1. Orders by Day of the Week: The first chart on the left shows the number of orders placed on each day of the week. The bars represent the days, from Thursday to Sunday, suggesting that the dataset might not include data for all days of the week or that these are the days with the most significant activity. The height of each bar indicates the total number of orders for that particular day. In this chart, Thursday seems to have the highest number of orders, followed closely by Monday and Tuesday, with Sunday having the fewest orders among the days displayed.

2. Orders by Hour of the Day: The second chart on the right illustrates the number of orders placed during different hours of the day. Each bar represents an hour, ranging from 6 AM to 8 PM. The chart suggests that the peak ordering time is around 12 PM, as indicated by the tallest bar, which then gradually decreases towards the evening. The lowest number of orders occur in the early hours of the morning, around 6 AM.

Together, these charts provide insights into customer purchasing behaviour, showing when customers are most and least likely to place orders. Businesses could use this information to optimize their operations, whether that means staffing, inventory management, or targeted marketing efforts to drive sales during slower periods.

3.4.2 Average order processing time :

Average Order Processing Time: 1 day 11:03:16.617549749

This indicates that the average time between consecutive orders for each customer in the dataset is approximately 1 day, 11 hours, 3 minutes, and roughly 17 seconds. This metric, often referred to as "average processing time," can be a critical indicator of customer re-engagement for a business. It suggests that on average, customers tend to make another purchase or interaction with the business just over one day after their previous order.

3.4.3 Seasonal Trends :

The graph below appears to show three separate graphs, each representing different order trends over time-based on a given dataset.

1. Weekly Order Trends: The top graph shows the number of orders per week for two different years – 2010 and 2011. It depicts fluctuations over the weeks, with some peaks and troughs indicating variations in the number of orders. In 2011, there was a notable increase in orders towards the later weeks, which sharply dropped off, possibly indicating incomplete data for the last week or a significant drop in orders.

2. Monthly Order Trends: The middle graph displays the number of orders per month, again comparing 2010 and 2011. It seems that in 2011 there was a significant peak around the 10th month (October), suggesting a higher volume of orders, potentially due to seasonal factors like holiday shopping.

3. Yearly Order Trends: The bottom graph compares the total number of orders for the entire years of 2010 and 2011. It clearly shows that the number of orders in 2011 was significantly higher than in 2010, indicating growth or an increase in sales volume.

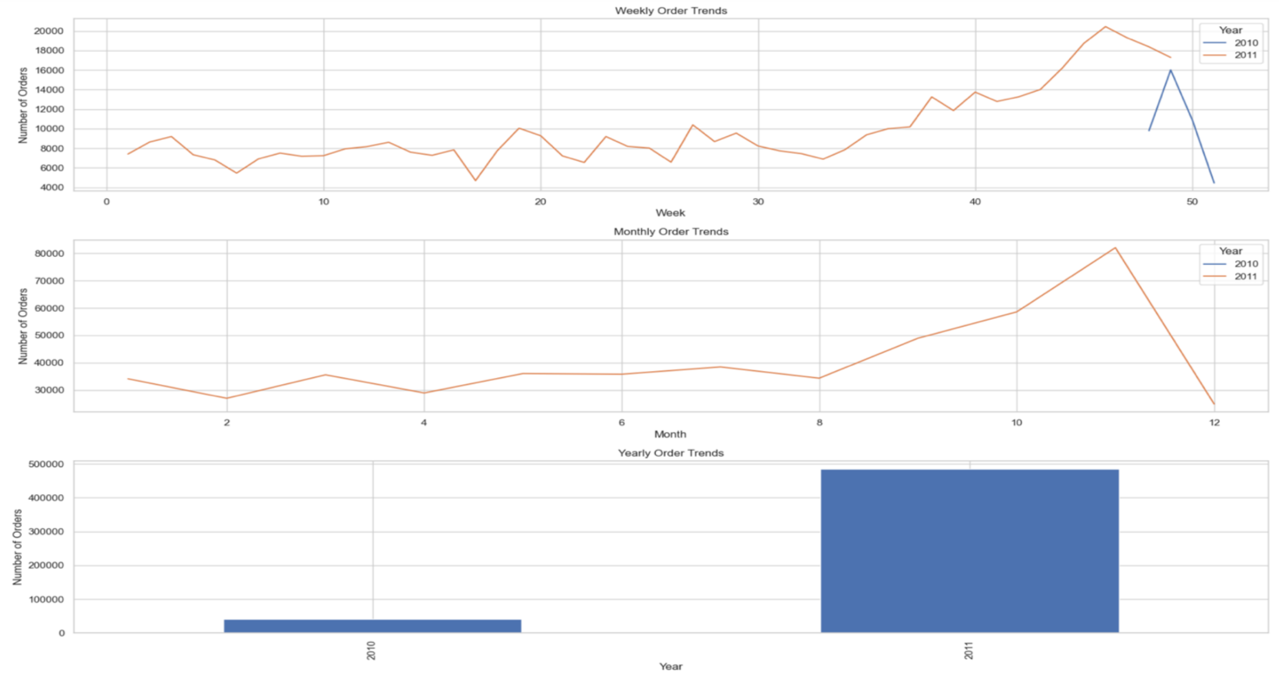


Figure 3.4.3.1 Seasonal Trends

3.5 Geographical Analysis:

3.5.1 The top 5 countries with the highest number of orders:

The United Kingdom has the highest number of orders at 320,247, showing a big customer base. Germany follows with 7,980 orders, indicating a strong presence in the market. France and EIRE also have substantial customer engagement, with 7,380 and 5,992 orders, respectively. Spain, though smaller in comparison, still contributes significantly with 2,175 orders. These numbers highlight the strong demand in the UK and consistent interest in Germany, France, EIRE, and Spain. The data is valuable for businesses to adjust their marketing and operational strategies to meet the varying demands in these key regions.

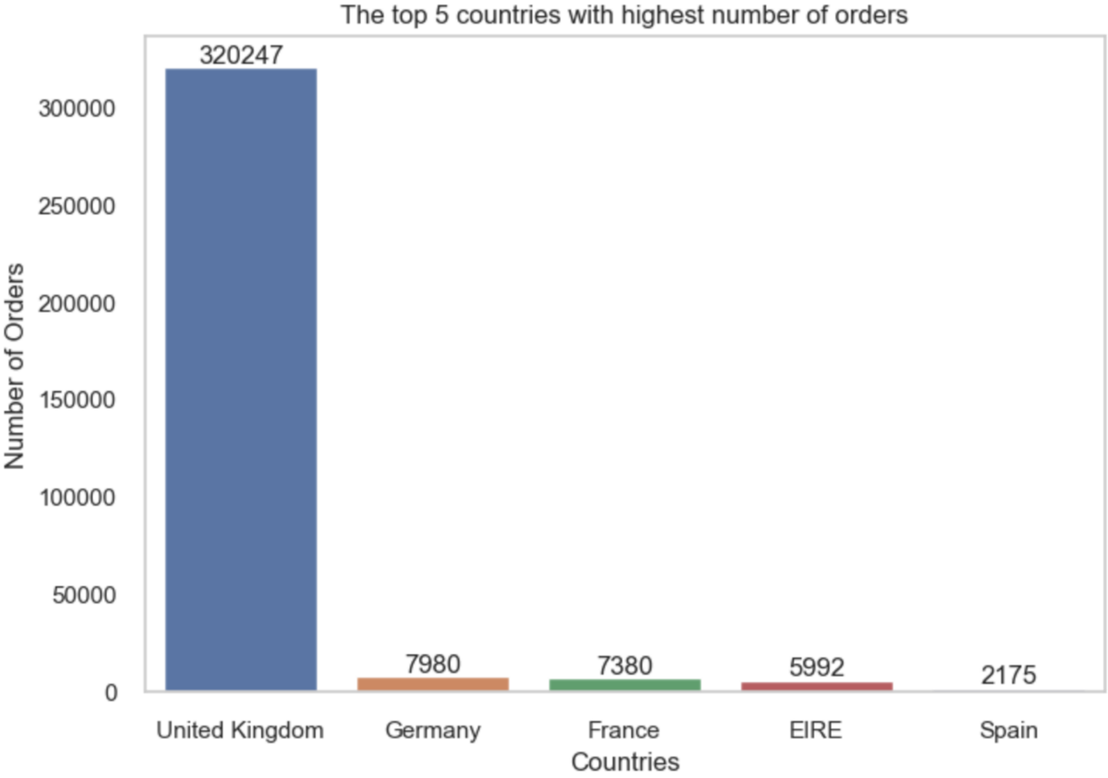


Figure 3.5.1 Top 5 countries with the highest number of orders

3.5.2 Correlation between the country of the customer and the average order value

Looking at how the country of the customer relates to the average order value gives us interesting insights from the data provided. Countries like Australia, Japan, and the Netherlands have higher average order values, meaning customers there tend to spend more per order. In contrast, the United Kingdom has a lower average order value, suggesting a focus on cheaper items or different spending habits. Lithuania stands out with an exceptionally high average order value, hinting at a preference for premium products in that market. These differences highlight the need for businesses to adjust their pricing and marketing strategies to match the spending habits in each region. Understanding these connections helps businesses offer the right products and cater to the diverse preferences of their customers.

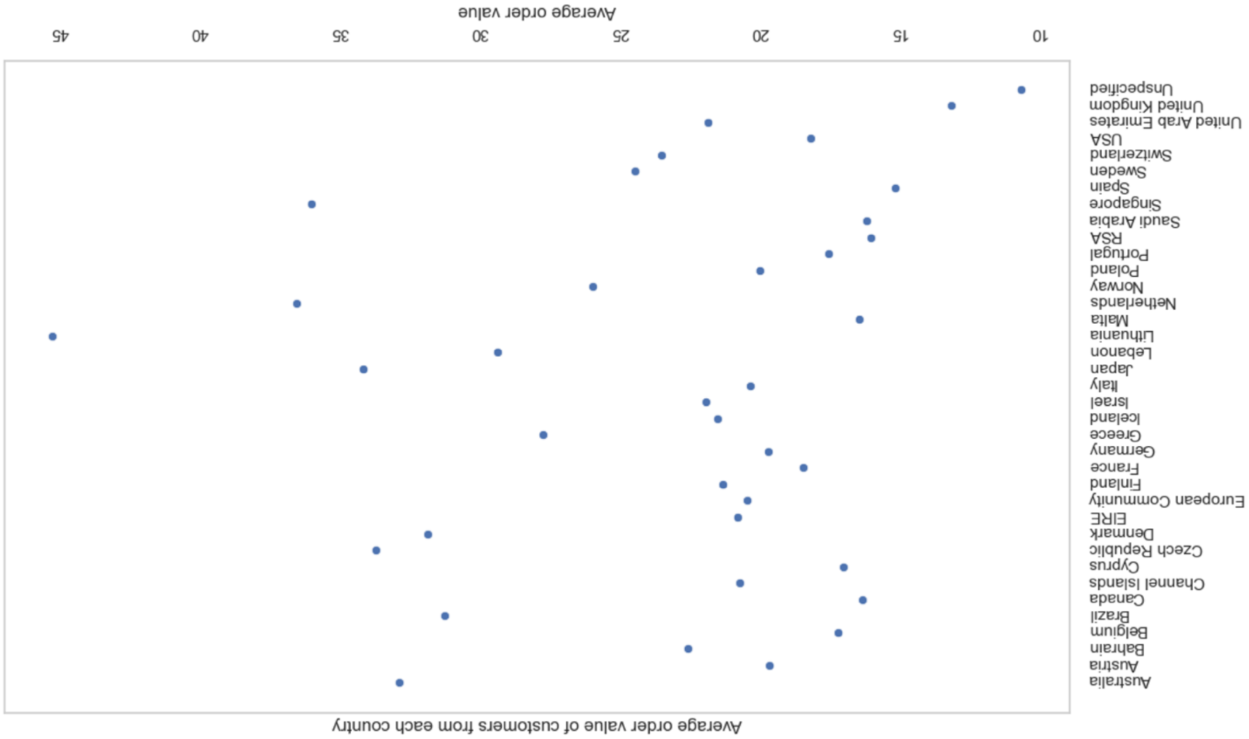


Figure 3.5.2 a) Country-wise Average total order

A scatter plot depicting the relationship between the country of the customer and their corresponding average order values has been generated to visually examine the potential correlation between these two variables.

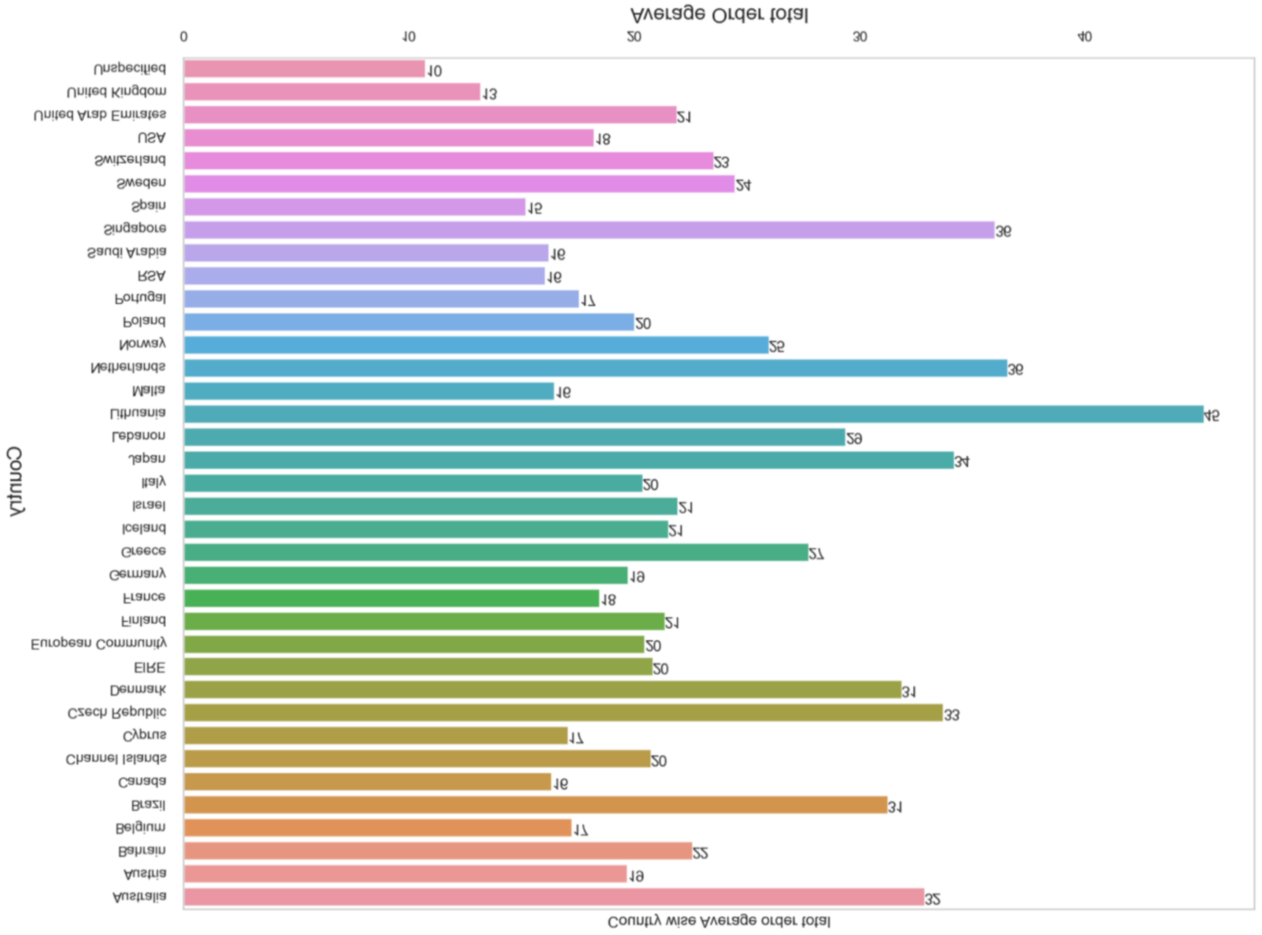


Figure 3.5.2 b) Scatter plot of Country wise Average total order

A visual representation called a heatmap was created to see if there's a connection between where customers are from and how much they typically spend per order. The correlation(-0.18) indicates that there isn't a clear relationship between a customer's country and their average order value.

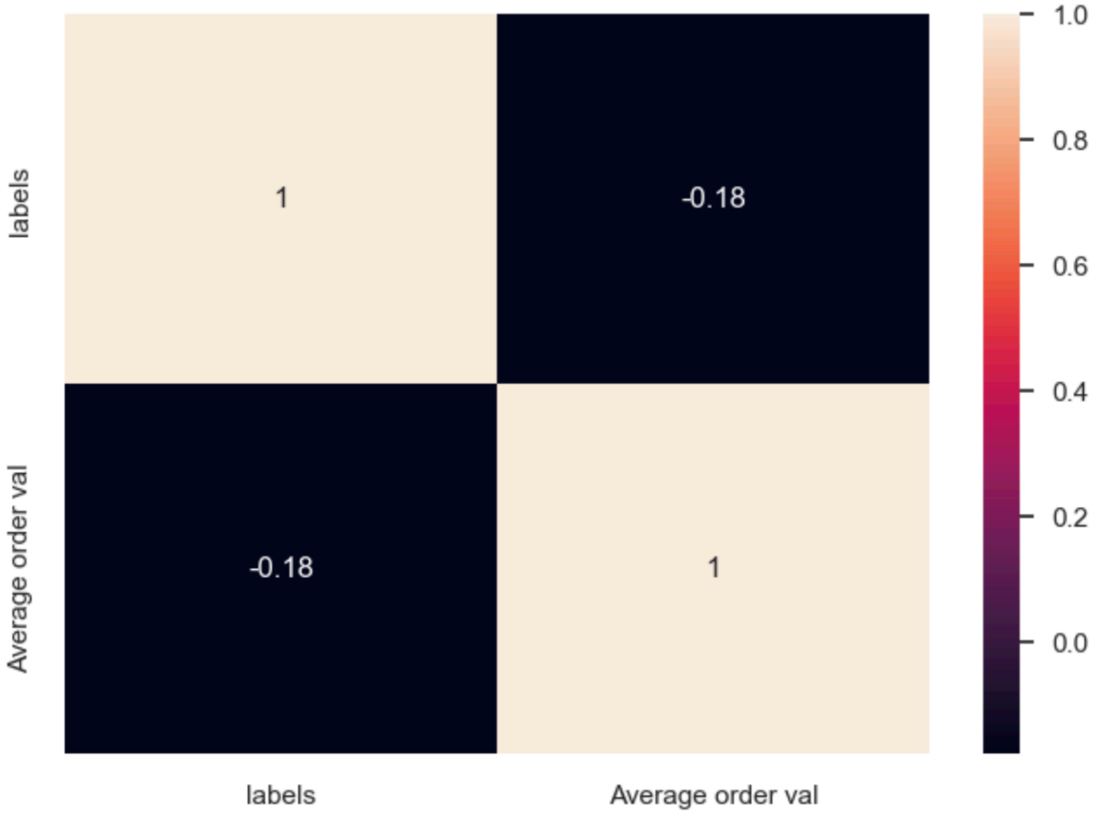


Figure 3.5.2 c) Heatmap of Country wise Average total order

3.6 Payment Analysis:

Unfortunately, there isn't enough data available to provide insights into questions related to payment analysis. The specific details about the most common payment methods used by customers and any potential relationship between payment methods and order amounts are not available for analysis. As a result, without adequate information on payment-related data, it's challenging to draw meaningful conclusions or explanations regarding these aspects of customer behaviour.

3.7 Customer behaviour:

3.7.1 Customer active duration

On average, customers stay engaged with the business for about 130 days from their first to last purchase. This duration is important for businesses to understand how long they typically keep customers interested. If the average duration is longer, it suggests a more lasting customer relationship. On the other hand, if it's shorter, it might mean customers are not sticking around as much, and the business may need to improve strategies to keep them interested. Knowing this average duration helps businesses adjust how they engage with customers, aiming to extend and make the most of these relationships for long-term success and loyalty.

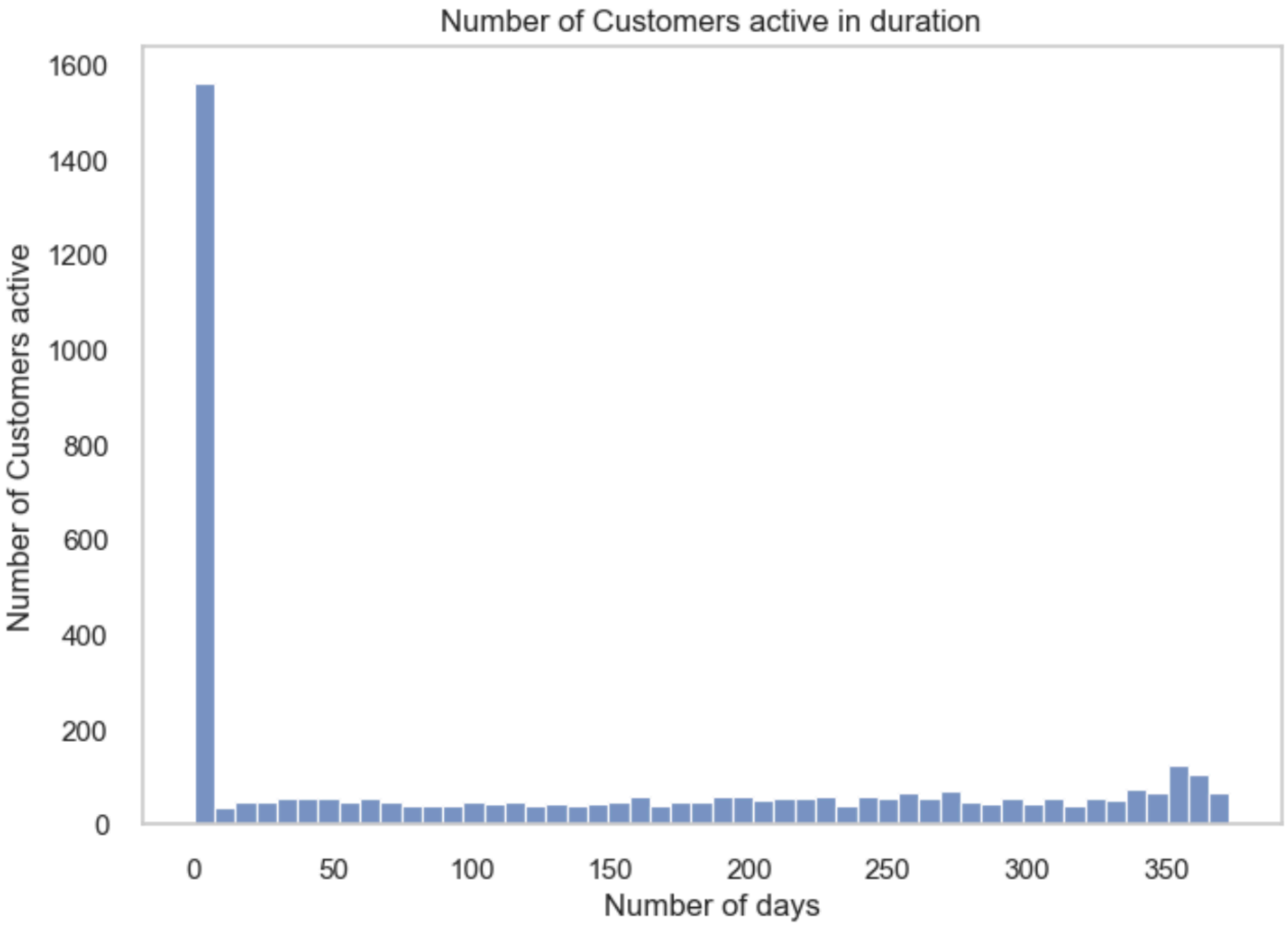


Figure 3.7.1 a) Number of Customers active in duration

The data seems uneven because most customers have made only one order. To better understand the distribution, we decided to exclude customers with just one order. This helps us focus on customers who have a more substantial ordering history. Analyzing those who order more frequently might uncover patterns or trends that get overshadowed by the large number of one-time buyers. Removing single-order customers allows for a more detailed analysis of customer behavior and ordering patterns, revealing insights that aren't influenced heavily by the characteristics of customers who buy only once.

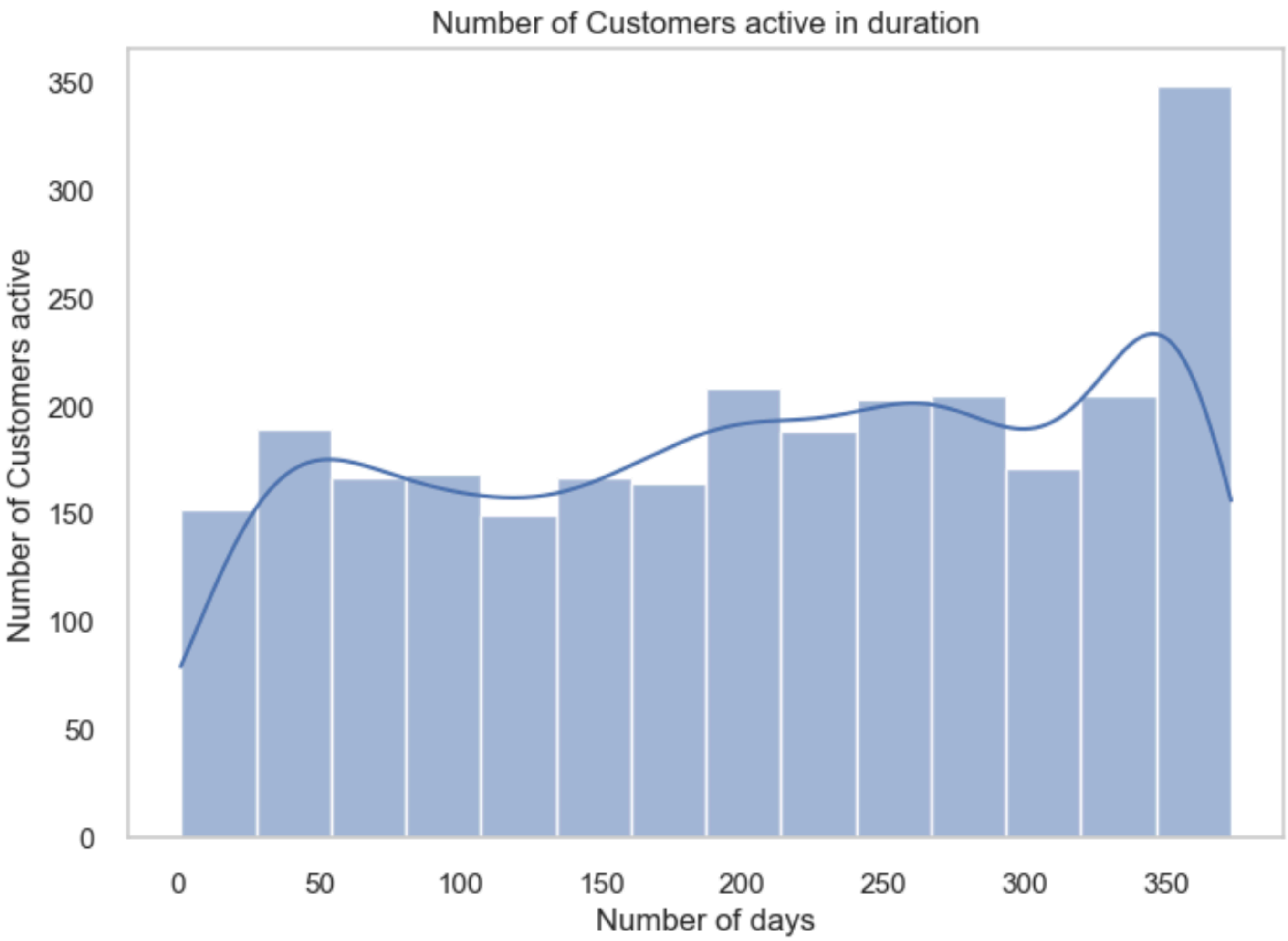


Figure 3.7.1 b) Number of Customers active in duration

Conclusion: On Average, The Customers remain active for 130 days.

3.7.2 customer segments based on their purchase behaviour

3.7.2 a) Segmenting customers according to their RFM score and subsequently assigning ranks based on the calculated RFM scores.

Customers were segmented based on their RFM (Recency, Frequency, Monetary) scores, which were then ranked from 0 to 10, with 10 being the highest. The customers were labeled as high, medium, or low value based on their RFM scores. High value customers, with scores greater than 25, were identified as those contributing the most, followed by medium value customers with scores between 15 and 25, and low value customers with scores below 15.

The normalization of the total customers and monetary columns was performed to provide a clearer interpretation of each customer segment and their contribution to overall spending.

Results show that the high value customer segment, constituting 17.84% of total customers, contributes 58.53% of the total spending. On average, each high value customer spends $3857. Medium value customers, representing 39.53% of total customers, contribute 32.46% of the total spending, with an average spending of $965 per customer. Low value customers, making up 42.63% of total customers, contribute 9.02% of the total spending, with an average spending of $248 per customer.

In terms of RFM characteristics:

- Low value customers are characterized by low frequency and high recency.
- Medium value customers have medium frequency and recency.
- High value customers exhibit high frequency and low recency.

The chart below illustrates the distribution of customers among the three segments.

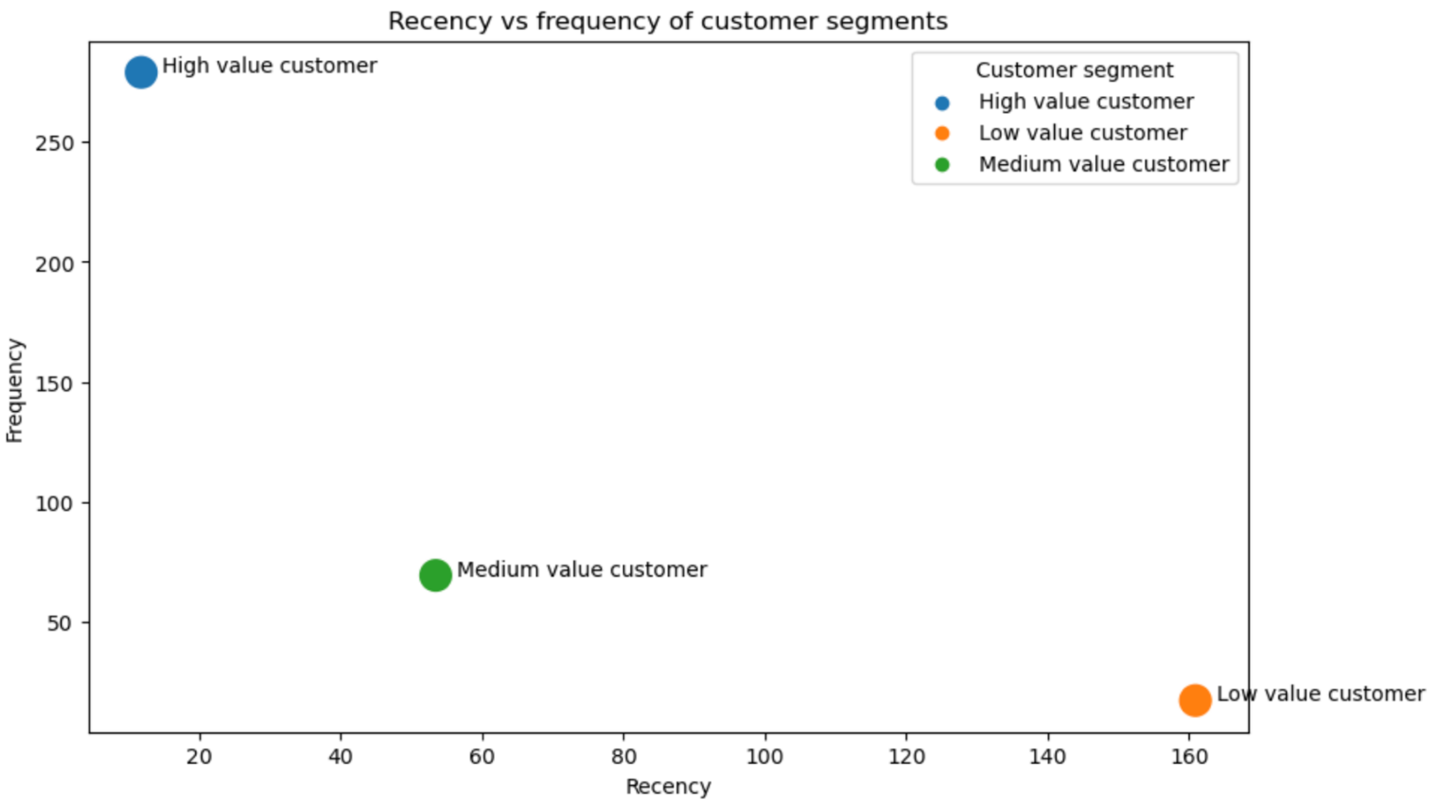


Figure 3.7.2 a) Recency vs frequency of customer segments

3.7.2 b) Clustering method to analyse the customer segments

K-Means Clustering: K-Means Clustering is a Non-parametric approach that groups the data points based on their similarity or closeness to each other and then forms K clusters from n observations.

In the process of K-means clustering, several key steps were followed. First, the data was standardized to ensure that all features had the same scale, preventing any particular variable from disproportionately influencing the clustering results.

Next, the optimal number of clusters was determined using two common techniques: the silhouette score and the elbow method. The silhouette score assesses the quality of clustering based on how similar an object is to its own cluster compared to other clusters. Meanwhile, the elbow method involves plotting the explained variation as a function of the number of clusters and looking for an "elbow" point, which signifies a suitable number of clusters where adding more does not significantly improve the model.

Upon analysis, it was determined that the optimal number of clusters for this dataset was found to be 4. Subsequently, the K-means algorithm was applied with k=4, forming distinct clusters within the data. Each cluster represents a group of data points that share similarities with each other, providing valuable insights into the underlying patterns within the dataset.

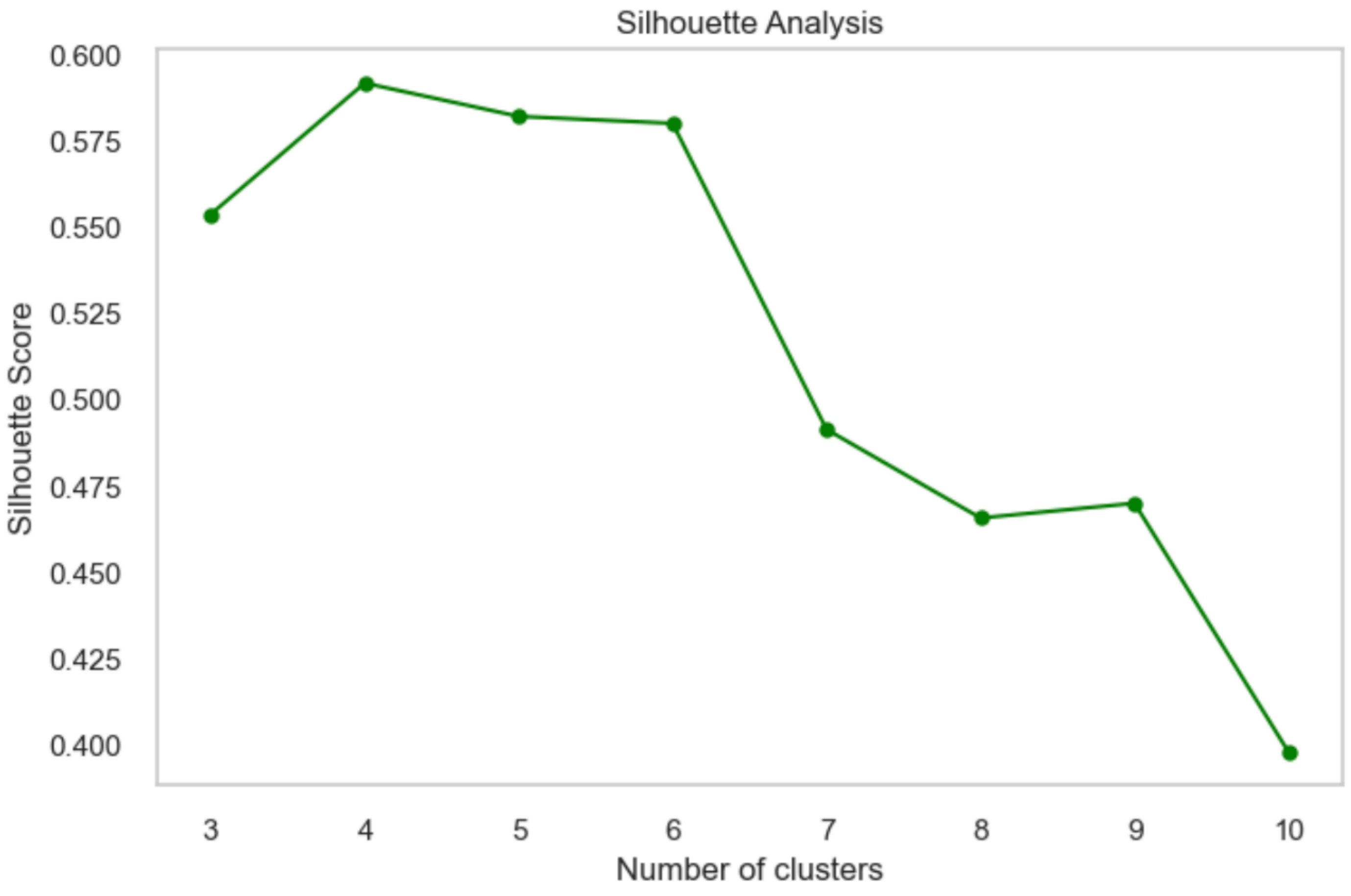


Figure 3.7.2 b) Silhouette score plot

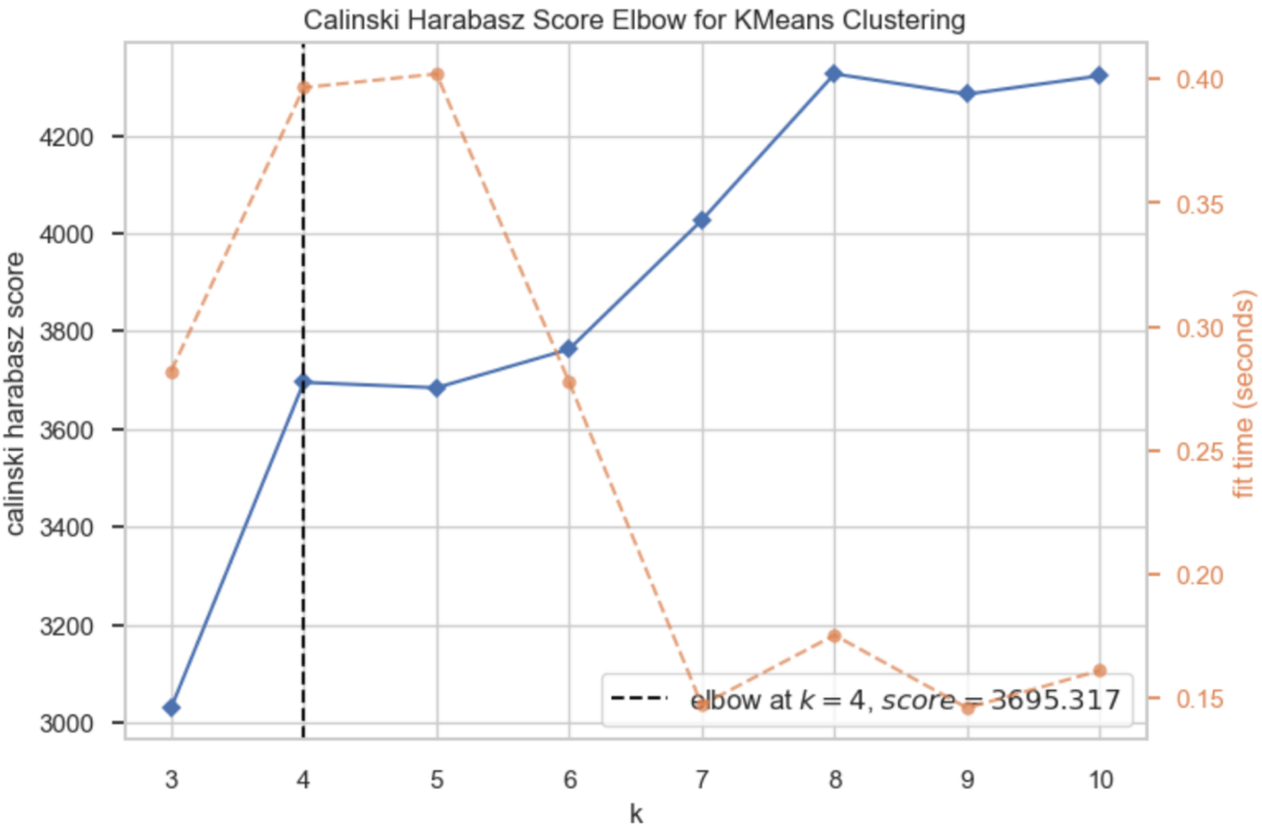


Figure 3.7.2 c) Elbow score plot

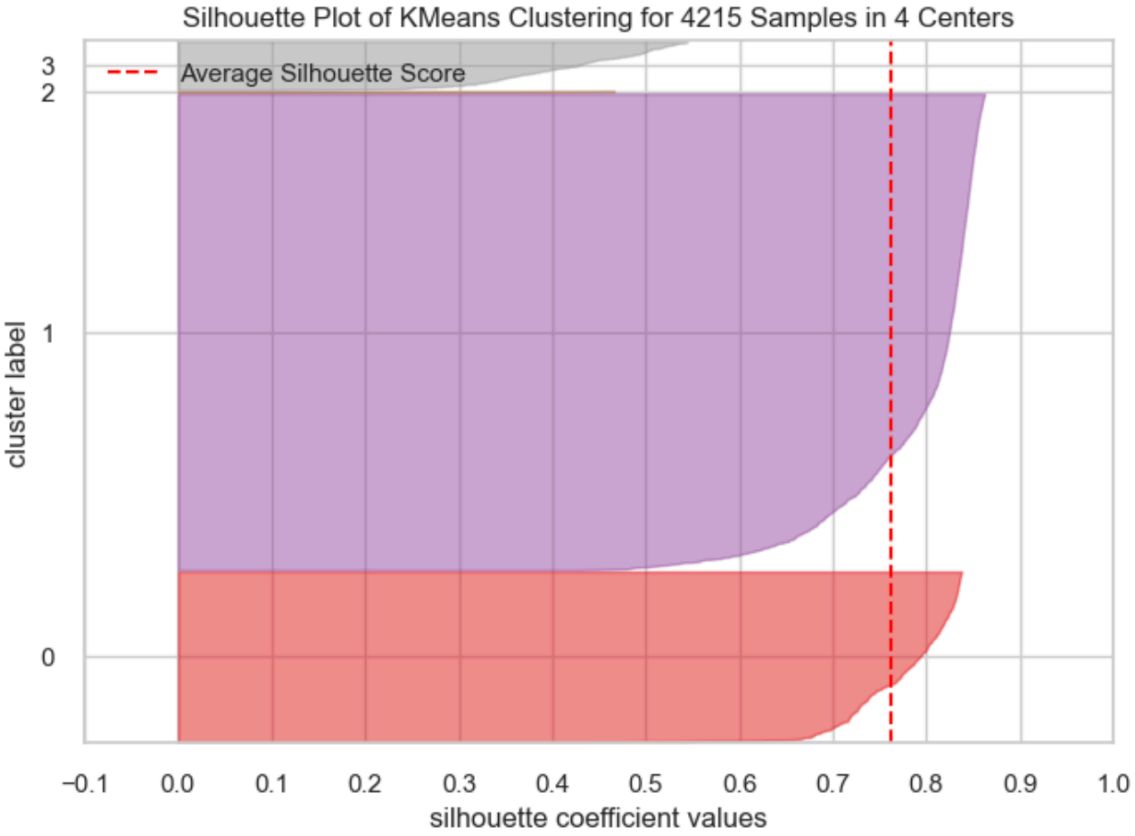


Figure 3.7.2 c) Silhouette score plot

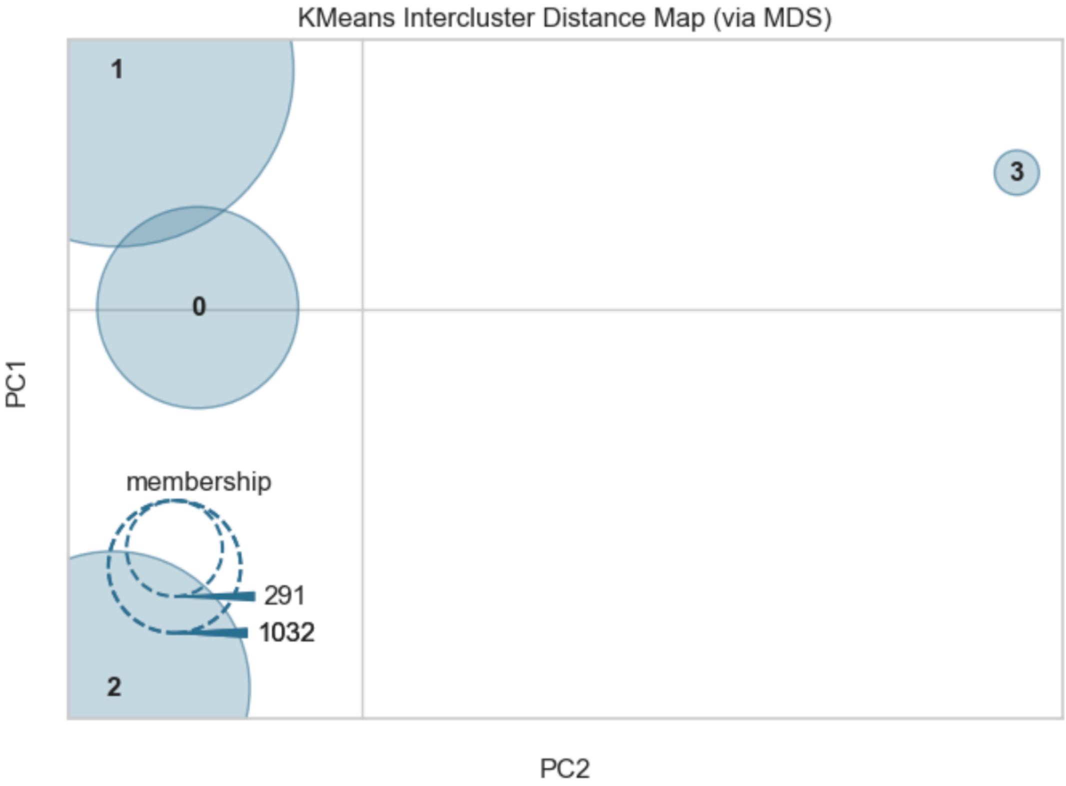


Figure 3.7.2 d) KMeans Intercluster distance map

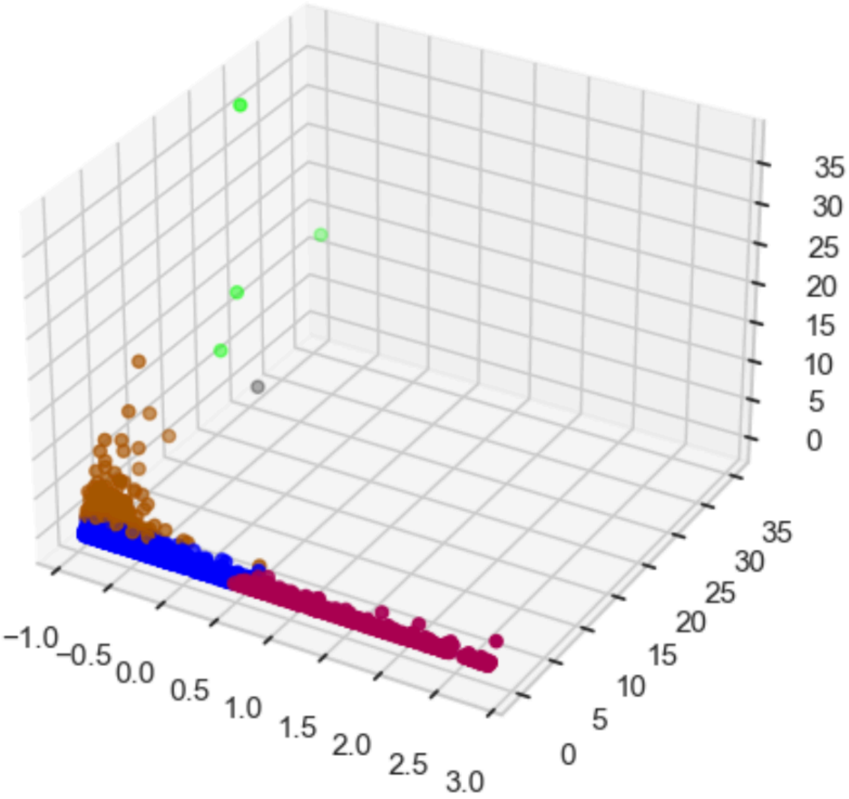


Figure 3.7.2 e) 3-D plot for segmented clusters

3.8 Returns and Refunds

3.8 a) What is the percentage of orders that have experienced returns or refunds?

This Python code utilizes a dataset loaded into a Pandas Data Frame to determine the percentage of orders that have experienced returns or refunds. The condition 'has_returns' checks if there are any rows in the dataset where the 'InvoiceNo' column starts with 'C', indicating returned or refunded orders. If such returns are present, it calculates the total number of unique orders and the number of unique orders with the 'C' prefix. The resulting percentage of orders with returns or refunds is then printed, conveying that 14.81% of the orders in the dataset have undergone returns or refunds. The significance of the 'C' prefix in 'InvoiceNo' serves as a criterion to identify returned or refunded orders within the dataset.

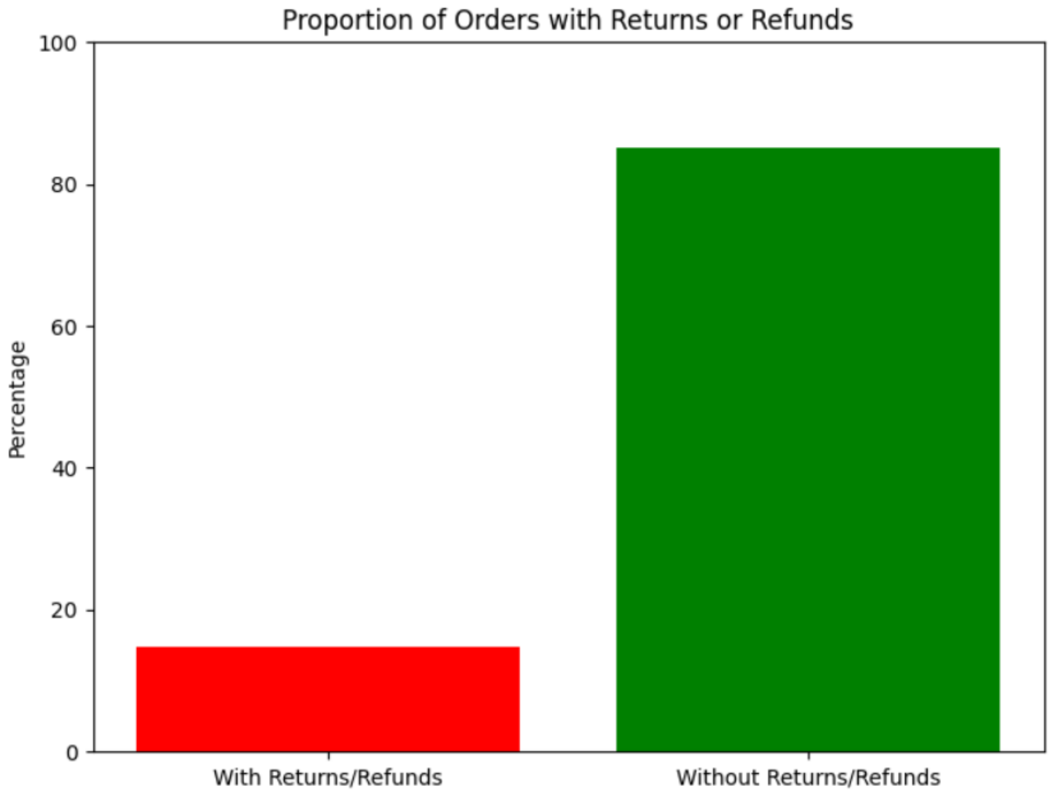


Figure 3.8. a) Bar Graph Portion of Orders with Returns or Refunds

3.8 b) Is there a correlation between the product category and the likelihood of returns?

The heatmap displayed is a visual representation of product returns by category. Each row corresponds to a different product category, such as 'Crafts and Hobbies', 'Fashion and Accessories', 'Home Decor', and so forth. There are two columns, '0' and '1', likely representing the status of returns, with '0' possibly denoting products not returned and '1' indicating products that were returned.

The colour intensity in each cell reflects the quantity related to that category and return status, with darker shades likely indicating higher numbers. For instance, 'Home Decor' has a very dark shade in the '0' column, suggesting a high number of products not returned, and a lighter shade in the '1' column, indicating fewer returns. This pattern varies across categories, with some like 'Fashion and Accessories' and 'Kitchen and Dining' also showing a higher number of products not returned, but with a comparatively higher number of returns as well, as indicated by the slightly darker colour in the '1' column.

The numbers within the cells provide exact counts for each category and return status. For example, there were 9,682 'Crafts and Hobbies' items not returned and 133 that were returned. This data can be used to analyse return rates across different categories, potentially informing decisions related to product quality, customer satisfaction, or return policy adjustments.

Given the context of the previously discussed dataset, the heatmap suggests a differential return behaviour across various product categories, which could be valuable for addressing the underlying reasons for returns and improving the product offering.

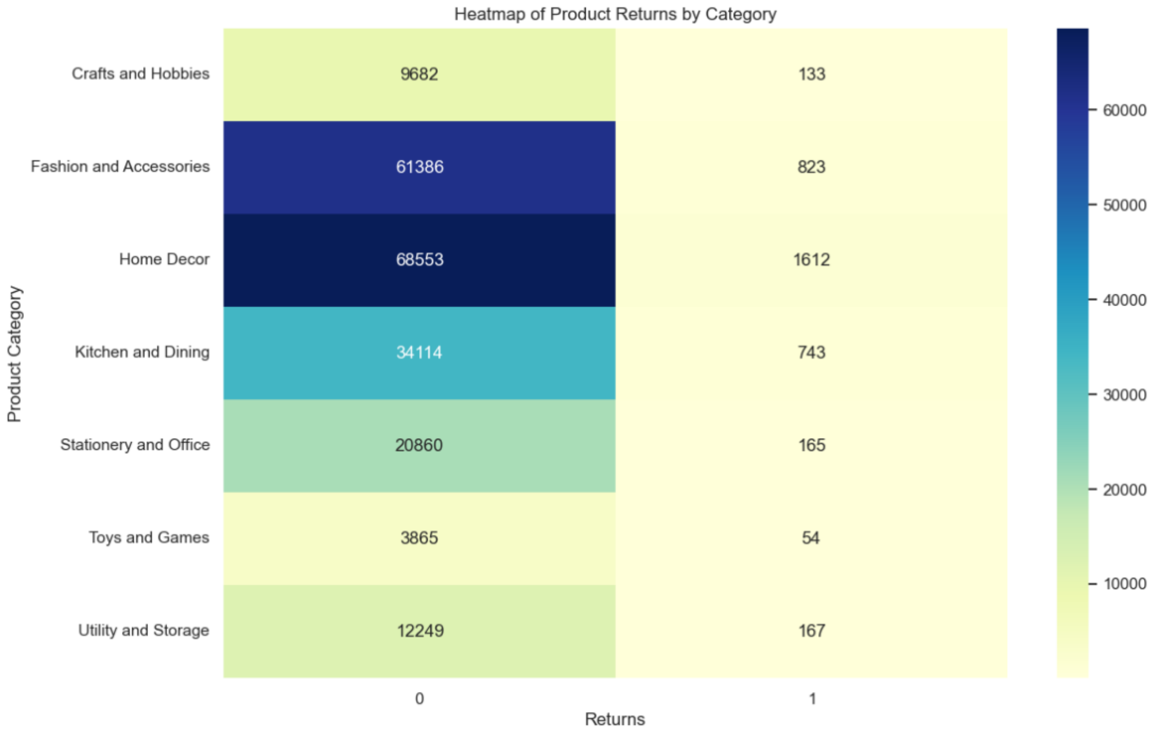


Figure 3.8. b) Heatmap of Product Returns by Category

3.9 Profitability Analysis

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

Due to the absence of comprehensive data on total profit earned by the company during the dataset's specified time period, the analysis focuses on the total revenue generated in this duration. The accompanying line graph illustrates the variations in total profit over time. Notably, a pronounced spike is observed in the month of November, implying a substantial increase in profits during this period. This surge aligns with expectations of heightened consumer activity typically associated with the peak holiday season shopping. While the analysis is limited to revenue rather than profit, the graph offers valuable insights into the notable fluctuations in financial performance, particularly during key periods of heightened consumer engagement, such as the holiday season.

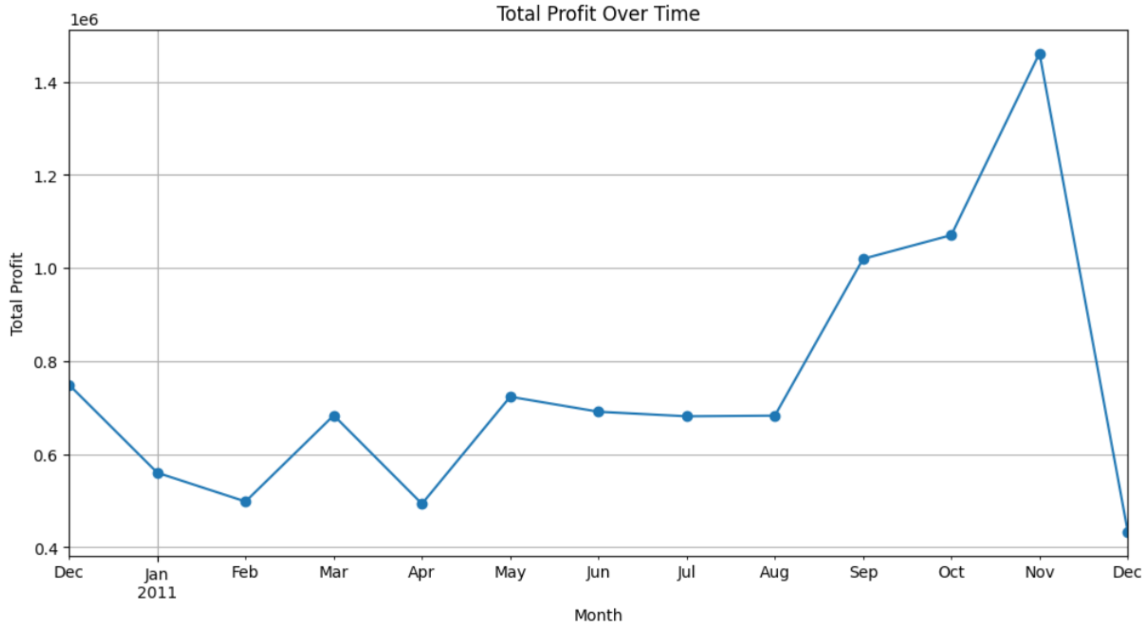


Figure 3.9. a) Total Profit Over Time

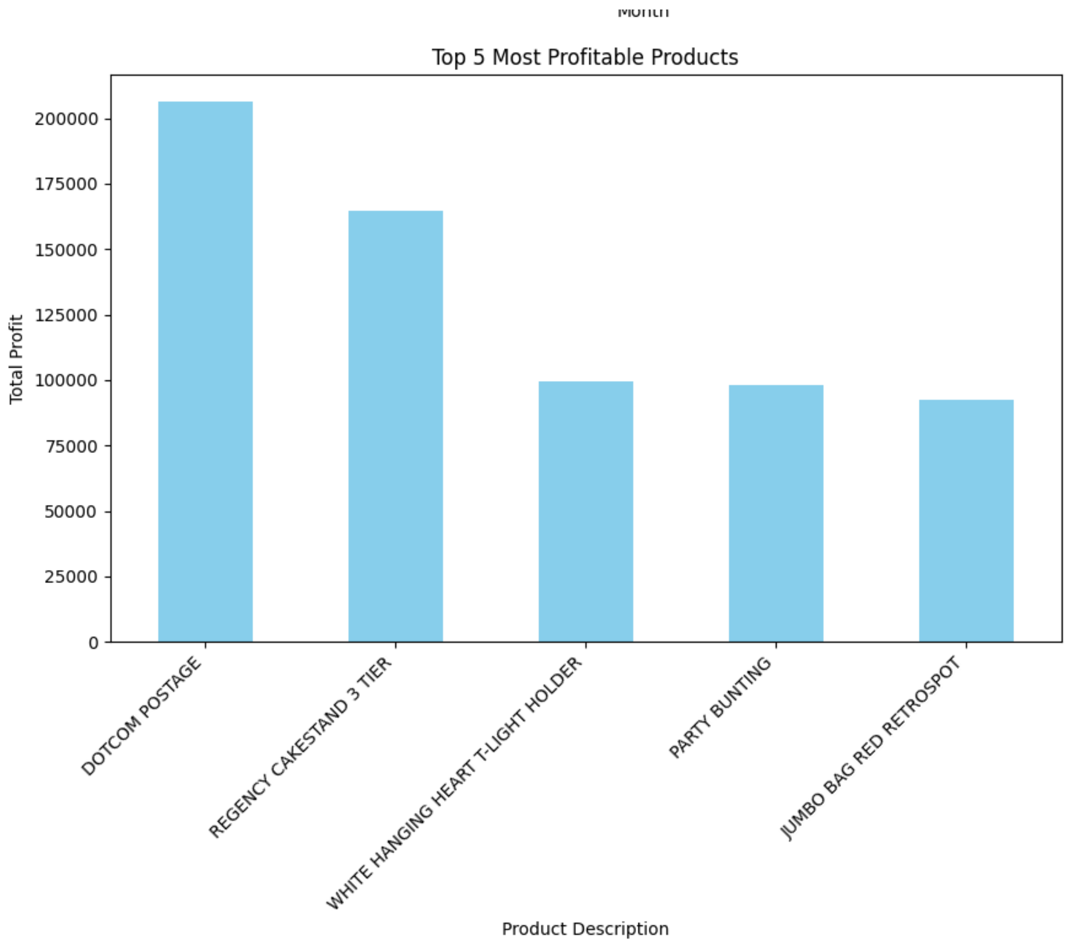


Figure 3.9. b) The top 5 most profitable products

The bar graph presents insights into the top 5 most profitable products, with DOTCOM POSTAGE emerging as the highest contributor to sales profitability. Following closely is

REGENCY CAKESTAND 3 TIER, while JUMBO BAG RED RETROSPOT appears to have the least profit among the highlighted products. The visualization provides a clear hierarchy of product profitability, allowing stakeholders to identify key contributors to overall revenue and potentially strategize around the sales performance of these specific items.

3.10 Customer Satisfaction

3.10 a) Is there any data available on customer feedback or ratings for products or services?

The absence of a dedicated column for customer feedback in the dataset precludes the derivation of any conclusive results related to customer opinions or sentiments. Customer feedback serves as a crucial element in understanding their satisfaction, preferences, and potential areas for improvement, but the dataset lacks this specific information. Without a designated column for customer feedback, it becomes challenging to gauge their experiences, sentiments, or overall satisfaction with the products or services. The limitation of not having this essential data hinders the comprehensive analysis of customer perspectives, which could otherwise offer valuable insights for enhancing the business's products and services or refining marketing and customer engagement strategies.

3.10 b) Can you analyse the sentiment or feedback trends, if available?

The dataset does not include a column labelled 'Customer Feedback,' thereby preventing the exploration and analysis of any trends or patterns related to customer sentiments or opinions. Customer feedback is a pivotal aspect of understanding consumer experiences and preferences, allowing businesses to identify strengths and areas for improvement in their products or services. Without this specific column, the dataset lacks the necessary information to delve into trends regarding customer feedback, hindering the ability to discern patterns that could inform strategic decision-making. In the absence of such data, comprehensive insights into customer satisfaction, preferences, or potential areas for enhancement remain inaccessible, limiting the scope of the analysis in understanding and responding to customer sentiments effectively.

CHAPTER 4 : LIMITATIONS

While the exploration of customer segmentation through RFM analysis offers valuable insights, it is essential to acknowledge certain limitations inherent in the dataset. Firstly, the absence of a column named 'Customer Feedback' precludes the opportunity to conduct a comprehensive trend analysis of customer sentiments or feedback trends. This limitation restricts our ability to gauge customer satisfaction and sentiment shifts over time, crucial elements in understanding the nuanced dynamics of customer experience.

Furthermore, the dataset's limitations become apparent in the realm of financial analysis. Due to insufficient data, the computation of the total profit earned by the company during the dataset's time period proves unfeasible. Instead, the analysis pivots towards calculating the total revenue generated during this timeframe. This adjustment underscores the importance of recognizing the dataset's constraints and emphasizes that the insights derived from this research are centred around revenue rather than encompassing a holistic view of the company's financial performance. As researchers and practitioners navigate these limitations, the findings presented in this project should be interpreted with a nuanced understanding of the dataset's constraints and the resultant implications for the depth of financial analysis and trend assessments.

CHAPTER 5: FUTURE WORK

1. Incorporate Customer Feedback Data: A critical avenue for future work involves obtaining and integrating customer feedback data into the analysis. Acquiring information on customer sentiments, reviews, and feedback trends would significantly enhance our ability to perform a thorough trend analysis. This addition would offer a holistic perspective on customer satisfaction and potentially unveil patterns that could inform strategic decision-making.

2. Enhance Financial Data Availability: To provide a more comprehensive financial analysis, future efforts should focus on securing additional financial data. This could involve obtaining details on operational costs, expenses, and other financial metrics beyond revenue. The inclusion of such data would enable a more nuanced exploration of the company's financial health, facilitating a more holistic understanding of its profitability over the specified time period.

Good-to-Have Enhancements:

1. Integration with External Datasets: To enrich the analysis, incorporating external datasets related to industry trends, economic indicators, or customer demographics could be advantageous. This integration would provide a broader context for understanding customer behaviour and market dynamics, leading to more informed decision-making.

2. Predictive Modelling for Customer Behaviour: Implementing predictive modelling techniques, such as machine learning algorithms, could offer insights into future customer behaviour. Forecasting trends in recency, frequency, and monetary aspects based on historical data may empower businesses to proactively tailor marketing strategies and enhance customer satisfaction.

3. Dynamic RFM Segmentation: Instead of relying on fixed quartiles or predefined bins for RFM segmentation, a dynamic approach that adapts to changing customer behaviour patterns could be explored. Continuous monitoring and adjustment of segmentation criteria based on evolving trends could provide a more accurate representation of customer segments over time.

4. Interactive Dashboards for Stakeholders: Developing interactive dashboards for stakeholders could enhance the accessibility and usability of the insights generated. Visualization tools that allow users to interact with and explore the data dynamically could empower decision-makers to derive real-time insights and make informed strategic choices.

By addressing these areas in future work, the analysis could evolve to provide a more nuanced, comprehensive, and forward-looking perspective, ultimately contributing to more effective decision-making in the dynamic landscape of eCommerce.