Customer Segmentation using RFM Analysis

IE6400 – Foundations for Data Analytics Eng

Project Two Report

Group 2

Jeffrey Johnston
Sanidhya Karnik
Himanshu Randad
Digvijay Ashish Raut
John Ayush Xavier

Introduction

In the fast-paced world of online commerce, the ability to comprehend and engage with customers on a nuanced level is critical to a company's long-term success. The eCommerce dataset at our disposal provides an opportunity to explore the complexities of customer behavior using the Recency, Frequency, and Monetary (RFM) analysis framework. RFM segmentation is a strategic cornerstone that allows businesses to categorize their customer base depending on the frequency of transactions and the monetary value associated with these transactions. This analytical methodology enables businesses to gain actionable insights, allowing them to customize marketing strategies and optimize customer retention initiatives.

As businesses struggle to compete in an ever-expanding digital marketplace, the need to personalize and target marketing efforts is more important than ever. This need is met by RFM analysis, which provides a systematic approach to understanding customer preferences and behaviors. We hope to uncover hidden patterns and trends in the eCommerce dataset by dissecting it through the lens of RFM, which can serve as a compass for businesses navigating the complex landscape of customer interactions.

The overarching goal of this project is to conduct a thorough RFM analysis, which will result in the creation of distinct customer segments. Each segment represents a distinct group of customers who have similar purchasing habits. Businesses can tailor their marketing strategies to align with the specific needs and preferences of each segment by identifying these patterns, thereby increasing the overall efficacy of their campaigns.

Monetary considerations add another level of granularity to the analysis, allowing businesses to assess each customer's economic importance. This three-part analysis serves as the foundation for assigning RFM scores, which, when combined, create a multifaceted profile for each customer. These scores serve as the foundation for subsequent customer segmentation into distinct groups, each representing a distinct combination of recency, frequency, and monetary attributes.

In conclusion, this project takes a journey into the heart of customer data, using RFM analysis as a powerful tool to uncover the nuances of customer behavior within the eCommerce dataset. The resulting segmentation will not only allow for a more in-depth understanding of customer dynamics, but it will also provide actionable insights that businesses can use to improve customer satisfaction, drive engagement, and ultimately optimize their market position in the ever-changing landscape of online commerce.

Data Overview

A data overview is a combination of Understanding the Landscape, Quality Assessment, Feature Exploration, Data Distribution, Resource Planning, Communication and Collaboration. Let's start by answering a few simple but important questions.

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

The dataset consists initially of 541,909 rows with 8 columns. The data is cleaned to obtain a refined dataset of 539,392 rows and 8 columns. This data is vital for measuring the amount of data involved, and knowing how much data cleaning impacts the dataset.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12-01-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12-01-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12-01-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12-01-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12-01-2010 08:26	3.39	17850.0	United Kingdom

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

There are eight columns that comprise the dataset which are InvoiceNo, Stock Code, Description, Quantity, Invoice Date, Unit Price CustomerID and Country. These include Integer for Quantity and float values for UnitPrice and CustomerID. Other columns are referred to as object columns.

```
InvoiceNo       object
StockCode       object
Description     object
Quantity         int64
InvoiceDate     object
UnitPrice      float64
CustomerID     float64
Country         object
dtype: object
Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'UnitPrice', 'CustomerID', 'Country'],
      dtype='object')
```

Basic Data description of the data frame is:

- The first column is the InvoiceNo. It is unique to an order and an order can have multiple items within it. InvoiceNo starts with a 'C' if it is a cancelled order.
- The second column is StockCode. It is unique to an item and is alphanumeric.
- The third column is Description. It gives a description of each item in an order.
- The fourth column is Quantity. It is the amount of any given item that's present in an order.
- The fifth column is InvoiceDate. It is the datetime value of when an order was created.
- The sixth column is UnitPrice. It is the price of any given item in an order.
- The seventh column is CustomerID. It is unique to a customer and helps in identifying who placed any given order.

- The eighth column is Country. It gives information about where the order was placed.

3. How long does this data series apply for?

The dataset covers a period from 2010-12-01 08:26:00 to 2011-12-09 12:50:00. There are 4372 unique customers in the dataset. It is crucial to consider such an understanding of the timespan within which the data was collected when it comes to understanding changes in customer behavior during those periods.

```
The dataset has 539392 rows and 8 columns after data cleaning
            Quantity      UnitPrice
count  539392.000000  539392.000000
mean        9.845904       4.673648
std       215.412652      94.614722
min    -80995.000000       0.001000
25%         1.000000       1.250000
50%         3.000000       2.080000
75%        10.000000       4.130000
max     80995.000000   38970.000000
```

A data overview lays the groundwork for successful data analysis by providing a clear understanding of what the data contains, its quality, and how it aligns with the objectives of the project.

Customer Analysis

For customer analysis we are shifting our aim to customers and a parameter any customer-oriented company looks at is Customer churn, also referred to as customer attrition, which refers to the loss of customers or subscribers for any reason at all. The Customer Churn percent calculated is 34.42%. General guidelines for churn detect that churn above 15% is a high churn rate. However, it's crucial to consider industry benchmarks and the context of the specific business. Some industries inherently have higher churn rates due to their nature or market dynamics. It's also important to analyze the trends over time, understand the reasons behind the churn, and compare the churn rate with competitors or industry standards. So, here we will answer some basic questions for customer analysis.

1. How many unique customers are there in the dataset?

There are 4372 unique customers in the dataset

2. What is the distribution of the number of orders per customer?

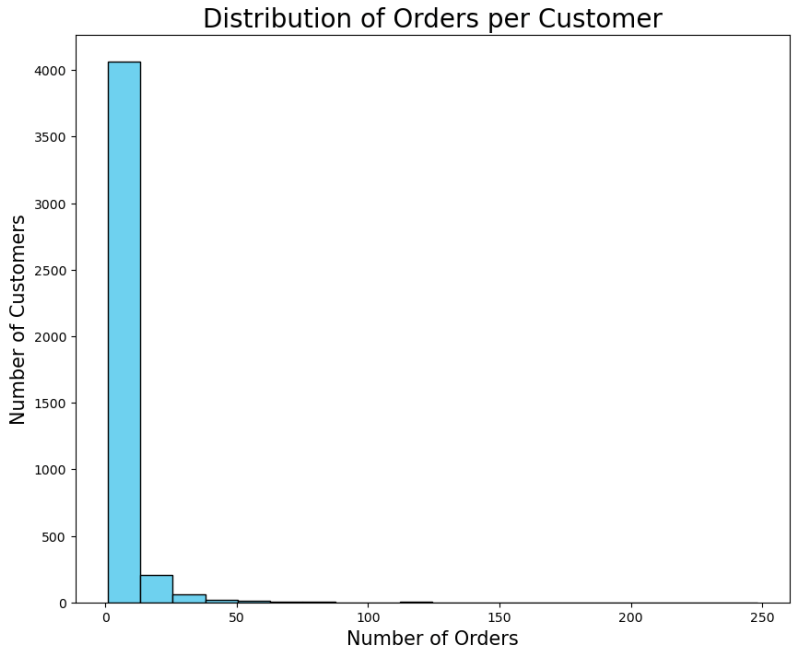
To understand the distribution of the number of orders per customer, we can infer the following:

Mean (Average Number of Orders per Customer): The mean value of 4.272015 represents the average number of orders per customer.

Standard Deviation (Variability): The standard deviation (7.697998) measures the amount of variation or dispersion in the number of orders per customer.

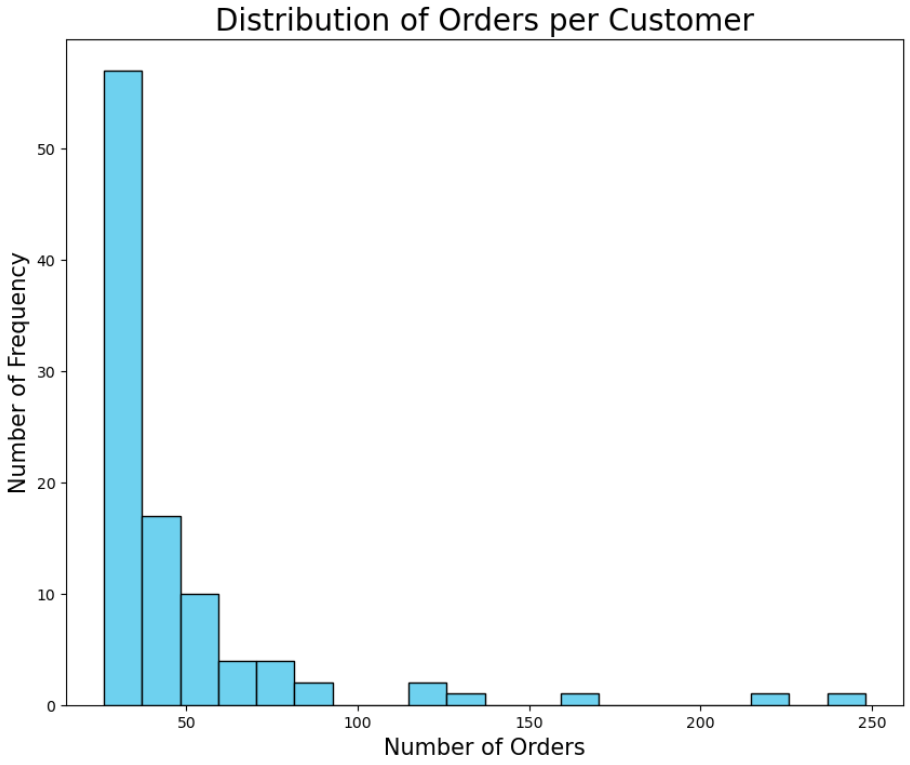
Percentiles (25%, 50%, 75%): These values provide insights into the distribution of orders. For example, 25% of customers have made 1 order or fewer (25th percentile), 50% of customers have made 2 orders or fewer (median), and 75% of customers have made 5 orders or fewer (75th percentile).

	Orders
count	4338.000000
mean	4.272015
std	7.697998
min	1.000000
25%	1.000000
50%	2.000000
75%	5.000000
max	209.000000

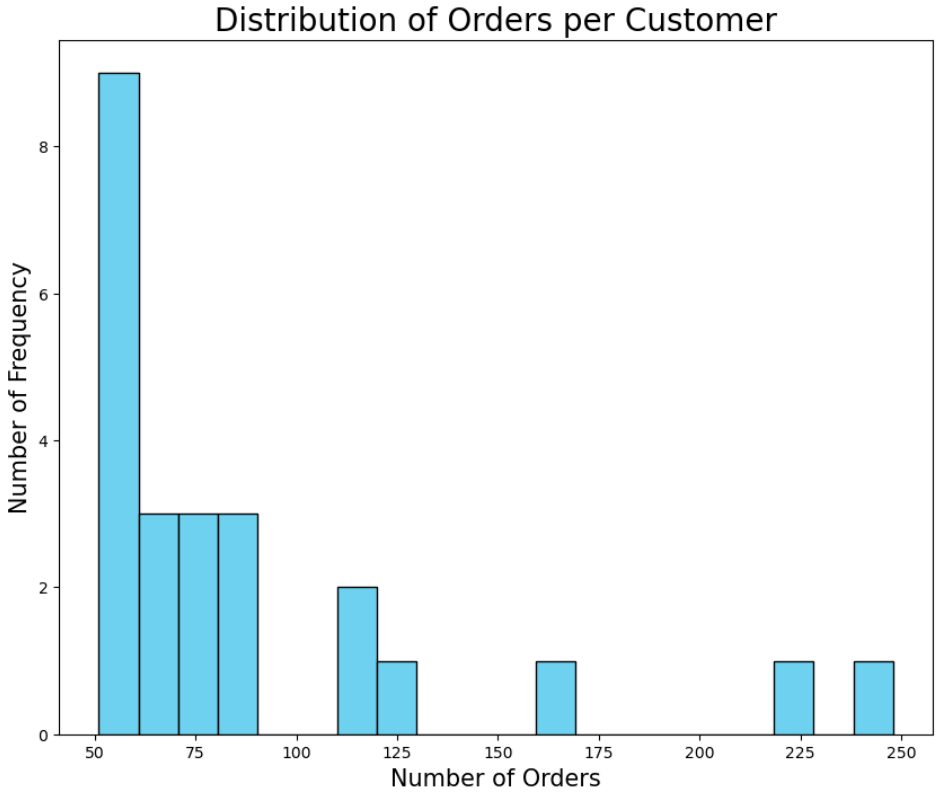


Understanding the distribution of orders per customer is valuable for businesses in tailoring marketing, especially considering the variations in customer behavior.

Order distribution among customers with greater than 25 orders



Order distribution among customers with greater than 50 orders



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

To identify most frequent buyers and making the platform more engaging for them by giving them special treatment and/or facilities like premium membership, free delivery, x% off on products, early access, etc., is a common practice to hold on to a customer by businesses. So, to identify the top 5 customers we sorted the data in descending order by “Orders” and displayed the top 5 rows.

	CustomerID	Orders
1894	14911.0	248
330	12748.0	223
4041	17841.0	169
1673	14606.0	128
568	13089.0	118

Product Analysis

Product analysis is a crucial aspect of business, it gives a business an understanding of how their product is performing, assessment of profit, inventory management, market trends, some of business decisions, etc.,

More often a business has to pick out top products which are selling. To find the top product we listed the top 10 products which have been ordered the most and printed the description and count of how many of those products were sold.

```
WHITE HANGING HEART T-LIGHT HOLDER    2365
REGENCY CAKESTAND 3 TIER              2198
JUMBO BAG RED RETROSPOT               2156
PARTY BUNTING                         1726
LUNCH BAG RED RETROSPOT               1638
ASSORTED COLOUR BIRD ORNAMENT         1501
SET OF 3 CAKE TINS PANTRY DESIGN      1473
PACK OF 72 RETROSPOT CAKE CASES       1385
LUNCH BAG  BLACK SKULL.               1350
NATURAL SLATE HEART CHALKBOARD        1280
Name: Description, dtype: int64
```

In a business it is very common that the same product will vary price in same company due to seasonal discounts, geographic location, promotions, membership discounts, customization options, etc., let's take the top selling product into consideration. The price ranges from 2.40 to 6.77.

	UnitPrice	count
0	2.40	1
1	2.55	373
2	2.95	1735
3	3.20	5
4	3.24	4
5	5.79	155
6	5.91	25
7	6.63	62
8	6.77	5

So, it's important to calculate the average price of a product. To calculate the average, we have grouped the data by description and calculated the mean of unit prices.

```
Description
 4 PURPLE FLOCK DINNER CANDLES       2.455366
 50'S CHRISTMAS GIFT BAG LARGE       1.426589
 DOLLY GIRL BEAKER                   1.502123
 I LOVE LONDON MINI BACKPACK         4.611364
 I LOVE LONDON MINI RUCKSACK         4.150000
                                       ...
ZINC T-LIGHT HOLDER STARS SMALL      0.943673
ZINC TOP  2 DOOR WOODEN SHELF       21.094167
ZINC WILLIE WINKIE  CANDLE STICK     1.089963
ZINC WIRE KITCHEN ORGANISER          9.929375
ZINC WIRE SWEETHEART LETTER TRAY     3.976522
Name: UnitPrice, Length: 4026, dtype: float64
```

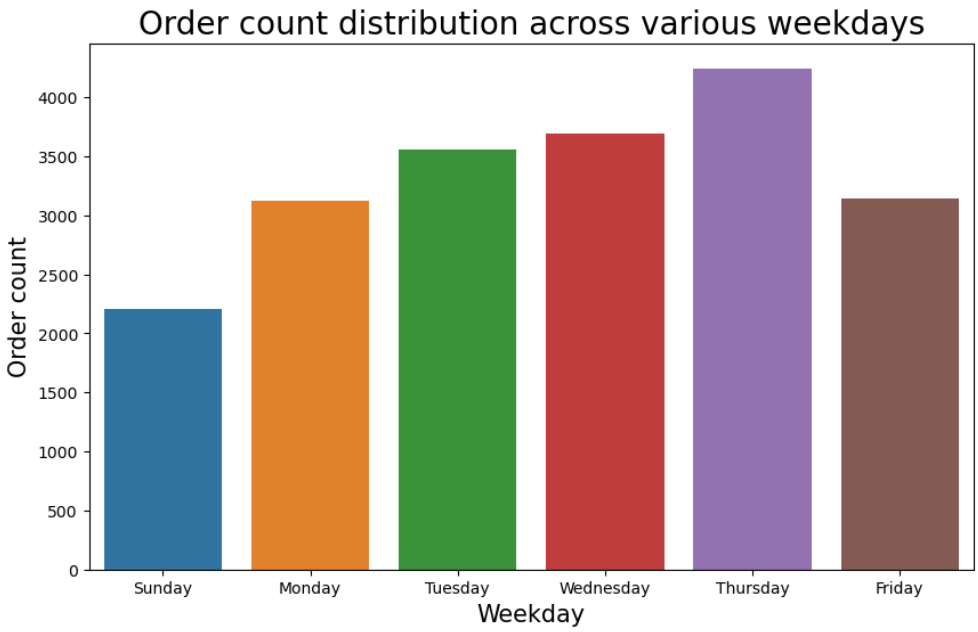
Another aspect of a business is revenue. To calculate the revenue, we have generated a new column revenue and filled it with values obtained from product of Quantity and UnitPrice. To calculate the top 10 products, we group them by description and calculate revenue by multiplying quantity by unit price then sorting them in descending order and displaying the first 10 rows.

	Description	Revenue
1074	DOTCOM POSTAGE	206245.48
2864	REGENCY CAKESTAND 3 TIER	164762.19
3859	WHITE HANGING HEART T-LIGHT HOLDER	99668.47
2422	PARTY BUNTING	98302.98
1825	JUMBO BAG RED RETROSPOT	92356.03
2752	RABBIT NIGHT LIGHT	66756.59
2703	POSTAGE	66230.64
2390	PAPER CHAIN KIT 50'S CHRISTMAS	63791.94
228	ASSORTED COLOUR BIRD ORNAMENT	58959.73
751	CHILLI LIGHTS	53768.06

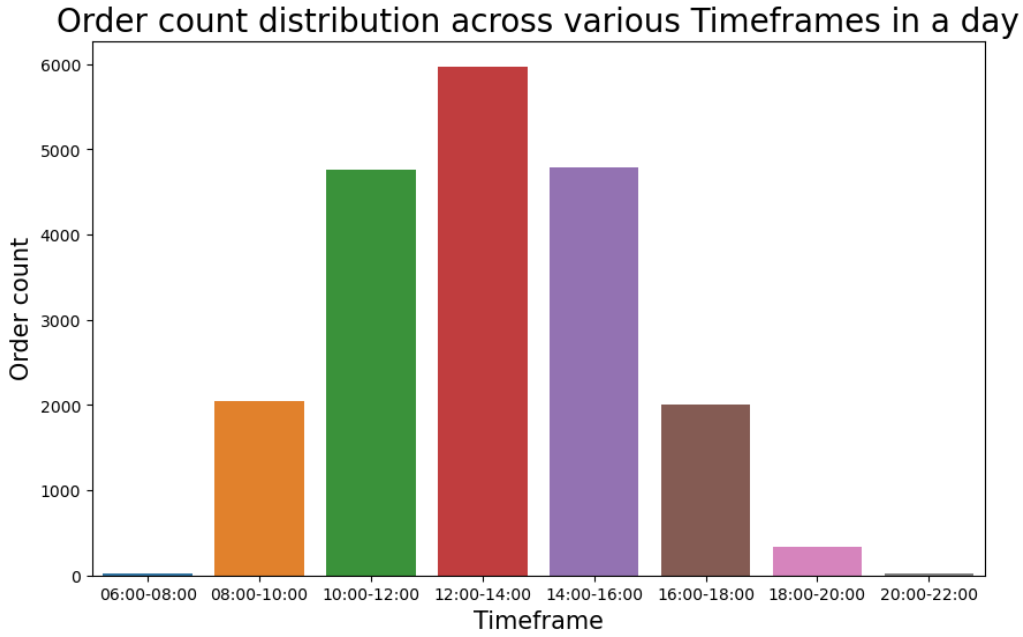
In summary, product analysis is a multifaceted process that contributes to strategic decision-making, resource optimization and overall business success. Not only do these statistics help a business to perform better but also this helps them to find all the flaws in products and help them to better understand their customers.

Timeframe Analysis

Conducting a time analysis of the purchases allows companies to visualize and understand when their customers are making the most purchases and when they are most likely to make future purchases. We start by separating the year, month, day, weekday and time from the InvoiceDate column into their own columns. By grouping the orders from the dataset by weekday we are able to determine that most orders are placed on Thursdays with the least number of orders placed on Sundays.

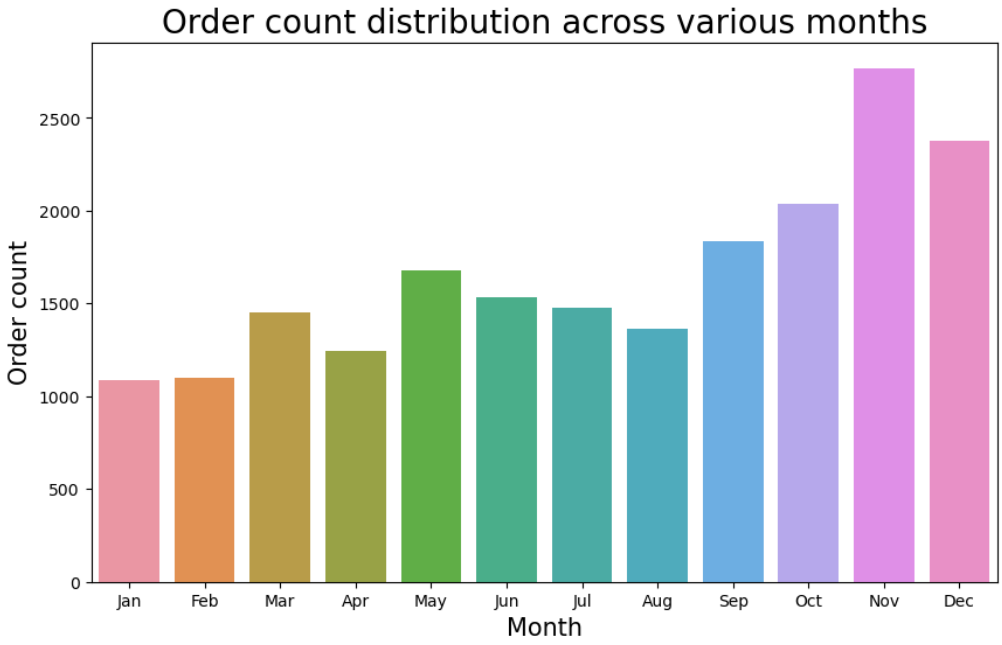


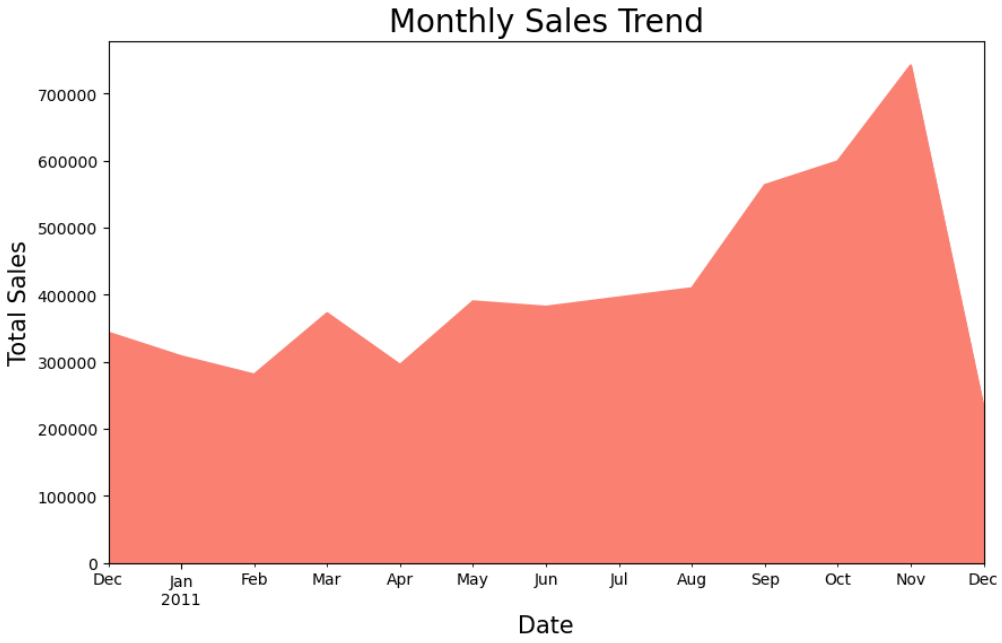
By ordering the invoices into bins of two-hour time blocks we determine most orders are placed between 10:00 am and 4:00 pm with the largest number of orders being placed specifically between 12:00 pm and 2:00 pm.



We are unable to determine the average order processing time because we do not have the information to answer that question. To determine the processing time, we would additionally need to know the day and time the order was shipped out from the company after the customer placed the order. Calculating the difference between when the order was placed and when it was shipped would give us the processing time.

To determine seasonal trends the dataset was grouped by month allowing us to determine the number of orders by month. The data shows a seasonal trend in which orders begin to increase towards the end of the year with November being the month with the most orders being placed (2769 orders). This is most likely due to customers purchasing gifts for the winter holidays as well as seasonal sales such as Black Friday sales after Thanksgiving. The graphs below show plots of the number of orders by month as well as total sales by month.



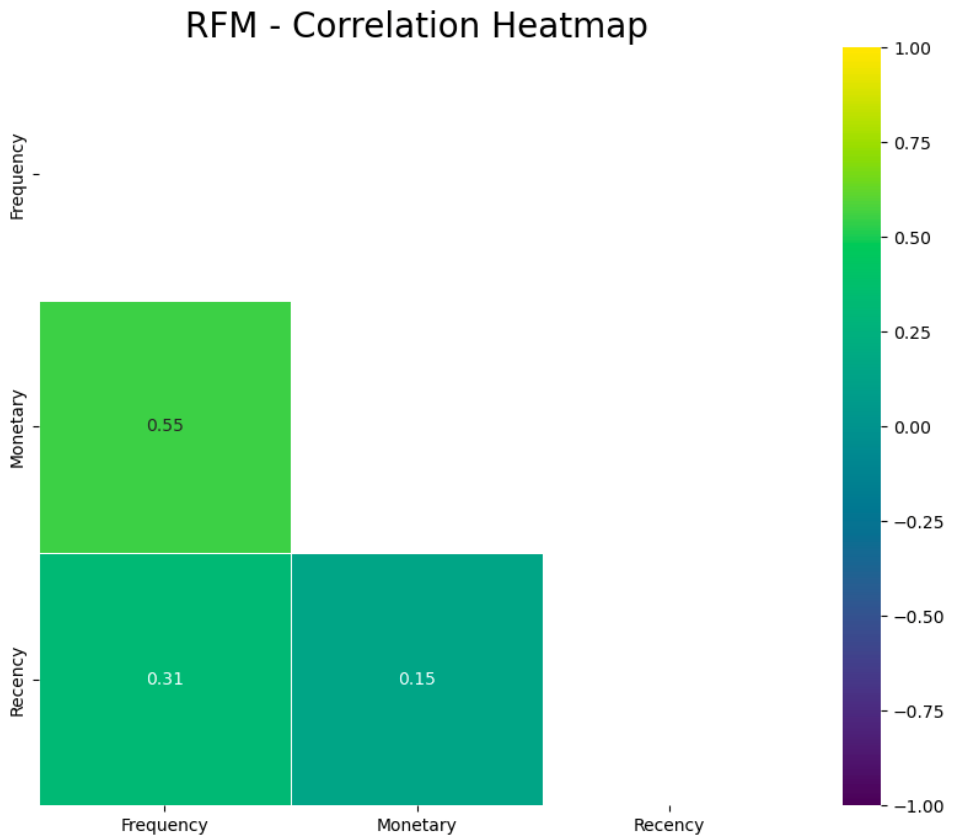


RFM Analysis

RFM stands for Recency, Frequency, and Monetary Value, and it is a marketing analysis technique used to understand and categorize customers based on their behavior. Recency refers to how recently a customer has made a purchase. For example, you may have segments like "Recent Customers" who have made a purchase within the last month, "Active Customers" within the last three months, and "Inactive Customers" who haven't made a purchase in the last six months. Similarly, frequency is the frequency of their transactions. This segment includes three types of customers "Frequent Shoppers" for those who make regular purchases, "Occasional Buyers" for those who buy infrequently, and "One-Time Customers" for those who have made only a single purchase. Monetary Value represents the total amount of money a customer has spent on purchases. For monetary values customers can be divided into three segments namely "High-Value Customers" for those who spend the most, "Medium-Value Customers" for moderate spenders, and "Low-Value Customers" for those with minimal spending.

Businesses often use a combination of these factors to create more detailed and targeted segments. For example, you might have a segment of "VIP Customers" who are recent, frequent, and high-spending customers, or a segment of "At-Risk Customers" who were once frequent buyers but haven't made a purchase in a while. All of these important calculations are derived by RFM analysis.

We have calculated and inferred that strongest correlation is between frequency-monetary, followed by correlation between frequency-recency and lastly correlation between monetary-recency, with values 0.55, 0.31 and 0.15 respectively.



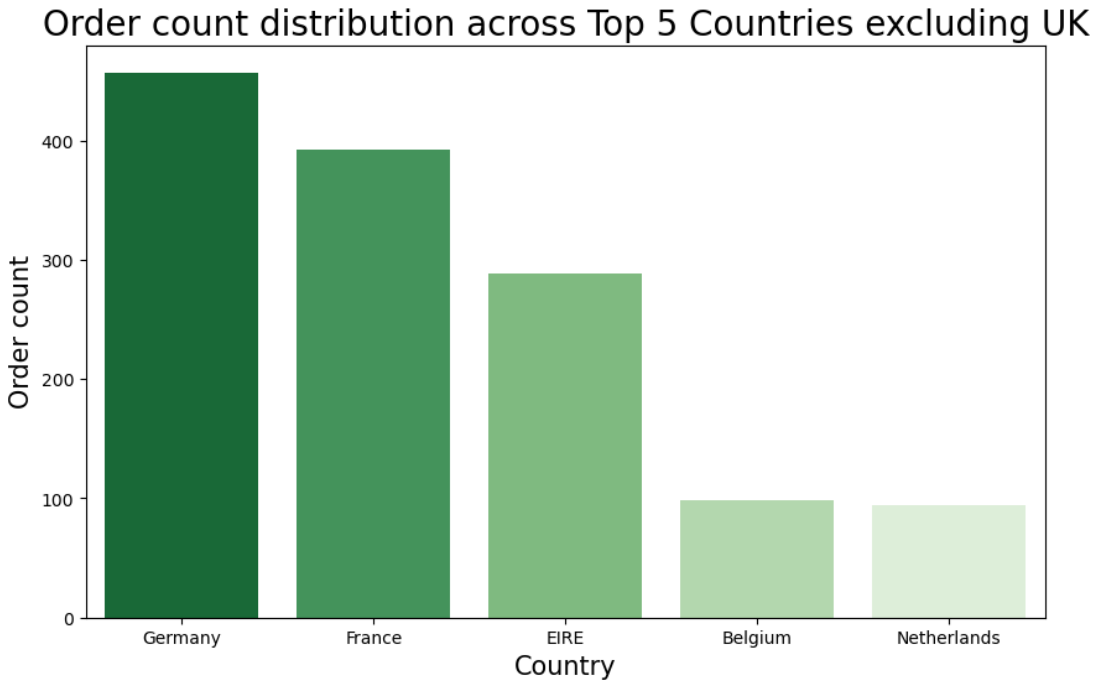
Geographical Analysis

The geographical analysis uses the country data from each invoice to determine where orders are being placed and can be used to help businesses determine where potential future orders could come from. In doing so they can more accurately use targeted advertising and keep customers active.

The five countries with the highest number of orders are the United Kingdom with 18019 orders, Germany with 457 orders, France with 392 orders, EIRE (Ireland) with 288 orders and Belgium with 98 orders.

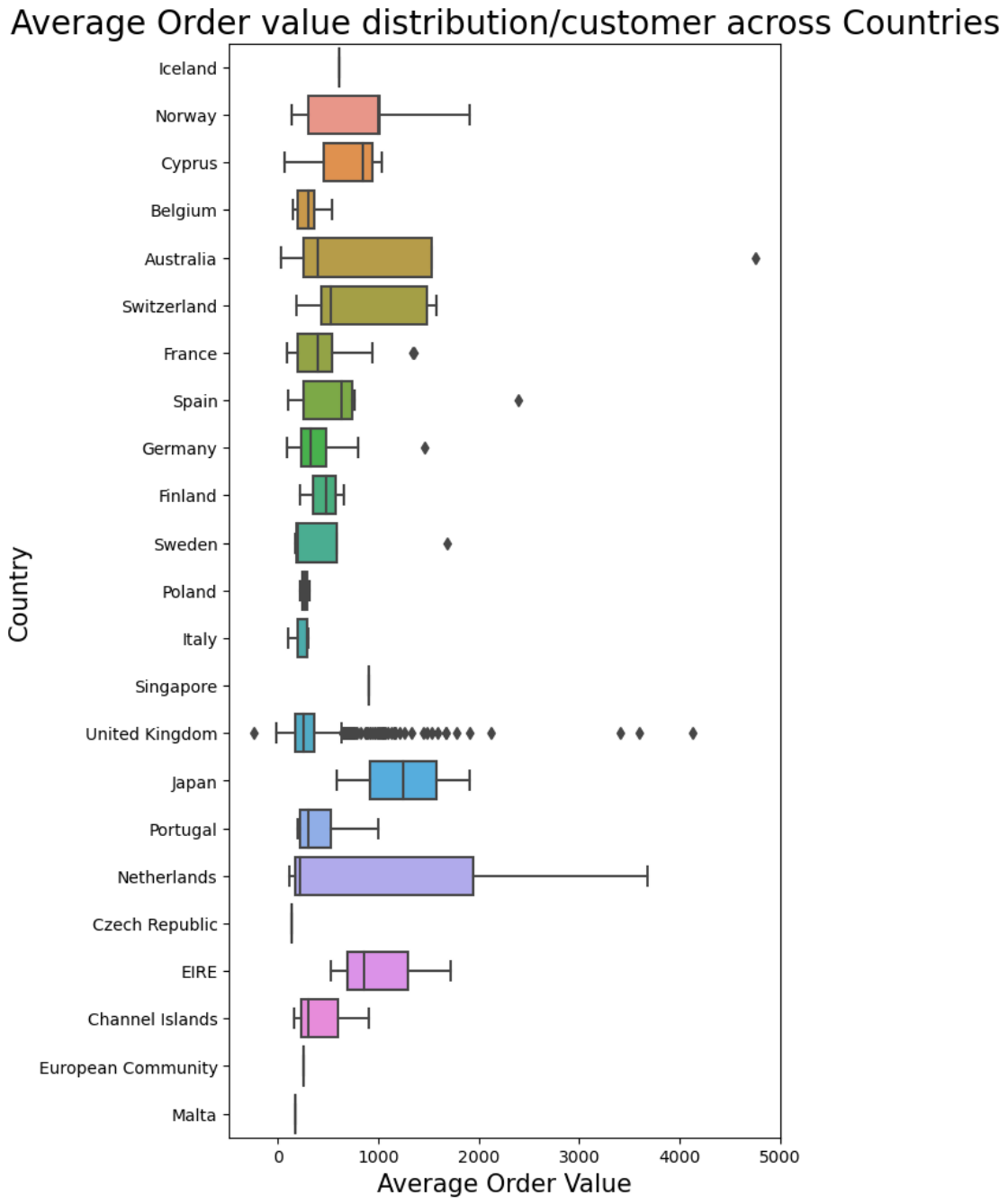
	Country	Orders
36	United Kingdom	18019
14	Germany	457
13	France	392
10	EIRE	288
3	Belgium	98

By removing the United Kingdom from consideration, the next five countries are much closer in number of total orders. The country with the next most orders is the Netherlands.



To determine correlation between the country of a customer and average order value the orders were grouped by CustomerID and the mean revenue for each customer's orders was calculated. The below

boxplot shows the distribution of customer's average order value by country. The plot shows us that even though the United Kingdom has the largest number of orders it has one of the lowest average order values. Germany, France and Belgium are all on the lower end of average order value as well. Generally, the more orders a country has, the lower the average order values are.



Payment Analysis

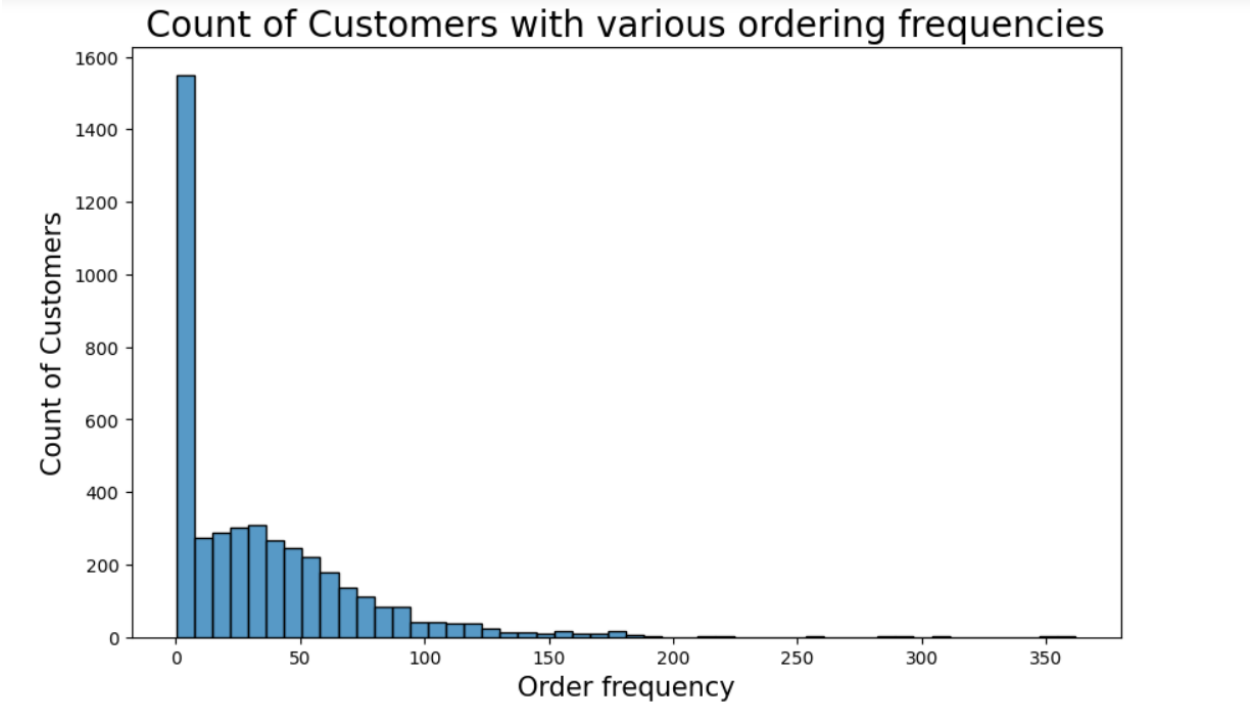
We do not have data to determine the most common payment methods used or to determine if there is a relationship between the payment method and the order amount. To determine this, we would need the dataset to include a column for payment method for each invoice.

Customer Behavior

Customer Order Activity

	CustomerID	First_order_date	Last_order_date	Orders	Days_active
0	17850.0	2010-12-01 08:26:00	2011-02-10 14:38:00	35	71 days 06:12:00
1	13047.0	2010-12-01 08:34:00	2011-11-08 12:10:00	18	342 days 03:36:00
2	12583.0	2010-12-01 08:45:00	2011-12-07 08:07:00	18	370 days 23:22:00
3	13748.0	2010-12-01 09:00:00	2011-09-05 09:45:00	5	278 days 00:45:00
4	15100.0	2010-12-01 09:09:00	2011-01-13 17:09:00	6	43 days 08:00:00
...	...	...	...	...	...
4366	13436.0	2011-12-08 10:33:00	2011-12-08 10:33:00	1	0 days 00:00:00
4367	15520.0	2011-12-08 10:58:00	2011-12-08 10:58:00	1	0 days 00:00:00
4368	13298.0	2011-12-08 13:11:00	2011-12-08 13:11:00	1	0 days 00:00:00
4369	14569.0	2011-12-08 14:58:00	2011-12-08 14:58:00	1	0 days 00:00:00
4370	12713.0	2011-12-09 12:16:00	2011-12-09 12:16:00	1	0 days 00:00:00

Customers and their order frequencies

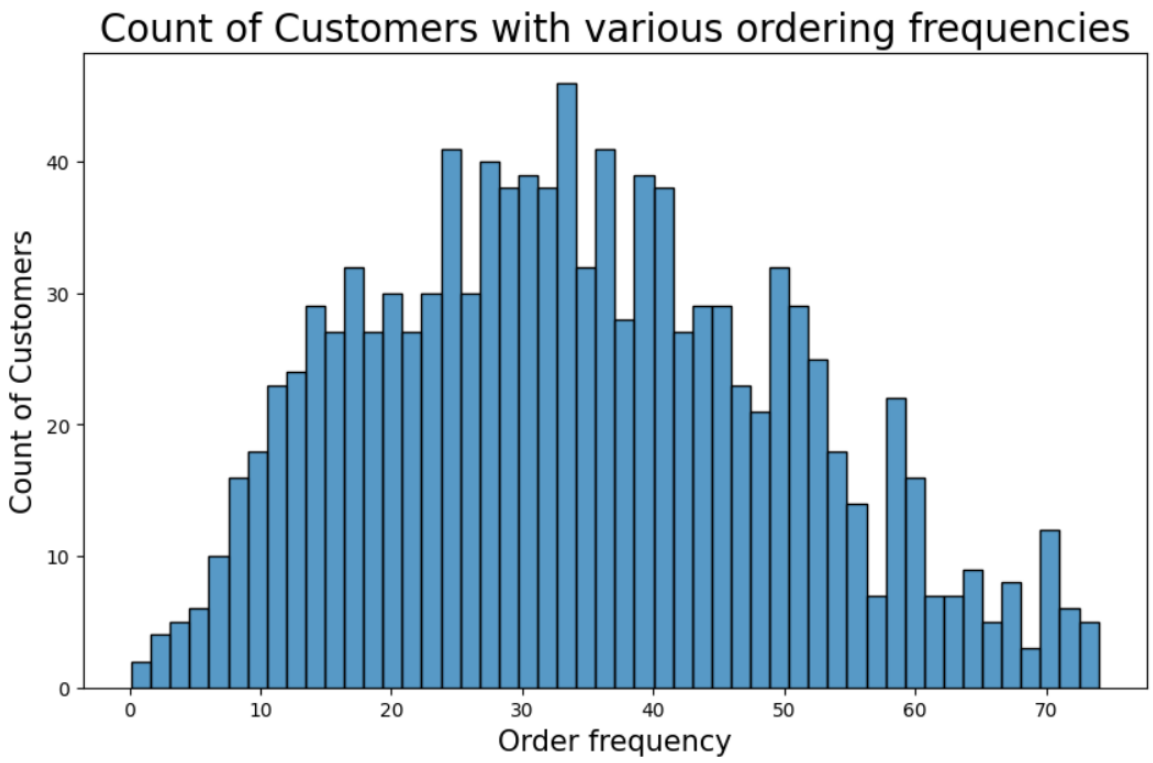


The data suggests that, on an average, there is a span of 134 days between a customer's initial order and their most recent one. This indicates a time lapse that, on average, customers take between making their first purchase and their latest one.

Furthermore, when focusing on customers who have a minimum of 2 orders, we observe that the average active status varies by 191 days. In other words, the time difference in the status of activity (perhaps indicating ongoing engagement or participation) for customers with at least two orders is, on average, 191 days.

Moreover, for customers who have made at least 5 orders, there is a substantial difference in active status, averaging 266 days. This suggests that customers who engage more frequently, with a higher order count, tend to exhibit a more prolonged and pronounced difference in their active status over time.

Customers with at least 5 orders and their frequencies



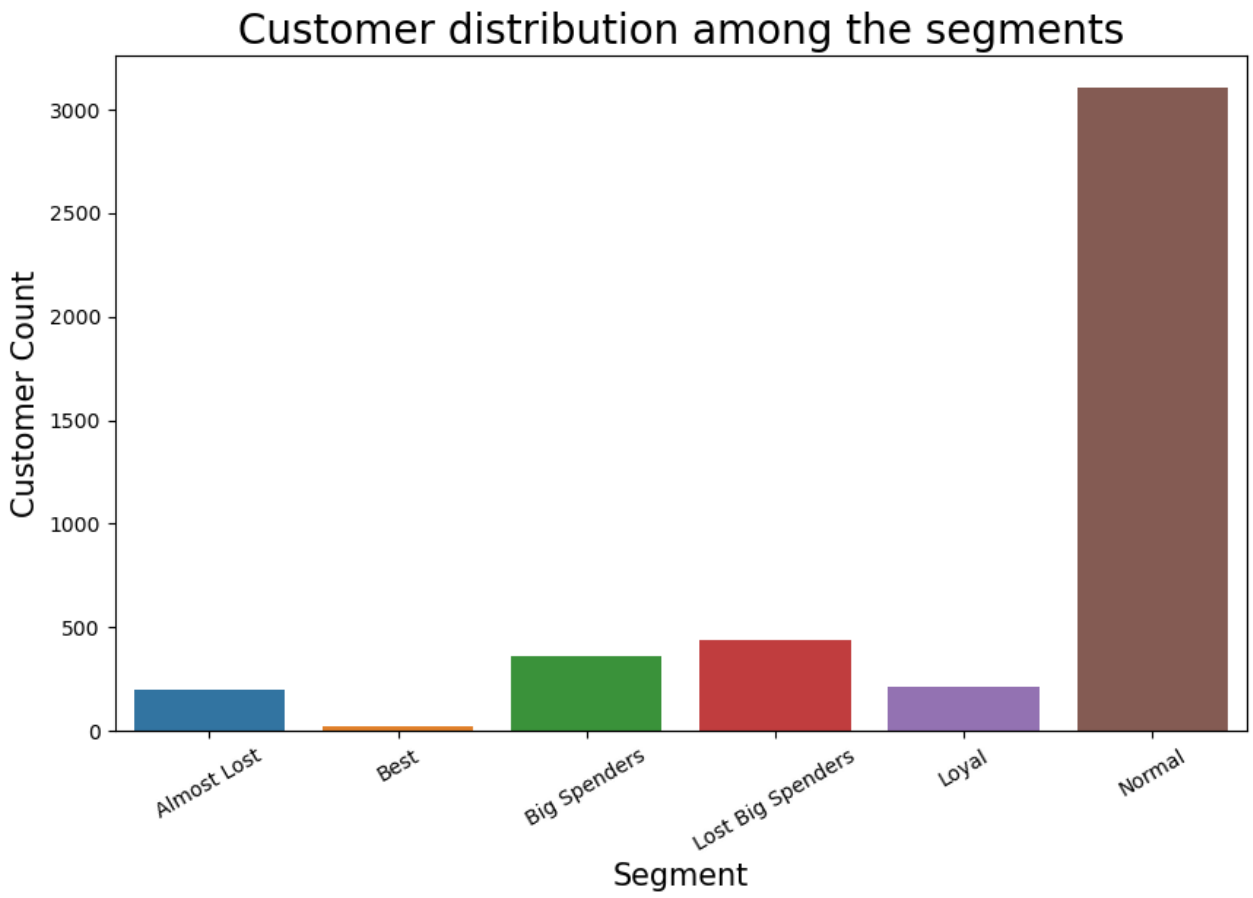
Through the use of a bar plot, we categorized customers into different types based on certain criteria to visualize customer segments based on purchase behavior. The dominant category, as indicated by the plot, is that of "normal customers." These are likely individuals whose purchasing behavior falls within a standard or typical range.

In contrast, the count of "best customers" is notably lower when compared to other customer types. This suggests that customers exhibiting the characteristics of the "best" category, which may include high spending or frequent purchases, are not as prevalent in the overall customer base.

Furthermore, the "lost" category, representing customers who may have discontinued their engagement with the business, shows a minimal count. This implies that there is a relatively low rate of customer churn or attrition.

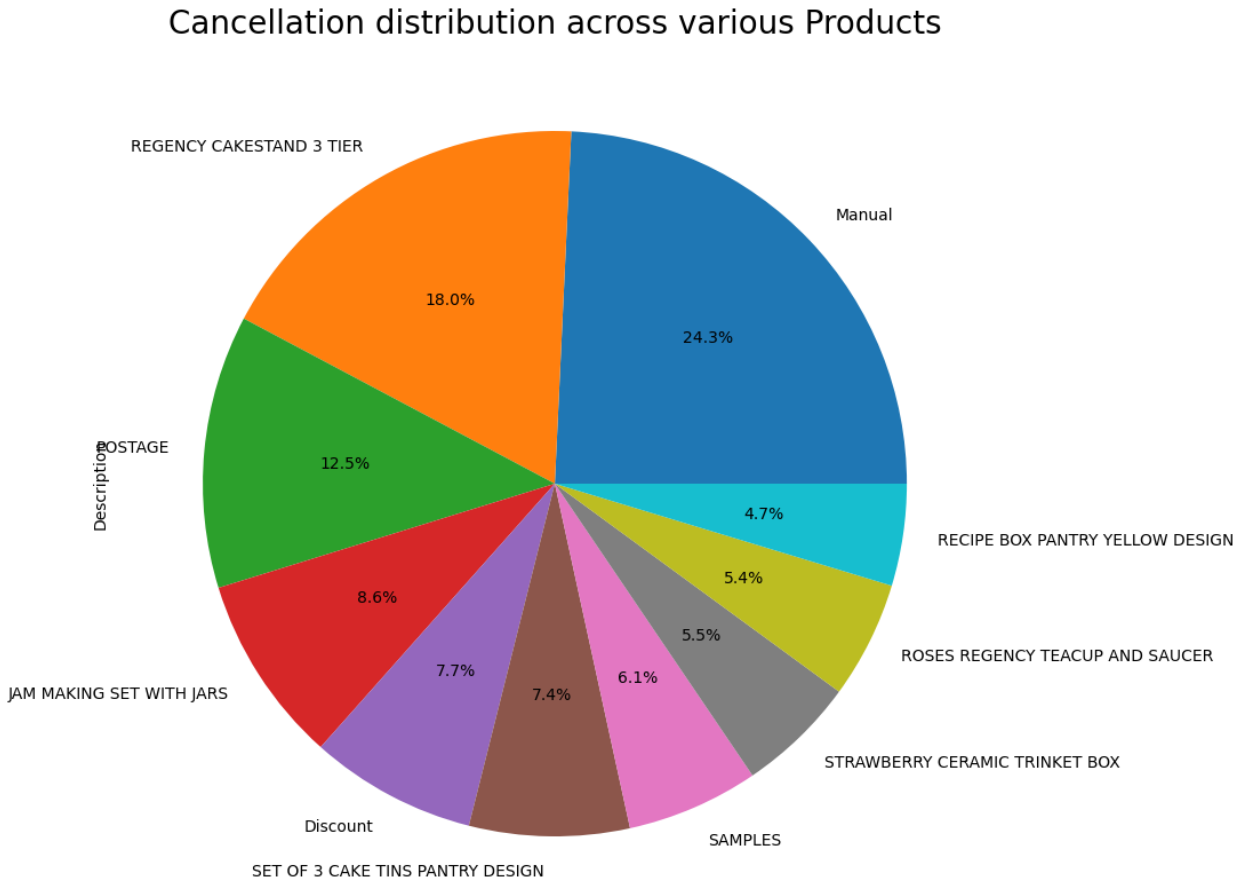
Interestingly, the analysis also reveals a decline in the count of customers associated with high spending. This decline in the number of customers who used to spend significantly on their purchases indicates a shift in the spending behavior of the customer base.

In summary, the bar plot offers a visual representation of customer distribution across different categories, shedding light on the prevalence of various customer types and changes in customer behavior, such as a decrease in high-spending customers and a low count of lost customers.

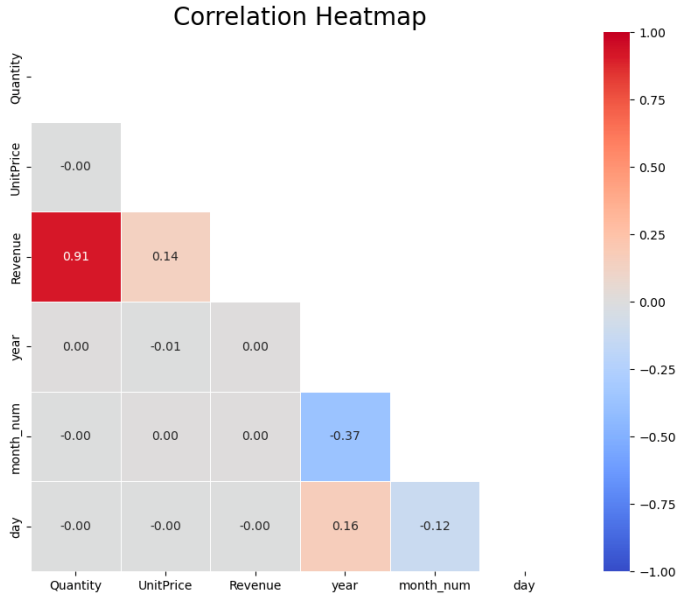


Returns and Refunds

Based on the dataset, the percentage of orders that have experienced return or refund is 19.22%. Based on the classification of products as shown in the pie chart below, orders with the item description of ‘manual’ comprise 24.3% of all returns and refunds followed by ‘Regency Cakestand 3 Tier’ comprising 18% of the refunds and returns.



The correlation heatmap below tells us that there is a strong positive correlation between revenue and quantity as expected and a weak positive correlation between revenue and unit price.



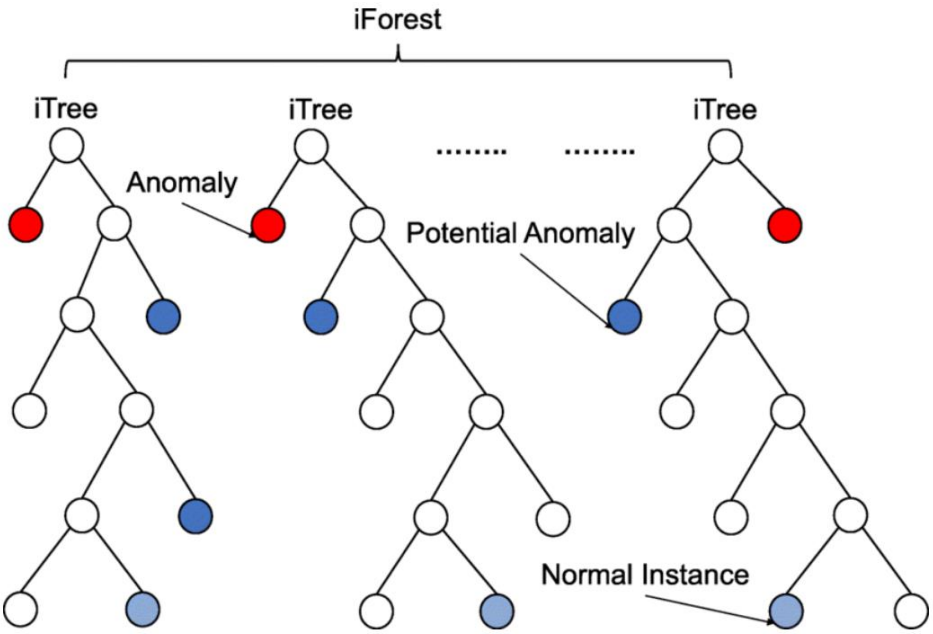
Clustering using K-Means

We have performed clustering on the dataset to find and classify comparable data points in the dataset.

Firstly we have Identified and eliminated the outliers using the Isolation Forest Method. Now, the isolation Forest (IF) is constructed using decision trees. Here, there are no pre-established labels. This approach for unsupervised learning isolates outliers from the data in order to identify anomalies.

Few - these represent the minority, with fewer occurrences and

Different - their attribute values differ significantly from those of typical instances.



This should make anomalies easier to spot because the Isolation Forest algorithm is predicated on the idea that anomalies are observations that are uncommon and unique.

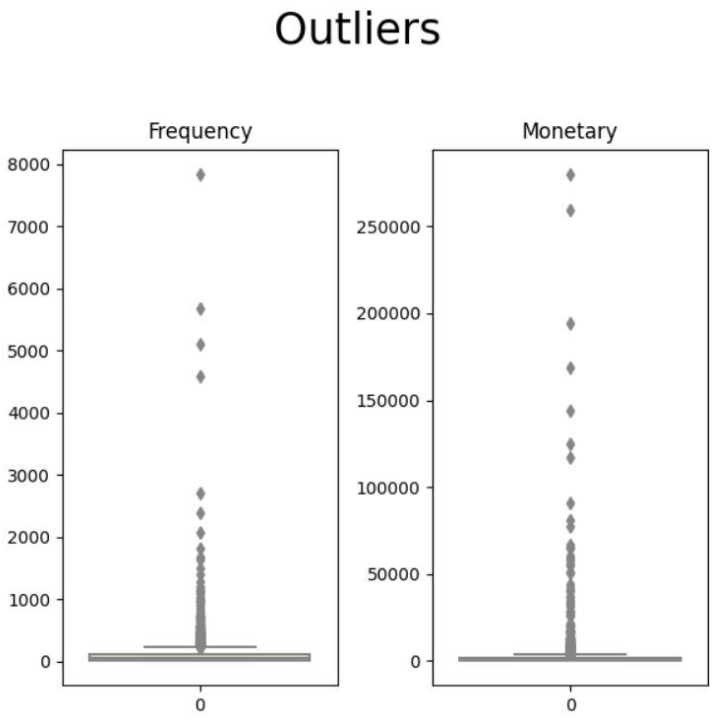
Followed by Isolation Forest, we then applied the Elbow method to calculate the optimum cluster count. The process consists of figuring out the Within-Cluster-Sum of Squared Errors (WSS) for various cluster counts (k) and identifying the k at which the WSS change begins to decline.

The elbow method's theory is that the explained variation changes quickly for a limited number of clusters before slowing down and forming an elbow in the curve. The number of clusters we can use for our clustering algorithm is known as the elbow point.

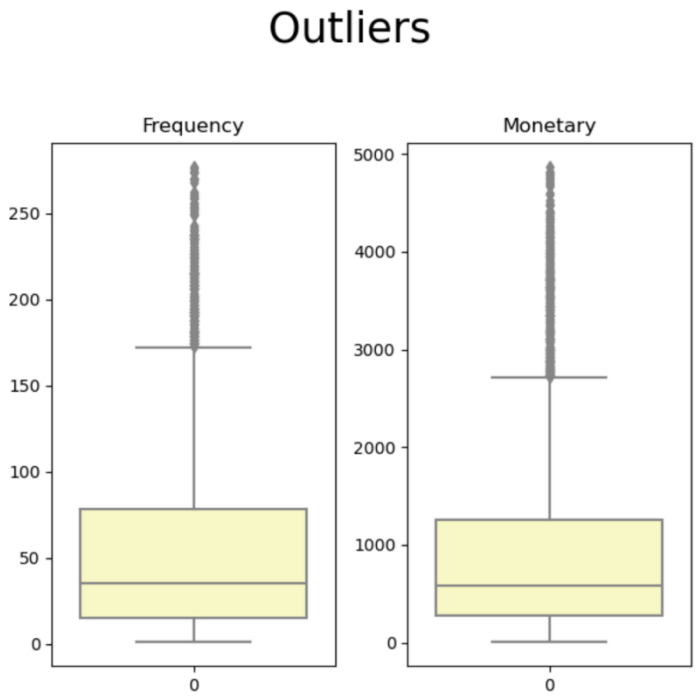
We performed clustering on two different sets of attributes, one on Frequency and Monetary and the second one on Recency and Monetary.

Frequency and Monetary:

Outliers before Isolation Forest:

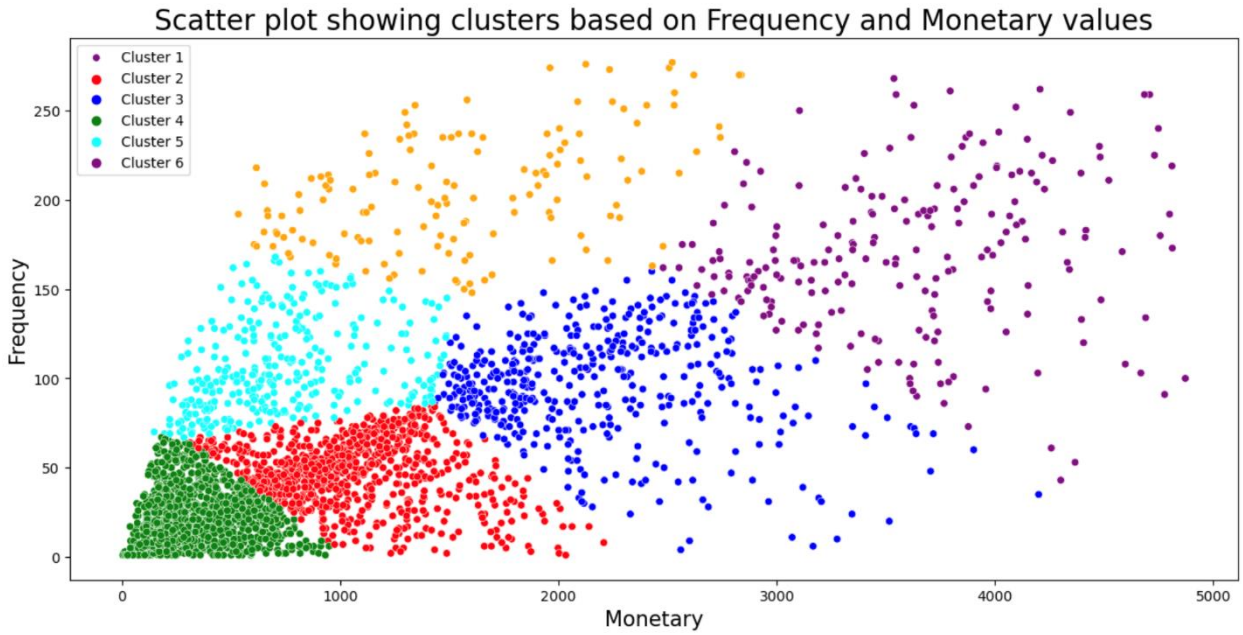


Outliers after applying Isolation Forest and eliminating the anomalies:



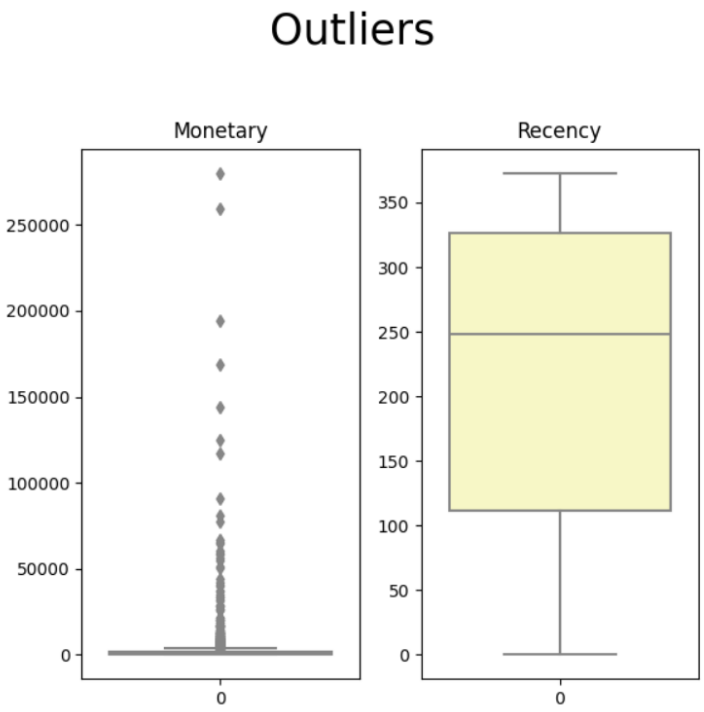
Plot showing Error score against various number of clusters

K-means clustering of customers based on Frequency and Monetary parameters

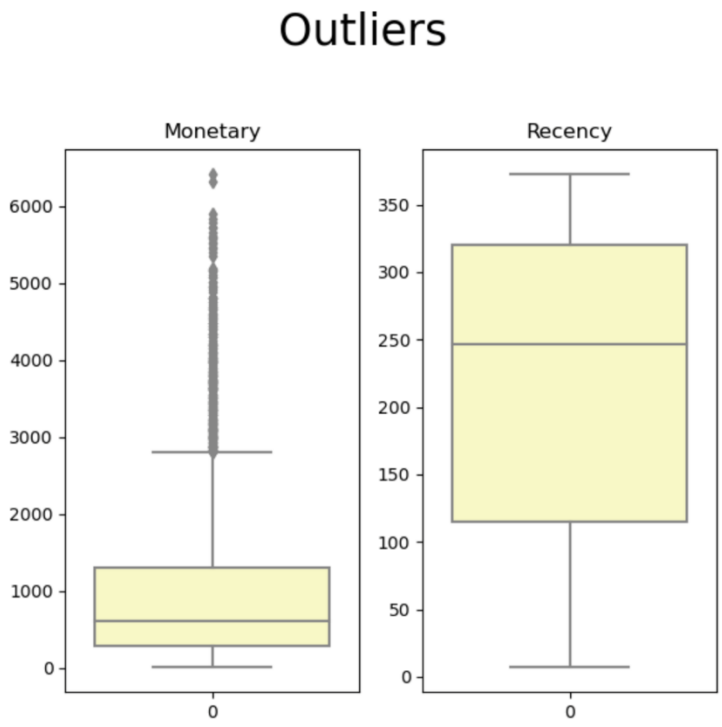


Monetary and Recency:

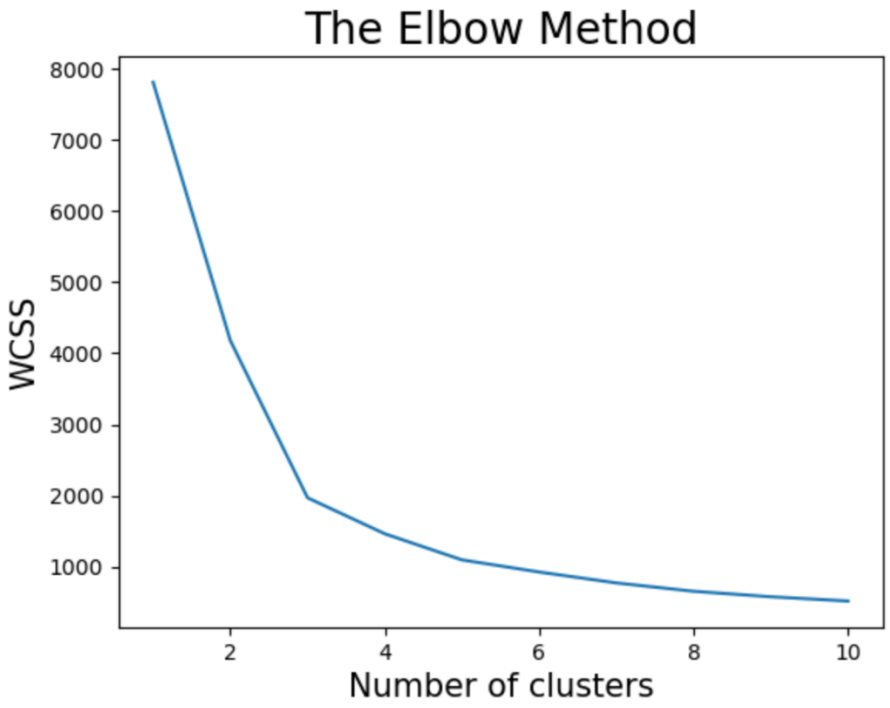
Outliers before Isolation Forest:



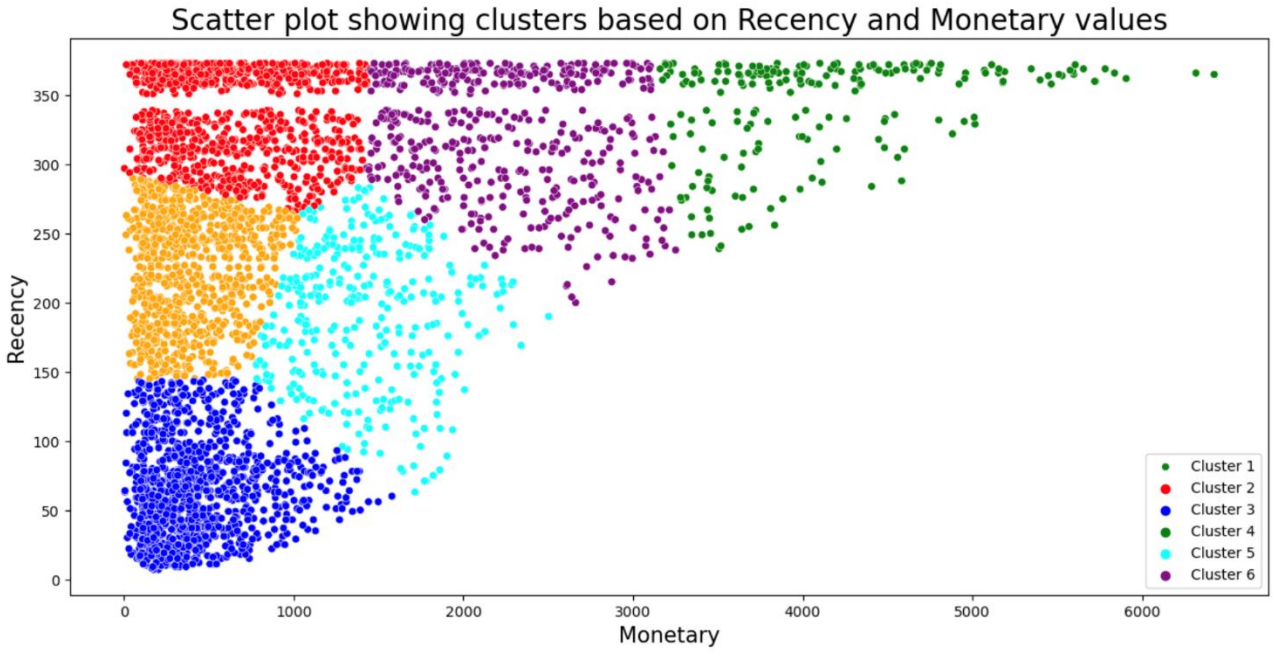
Outliers after applying Isolation Forest and eliminating the anomalies:



Plot showing Error score against various number of clusters



K-means clustering of customers based on Recency and Monetary parameters



Profitability Analysis

The profitability analysis helps the company understand the gains it is receiving from its sales relative to its expenses. Although we do not have data of the company's expenses, we are able to calculate that their revenue based solely on order price multiplied by quantity is roughly $9,747,748 after approximately $918,937 in returns and refunds.

We are also unable to calculate the profit margins of each product because we do not have the data that tells us the expenses that go into each product. To calculate the profit margin, we would subtract the item expense from its revenue. The top five products that generated the highest revenue are displayed in the chart below.

Top Five Products with Highest Revenue

	Description	Revenue
1074	DOTCOM POSTAGE	206245.48
2864	REGENCY CAKESTAND 3 TIER	164762.19
3859	WHITE HANGING HEART T-LIGHT HOLDER	99668.47
2422	PARTY BUNTING	98302.98
1825	JUMBO BAG RED RETROSPOT	92356.03

Customer Satisfaction

There is no data available on customer feedback or ratings for products or services therefore we are unable to analyze the sentiment or feedback trends.

Conclusion

The project emphasizes the value of RFM (Recency, Frequency, Monetary) analysis as an effective strategy for understanding and interacting with consumers in the ever-changing world of e-commerce. RFM analysis exceeds the levels of just data examination. It goes beyond in a way such that it behaves as a strategic model for businesses to adapt and enhance their operations in the world of online commerce. The insights gained can prove monumental to the way businesses operate as it enables them to adapt, strategize and position themselves to meet the demands of the varying customers and the ever-changing market.

Based on the data analysis we found that there are 4372 unique customers in the dataset with a churn rate of 34.42%. We found that most orders were placed on Thursday with the busiest time being between 12:00pm and 2:00pm. Due to the possibility of it being the holiday season, there is a seasonal trend of a spike in orders towards the end of the year. After geographically analyzing the dataset, the top five countries by the number of orders were UK, Germany, France Ireland and Belgium. The total revenue approximately totaled up to $9,747,748 after accounting for returns and refunds. About 19.22% of orders were returned with 'Manual' and 'Regency Cakestand 3 Tier' were concluded to be the product categories most associated with returns. Due to limitations in the dataset we were not able to analyze the payment trends and customer satisfaction aspects of the data.

From the analysis above a good way to battle the risk of churning could be to offer exclusive deals or personalized incentives to encourage repetition of purchases and also long-term engagement. The business also could capitalize on the time analysis insights to schedule marketing efforts during peak purchasing periods. A major emphasis should be placed on continuous monitoring and adaptation to the shifting customer behavior. Overall, the RFM analysis provides valuable insights into customer dynamics, product performance, time-based trends and geographical influences on the dataset.