RFM ANALYSIS

IE6400 FOUNDATIONS FOR DATA ANALYTICS

PROJECT 2 REPORT

GROUP NUMBER 26

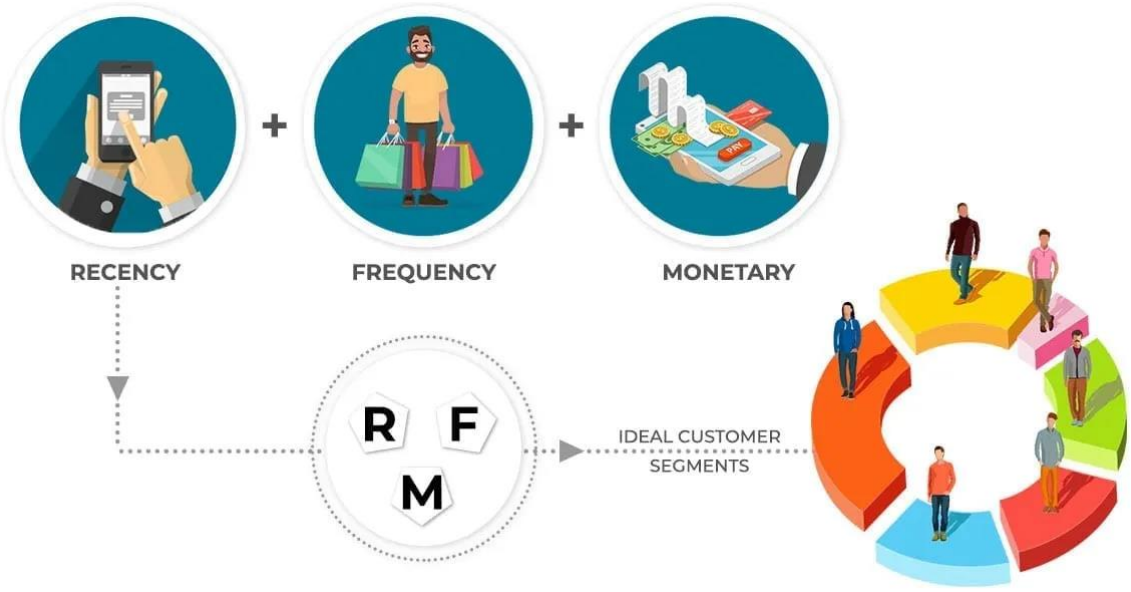
VIKRAMADITHYA PABBA (002853240)

DHEERAJ KUMAR GOLI (002897655)

LAAWANYAA SAI THOTA (002208176)

ROHAN REDDY PATHI (002832891)

UDHAY CHITYALA (002830533)



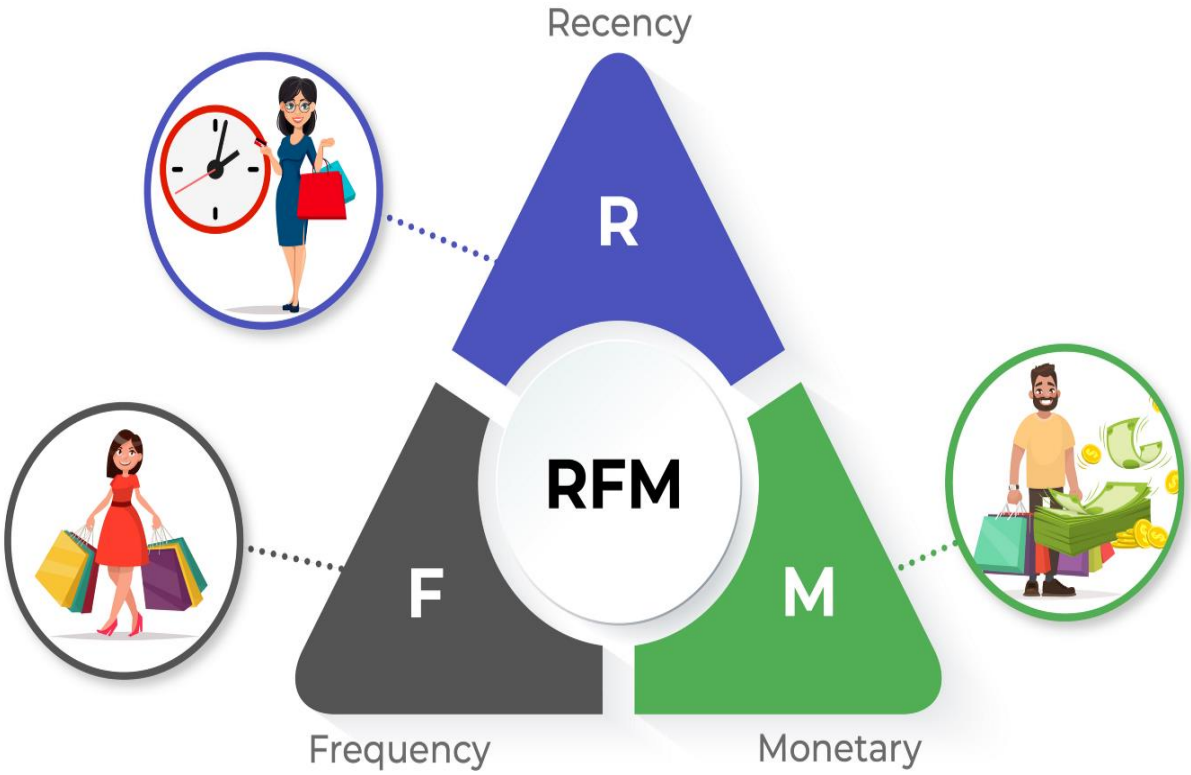
What is RFM Analysis?

RFM analysis is an analysis method that allows you to segment customers by frequency and number of purchases and identify those that bring more income.

RFM stands for **R**ecency, **F**requency, and **M**onetary value, each corresponding to some key customer trait.

- Recency, meaning a period of time during which your customers bought something from you. A high recency rate implies that the customer already has a good enough impression of your brand to have paid a visit to you recently.
- Frequency, meaning how often your customers buy from you. A high frequency indicates that the customer likes your brand, your products, and services, so he/she/they often come back to you.
- Monetary value, meaning total amount of purchases. A high level of this indicator implies that the client likes to spend money at your shop.

These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement.



RFM ANALYSIS

Data Cleaning:

The given dataset contains 541,909 rows and 8 columns.

The dataset contains missing values in two columns:

- Description: 1,454 missing values.
- CustomerID: 135,080 missing values.

Handling these missing values is crucial for accurate analysis. The approach to dealing with them depends on the context:

Description: Since this is a text field, the missing descriptions might not significantly impact the RFM analysis. However, they could be relevant for product-based insights. Since the number of rows without description is <0.3% (1,454 out of 541,909 rows) of the dataset we can remove these rows from the analysis as it might not cause significant impact.

CustomerID: This field is essential for RFM analysis since it identifies unique customers. Missing customer IDs could indicate guest purchases or data entry errors. We cannot tag an invoice to a customer in our case. So, we need to remove them from the analysis.

There are a few duplicate rows in the dataset. We have removed them from the analysis.

When calculating RFM (Recency, Frequency, Monetary) metrics, the inclusion of cancelled orders in the Frequency calculation depends on the specific context and objectives of the analysis. Here are some considerations:

Include Cancelled Orders:

Pros:

Provides a complete view of customer engagement, including both successful and unsuccessful transactions.

Can be useful for understanding customer behavior, including indecisiveness or issues leading to cancellations.

Cons:

Might inflate the Frequency metric with transactions that did not result in actual sales.

Can skew the analysis if a significant proportion of orders are cancelled.

Exclude Cancelled Orders:

Pros:

Focuses on successful transactions, which might be more relevant for understanding valuable customer behavior.

Results in a cleaner, more sales-focused analysis.

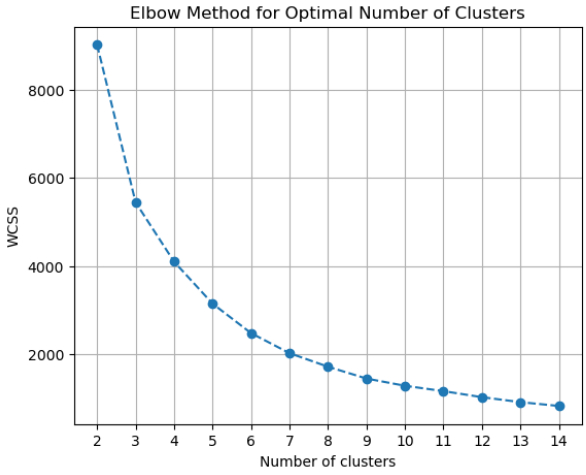
Cons:

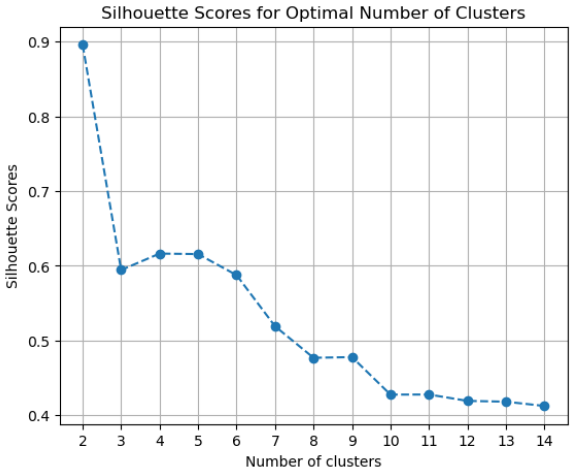
Potentially overlooks part of the customer's interaction history.

For most business analyses, excluding cancelled orders is preferred because it provides a clearer picture of actual sales activity and customer value. Hence, we plan to exclude the return orders from the analysis (quantity <0)

After data cleaning we have 392,732 rows and 8 columns

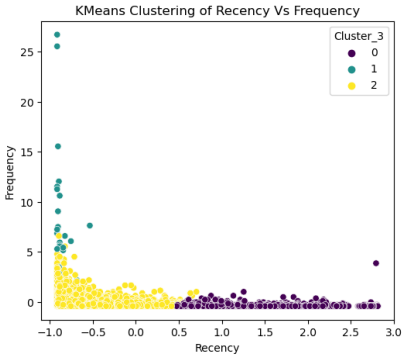
Using Elbow Method and Silhouette Scores to determine the optimal number of clusters.

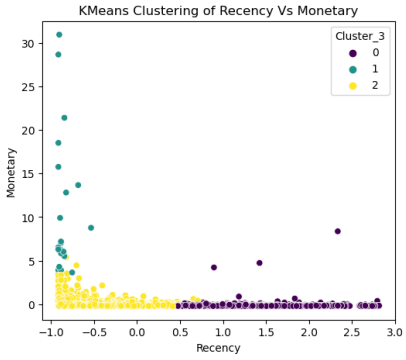


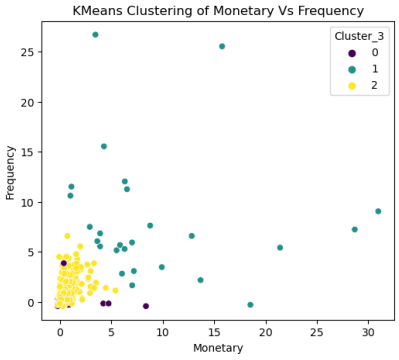


Based on the above two methods i.e., Elbow and Silhouette methods, the optimal number of clusters to be considered for this analysis is 3.

Here is the scatter plot displaying the scaled RFM data, specifically focusing on Recency and Frequency dimensions, with the hue parameter indicating the cluster each point belongs to.







This visualization helps in understanding how the customers are distributed across different clusters based on their recent engagement and frequency of purchases.

The profile for each customer segment, based on the mean values of their Recency, Frequency, and Monetary metrics, is as follows:

Cluster 0:

Recency: Customers in this cluster have not made a purchase in a while, with an average recency of 247 days.

Frequency: They have a lower frequency of orders, averaging around 1.58.

Monetary: Their average total spending is relatively low, around 630.

Cluster 1:

Recency: This group consists of very recent customers, with an average last purchase made just 6 days ago.

Frequency: They are the most frequent shoppers, with a very high average order count of 66.5.

Monetary: They also spend the most, with an average total spending of a staggering 85,826.

Cluster 2:

Recency: On average, customers in this cluster made their last purchase around 41 days ago.

Frequency: They make purchases relatively frequently, averaging about 4.67 orders.

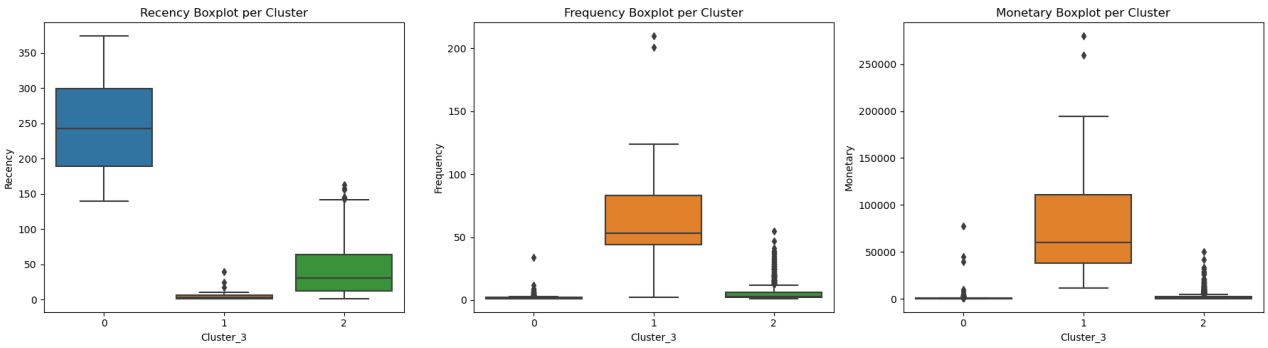
Monetary: The average total spending of these customers is approximately 1,849.

Segment Profiling Summary:

Cluster 0 includes customers who have not shopped recently, are infrequent buyers, and spend less.

Cluster 1 consists of highly engaged and high-spending customers who have made purchases very recently and do so frequently.

Cluster 2 represents a segment of regular and moderately recent customers with moderate spending.



These boxplots display the distribution of Recency, Frequency, and Monetary values within each cluster. The spread and outliers in each metric can be observed, indicating the variability within each segment.

Recency Boxplot:

- This shows how recently customers in each cluster have made a purchase.
- A smaller median (the line in the middle of the box) indicates more recent activity.
- A larger box or longer whiskers indicate greater variation in the recency of purchases within the cluster.
- Outliers, if any, are shown as points beyond the whiskers and represent customers whose purchasing behavior is significantly different from the majority in the cluster.

Frequency Boxplot:

- This illustrates how often customers in each cluster make purchases.
- A higher median suggests that customers in that cluster purchase more frequently.
- A wider box or longer whiskers indicate a greater variety in purchasing frequency.
- Outliers here would represent customers who are either unusually frequent or infrequent shoppers compared to their cluster.

Monetary Boxplot:

- This plot indicates the distribution of the total amount spent by customers in each cluster.
- A higher median shows a higher average spend in the cluster.
- The range of the box and the length of the whiskers reveal the spread of spending within the cluster, with a wider range indicating more variability.
- Outliers in this context would be customers who spend significantly more or less than others in their cluster.

6.Market Recommendations

Cluster 0

Characteristics: Regular and moderately recent customers with moderate spending.

Marketing Strategies:

Loyalty Programs: Encourage repeat purchases through loyalty rewards or points.

Targeted Promotions: Send promotions on products related to their past purchases.

Engagement Campaigns: Use email or social media campaigns to keep the brand top-of-mind.

Feedback and Surveys: Engage them in product feedback to enhance their sense of involvement.

Cluster 1

Characteristics: Customers who have not shopped recently are infrequent buyers, and spend less.

Marketing Strategies:

Reactivation Campaigns: Send "We miss you" messages with special discounts to re-engage.

Personalized Offers: Make them feel special with personalized offers based on past purchases or browsing history.

Market Research: Understand their lack of engagement - price, product range, or service issues.

Adjustment in Product/Service Offerings: Consider diversifying the product range or adjusting pricing strategies to meet their needs.

Cluster 2

Characteristics: Highly engaged, high-spending customers who have made purchases very recently.

Marketing Strategies:

Exclusive Membership: Offer them an exclusive membership or VIP status with special benefits.

Upselling and Cross-Selling: Recommend premium products or complementary items to their usual purchases.

Early Access to New Products: Give them early access to new products or sales events.

Personalized Communication: Maintain regular, personalized communication and recognize their value to your business.

General Strategies

Data Analysis: Continuously analyze customer data to stay updated on changing behaviors and preferences.

Omnichannel Presence: Ensure a consistent, high-quality experience across all channels - online, offline, social media.

Customer Service Excellence: Provide exceptional customer service to enhance satisfaction and loyalty.

Further refining the segmentation criteria for a more balanced approach

We have successfully categorized customers into different segments. Here's the distribution of customers in each segment:

At Risk: 480 customers

Best Customers: 1809 customers

Big Spenders: 360 customers

Lost Cheap Customers: 1330 customers

Loyal Customers: 360 customers

These segments now offer a more balanced view of the customer base, allowing for targeted marketing strategies:

At Risk: Focus on re-engagement campaigns and personalized offers.

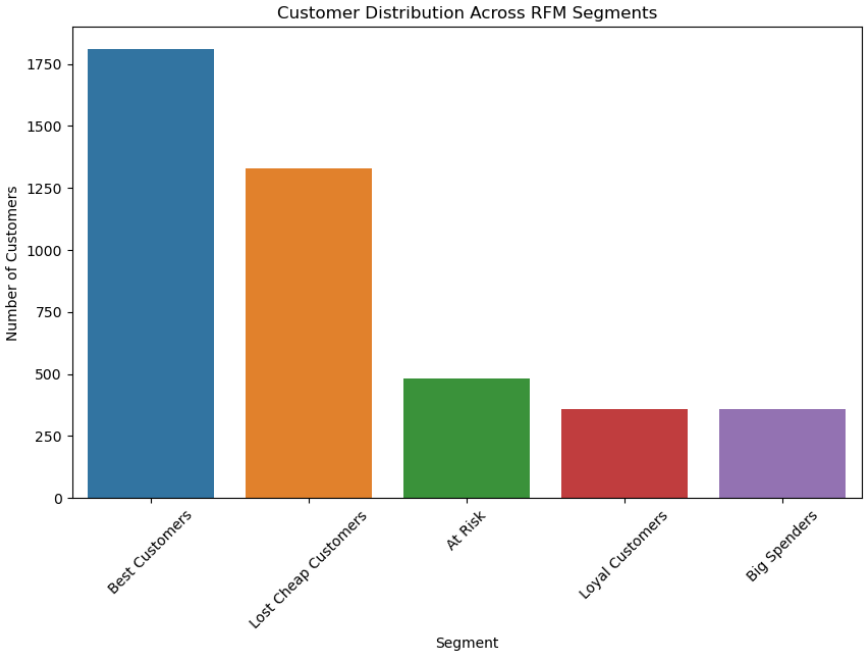
Best Customers: Prioritize loyalty programs and exclusive benefits.

Big Spenders: Target with upselling opportunities and premium offers.

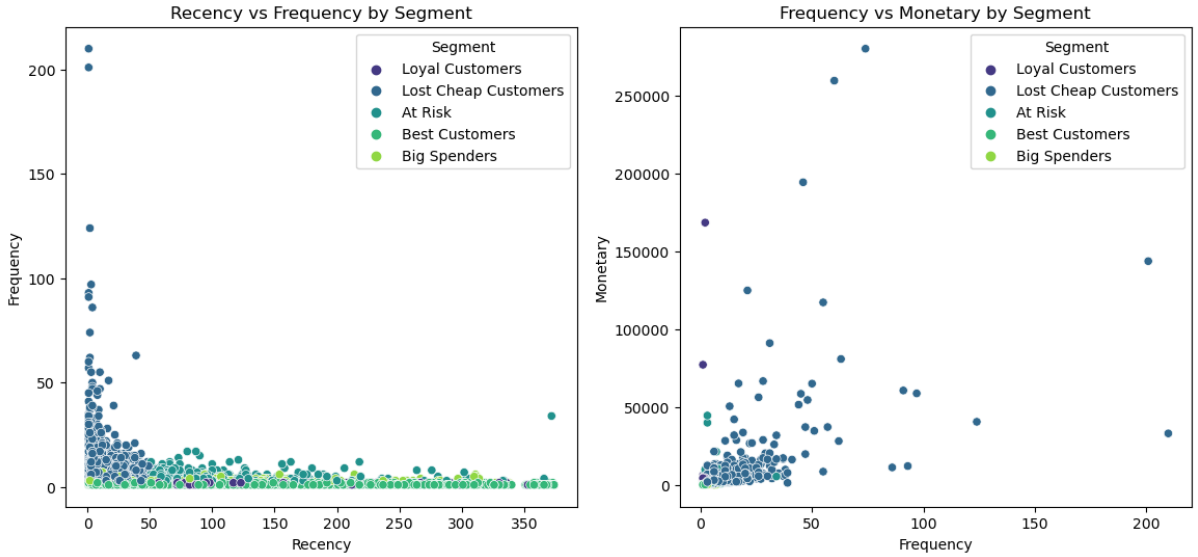
Lost Cheap Customers: Implement occasional reactivation tactics.

Loyal Customers: Continue to nurture with rewards for frequent purchases.

Here are the visualizations illustrating the RFM distribution and the clusters formed:



- This chart shows the number of customers in each RFM segment, providing a clear view of how customers are distributed across different categories.
- The largest segment is 'At Risk', indicating a significant number of customers have not made recent purchases, which might be a concern for customer retention.
- 'Best Customers' and 'Big Spenders' segments also have a substantial number of customers, suggesting a strong base of loyal and high-spending customers.
- The 'Loyal Customers' segment is relatively small, suggesting a potential area to increase customer loyalty initiatives.



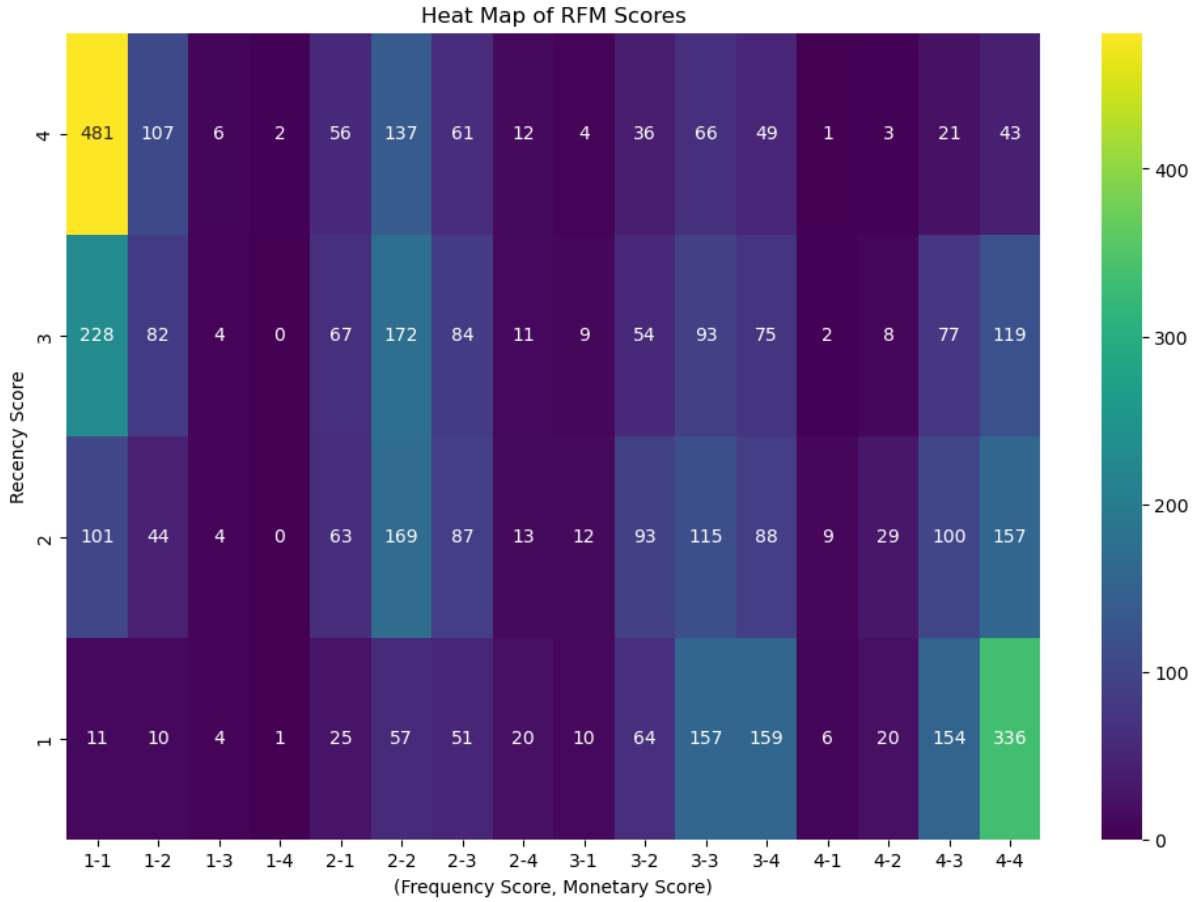
Recency vs Frequency:

- Customers in the 'Best Customers' segment tend to cluster towards lower recency and higher frequency, indicating they are recent and frequent shoppers.

- 'At Risk' customers are spread across high recency values, showing their lack of recent engagement.
- The spread of data points suggests varying degrees of engagement across segments.

Frequency vs Monetary:

- 'Big Spenders' and 'Best Customers' show higher frequency and monetary values, indicating they are valuable customers with frequent and high-value purchases.
- The variation in monetary value across different frequencies indicates diverse spending behaviors even among similarly frequent shoppers.



- This heat map presents the concentration of customers across different combinations of RFM scores. The intensity of the color indicates the number of customers in each combination of Recency, Frequency, and Monetary scores.
- The heat map displays the distribution of customers across combinations of Recency, Frequency, and Monetary scores.
- High concentration areas (warmer colors) indicate common customer profiles, while cooler colors represent less common profiles.
- The heatmap can be particularly useful in identifying common customer behaviors and trends, as well as rare but potentially significant customer profiles

Conclusions and Recommendations:

- Focus on 'At Risk' Customers: Develop re-engagement strategies like personalized emails or special offers to bring these customers back.
- Nurture 'Best Customers' and 'Big Spenders': These segments are crucial for revenue. Offer loyalty programs, exclusive deals, and premium services to retain them.
- Expand the 'Loyal Customers' Base: Since this segment is small, consider strategies to increase customer loyalty among other segments.
- Tailored Marketing Strategies: Use the insights from the scatter plots and heat map to develop targeted marketing campaigns based on customer purchasing behavior and value to the business

Solutions to problem statement

1. Data Overview

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

  The given dataset contains 541,909 rows and 8 columns.

  Data Cleaning:

  The dataset contains missing values in two columns:

  - Description: 1,454 missing values.
  - CustomerID: 135,080 missing values.

  Handling these missing values is crucial for accurate analysis. The approach to dealing with them depends on the context:

  Description: Since this is a text field, the missing descriptions might not significantly impact the RFM analysis. However, they could be relevant for product-based insights. Since the number of rows without description is <0.3% (1,454 out of 541,909 rows) of the dataset we can remove these rows from the analysis as it might not cause significant impact.

  CustomerID: This field is essential for RFM analysis since it identifies unique customers. Missing customer IDs could indicate guest purchases or data entry errors. We cannot tag an invoice to a customer in our case. So, we need to remove them from the analysis.

  There are a few duplicate rows in the dataset. We have removed them from the analysis.

  After data cleaning we have 392,732 rows and 8 columns

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

  InvoiceNo: The invoice number for each transaction

  StockCode: Code for each item

  Description: Description of the item

Quantity: The quantity of each item purchased

InvoiceDate: The date and time of the transaction

UnitPrice: Price per unit of the item

CustomerID: ID of the customer

Country: Country of the customer

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

  After cleaning the dataset, we have the data from 01 December,2010 to 09 December,2011.

2. Customer Analysis

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

  There are 4,339 unique customers in the dataset.

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

  We have only considered the orders which have a positive value for the “Quantity” column. This would ensure that there are no returns and refunds considered while finding the number of orders per customer.

  Based on the above assumption,

  On an Average, a customer placed approximately 4 orders in the given time frame:

```
Order per Customer:

CustomerID
12346     1
12347     7
12348     4
12349     1
12350     1
         ..
18280     1
18281     1
18282     2
18283    16
18287     3
Name: InvoiceNo, Length: 4339, dtype: int64

count    4339.000000
mean        4.271952
std         7.705493
min         1.000000
25%         1.000000
50%         2.000000
75%         5.000000
max       210.000000
Name: InvoiceNo, dtype: float64
```

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

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

CustomerID
12748    210
14911    201
17841    124
13089     97
14606     93
```

3. Product Analysis

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

  Below are the top 10 frequently purchased products and number of times they have been purchased during the given timeframe:

```
WHITE HANGING HEART T-LIGHT HOLDER    2016
REGENCY CAKESTAND 3 TIER              1714
JUMBO BAG RED RETROSPOT               1615
ASSORTED COLOUR BIRD ORNAMENT         1395
PARTY BUNTING                         1390
LUNCH BAG RED RETROSPOT               1303
SET OF 3 CAKE TINS PANTRY DESIGN      1152
POSTAGE                               1099
LUNCH BAG  BLACK SKULL.               1078
PACK OF 72 RETROSPOT CAKE CASES       1050
```

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

  Weighted average price of each product is as below:

```
Description
 4 PURPLE FLOCK DINNER CANDLES       1.925072
 50'S CHRISTMAS GIFT BAG LARGE       1.205438
 DOLLY GIRL BEAKER                   1.150585
 I LOVE LONDON MINI BACKPACK         4.038579
 I LOVE LONDON MINI RUCKSACK         4.150000
                                      ...
ZINC T-LIGHT HOLDER STARS SMALL      0.792803
ZINC TOP  2 DOOR WOODEN SHELF       16.950000
ZINC WILLIE WINKIE  CANDLE STICK     0.835035
ZINC WIRE KITCHEN ORGANISER          6.272000
ZINC WIRE SWEETHEART LETTER TRAY     3.165500
```

  Weighted average price is calculated as sum(revenue)/sum(quantity) for each product

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

  The product that generates the highest revenue is ‘Regency Cakestand 3 tier’. The total revenue generated by the product is 132567.7 currency units.

4. Time Analysis

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

  The day when most orders are placed is Wednesday and the number of orders placed is 4033. The hour of the day when most orders are placed is at 12 Noon.

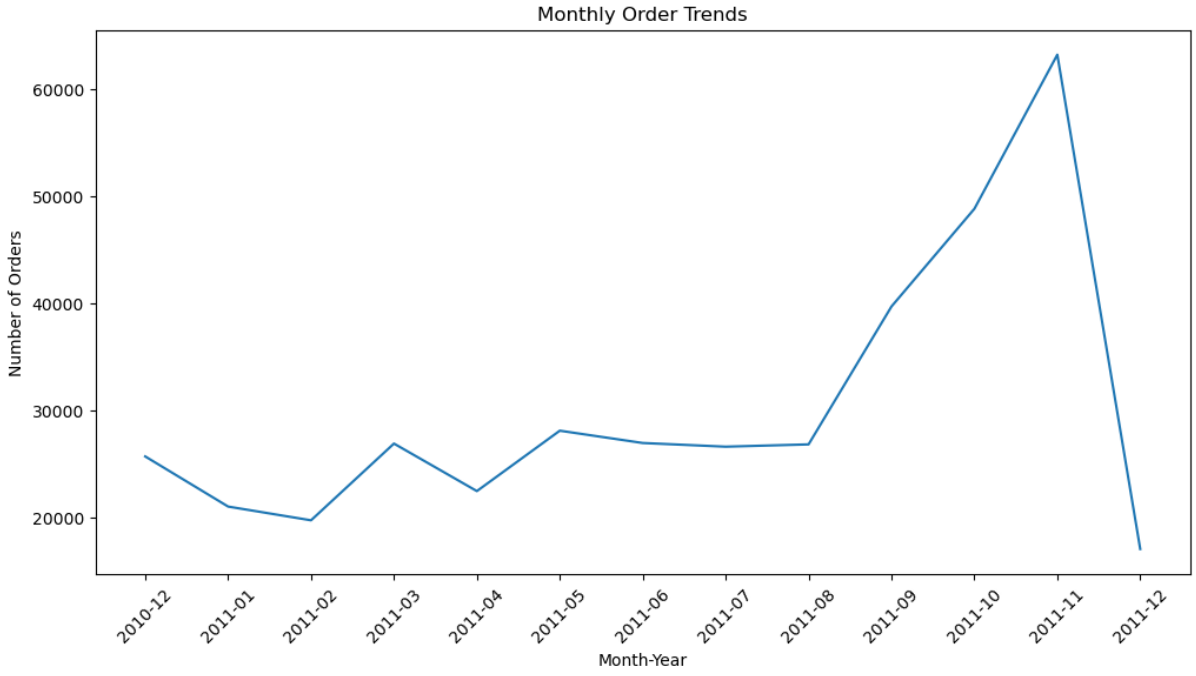
  This tells us that there are a greater number of customers who want to buy products on a Wednesday. Further, there are more customers buying products at 12’O Clock in the noon.

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

  We cannot find the average order processing time with the given dataset as we do not have the order date and delivered data.

  If we have these columns in the data, we can calculate the average processing time by calculating the average of (difference between order date and delivered date for each invoice)

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



On the overall there seems to be an increase in sales from August to November. However, we need data for more years to determine any seasonality.

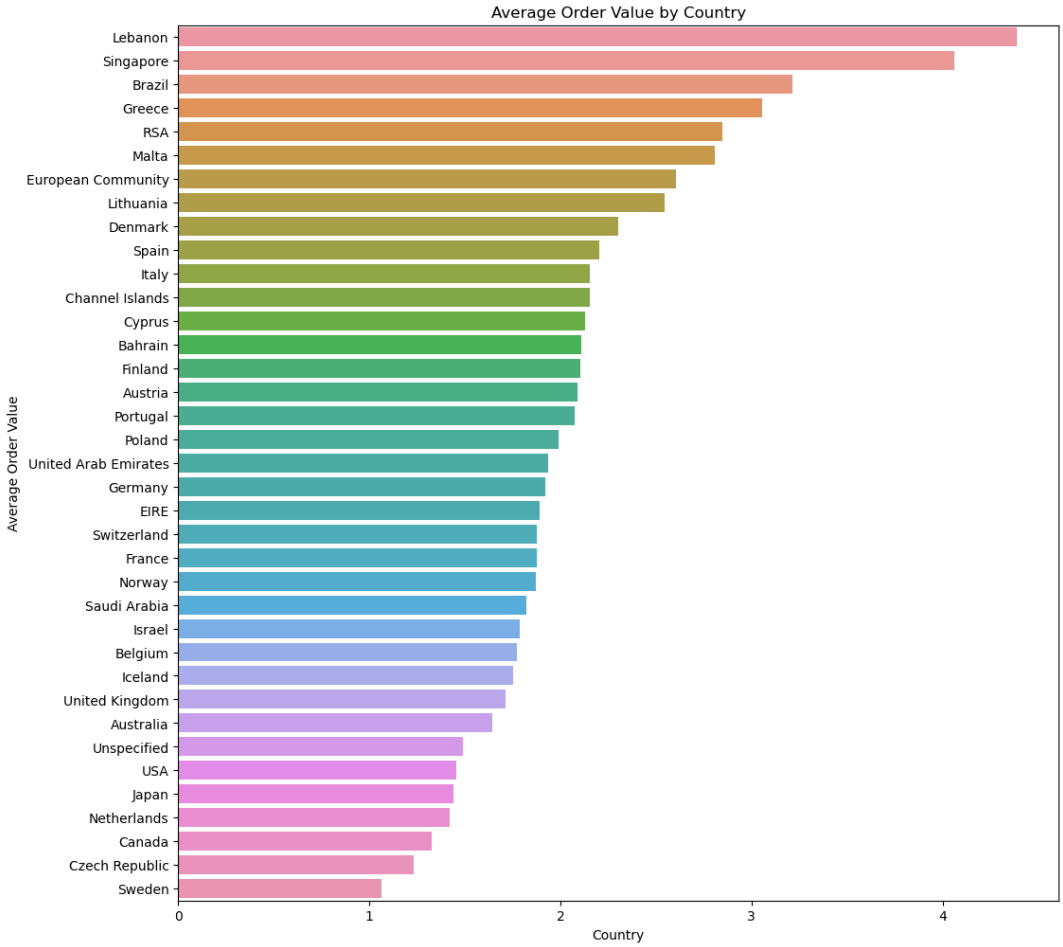
5. Geographical Analysis

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

  United Kingdom has the highest number of orders (16649 orders) followed by Germany (457), France (389), EIRE (260) and Belgium (98)

```
Country
United Kingdom    16649
Germany             457
France              389
EIRE                260
Belgium              98
```

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



Correlation

Certain countries, like the Lebanon,Singapore and Brazil have higher average order values compared to others.

There is a wide range of average order values, from over 4.4 in the Lebanon to under 1.06 in the Sweden.

The United Kingdom, which may be where the majority of customers are based (assuming from a common e-commerce dataset pattern), has a relatively lower average order value compared to the top countries listed

6. Payment Analysis

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

  We do not have the payment details in the given data set.

  If we have the payment details, we can find the most common payment method by calculating the number of unique invoices for each payment method.

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

  We do not have the payment details in the given data set.

  If we have the payment details, we can for any correlation between payment method and order amount

7. Customer Behavior

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

  On an average, customers remain active for 130 days (duration between their first and last purchase)

```
Average duration customers remain active: 130 days 17:48:44.194514866
```

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

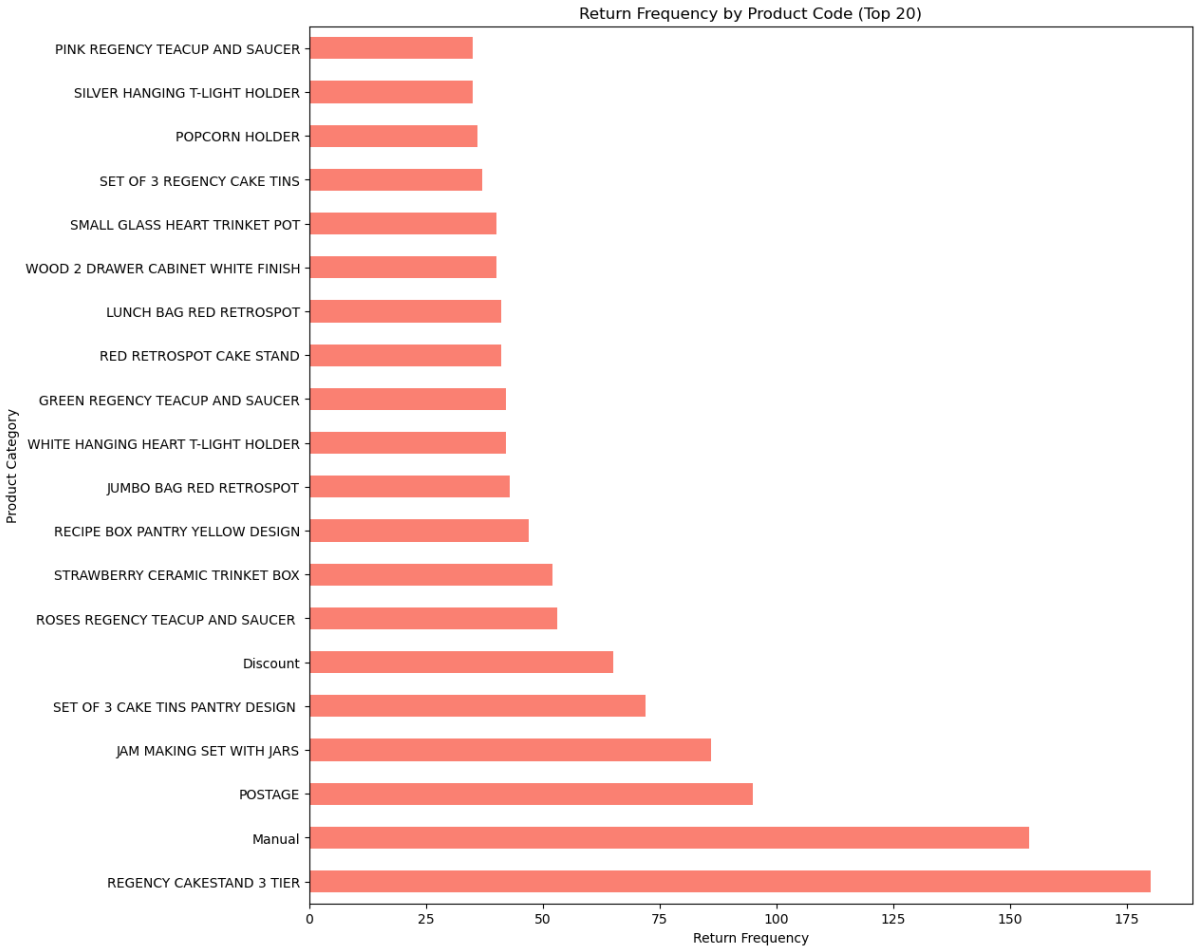
  Yes, There are customer segments based on their purchase behaviour. The segments are clearly explained in the RFM Analysis section

8. Returns and Refunds

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

  Percentage of orders with returns or refunds: 16.47%

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



9. Profitability Analysis

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

  We do not have the cost price of the products in the given dataset. We only have the selling price of the product. Hence, we cannot find the profit generated by the company.

  If we have the cost price of every product, we can calculate the profit of each item by creating a new column (profit) using the formula (cost price – selling price) *quantity. This gives the profit at a row level. We can then aggregate it (sum(profit)) to get the total profit generated by the company during the time-period of the dataset

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

  As discussed, we do not have the required data to get the top 5 products with highest profit margins

If we have the cost price of every product, we can get the profit margin of each product by creating a new column (revenue) using the formula quantity*selling price

We can then group by product and aggregate it using the formula sum(profit)/sum(revenue) to get the average profit margin of each product throughout the timeframe of the dataset.

10. Customer Satisfaction

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

  We don't have any data regarding customer feedback/ratings

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

  Analyzing sentiment or feedback trends typically involves using natural language processing (NLP) techniques to understand the sentiment expressed in customer reviews or feedback. Unfortunately, sentiment analysis requires textual data associated with customer feedback, and the provided dataset does not include a column specifically for customer comments or reviews.

  If we have a column containing customer feedback or comments, you can use NLP tools and techniques to analyze sentiment.