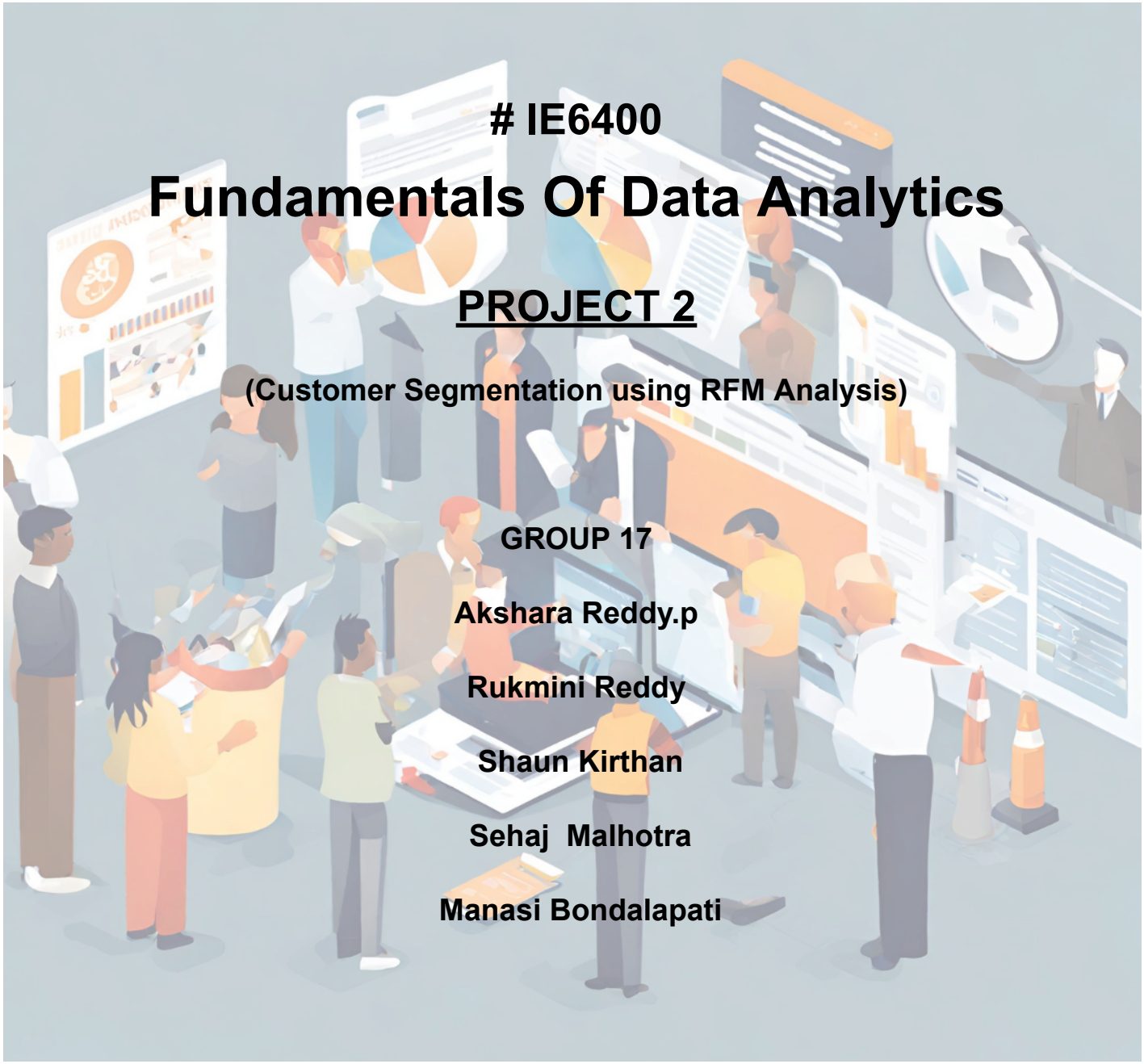
Northeastern University



IE6400

Fundamentals Of Data Analytics

PROJECT 2

(Customer Segmentation using RFM Analysis)

GROUP 17

Akshara Reddy.p

Rukmini Reddy

Shaun Kirthan

Sehaj Malhotra

Manasi Bondalapati

College of Engineering, Northeastern University

Boston, MA 02115

Introduction

The aim of this Project is to work on a E-Commerce dataset provided to create a Customer Segmentation model using the RFM analysis method (Recency,Frequency,Monetary) ,which is a powerful technique used by business to grow customers based on their recent purchasing behavior, purchase frequency and monetary value enabling more targeted marketing and customer engagement strategies.

The project initiation involved the importation of the dataset and a comprehensive exploration of its contents to gain a thorough understanding of the data. Subsequently, we delved into the crucial phase of data preprocessing, where essential steps, including data cleaning, handling missing values, and adjusting data types, were meticulously performed.

Following the data preprocessing stage, we seamlessly transitioned into RFM calculations and segmentation. This involved the meticulous computation of RFM metrics for each customer, focusing on evaluating recency, frequency, and monetary aspects of their purchasing behavior. The subsequent task entailed the assignment of RFM scores to individual customers based on either quartiles or custom-defined bins. Ultimately, a unified RFM score was derived for each customer, paving the way for further analysis.

Following the completion of RFM calculations and segmentation, the project proceeds to customer segmentation and segment profiling. In this phase, we leverage clustering techniques, with a specific focus on K-Means clustering, to effectively segment customers based on their RFM scores. An essential part of this process involves experimenting with various cluster numbers to discern and identify meaningful segments among the customer base.Subsequently, detailed analysis and profiling are conducted for each customer segment. This entails a thorough examination of the characteristics of each segment, including their unique RFM scores and any other relevant attributes that contribute to a comprehensive understanding of their behavior and preferences.

Moving forward, the subsequent phase involves presenting actionable marketing recommendations tailored to each customer segment. This entails a comprehensive exploration of strategies aimed at improving customer retention and maximizing revenue for each distinct group identified through segmentation.

Following the strategic marketing recommendations, the project concludes with the creation of visualizations. This includes the development of visual representations such as bar charts, scatter plots, or heat maps, designed to vividly illustrate the distribution of RFM scores and highlight the clusters formed during the segmentation process.

TASKS

1)Data pre-processing

The data preprocessing phase commenced with an initial examination of null values in the columns. This scrutiny revealed the presence of null values in two columns: Description and Customer ID. To address this issue, a decision was made to resolve it by employing the strategy of dropping the rows containing null values in these specific columns. This approach ensures a cleaner dataset, devoid of instances where essential information is missing, thereby facilitating subsequent analyses and modeling.

Following the resolution of null values, the next step in the data preprocessing involved a comprehensive examination of the data types across all columns. Notably, it was identified that the "Invoice Date" column was initially classified as an object type. To enhance the accuracy and facilitate temporal analyses, a pivotal decision was made to transform the data type of the "Invoice Date" column from object to datetime. This adjustment ensures that the date-related information is appropriately represented, allowing for more precise temporal analyses and computations in subsequent stages of the project.

2)RFM Calculations

We calculate the recency for each customer in the dataset. It does so by finding the number of days between the latest invoice date for each customer and a specified reference date (December 10, 2011) which is the date after the time period seen in the dataset. The calculated recency values are then added back to the original dataset, allowing for an analysis of how recently each customer made a purchase.

Next, the calculation of frequency for each customer is done. This is achieved by grouping the data based on the "CustomerID" and counting the occurrences of "InvoiceNo" for each customer. Essentially, this step yields the number of purchases (frequency) made by each customer. The calculated frequency values are then merged back into the original dataset, providing a comprehensive view of how often each customer has engaged in purchases.

In the concluding step of our analysis, we derive the monetary value for each customer. This is achieved by introducing a new column named "Total order value," formed through the multiplication of the "UnitPrice" and "Quantity" columns. Subsequently, the dataset undergoes a grouping operation based on "CustomerID," and the sum of the "Total order value" is computed for each customer. This summation represents the total amount spent by individual customers, effectively quantifying their monetary contribution. The resultant values, indicative of each customer's overall expenditure, are then incorporated back into the original dataset under the column heading "Monetary."

This gives us the RFM values which are going to play a major role in our RFM Segmentation later.In the RFM calculation we have achieved a comprehensive understanding of customer behavior based on recency, frequency, and monetary factors.

3)RFM Segmentation

Recency scores were assigned on a scale from 5 to 1, with a higher score denoting a more recent purchase. This implies that customers who have made purchases more recently are allocated higher recency scores.
Similarly, frequency scores were assigned on a scale from 1 to 5, with a higher score indicating a greater purchase frequency. Customers who made more frequent purchases received higher frequency scores.
For monetary scores, a scale from 1 to 5 was employed, with a higher score signifying a greater amount spent by the customer. This approach ensures that customers who have contributed higher monetary value to the business are assigned higher monetary scores.

In computing RFM scores, we applied the pd.cut() function to segment recency, frequency, and monetary values into distinct bins. Each of these values was divided into four bins, and corresponding scores were assigned to each bin based on defined criteria.

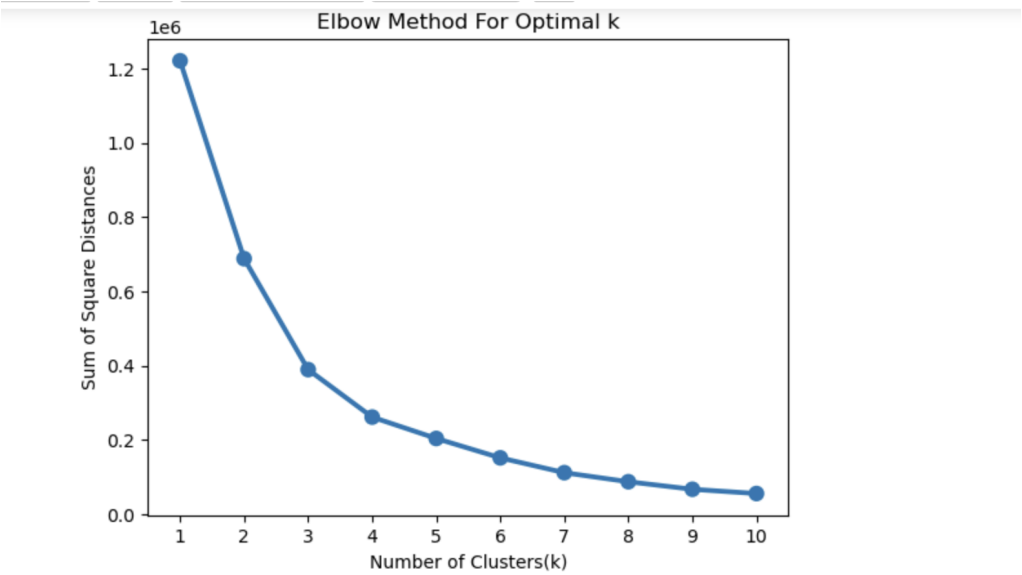
In the code segment, the 'RecencyScore,' 'FrequencyScore,' and 'MonetaryScore' columns in the dataset are converted to integer data types, enhancing their utility for numerical calculations. Subsequently, a new column named 'RFM_Score' is created by summing the individual scores for recency, frequency, and monetary values. This collective RFM score serves as a comprehensive metric, consolidating information on customer recency, purchase frequency, and monetary contribution. The conversion to integer types ensures that these scores can be effectively utilized in subsequent analyses.

4)Customer Segmentation

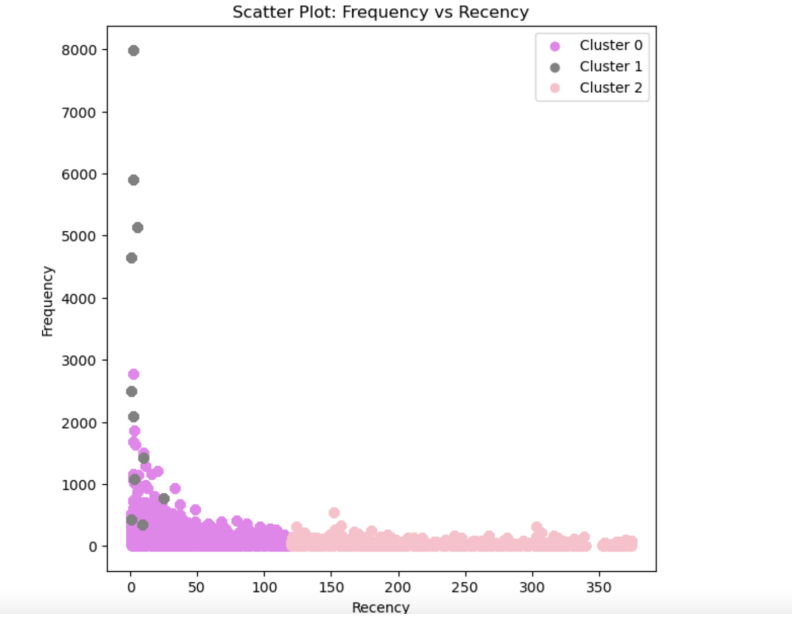
The initial phase involves standard scaling, wherein the 'Recency,' 'Frequency,' and 'Monetary' columns of the dataset undergo standardization using the StandardScaler. This crucial preprocessing step aims to achieve a mean of 0 and a standard deviation of 1 for the features. By doing so, it prevents features with larger scales from disproportionately influencing the subsequent K-Means clustering process, ensuring fair and meaningful comparisons between different variables.

Subsequently, the Elbow Method is implemented to ascertain the optimal number of clusters (k) for the K-Means algorithm. A systematic loop is employed to fit K-Means models for varying values of k, ranging from 1 to 10. Within each iteration, the KMeans algorithm is applied with k clusters, utilizing the 'k-means++' initialization strategy, allowing a maximum of 300 iterations,

and performing 10 initializations. The key metric employed to assess the performance of each model is the sum of squared distances from each data point to its assigned center, commonly known as inertia. The results are systematically recorded in the `sum_of_sq_dist` dictionary, facilitating the subsequent identification of the "elbow" point in the graphical representation. This method aids in determining the optimal number of clusters, striking a balance between capturing meaningful patterns in the data and avoiding overfitting.



In this code segment, the K-Means clustering algorithm is employed to group customers into three clusters based on their scaled 'Recency' and 'Frequency' values. The choice of three clusters aligns with the optimal number determined through the earlier Elbow Method analysis. The resulting clusters are visualized using a scatter plot, where each cluster is distinctly color-coded. The x-axis represents the 'Recency' of customer purchases, the y-axis represents the 'Frequency' of purchases, and each point on the plot corresponds to a customer. The plot offers a clear depiction of how customers are distributed among the clusters, providing valuable insights into distinct patterns of recency and frequency behaviors within the dataset. The visual representation enhances our understanding of customer segmentation and sets the stage for further analysis and targeted marketing strategies tailored to each cluster.



5)Segment profiling

Segment profiling is a strategic process that involves analyzing and describing specific customer segments based on various characteristics, behaviors, and attributes. The goal of segment profiling is to gain a deeper understanding of each segment's unique features, enabling businesses to tailor their marketing strategies, product offerings, and customer engagement approaches to better meet the needs and preferences of each group

The provided code initiates a comprehensive analysis of customer behavior and segmentation based on RFM (Recency, Frequency, Monetary) metrics. The first set of visualizations employs bar charts to compare the average values of Recency, Frequency, and Monetary metrics among different customer clusters. This enables a clear understanding of how these key metrics vary across identified clusters.

Subsequently, a second set of bar charts is generated to showcase the average RFM scores—RecencyScore, FrequencyScore, and MonetaryScore—across the clusters. This visual representation provides insights into the overall scoring patterns within each cluster.

Following the metric comparisons, the code calculates the Customer Lifetime Value (CLV) using a straightforward formula that takes into account the average Monetary, Frequency, and Recency metrics. This CLV estimation offers a projection of the average value a customer contributes to the business over a specific duration.

Moreover, the code undertakes customer segmentation based on RFM scores, categorizing customers into distinct segments such as "Most_valued_customers," "Loyal Customers," and others. This segmentation strategy facilitates targeted marketing and engagement approaches for each identified customer group.

Finally, the code filters the dataset to specifically examine customers falling into the "Most_valued_customers" segment, allowing for a focused analysis of this high-value customer group. Overall, the code presents a holistic approach to customer analysis, segmentation, and strategic decision-making, leveraging RFM metrics and scores for actionable insights.

6)Marketing Recommendations

The recommended strategies are tailored to three distinct customer clusters identified through RFM analysis:
Cluster 0 - Potential Loyalists:

- Introduce exclusive loyalty programs or VIP memberships.
- Implement personalized communication strategies based on preferences.
- Create incentives for repeat purchases to highlight long-term value.

Cluster 1 - Recent Customers:

- Encourage social media engagement for updates and promotions.
- Implement re-engagement campaigns with time-sensitive discounts.
- Share relevant content beyond transactions to keep recent customers engaged.

Cluster 2 - Loyal Customers:

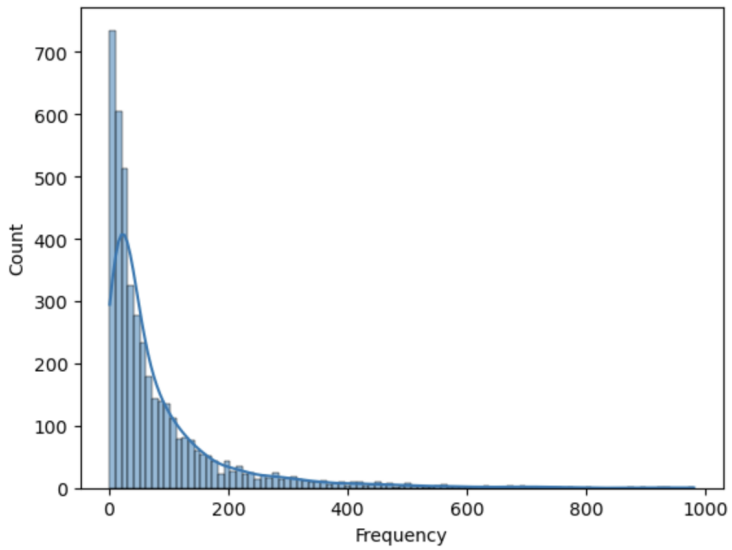
- Recommend complementary products or upgrades based on purchase history.
- Introduce VIP memberships with additional benefits for exclusivity.
- Showcase loyal customers on social media for community engagement.
- Encourage referrals through referral programs and surprise loyal customers with personalized gestures.

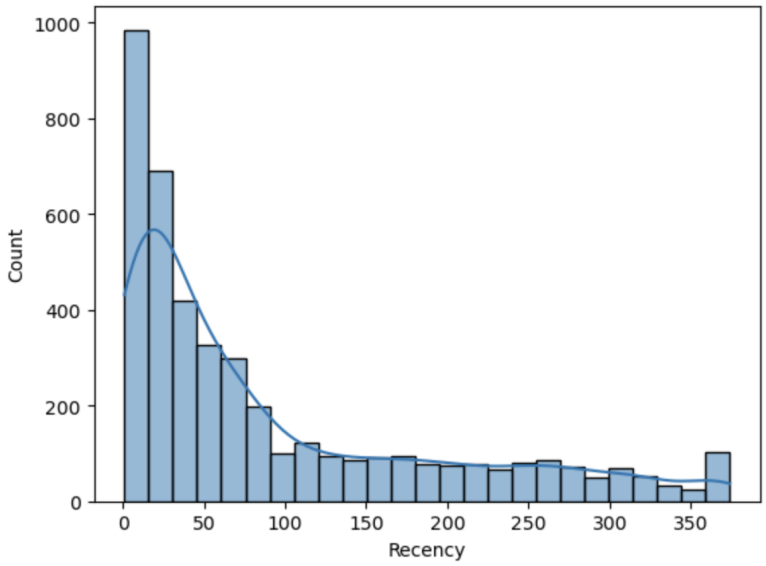
General Recommendations:

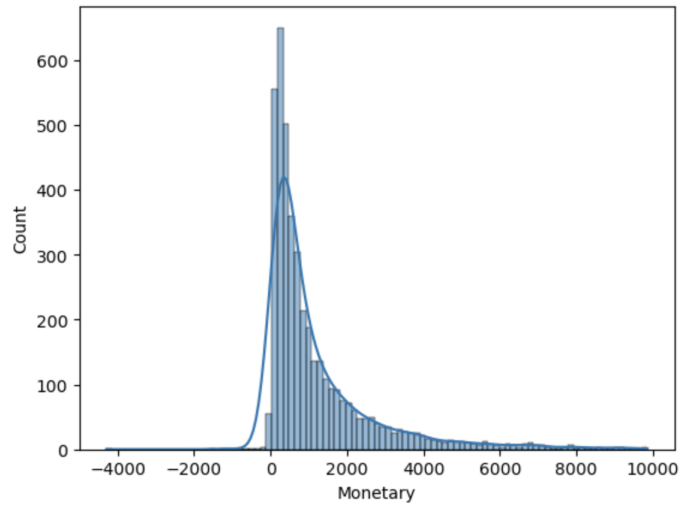
- Implement a comprehensive customer feedback system to understand preferences and pain points.
- Utilize data analytics for continuous refinement and optimization of marketing strategies.
- Personalize communication channels, such as email marketing, based on individual customer preferences.
- Regularly update product offerings to keep customers excited and engaged.

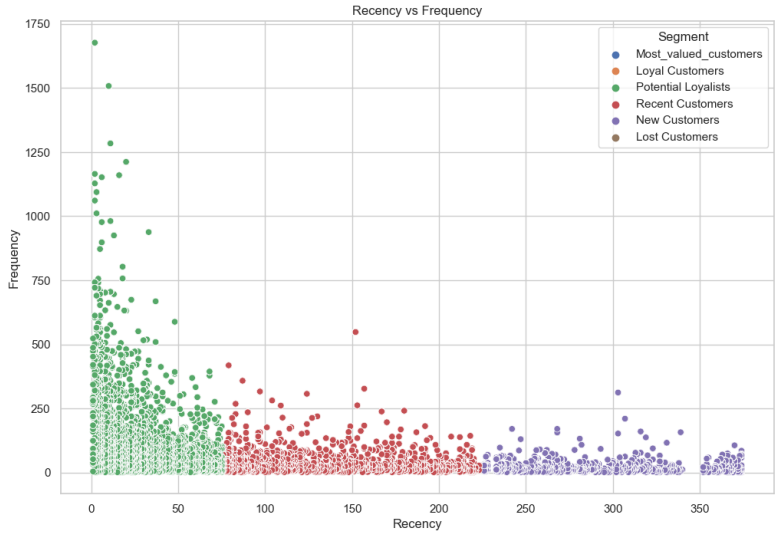
These strategies aim to enhance customer engagement, retention, and satisfaction by addressing the specific needs and behaviors of each identified customer segment.

7)Visualizations




1000
800
600
400
200
0
Count
0
50
100
150
200
250
300
350
Recency


600
500
400
300
200
100
0
Count
−4000
−2000
0
2000
4000
6000
8000
10000
Monetary


Recency vs Frequency
Segment
Most_valued_customers
Loyal Customers
Potential Loyalists
Recent Customers
New Customers
Lost Customers
Frequency
Recency


Frequency vs Monetary
Segment
Most_valued_customers
Loyal Customers
Potential Loyalists
Recent Customers
New Customers
Lost Customers
Monetary
Frequency


Recency vs Monetary
Segment
Most_valued_customers
Loyal Customers
Potential Loyalists
Recent Customers
New Customers
Lost Customers
Monetary
Recency

Questions

1)Data Overview
1.a)What is the size of the dataset in terms of the number of rows and columns?

```
In [12]: print("Number of rows:", data.shape[0])
         print("Number of columns:", data.shape[1])
```

```
Number of rows: 406829
Number of columns: 12
```

1.b)Can you provide a brief description of each column in the dataset?

1. `InvoiceNo` : Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
2. `StockCode` : Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
3. `Description` : Product (item) name. Nominal.
4. `Quantity` : The quantities of each product (item) per transaction. Numeric.
5. `InvoiceDate` : Invoice date and time. Numeric. The day and time when a transaction was generated.
6. `UnitPrice` : Unit price. Numeric. Product price per unit in sterling (Â£).
7. `CustomerID` : Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
8. `Country` : Country name. Nominal. The name of the country where a customer resides.

1.c)What is the time period covered by this dataset?

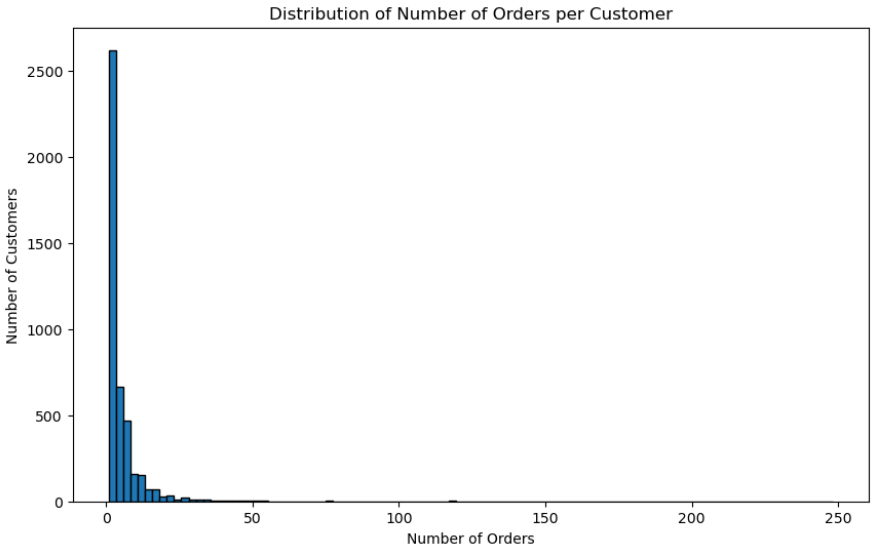
```
In [13]: data['InvoiceDate'] = pd.to_datetime(data['InvoiceDate'])

         print("Time period covered by this dataset:")
         print("Start date:", data['InvoiceDate'].min())
         print("End date:", data['InvoiceDate'].max())
```

```
Time period covered by this dataset:
Start date: 2010-12-01 08:26:00
End date: 2011-12-09 12:50:00
```

2)Customer Analysis
2.a)How many unique customers are there in the dataset?

```
In [14]: print('Number of unique customers: ', data['CustomerID'].nunique())
         Number of unique customers:  4372
```

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



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

Out[14]:

	CustomerID	OrderCount
1895	14911	248
330	12748	224
4042	17841	169
1674	14606	128
568	13089	118

3)Product Analysis

3.a)What are the top 10 most frequently purchased products?

]:

	StockCode	Description	Quantity
3028	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	53215
3460	85099B	JUMBO BAG RED RETROSPOT	45066
3288	84879	ASSORTED COLOUR BIRD ORNAMENT	35314
3475	85123A	WHITE HANGING HEART T-LIGHT HOLDER	34147
434	21212	PACK OF 72 RETROSPOT CAKE CASES	33409
1112	22197	POPCORN HOLDER	30504
2010	23084	RABBIT NIGHT LIGHT	27094
1387	22492	MINI PAINT SET VINTAGE	25880
1509	22616	PACK OF 12 LONDON TISSUES	25321
930	21977	PACK OF 60 PINK PAISLEY CAKE CASES	24163

3.b)What is the average price of products in the dataset?

```
Average Unit price of product:  3.46
```

3.c)Can you find out which product category generates the highest revenue?

Out[50]:

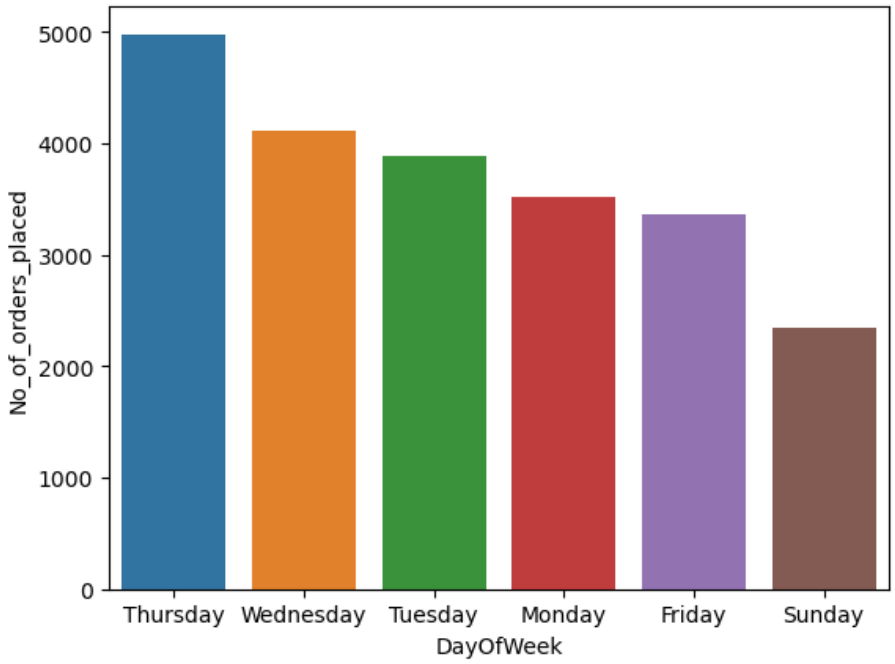
	StockCode	Description	Total order value
1322	22423	REGENCY CAKESTAND 3 TIER	132870.4

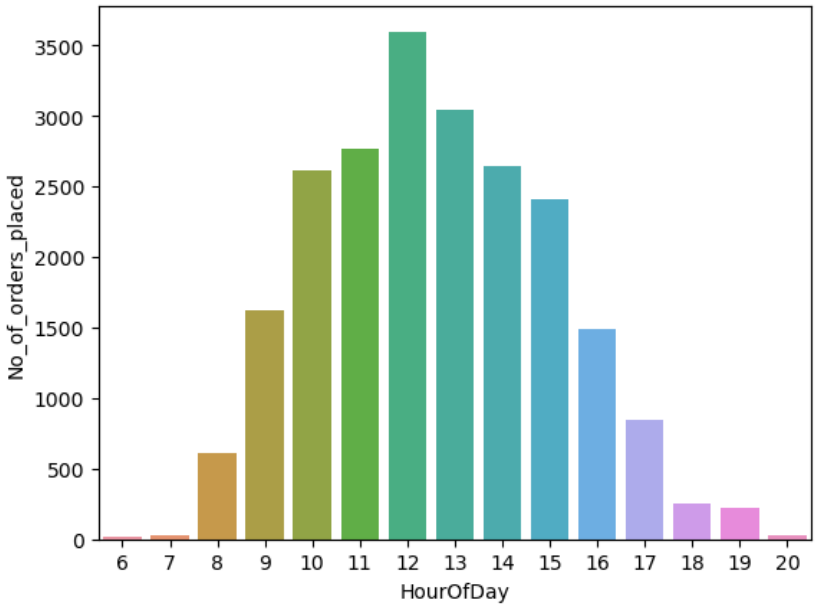
Out[54]:

	Category	Total order value
20	cakestand	132870.4

4)Time Analysis

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

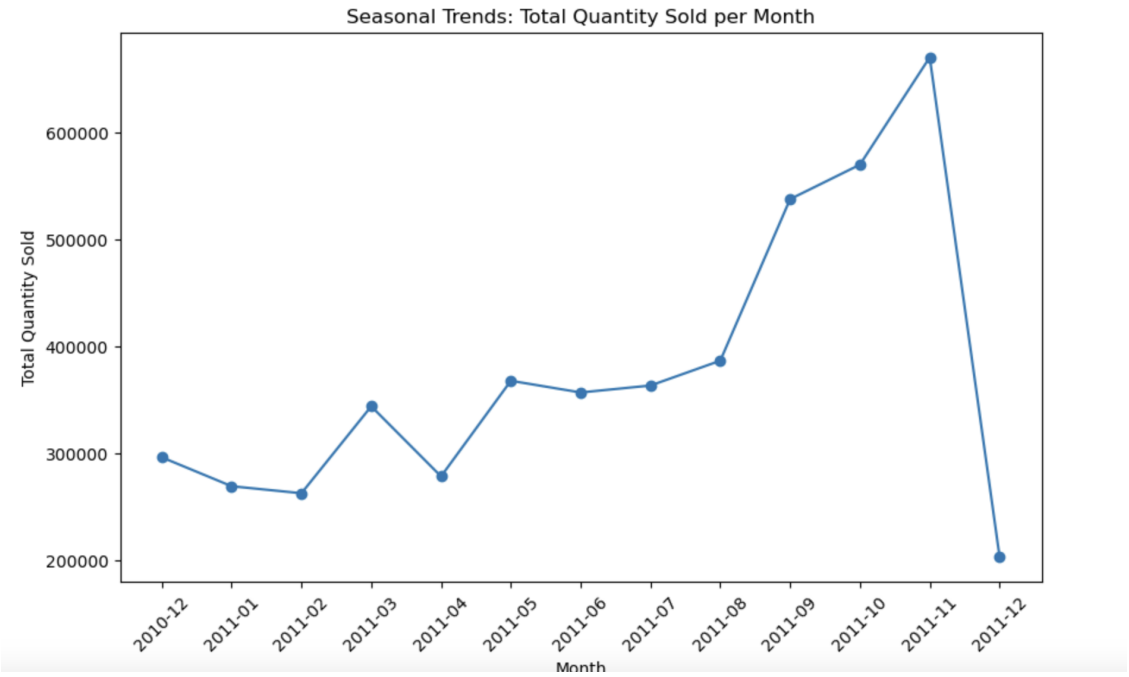




4.b)What is the average order processing time?

It appears that the necessary data to determine the average order processing time is not available. In the absence of the relevant information, calculating the average order processing time becomes infeasible. To gain insights into order processing efficiency, it is essential to have data on order creation timestamps and order fulfillment or shipment timestamps. If this information is not present in the dataset or has not been provided, obtaining the required data points is crucial for conducting a meaningful analysis of average order processing time.

4.c)Are there any seasonal trends in the dataset?



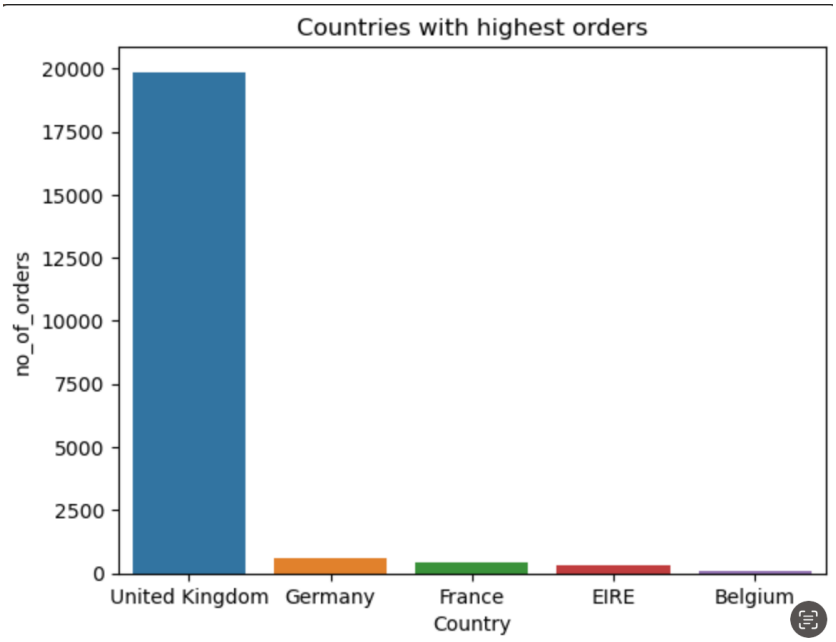
The examination of seasonal trends within the dataset reveals that the highest total quantity sold occurs in November 2011, while the lowest point is observed in December 2011. This discernment highlights notable fluctuations in sales volume during the specified time intervals, emphasizing the need for a thorough investigation into potential contributing factors. Such an analysis holds implications for refining business strategies to effectively address and leverage these observed seasonal variations.

5)Geographical Analysis

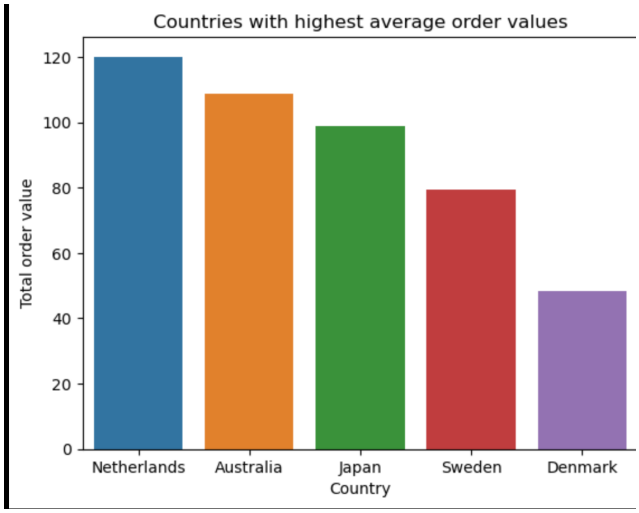
5.a)Can you determine the top 5 countries with the highest number of orders?

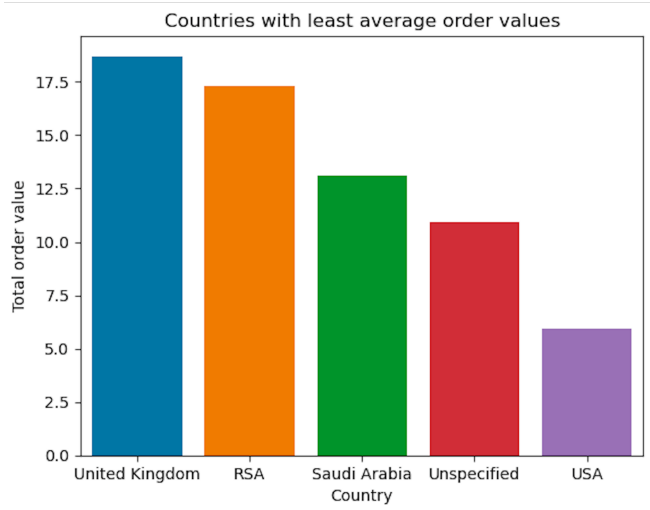
Out[24]:

	Country	InvoiceNo
35	United Kingdom	19857
14	Germany	603
13	France	458
10	EIRE	319
3	Belgium	119



5.b)Is there a correlation between the country of the customer and the average order value?

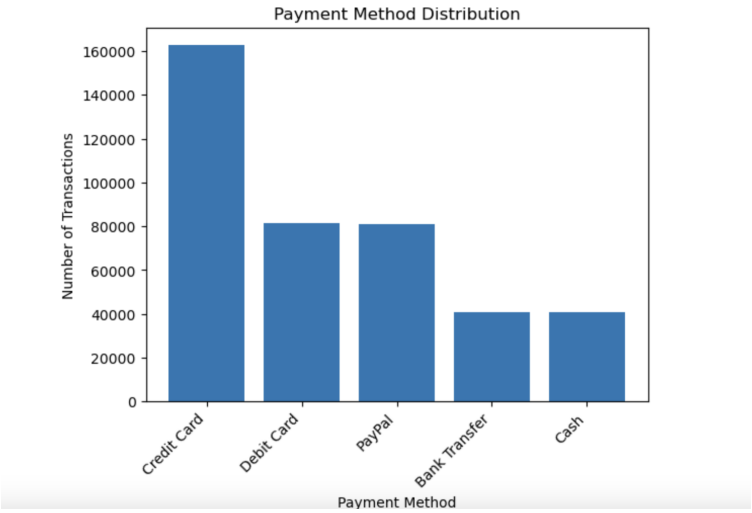




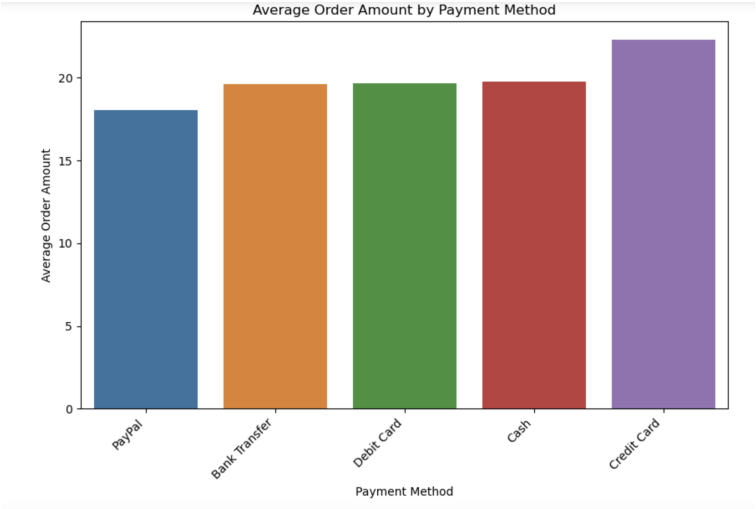
A correlation coefficient close to zero (-0.01) suggests a very weak or negligible linear relationship between the 'Country' and 'Total order value' columns. In this case, there is almost no discernible pattern or trend in the data that indicates how the 'Total order value' varies with different countries. The negative sign of the correlation indicates a very slight inverse relationship, but the magnitude is too small to draw any meaningful conclusions.
In practical terms, the correlation of -0.01 suggests that changes in the 'Total order value' are not significantly associated with changes in the 'Country' column. Other factors or variables may have a more substantial impact on the total order value, and the country alone does not provide a clear indication of its variation.

6)Payment Analysis

6.a)What are the most common payment methods used by customers?



6b)Is there a relationship between the payment method and the order amount?



In the absence of sufficient payment data in the provided dataset, we employed a random generation approach to simulate payment methods. Utilizing a randomization process, we generated hypothetical payment methods to facilitate analysis. The outcome of this random generation revealed specific results, including the identification of credit cards as the payment method with the highest total order value. It's important to note that these findings are based on

hypothetical data introduced for analytical purposes, as the original dataset lacked specific payment information. The utilization of randomization allows us to explore potential scenarios and draw insights, keeping in mind the hypothetical nature of the generated payment methods.

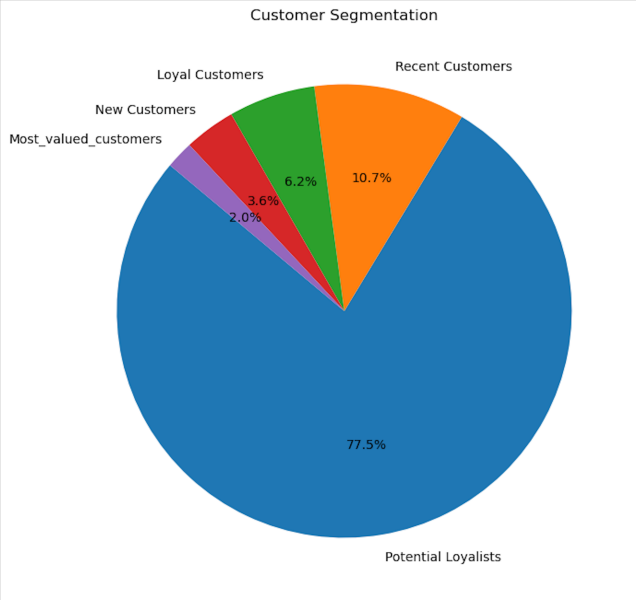
7)Customer Behavior

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

```
Average active days for customers: 133.38586459286367
```

The average active days for customers, which is approximately 133.39 days in our case, refers to the average length of time between the first and last purchase made by each customer. This means, on average, customers in your dataset make purchases over a period of about 133.39 days from their first to their last purchase.

7.b)Are there any customer segments based on their purchase behavior?



From the data available we have recognized 5 customer segments namely Loyal Customers,New Customers,Recent Customers,Most_valued_customres and Potential Loyalists based on purchase behavior.

8)Returns and Refunds
8.a)What is the percentage of orders that have experienced returns or refunds?

```
Out[27]: 2.188880340388714
```

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

```
Product categories with highest return rates:
```

	Description	ReturnRate
199	BLUE FLYING SINGING CANARY	1.0
1806	VINTAGE GOLD TINSEL REEL	1.0
434	CRUK Commission	1.0
432	CREAM SWEETHEART TRAYS	1.0
431	CREAM SWEETHEART SHELF + HOOKS	1.0
429	CREAM SWEETHEART MAGAZINE RACK	1.0
91	ANTIQUE LILY FAIRY LIGHTS	1.0
115	ASSORTED TUTTI FRUTTI ROUND BOX	1.0
530	Discount	1.0
947	LARGE ROUND CUTGLASS CANDLESTICK	1.0
1707	SWEETHEART KEY CABINET	1.0
1650	SMALL TAHITI BEACH BAG	1.0
1723	TEA TIME CAKE STAND IN GIFT BOX	1.0
1855	WHITE CHERRY LIGHTS	1.0
1252	PINK POODLE HANGING DECORATION	1.0
1240	PINK LARGE JEWELED PHOTOFRAME	1.0

```
Product categories with least return rate:
```

	Description	ReturnRate
48	4 TRADITIONAL SPINNING TOPS	0.002653
1574	SET/20 RED RETROSPOT PAPER NAPKINS	0.002649
1634	SMALL DOLLY MIX DESIGN ORANGE BOWL	0.002632
1659	SPACEBOY BIRTHDAY CARD	0.002513
1143	PACK OF 72 SKULL CAKE CASES	0.001953

9)Profitability Analysis

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

```
Out[32]: 8300065.814000001
```

9.b)What are the top 5 products with the highest profit margins?

```
Out[33]: StockCode
         22423     132870.40
         85123A     93979.20
         85099B     83236.76
         47566      67687.53
         POST       66710.24
         Name: Total order value, dtype: float64
```

10)Customer Satisfaction

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

The dataset does not contain any information about customer feedback or ratings for products or services. Consequently, we are unable to conduct an analysis on customer sentiments and satisfaction due to the absence of relevant data.

10.b) Can you analyze the sentiment or feedback trends, if available?

We cannot provide sentiment or feedback trends also due to lack of data.

Conclusion

The inception of this project was marked by the exploration of an E-Commerce dataset, with a distinct goal in mind – the implementation of RFM analysis. RFM, standing for Recency, Frequency, and Monetary, is a potent method utilized by businesses to gain insights into customer behavior. By scrutinizing recent purchasing patterns, frequency of transactions, and the monetary value of interactions, RFM analysis facilitates the creation of a robust Customer Segmentation model.

The project journey commenced with the importation of the dataset, followed by an in-depth exploration to unravel the nuances of the data. Transitioning into the pivotal phase of data preprocessing, meticulous steps were taken to ensure data cleanliness, address missing values, and align data types for optimal analysis.

Upon completing the data preparation, the focus shifted to the heart of the project – RFM analysis. The calculation of Recency, Frequency, and Monetary metrics for each customer laid the foundation for understanding their purchasing behavior. The subsequent step involved RFM segmentation, where customers were assigned scores based on quartiles or custom-defined bins, culminating in a unified RFM score for comprehensive insights.

The project's evolution continued with customer segmentation using K-Means clustering, seeking an optimal cluster configuration that revealed meaningful customer segments. These segments were then thoroughly analyzed and profiled to unveil distinctive characteristics.

In the pursuit of strategic excellence, actionable marketing recommendations were crafted for each customer segment, tailor-made to enhance retention and maximize revenue. Visualizations, such as bar charts and scatter plots, were harnessed to visually articulate the distribution of RFM scores and clusters, adding a layer of clarity to the analysis.

In essence, this project not only showcased the application of RFM analysis in customer segmentation but also underscored the pivotal role of targeted marketing strategies informed by customer behavior. The harmonious integration of data preprocessing, RFM analysis, clustering, and strategic recommendations establishes a robust framework, empowering businesses to optimize their customer engagement strategies within the dynamic landscape of e-commerce. If sufficient data pertaining to payment methods and customer satisfaction were available, the analysis could have achieved a higher level of robustness and depth.