



Customer Segmentation Using RFM Analysis

IE6400 - Foundations of Data Analytics Engineering Project Report

Group 1

Kruthika Srinivas Vasisht (002798505)

Ruthvika Reddy Tangirala (002293262)

Sabarish Subramaniam A V (002243373)

Sarvesh Selvam (002874621)

Sneha Manjunath Chakrabhavi (002836841)

INTRODUCTION

For any business, understanding customer behavior and product sales is crucial for survival and success, whether it operates online or has a physical location. In this project report, we utilized Jupyter Notebook and Python to help an online retail business increase its revenue and profit margin by providing valuable insights through customer segmentation and an RFM (Recency, Frequency, Monetary) analysis on the dataset. We obtained the dataset from a Kaggle notebook sourced through the UCI Machine Learning Repository.


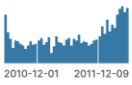

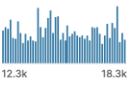

Analyzing customer data is crucial for businesses to identify the most suitable target audience for different advertising strategies. To begin with, we sorted the data and conducted an RFM analysis, which helped us extract valuable insights into the ideal customer base. We analyzed their activity time, purchase frequency, and spending patterns to identify the most relevant customer groups. Finally, we assigned each group an RFM score based on the above parameters and determined the appropriate approach.

We have meticulously cleaned and prepared the dataset obtained from the Kaggle notebook to complete this project analysis. Additionally, we have followed a structured and reliable method for exploring the dataset using Python's Pandas library on the Jupyter Notebook platform. Our ultimate goal is to offer comprehensive and actionable insights on marketing and advertising strategies to help the business increase profitability and achieve overall success.

DATA SOURCE

data.csv (45.58 MB) 📄 🔄 ⏪

Detail Compact Column 8 of 8 columns ▾

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
25900 unique values	4070 unique values	4224 unique values	 -80995 81.0k	 2010-12-01 2011-12-09	 -11.1k 39k	 12.3k 18.3k	
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom

The customer dataset in the analysis was obtained from a Kaggle notebook sourced through the UCI Machine Learning Repository. The dataset contains information on various products, their cost and description, quantity purchased, customer ID, and country of purchase. The initial exploration of the dataset involved:

- checking and converting data types
- checking for and removing null values
- dropping duplicate rows
- feature extraction and clustering
- recency, frequency, and monetary calculations

The dataset provides valuable insights on when the customer had purchased the order and the number of orders placed which allows us to perform an RFM analysis with ease and create customer segmentation based on that analysis.

RESULTS AND METHODS

1) Data Preprocessing

Data after replacing missing values for string feature and CustomerID

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680	France

401604 rows × 8 columns

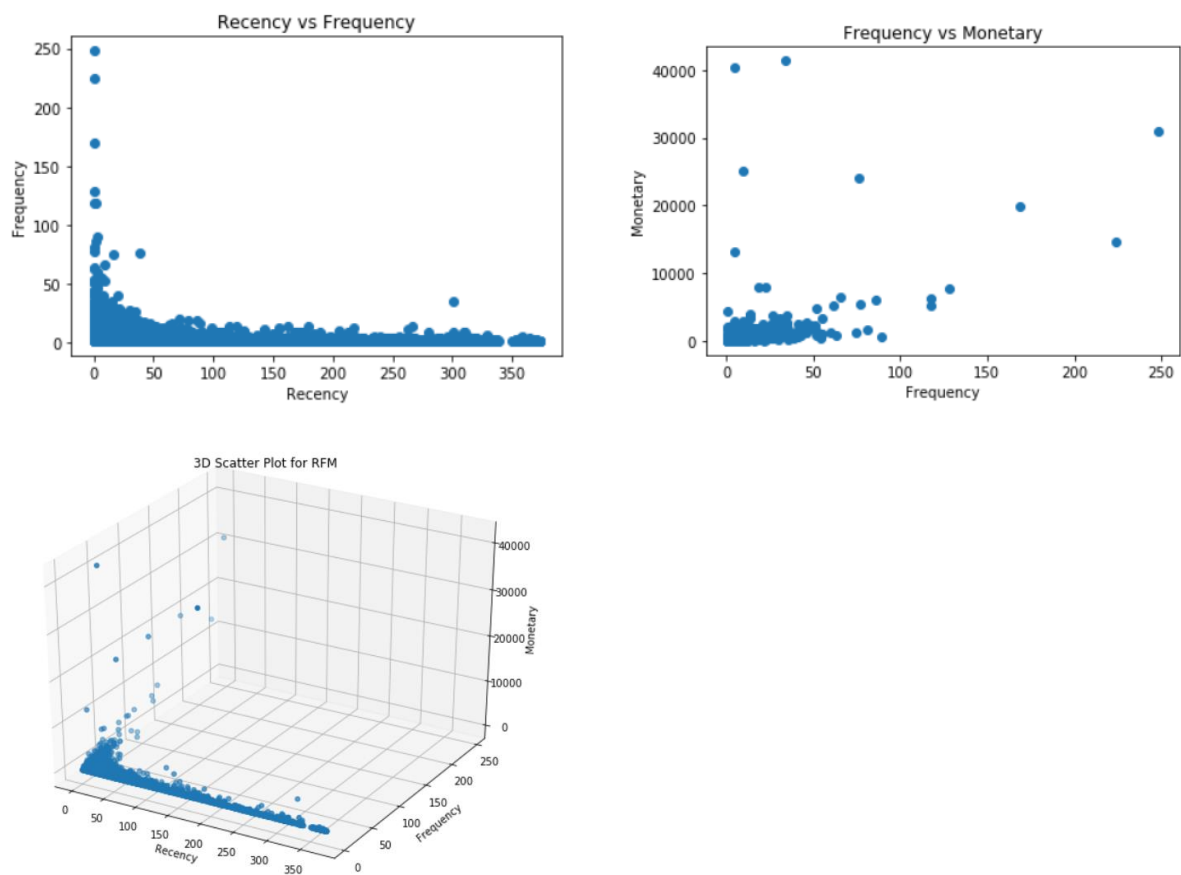
2) RFM Calculation

The RFM metrics are:

	CustomerID	Recency	Frequency	Monetary
0	17850	301	35	1209.66
9	13047	31	18	798.30
26	12583	2	18	791.28
46	13748	95	5	111.90
65	15100	329	6	65.70
...
536969	13436	1	1	69.96
537255	15520	1	1	31.04
538064	13298	0	1	7.50
538812	14569	0	1	47.04
541768	12713	0	1	95.13

4372 rows × 4 columns

Visualization of RFM metrics:



3) RFM Segmentation

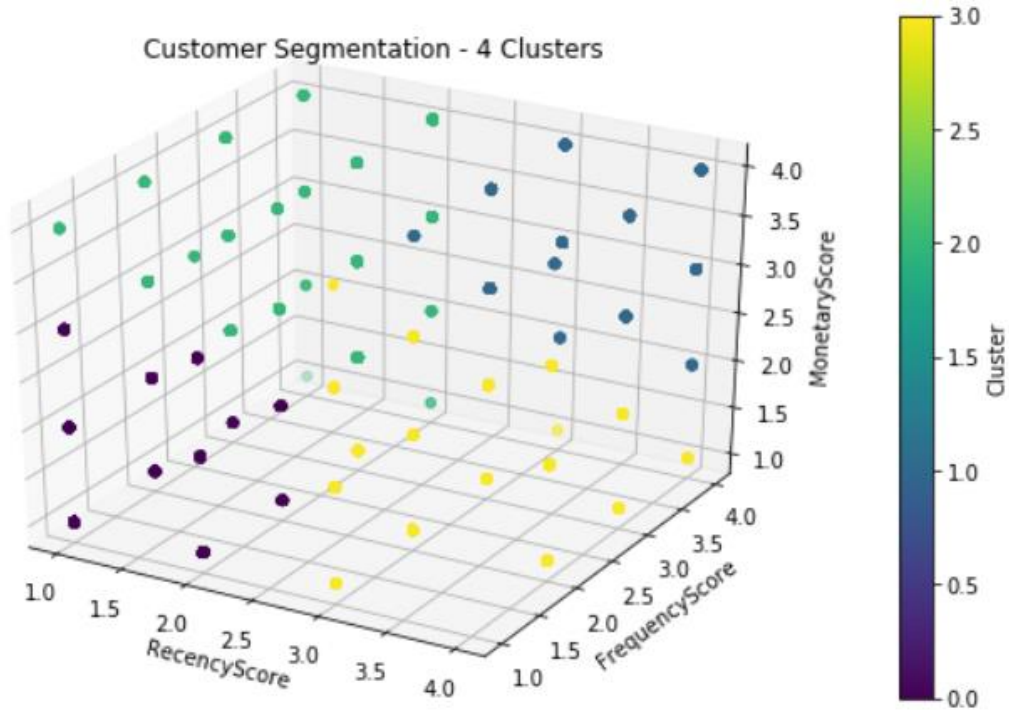
Scoring RFM Metrics:

	CustomerID	RecencyScore	FrequencyScore	MonetaryScore	RFMScore
17972	14051	4	4	4	444
14167	14907	4	4	4	444
33167	14309	4	4	4	444
214262	15152	4	4	4	444
86540	17686	4	4	4	444
...
180312	12700	4	4	4	444
8933	17858	4	4	4	444
8953	16393	4	4	4	444
34427	14769	4	4	4	444
9130	15023	4	4	4	444

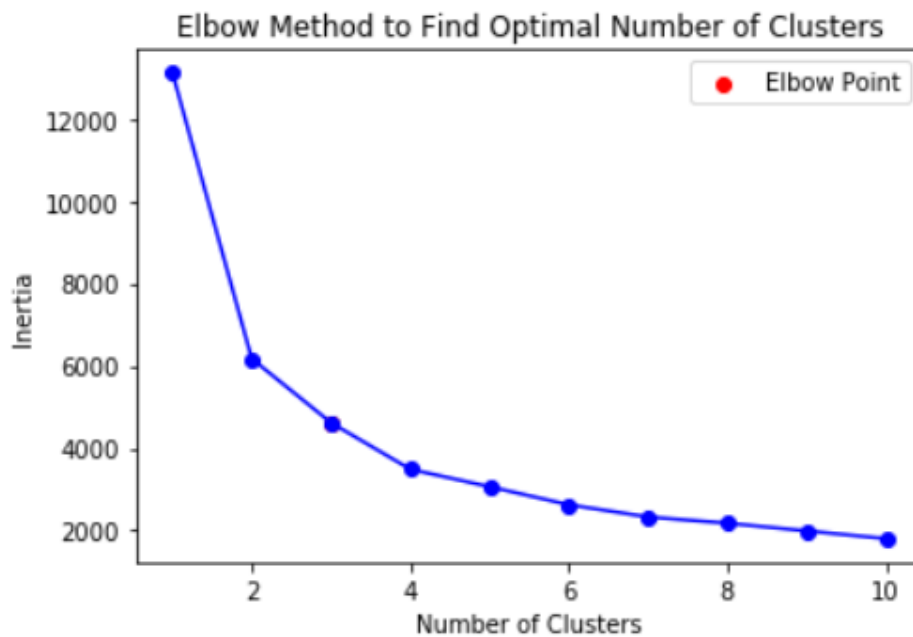
100 rows × 5 columns

4) Customer Segmentation

Segmenting customers based on their RFM scores using k-means clustering:



Experimenting with different numbers of clusters:



5) Segment Profiling

Segment profiles:

	Cluster	RecencyMean	RecencyMin	RecencyMax	FrequencyMean	FrequencyMin	\
0	0	1.382353	1	2	1.299265	1	
1	1	3.601757	3	4	3.632504	2	
2	2	1.675393	1	2	2.981675	1	
3	3	3.301587	2	4	2.180272	1	

	FrequencyMax	MonetaryMean	MonetaryMin	MonetaryMax	CustomerCount
0	3	1.536029	1	3	1360
1	4	3.608346	2	4	1366
2	4	3.075916	1	4	764
3	4	1.769841	1	4	882

Cluster 0:

RecencyMean: The average recency score is approximately 2.05, suggesting that customers in this segment made purchases recently.

FrequencyMean: The average frequency score is around 2.84, indicating that customers in this segment make purchases moderately frequently.

MonetaryMean: The average monetary score is 3.13, suggesting that customers in this segment contribute a relatively high monetary value.

CustomerCount: This segment contains 959 customers.

Cluster 1:

RecencyMean: The average recency score is approximately 1.38, suggesting that customers in this segment made very recent purchases.

FrequencyMean: The average frequency score is around 1.25, indicating that customers in this segment make purchases less frequently.

MonetaryMean: The average monetary score is 1.58, suggesting that customers in this segment contribute a relatively low monetary value.

CustomerCount: This segment contains 1407 customers.

Cluster 2:

RecencyMean: The average recency score is approximately 3.68, suggesting that customers in this segment made purchases less recently.

FrequencyMean: The average frequency score is around 3.72, indicating that customers in this segment make purchases quite frequently.

MonetaryMean: The average monetary score is 3.72, suggesting that customers in this segment contribute a relatively high monetary value.

CustomerCount: This segment contains 1167 customers.

Cluster 3:

RecencyMean: The average recency score is approximately 3.36, suggesting that customers in this segment made purchases less recently.

FrequencyMean: The average frequency score is around 2.50, indicating that customers in this segment make purchases moderately frequently.

MonetaryMean: The average monetary score is 1.60, suggesting that customers in this segment contribute a relatively low monetary value.

CustomerCount: This segment contains 805 customers.

Interpretation:

Cluster 1 represents recently active but less frequent and lower-value customers.

Cluster 2 represents active and frequent customers with higher monetary contributions.

Cluster 3 represents less recent, moderately frequent, and lower-value customers.

This interpretation is based on the average scores for recency, frequency, and monetary values within each cluster.

6) Marketing Recommendations

Cluster 0: Recent and High-Value Customers

Recommendations:

Promotional Offers: Offer exclusive promotions or discounts to incentivize repeat purchases from this segment.

Loyalty Programs: Introduce a loyalty program to reward these customers for their high-value contributions.

New Product Releases: Inform this segment about new product releases to encourage them to make additional purchases.

Cluster 1: Very Recent but Lower-Value Customers

Recommendations:

Engagement Campaigns: Implement targeted engagement campaigns to encourage more frequent purchases.

Upselling Opportunities: Identify opportunities for upselling or cross-selling to increase the average transaction value.

Personalized Recommendations: Provide personalized product recommendations based on their recent purchases to increase relevancy.

Cluster 2: Active and High-Value Customers

Recommendations:

Exclusive Access: Provide early access to sales or exclusive products to reward their loyalty.

VIP Programs: Establish a VIP program with premium benefits for this segment to enhance their loyalty.

Cross-Sell Complementary Products: Suggest complementary products to increase the average transaction value.

Cluster 3: Less Recent and Moderate-Value Customers

Recommendations:

Reactivation Campaigns: Implement reactivation campaigns to bring these customers back with special offers.

Retention Discounts: Offer special discounts for their next purchase to encourage repeat business.

Feedback Surveys: Gather feedback to understand reasons for reduced activity and tailor offerings accordingly.

General Recommendations:

Segment-Specific Communication: Tailor marketing communication to each segment's preferences and behaviors.

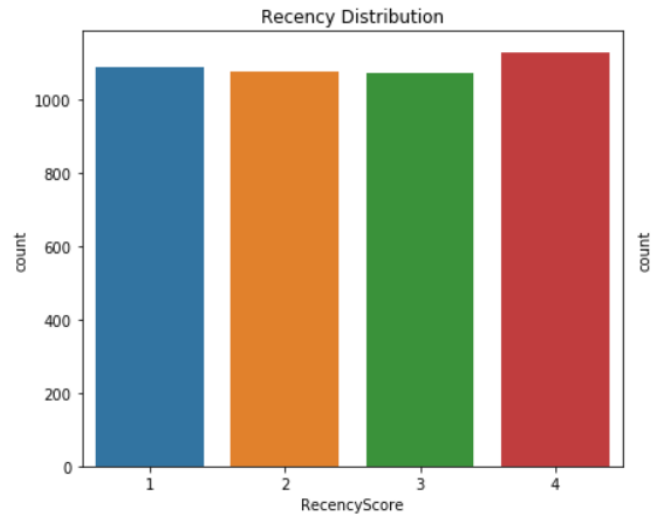
Multichannel Engagement: Utilize various channels such as email, social media, and targeted advertising to reach customers where they are most active.

Data-Driven Personalization: Leverage customer data to personalize marketing messages, recommendations, and promotions for each segment.

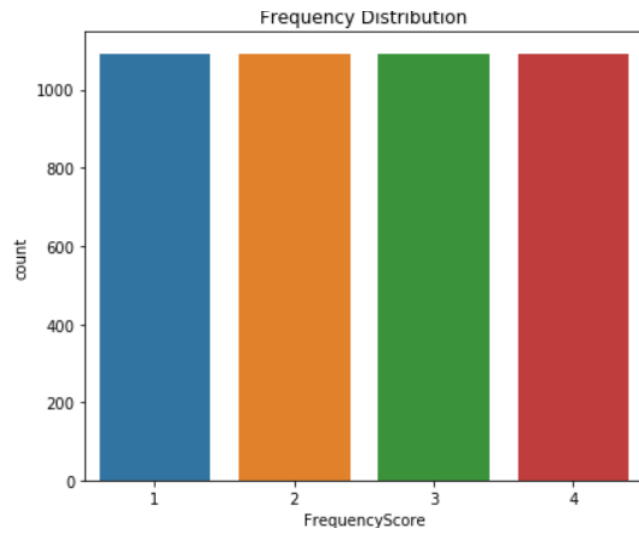
Customer Feedback: Collect feedback from each segment to continuously improve products, services, and overall customer experience.

By implementing these tailored strategies, the business can build stronger relationships with each customer segment, enhance customer satisfaction, and optimize revenue generation. Regularly analyzing and adjusting these strategies based on customer feedback and evolving market trends will further contribute to the success of the business.

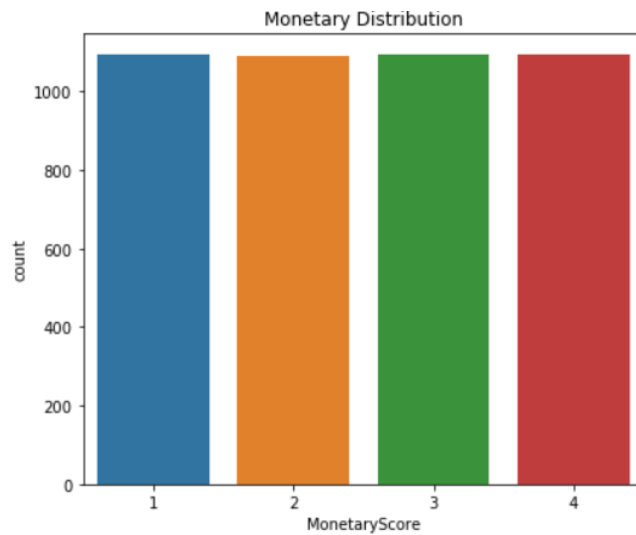
7) Visualization



Bar chart for Recency distribution

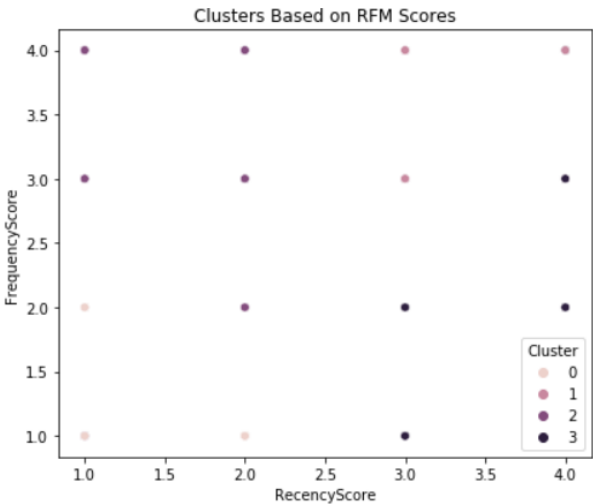


Bar chart for Frequency distribution

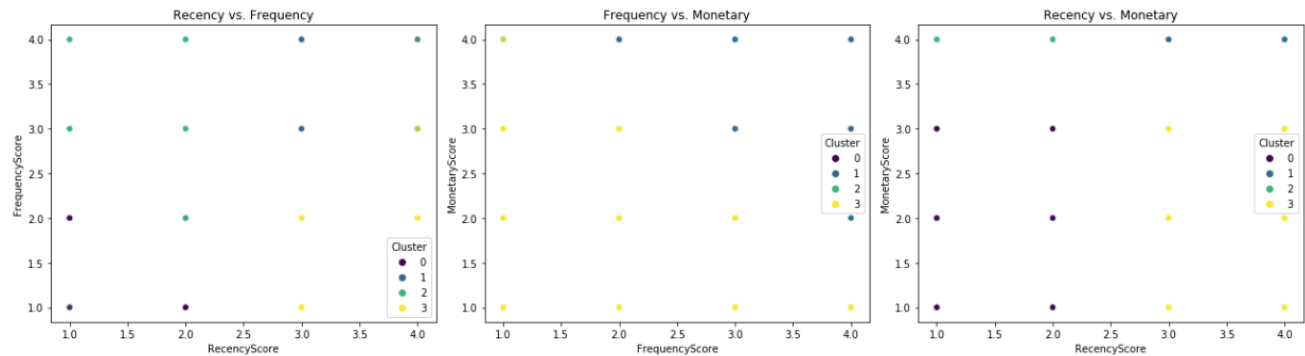


Bar chart for Monetary distribution

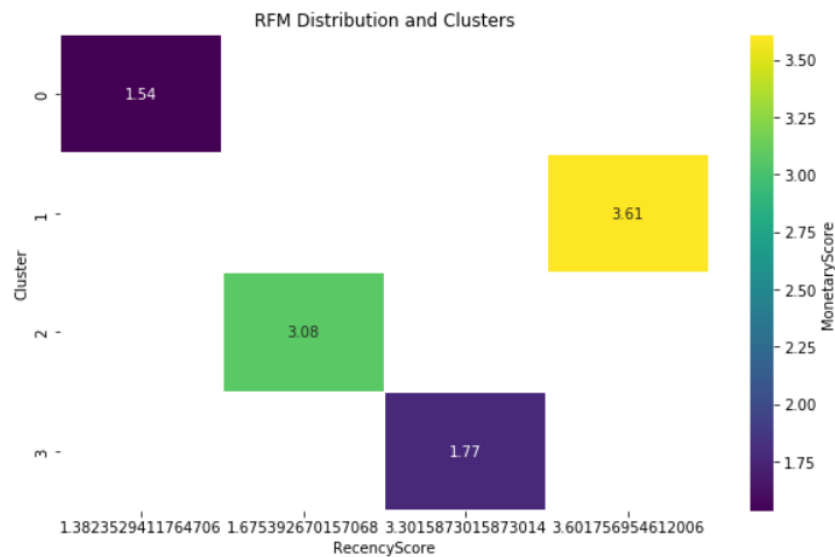
Scatter plot for Clusters



Scatter plot for RFM distribution



RFM Distribution and Clusters Heatmap



1) Data information

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

The size of the dataset is 541909 rows and 8 columns.

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

- InvoiceNo: It is a unique number identified for each transaction.
- StockCode: It is a unique code identified for each product.
- Description: The description regarding each product.
- Quantity: No.of units of products that are associated with each transaction.
- InvoiceDate: The point of date and time when the transaction was made.
- UnitPrice: It's the price for each unit.
- CustomerID: A unique id associate with each customer.
- Country: The country where the customer stays.

1.3 What is the period covered by this dataset?

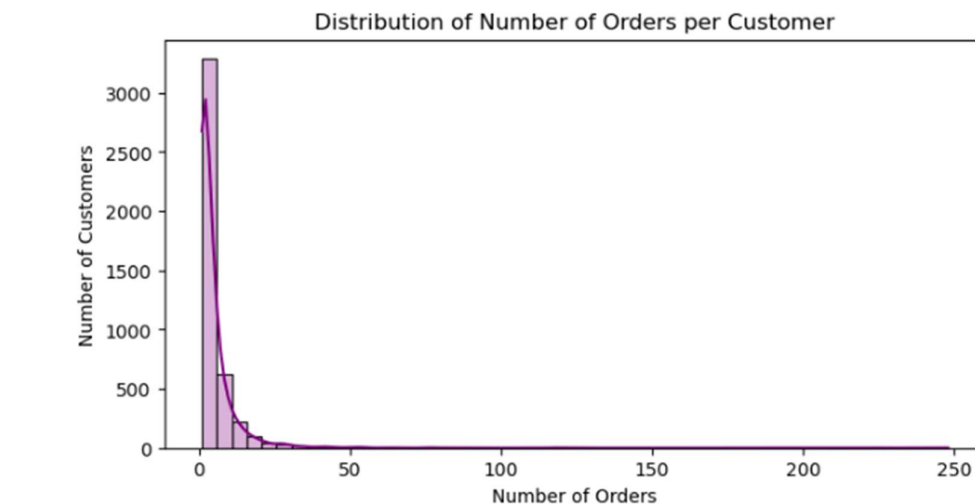
The dataset covers a time from 2010-12-01 08:26:00 A.M. to 2011-12-09 12:50:00 P.M.

2) Customer Analysis

2.1 How many unique customers are there in the dataset?

The number of unique customers are: 4372

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



```
count    4372.000000
mean       5.075480
std        9.338754
min         1.000000
25%         1.000000
50%         3.000000
75%         5.000000
max        248.000000
Name: InvoiceNo, dtype: float64
```

The average number of orders per customer is 4. The minimum no.of orders per customer is 1 and the maximum no.of orders per customer is 209.

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

The top 5 customers by orders are:

```
CustomerID
14911      248
12748      224
17841      169
14606      128
13089      118
Name: InvoiceNo, dtype: int64
```

Customer with ID 12748 has the highest no.of orders which is 209.

3) Product Analysis

3.1 What are the top 10 most frequently purchased products?

The top 10 purchased products are:

```
WHITE HANGING HEART T-LIGHT HOLDER    2058
REGENCY CAKESTAND 3 TIER                1894
JUMBO BAG RED RETROSPOT                 1659
PARTY BUNTING                          1409
ASSORTED COLOUR BIRD ORNAMENT           1405
LUNCH BAG RED RETROSPOT                  1345
SET OF 3 CAKE TINS PANTRY DESIGN         1224
POSTAGE                                 1196
LUNCH BAG BLACK SKULL.                   1099
PACK OF 72 RETROSPOT CAKE CASES          1062
Name: Description, dtype: int64
```

WHITE HANGING HEART T-LIGHT HOLDER is the highest purchased product.

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

The average price of products is: 3.4740636398043865

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

```
Description
REGENCY CAKESTAND 3 TIER          132567.70
WHITE HANGING HEART T-LIGHT HOLDER 93767.80
JUMBO BAG RED RETROSPOT           83056.52
PARTY BUNTING                     67628.43
POSTAGE                            66710.24
ASSORTED COLOUR BIRD ORNAMENT      56331.91
RABBIT NIGHT LIGHT                 51042.84
CHILLI LIGHTS                      45915.41
PAPER CHAIN KIT 50'S CHRISTMAS     41423.78
PICNIC BASKET WICKER 60 PIECES      39619.50
Name: TotalRevenue, dtype: float64
```

REGENCY CAKESTAND 3 TIER is the product that generated the highest revenue.

4) Time Analysis

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

The count of the orders of a particular hour

```
6      41
7     383
8    8789
9   22446
10  38725
11  49525
12  72213
13  64051
14  54194
15  45641
16  24618
17  13604
18   3104
19   3423
20    847
Name: HourOfDay, dtype: int64
```

At 12'o clock maximum no.of orders has been placed.

The count of the orders of a particular day

```
Thursday    81575
Wednesday   69753
Tuesday     67376
Monday      65715
Sunday      61673
Friday      55512
Name: DayOfWeek, dtype: int64
```

Thursdays has the highest number of orders placed.

4.2 What is the average order processing time?

The average order processing time is: 0 days 00:01:20.285854438. The result "0 days 00:01:20.285854438" indicates that, on average, there is approximately 1 minute and 20 seconds of processing time between consecutive orders based on the assumption that the processing time is the time between placing the current order and placing the next one.

4.3 Are there any seasonal trends in the dataset?



- As the data contains mostly one year of the data it is hard to determine if there are any seasonalities.
- From the given data it can be seen that orders has increased towards the end of the year.
- It has increased from the fall season, may be due to start of holiday season.
- It has peaked in the month of November, which can be explained with the heavy purchasing during thanksgiving and black Friday season.

5) Geographical Analysis

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

Average Order Value by Country:

Country			
Australia	1985.648841	Japan	1262.165000
Austria	534.437895	Lebanon	1693.880000
Bahrain	274.200000	Lithuania	415.265000
Belgium	343.789580	Malta	250.547000
Brazil	1143.600000	Netherlands	2818.431089
Canada	611.063333	Norway	879.086500
Channel Islands	608.375455	Poland	300.547500
Cyprus	642.938000	Portugal	414.225143
Czech Republic	141.544000	RSA	1002.310000
Denmark	893.720952	Saudi Arabia	65.585000
EIRE	783.704639	Singapore	912.039000
European Community	258.350000	Spain	521.486000
Finland	465.140417	Sweden	795.335000
France	429.314520	Switzerland	785.061972
Germany	367.345721	USA	247.274286
Greece	785.086667	United Arab Emirates	634.093333
Iceland	615.714286	United Kingdom	339.787287
Israel	1164.733333	Unspecified	332.596250
Italy	307.100182	Name: TotalOrderValue, dtype: float64	

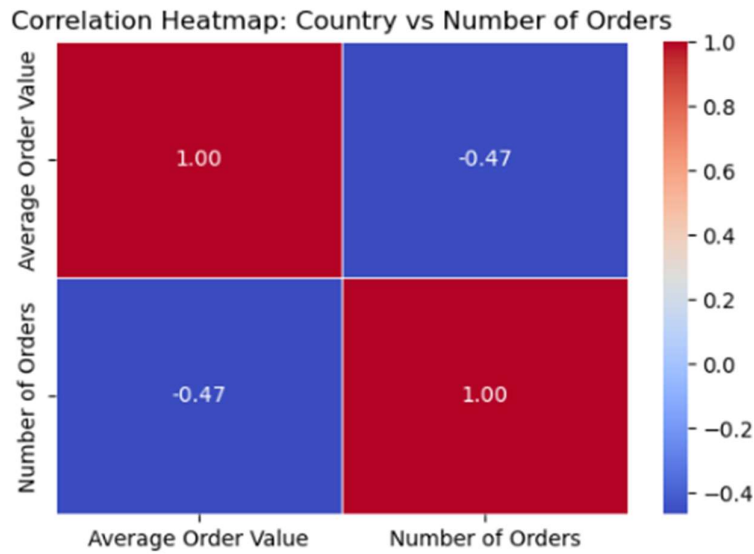
Top 5 countries with highest no.of orders:

United Kingdom	356728
Germany	9480
France	8475
EIRE	7475
Spain	2528
Name: Country, dtype: int64	

United Kingdom has the highest no.of orders.

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

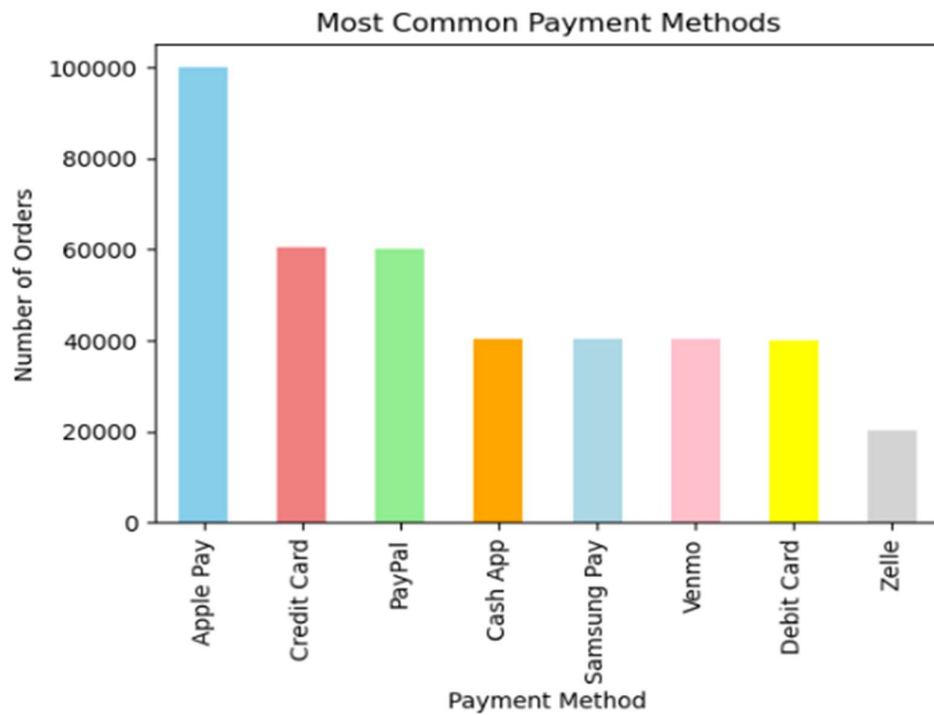
Correlation coefficient of -0.47 suggests a moderate negative correlation between the country of the customer and the number of orders. This implies that, on average, as the number of orders increases for a particular country, the average order value tends to decrease.



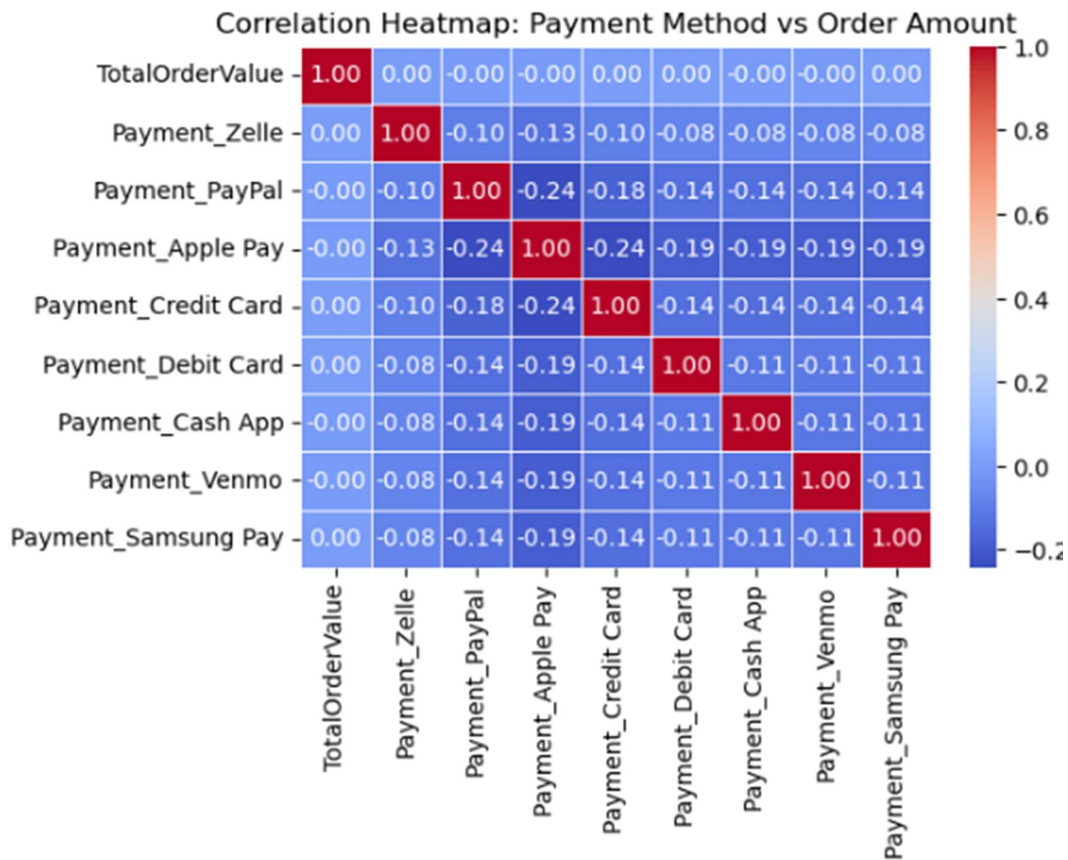
6) Payment Analysis

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

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Recency	Frequency	...	Year	Month	TotalOrderValue
0	536365	85123A WHITE HANGING HEART T- LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	301	35	...	2010	12	15.30
1	536365	71053 WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	301	35	...	2010	12	20.34
2	536365	84406B CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom	301	35	...	2010	12	22.00
3	536365	84029G KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	301	35	...	2010	12	20.34
4	536365	84029E RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	301	35	...	2010	12	20.34
...
541896	581587	22555 PLASTERS IN TIN STRONGMAN	12	2011-12-09 12:50:00	1.65	12680	France	0	4	...	2011	12	19.80
541895	581587	22556 PLASTERS IN TIN CIRCUS PARADE	12	2011-12-09 12:50:00	1.65	12680	France	0	4	...	2011	12	19.80



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



Overall Correlation between Payment Methods and Order Amount: 0.1120

The overall correlation value between payment methods and the order amount is 0.1121. This value indicates a very weak positive correlation on average.

This correlation coefficient is very small, suggesting that there is no significant linear relationship between the payment method and the order amount.

In other words, the choice of payment method does not appear to have a substantial impact on the total order amount based on the linear correlation analysis.

7) Customer Behaviour

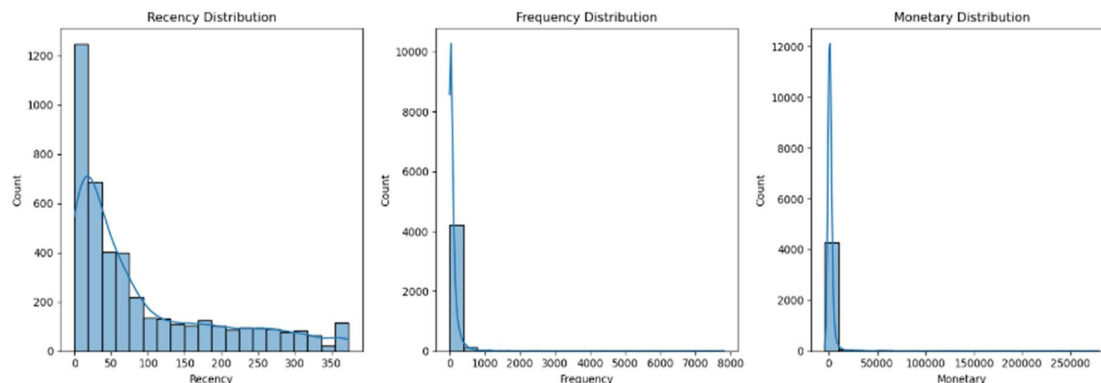
7.1 How long, on average, do customers remain active (between their first and last purchase)?

Timedelta ('133 days 17:25:29.204025618')

The average time that the customer are being active is 133 days 17hours

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

	Recency	Frequency	Monetary
count	4372.000000	4372.000000	4372.000000
mean	91.047118	91.858188	1893.531433
std	100.765435	229.223566	8218.696204
min	0.000000	1.000000	-4287.630000
25%	16.000000	17.000000	291.795000
50%	49.000000	41.000000	644.070000
75%	142.000000	99.250000	1608.335000
max	373.000000	7812.000000	279489.020000



Recency:

- The average recency (mean) is approximately 91 days, suggesting that, on average, customers made their most recent purchase around 91 days ago.
- The minimum recency is 0, indicating that some customers made a purchase very recently.
- The maximum recency is 373, indicating that some customers made their last purchase a considerable time ago.

Frequency:

- The average frequency (mean) is around 91.86, indicating that, on average, customers made around 92 purchases.
- The minimum frequency is 1, indicating that some customers made only one purchase.

- The maximum frequency is 7812, indicating that some customers made a very high number of purchases.

Monetary:

- The average monetary value (mean) is approximately 1893.53, suggesting that, on average, customers spent around 1893.53 dollars .
- The minimum monetary value is negative (-4287.63), indicating that some customers have negative order values (possibly due to refunds or returns).
- The maximum monetary value is 279,489.02 dollars, indicating that some customers have made very high-value purchases.

Inferences:

- There is a wide range of recency, suggesting that there are both recent and long-time customers.
- The distribution of frequency indicates that while many customers make a moderate number of purchases, there are also customers who make a very high number of purchases.
- The monetary values vary widely, with some customers making high-value purchases.

8) Returns and Refunds

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

22190 3654 16.466876971608833

Total no.of returns and refunds orders are: 3654

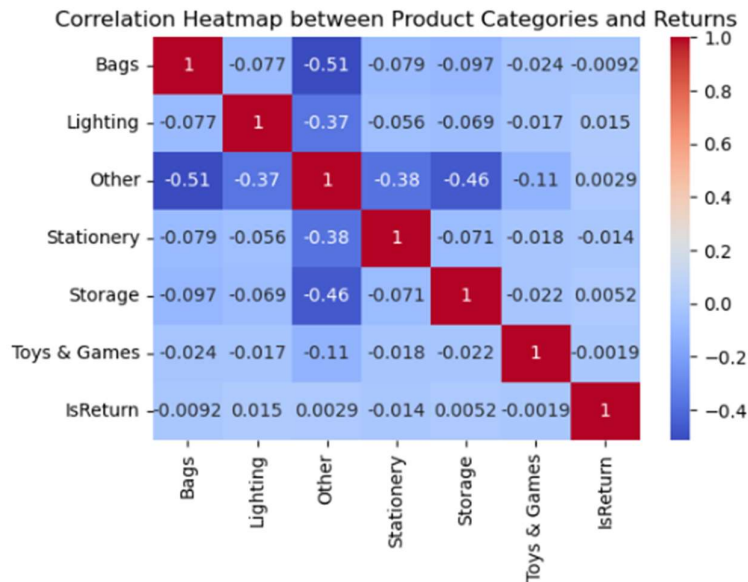
Total percentage returns and refunds orders are: 16.466876971608833

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

```
Category
Other      17.075509
Lighting   6.001396
Storage    5.726826
Bags       4.543222
Stationery 3.036908
Toys & Games 2.135922
Name: ReturnRate, dtype: float64
```

The return rate of the other category is high compared to the rest of the categories, which shows that they are correlated to an extent.

We can see that there is very less positive and negative correlation between returns and the products which says that there is not much high chance of returning the product based on the categories.



9) Profitability Analysis

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

Total Revenue generated from the products: 8278519.423999998

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

Top 5 products with the highest profit margins are

Description	TotalRevenue
REGENCY CAKESTAND 3 TIER	132567.70
WHITE HANGING HEART T-LIGHT HOLDER	93767.80
JUMBO BAG RED RETROSPOT	83056.52
PARTY BUNTING	67628.43
POSTAGE	66710.24

Name: TotalRevenue, dtype: float64

10) Customer Satisfaction

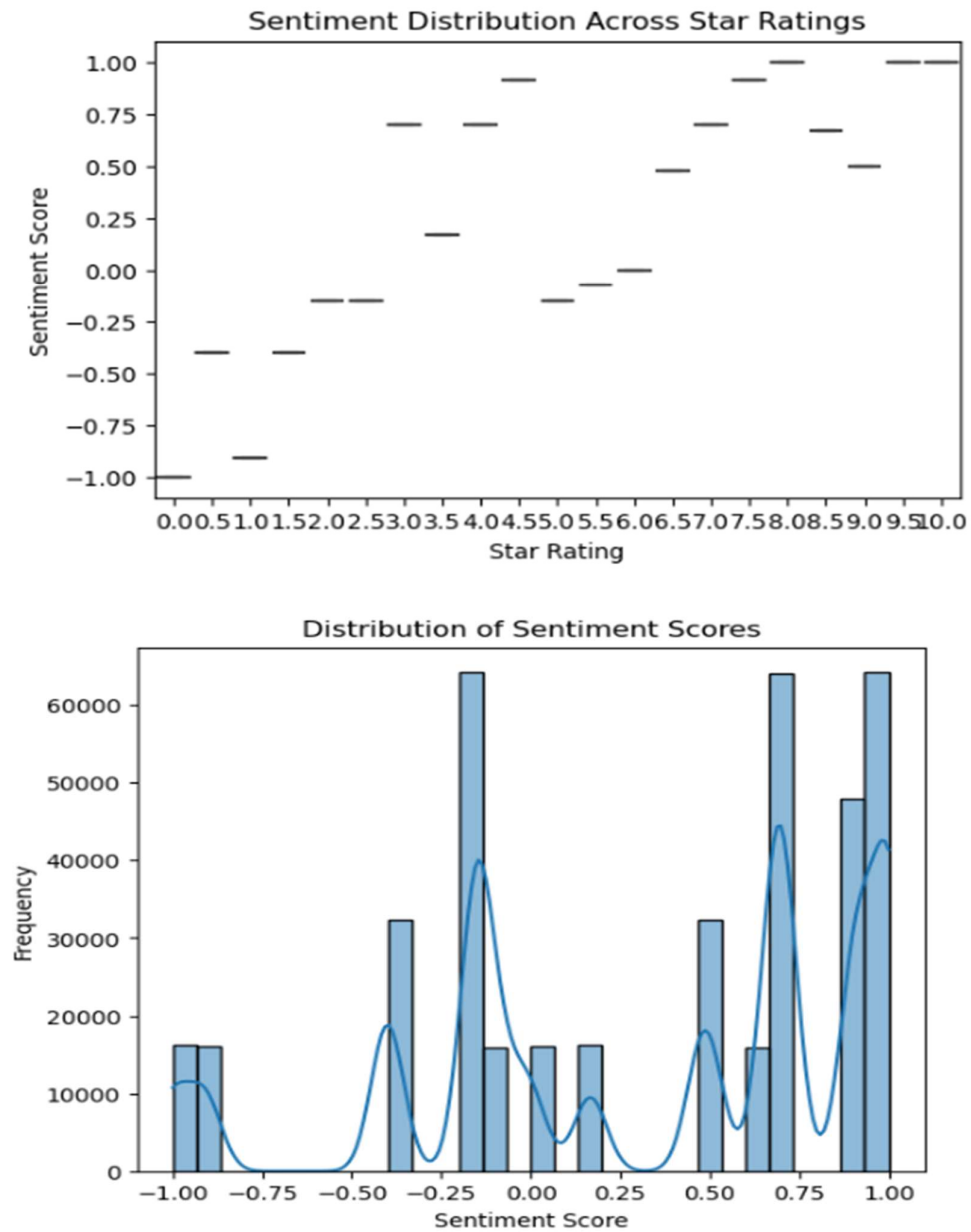
10.1 Is there any data available on customer feedback or ratings for products or services?

	StarRating	StarRatingComment
0	6.0	Satisfactory
1	3.0	Fair
2	4.0	Good
3	5.5	Above Average
4	3.5	Decent
...
541896	6.5	Pretty Good
541895	2.0	Below Average
541907	8.0	Excellent
541900	4.5	Very Good
541908	6.0	Satisfactory

[401604 rows x 2 columns]

Product with the Best Rating: BLUE/NAT SHELL NECKLACE W PENDANT
Product with the Worst Rating: ASSORTED COLOUR SILK GLASSES CASE

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



Average Sentiment: 0.316849944223663

The average sentiment score of approximately 0.32 suggests that, on average, the sentiment expressed in the feedback column is positive.

With an average sentiment score of 0.32:

The majority of the predefined comments associated with star ratings are leaning towards positive expressions. Customers, on average, use language in the comments that reflects a positive sentiment or satisfaction.

SUMMARY

Customer cluster analysis reveals three segments based on recency, frequency, and value. Cluster 1: recently active but less frequent/lower value customers - needs targeted engagement. Cluster 2: highly valuable, active, regular customers - need personalized strategies. Cluster 3: less recent, moderately frequent, lower-value customers - need reactivation efforts.

To make the most of insights:

1. Tailor marketing communication for each segment and use a multichannel approach.
2. Personalize messages, recommendations, and promotions with customer data.
3. Collect feedback from each segment for continuous improvement.

"WHITE HANGING HEART T-LIGHT HOLDER" is the most purchased product among the top 10, while "PAPER CRAFT, LITTLE BIRDIE" generates the highest revenue. Use this information to guide inventory management and promotions for better business success. Order processing time takes around 1 minute and 20 seconds between consecutive orders. Apple Pay is the most common payment method, but it has a weak correlation with order amount.

On average, customers are active for about 13 days and 18 hours. The customer base is diverse, with moderate and high purchase frequency and varying monetary values. Finally, the average sentiment score of 0.32 from customer feedback suggests a generally positive tone in the comments associated with star ratings. This indicates that customers express satisfaction and positive experiences in their feedback. These insights help understand customer behavior, preferences, and satisfaction levels, offering guidance for business strategies and improvements.

LIMITATIONS AND FUTURE WORK

It is important to take into account the various constraints of the analysis. First off, the caliber of the dataset used has a significant impact on the caliber of the outcomes. Any errors, omissions, or discrepancies in the dataset might jeopardize the reliability of the conclusions. Furthermore, the dataset's very short time span—December 1, 2010 to December 9, 2011—may make it difficult to identify long-term patterns or seasonality, which might restrict the capacity to gather thorough observations on consumer behavior. In addition, a complete profitability analysis is impeded by the lack of cost information for the items, which makes it difficult to compute profit margins and evaluate the products' overall financial health.

In order to improve the study and overcome these constraints, future work should concentrate on enhancing data quality by means of stringent cleaning and validation procedures. A more comprehensive assessment of trends and patterns might be possible by extending the dataset's coverage span. To ensure a thorough profitability analysis, real product cost information must be included. It is best to use actual payment method data rather than estimated possibilities in order to comprehend client payment preferences. In addition, adding real customer evaluations and comments might help future analyses do sentiment analysis more accurately. More sophisticated client segmentation strategies might be investigated, and dynamic product categorization techniques like natural language processing can take the place of the conventional keyword-based strategy.