Customer Segmentation using RFM Analysis

Foundations of Data Analytics - 6400

Project Report

Group - 2

Akshat Jain

Vamsi Krishna Jilla

Manoghn Kandiraju

Dhiksha Mathanagopal

Ananya Mahesh Shetty

Introduction:

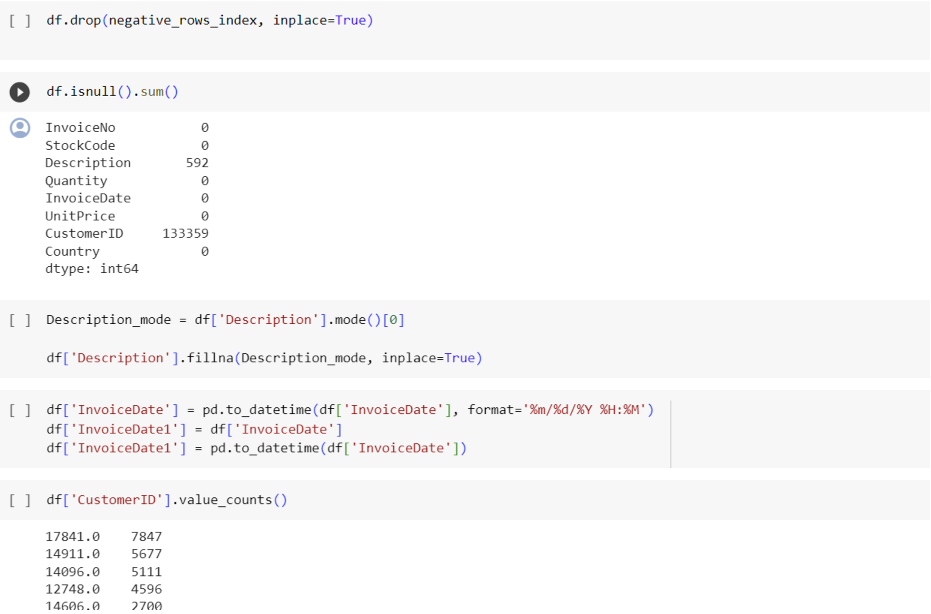
Customer Segmentation through RFM Analysis is a strategic approach that enables businesses to categorize their customer base based on three key dimensions: Recency, Frequency, and Monetary value. This method, widely used in marketing and customer relationship management, offers a nuanced understanding of customer behavior and preferences. "Recency" evaluates the time since the last customer transaction, highlighting the freshness of engagement. "Frequency" measures the number of transactions within a specific period, indicating customer loyalty and engagement level. Lastly, "Monetary" reflects the total value of a customer's transactions, emphasizing their overall contribution to revenue. By analyzing these RFM dimensions collectively, businesses can create distinct customer segments, allowing for targeted marketing strategies and personalized communication. This data-driven segmentation approach empowers organizations to tailor their efforts to meet the unique needs of each customer segment, fostering stronger customer relationships and maximizing the effectiveness of marketing initiatives.

In this project, our objective has been to perform RFM analysis on the given dataset and segment the customers into distinct groups based on their RFM scores to see if the segments will provide valuable insights for marketing and customer retention strategies.

Methods and Results:

Data Preprocessing:

After downloading the data from https://www.kaggle.com/datasets/carrie1/ecommerce-data, the data was preprocessed to remove invalid data. This included removing rows that contained negative values for Quantity and Unit Price, replacing the null values in the description column with the mode of the value of the column, changing the data type of the InvoiceDate column to an appropriate format and dropping the row with CustomerID as null.



```
[ ]  df.drop(negative_rows_index, inplace=True)
```

```
df.isnull().sum()
```

```
InvoiceNo           0
StockCode           0
Description       592
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     133359
Country             0
dtype: int64
```

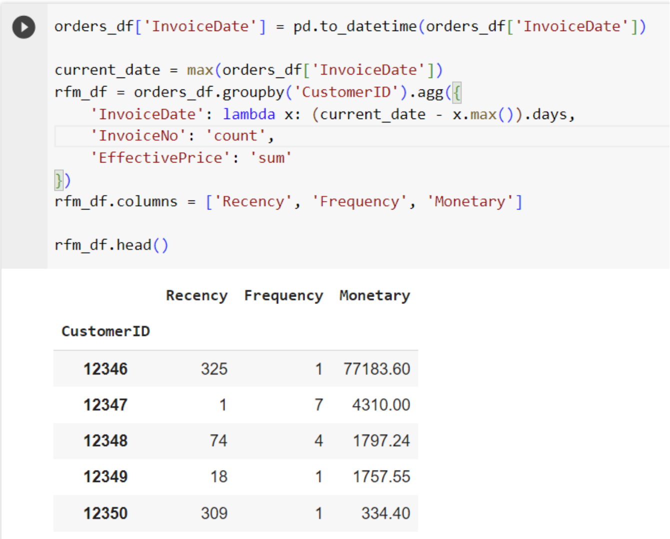
```
[ ]  Description_mode = df['Description'].mode()[0]

     df['Description'].fillna(Description_mode, inplace=True)
```

```
[ ]  df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], format='%m/%d/%Y %H:%M')
     df['InvoiceDate1'] = df['InvoiceDate']
     df['InvoiceDate1'] = pd.to_datetime(df['InvoiceDate'])
```

```
[ ]  df['CustomerID'].value_counts()

     17841.0    7847
     14911.0    5677
     14096.0    5111
     12748.0    4596
     14606.0    2700
```

RFM Calculation:

RFM metrics are calculated on the dataset, with Recency being represented using InvoiceDate, Frequency using InvoiceNo and Monetary using EffectivePrice, which is calculated by multiplying Unit Price and Quantity.

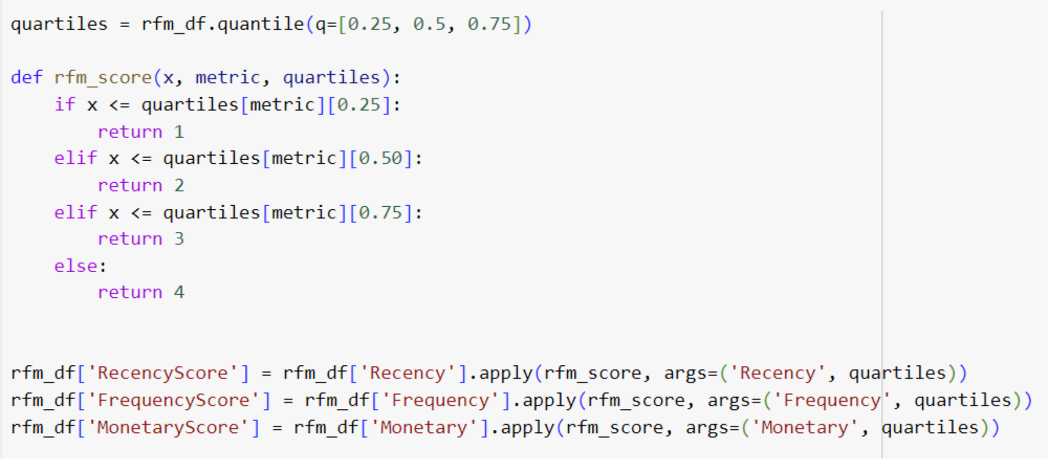


```python
orders_df['InvoiceDate'] = pd.to_datetime(orders_df['InvoiceDate'])

current_date = max(orders_df['InvoiceDate'])
rfm_df = orders_df.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (current_date - x.max()).days,
    'InvoiceNo': 'count',
    'EffectivePrice': 'sum'
})
rfm_df.columns = ['Recency', 'Frequency', 'Monetary']

rfm_df.head()
```

	Recency	Frequency	Monetary
CustomerID			
12346	325	1	77183.60
12347	1	7	4310.00
12348	74	4	1797.24
12349	18	1	1757.55
12350	309	1	334.40

RFM Segmentation:

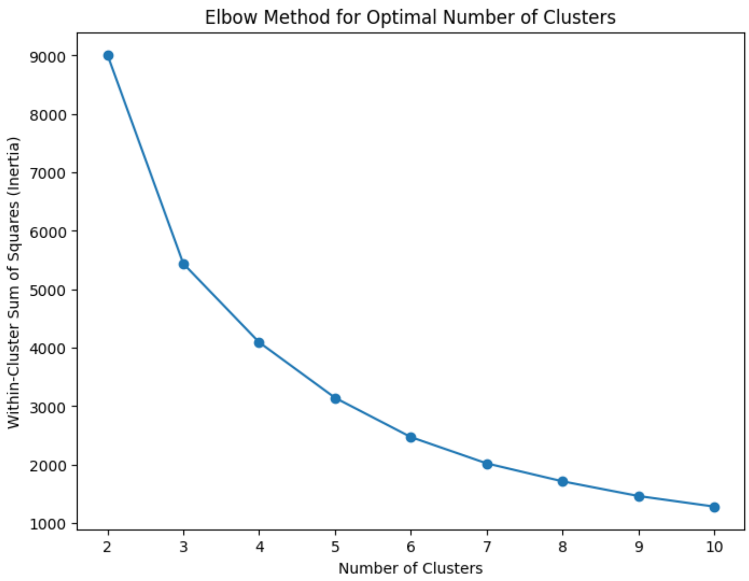
RFM Scores were assigned to each customer based on their quartiles



```python
quartiles = rfm_df.quantile(q=[0.25, 0.5, 0.75])

def rfm_score(x, metric, quartiles):
    if x <= quartiles[metric][0.25]:
        return 1
    elif x <= quartiles[metric][0.50]:
        return 2
    elif x <= quartiles[metric][0.75]:
        return 3
    else:
        return 4


rfm_df['RecencyScore'] = rfm_df['Recency'].apply(rfm_score, args=('Recency', quartiles))
rfm_df['FrequencyScore'] = rfm_df['Frequency'].apply(rfm_score, args=('Frequency', quartiles))
rfm_df['MonetaryScore'] = rfm_df['Monetary'].apply(rfm_score, args=('Monetary', quartiles))
```

Customer Segmentation:

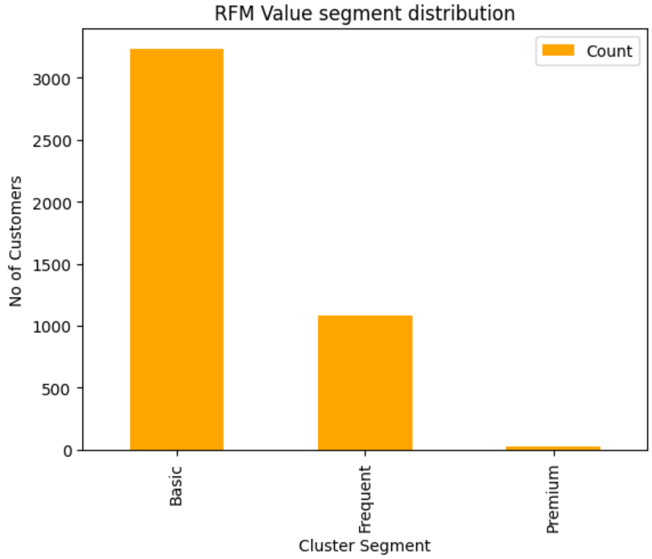
Using the Elbow Method, it was determined that 3 was that ideal number of clusters for this dataset. After that the customers were clustered into 3 clusters- Basic, Frequent and Premium based on their score.

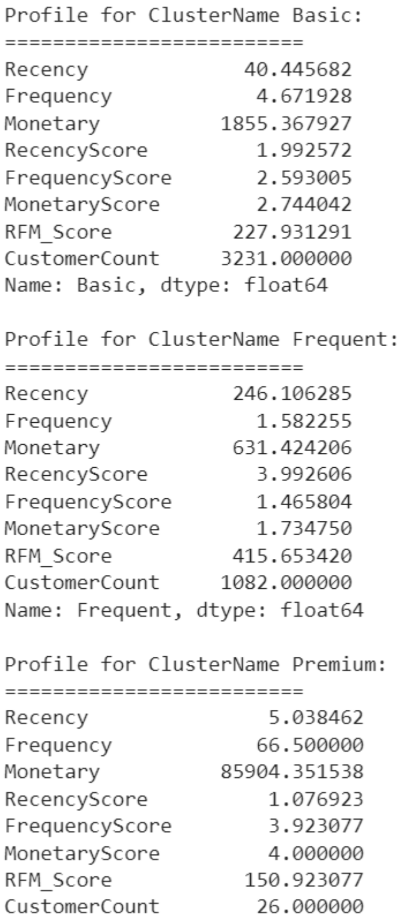


Silhouette Score for 3 clusters was found to be 0.594

Segment Profiling:

The data points in the 3 clusters were explored, and the mean RFM scores for each cluster was calculated. Each segments statistics are as followed:



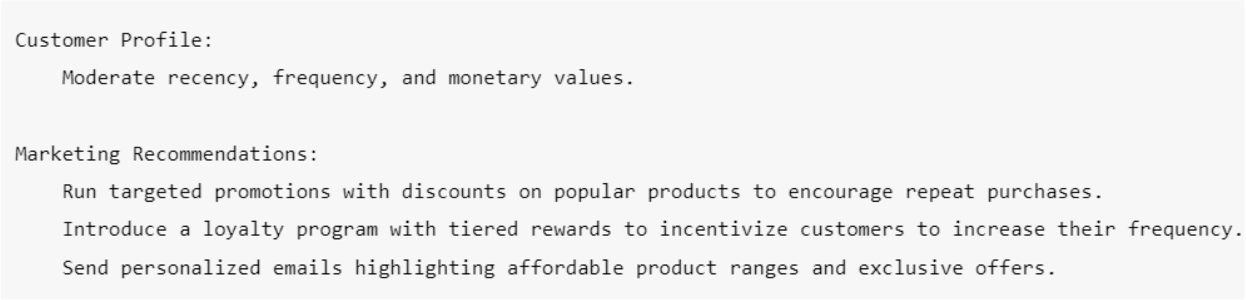


```
Profile for ClusterName Basic:
=========================
Recency              40.445682
Frequency             4.671928
Monetary           1855.367927
RecencyScore          1.992572
FrequencyScore        2.593005
MonetaryScore         2.744042
RFM_Score           227.931291
CustomerCount      3231.000000
Name: Basic, dtype: float64

Profile for ClusterName Frequent:
=========================
Recency             246.106285
Frequency             1.582255
Monetary            631.424206
RecencyScore          3.992606
FrequencyScore        1.465804
MonetaryScore         1.734750
RFM_Score           415.653420
CustomerCount      1082.000000
Name: Frequent, dtype: float64

Profile for ClusterName Premium:
=========================
Recency                5.038462
Frequency             66.500000
Monetary           85904.351538
RecencyScore           1.076923
FrequencyScore         3.923077
MonetaryScore          4.000000
RFM_Score            150.923077
CustomerCount         26.000000
```

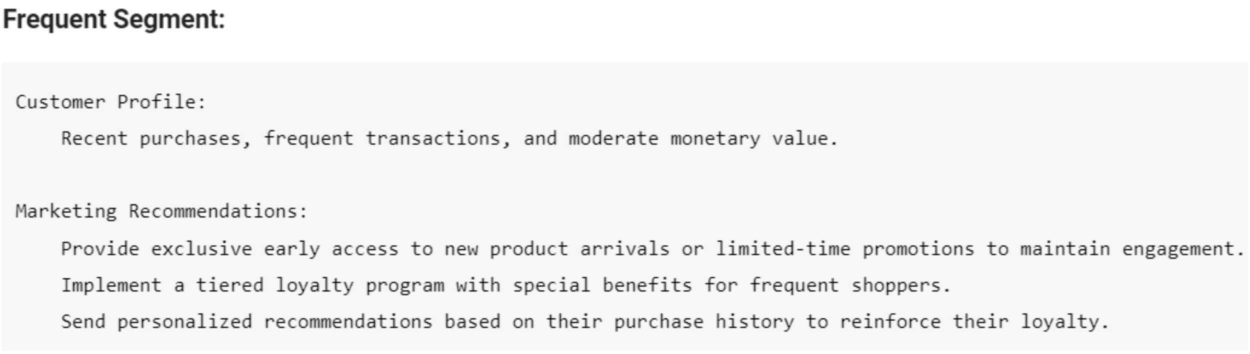
Marketing Recommendations:

Based on previous results these are the Marketing Recommendations for the 3 types of Customers:

Basic Segment:

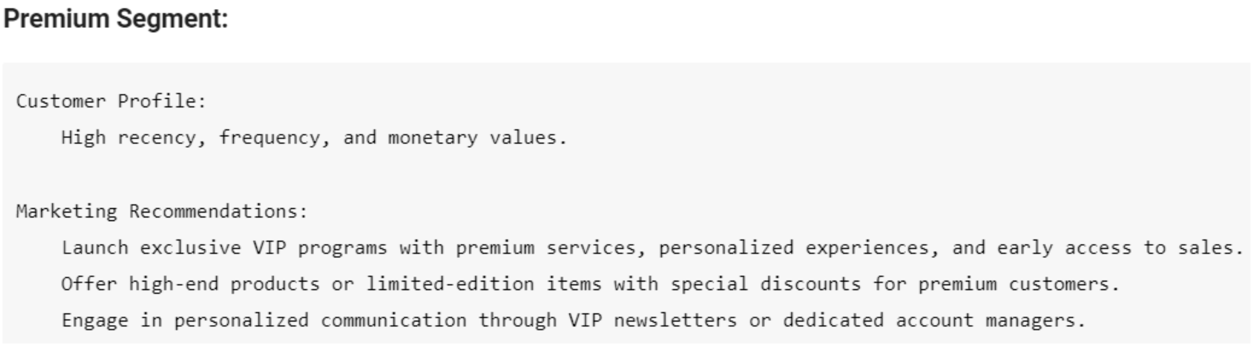


```
Customer Profile:
    Moderate recency, frequency, and monetary values.

Marketing Recommendations:
    Run targeted promotions with discounts on popular products to encourage repeat purchases.
    Introduce a loyalty program with tiered rewards to incentivize customers to increase their frequency.
    Send personalized emails highlighting affordable product ranges and exclusive offers.
```



Frequent Segment:

```
Customer Profile:
    Recent purchases, frequent transactions, and moderate monetary value.

Marketing Recommendations:
    Provide exclusive early access to new product arrivals or limited-time promotions to maintain engagement.
    Implement a tiered loyalty program with special benefits for frequent shoppers.
    Send personalized recommendations based on their purchase history to reinforce their loyalty.
```

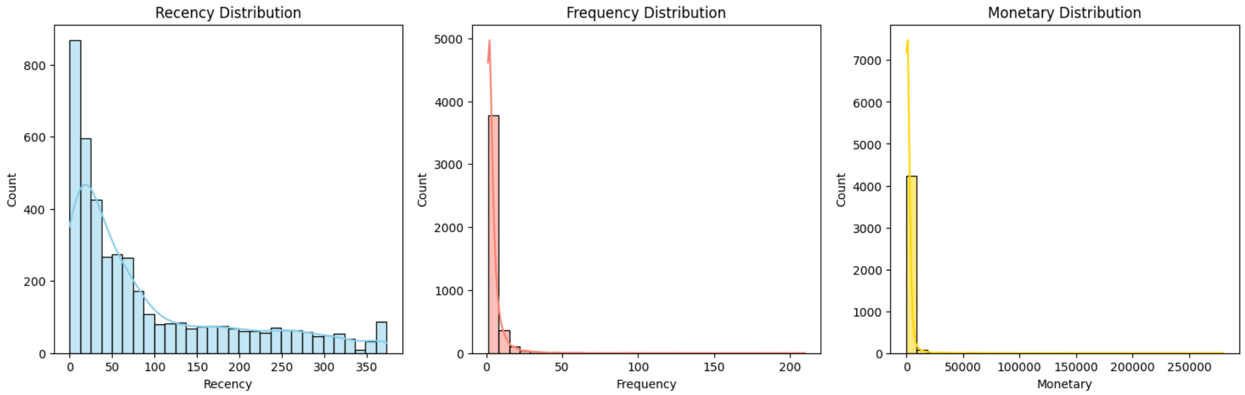


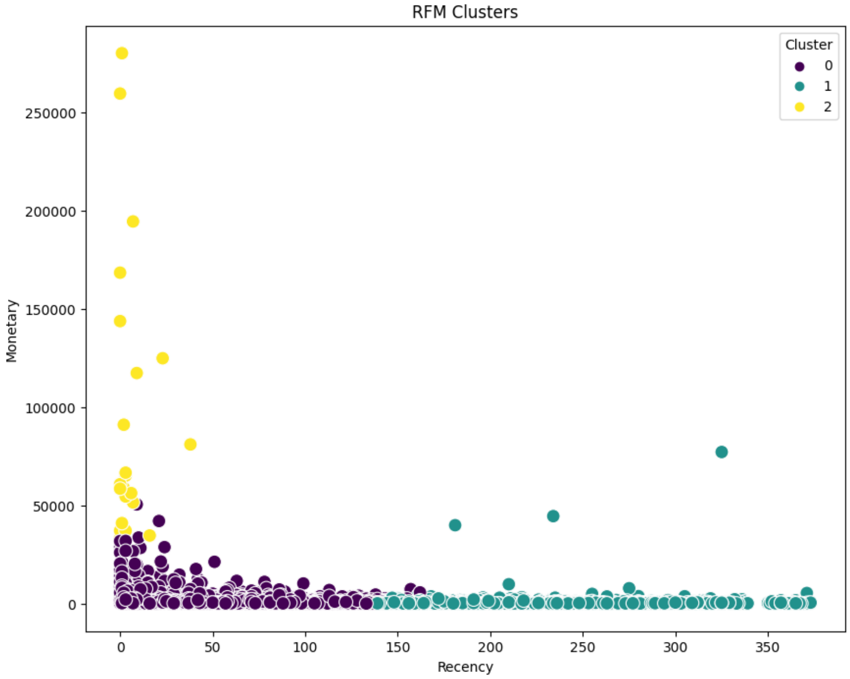
Premium Segment:

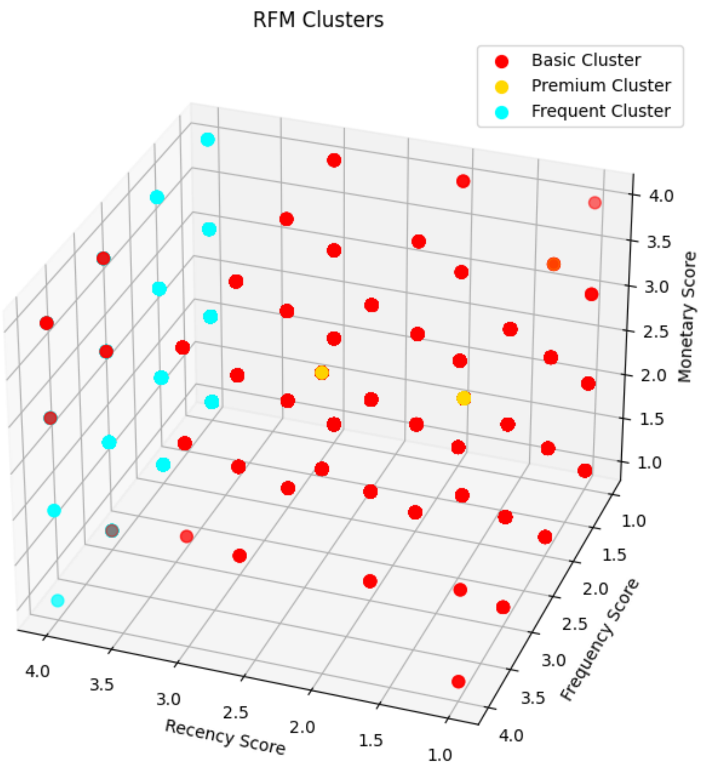
```
Customer Profile:
    High recency, frequency, and monetary values.

Marketing Recommendations:
    Launch exclusive VIP programs with premium services, personalized experiences, and early access to sales.
    Offer high-end products or limited-edition items with special discounts for premium customers.
    Engage in personalized communication through VIP newsletters or dedicated account managers.
```

Visualization:

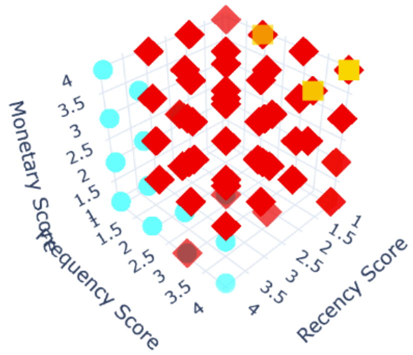
The following graphs and visualizations explore the data and the findings in various methods:

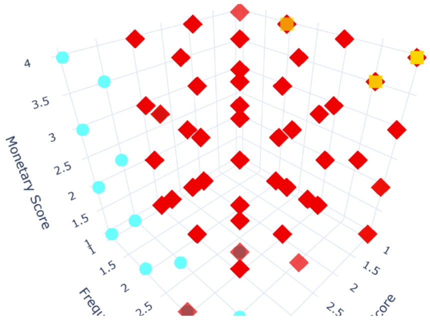

Recency Distribution
Count
800
600
400
200
0
0
50
100
150
200
250
300
350
Recency
Frequency Distribution
Count
5000
4000
3000
2000
1000
0
0
50
100
150
200
Frequency
Monetary Distribution
Count
7000
6000
5000
4000
3000
2000
1000
0
0
50000
100000
150000
200000
250000
Monetary


RFM Clusters
Monetary
250000
200000
150000
100000
50000
0
0
50
100
150
200
250
300
350
Recency
Cluster
0
1
2



The 3 Dimensional plot of the clusters was also explored:





Questions:

Data Overview:

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

```
Number of rows: 541909
Number of columns: 18
```

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

```
Column Descriptions:
InvoiceNo                   object
StockCode                   object
Description                 object
Quantity                     int64
InvoiceDate         datetime64[ns]
UnitPrice                  float64
CustomerID                 float64
Country                     object
InvoiceDate1        datetime64[ns]
EffectivePrice             float64
TotalPrice                 float64
TotalRevenue               float64
DayOfWeek                   object
HourOfDay                    int64
Month                        int64
Season                      object
TotalOrderValue            float64
Payment Method              object
dtype: object
```

3) What is the time period covered by this dataset?

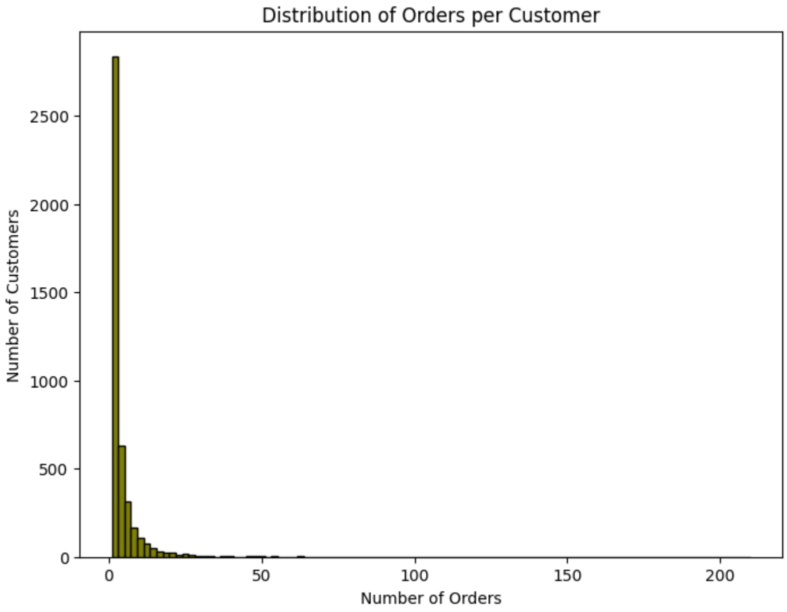
```
Time period covered by the dataset:
Start date: 2010-12-01 08:26:00
End date: 2011-12-09 12:50:00
Time Period Covered: 373 days 04:24:00
```

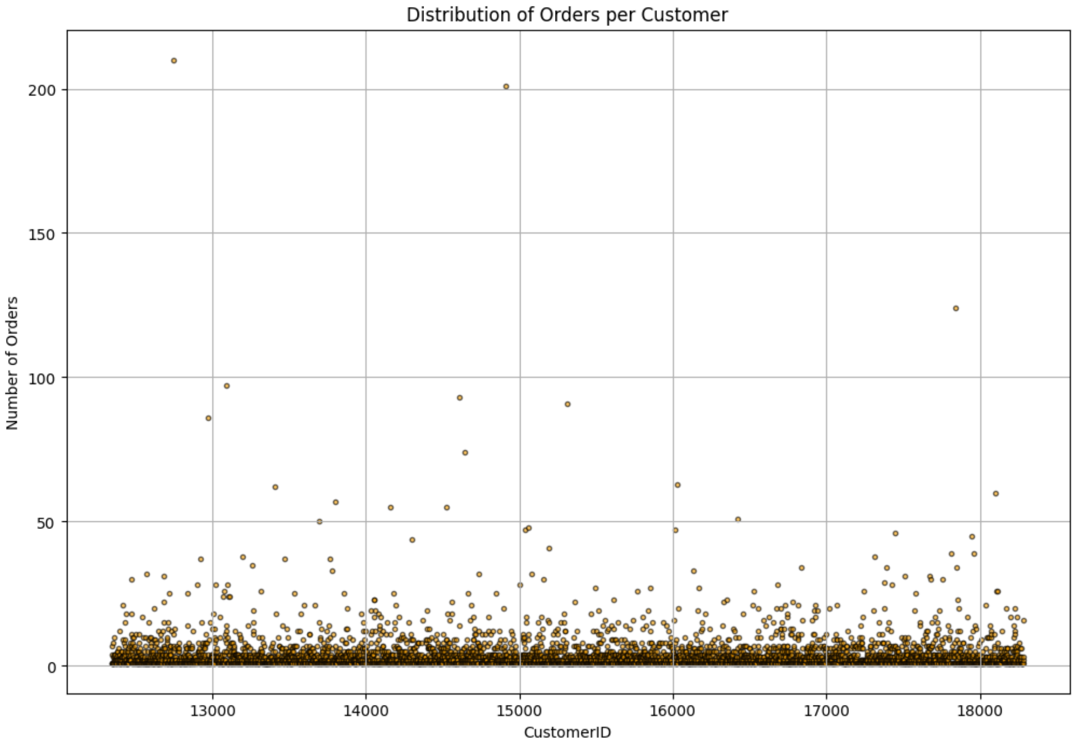
Customer Analysis:

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

```
Number of unique customers is 4339
```

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





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

```
Top 5 Customers with the Most Purchases by Order Count:
CustomerID
12748    210
14911    201
17841    124
13089     97
14606     93
```

Product Analysis:

1) What are the top 10 most frequently purchased products?

```
Top 10 Most Frequently Purchased Products:
                          Description  TotalQuantity
0         PAPER CRAFT , LITTLE BIRDIE          80995
1      MEDIUM CERAMIC TOP STORAGE JAR          77916
2   WORLD WAR 2 GLIDERS ASSTD DESIGNS          54415
3             JUMBO BAG RED RETROSPOT          46181
4  WHITE HANGING HEART T-LIGHT HOLDER          36725
5       ASSORTED COLOUR BIRD ORNAMENT          35362
6     PACK OF 72 RETROSPOT CAKE CASES          33693
7                      POPCORN HOLDER          30931
8                  RABBIT NIGHT LIGHT          27202
9             MINI PAINT SET VINTAGE           26076
```

2) What is the average price of products in the dataset?

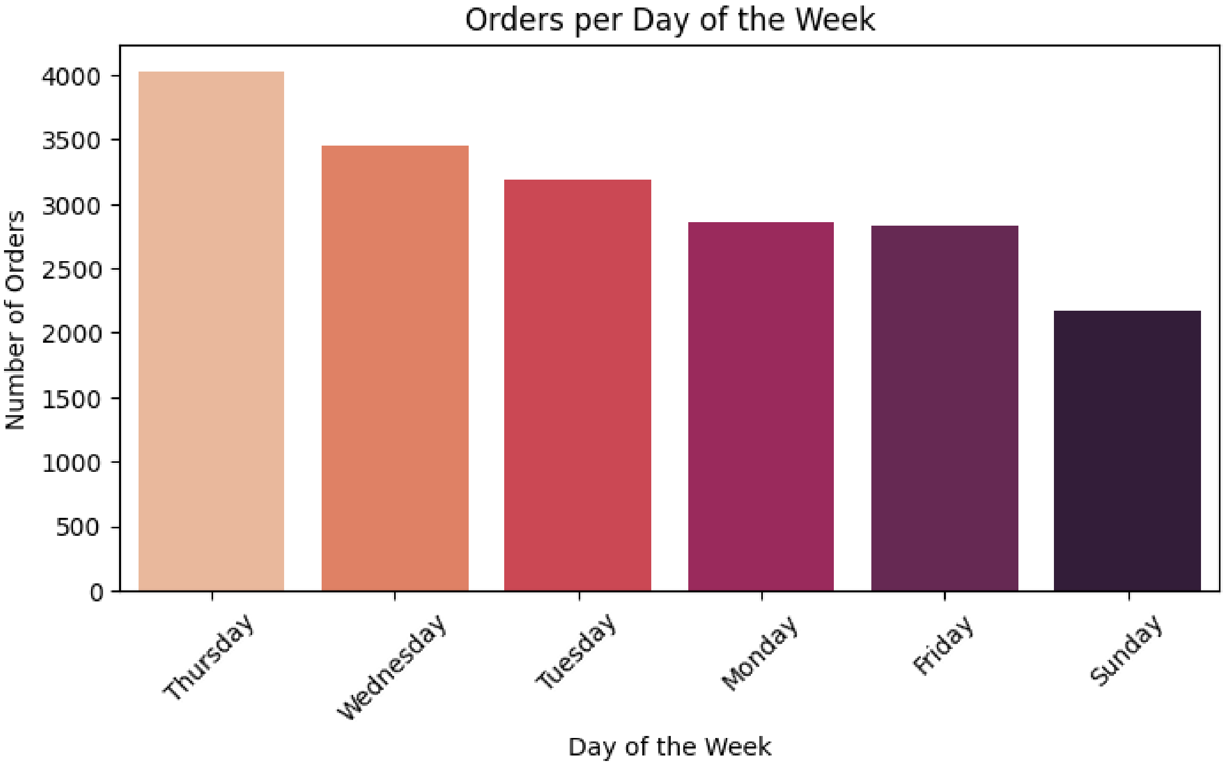
```
Average price of products: 3.1161744805540756
```

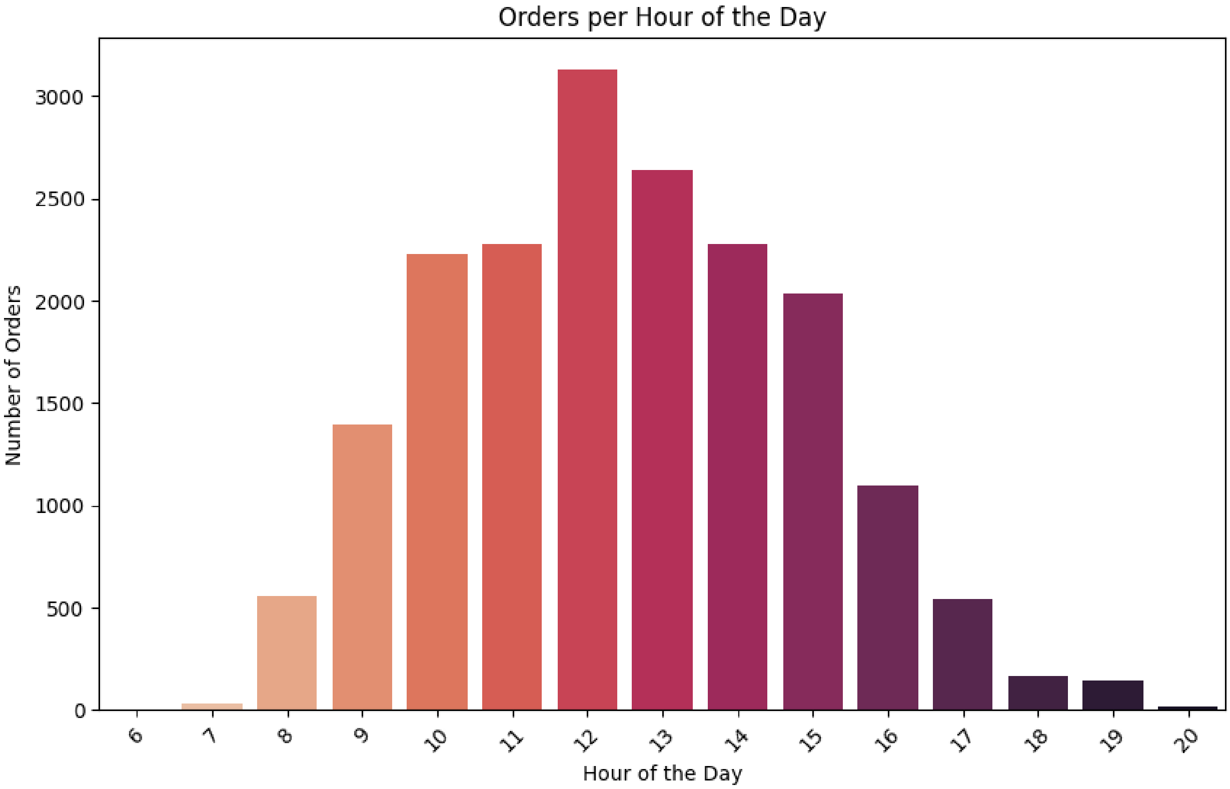
3) Can you find out which product category generates the highest revenue?

```
The product 'PAPER CRAFT , LITTLE BIRDIE' generates the highest revenue
with a total of 168469.60 $
```

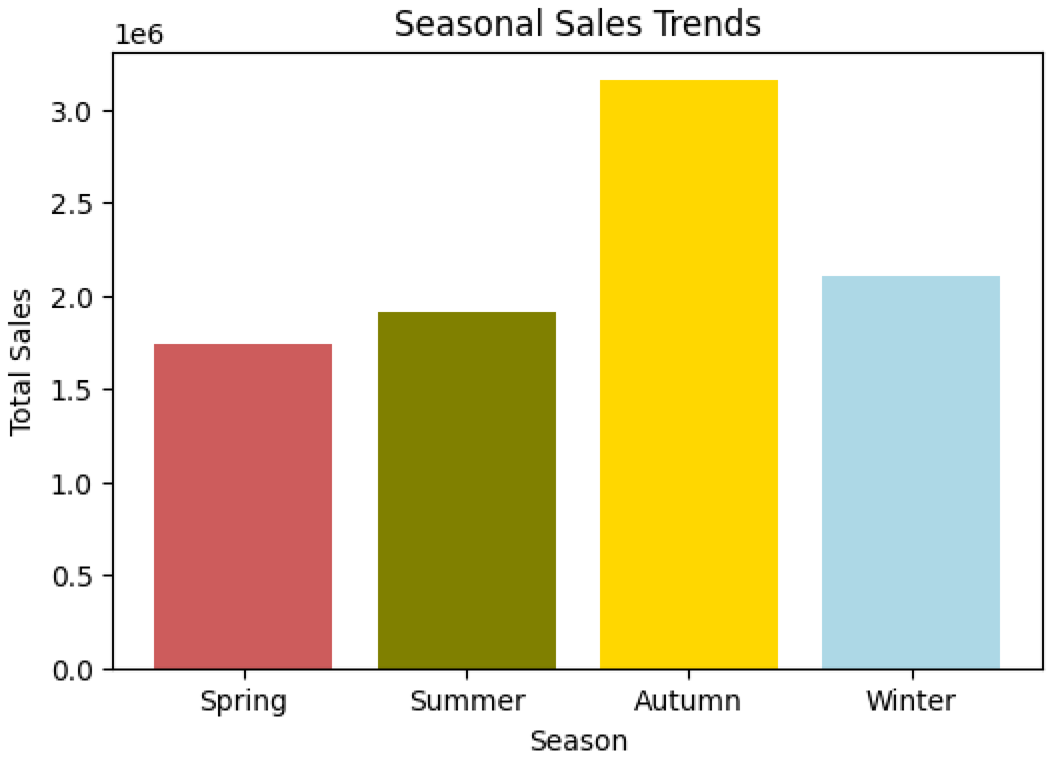
Time Analysis:

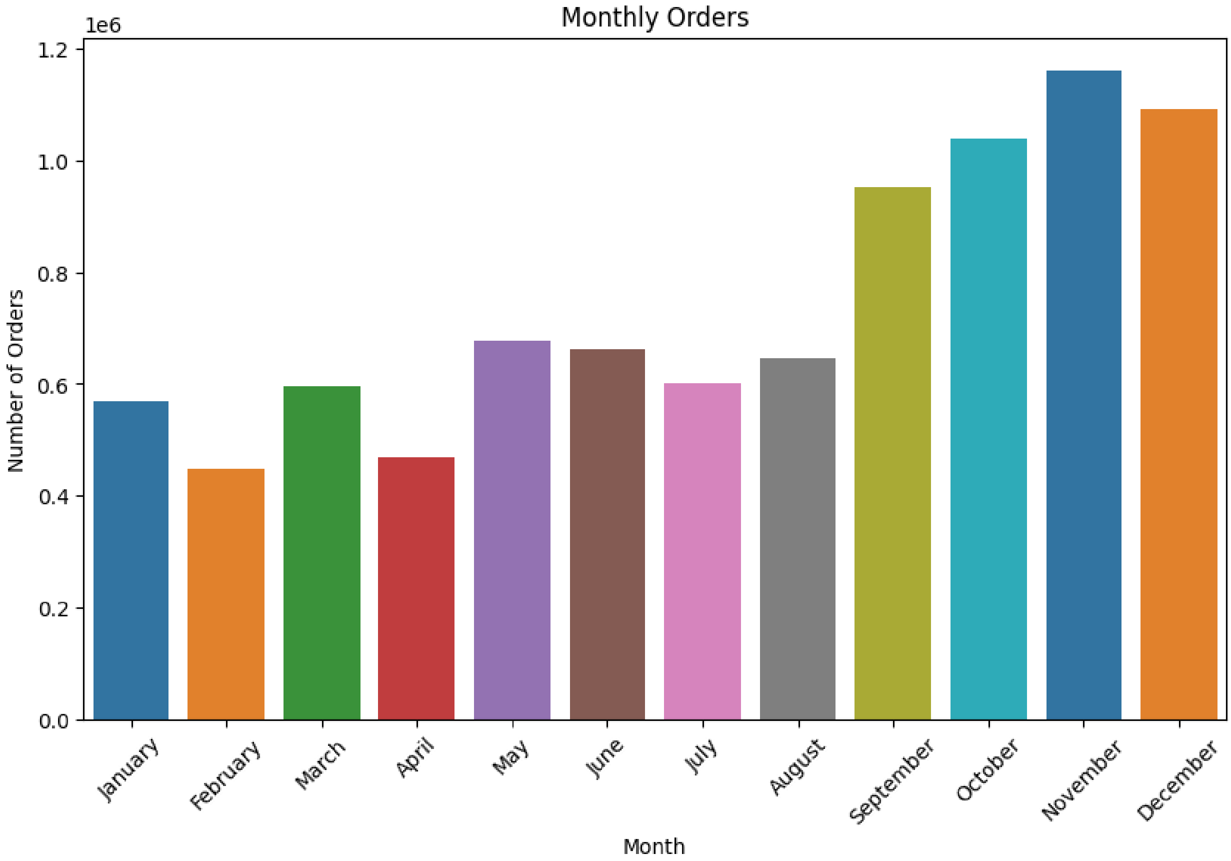
1) Is there a specific day of the week or time of day when most orders are placed?


Orders per Day of the Week
Number of Orders
4000
3500
3000
2500
2000
1500
1000
500
0
Thursday
Wednesday
Tuesday
Monday
Friday
Sunday
Day of the Week


Orders per Hour of the Day
Number of Orders
3000
2500
2000
1500
1000
500
0
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Hour of the Day

2) Are there any seasonal trends in the dataset?





Geographical Analysis:

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

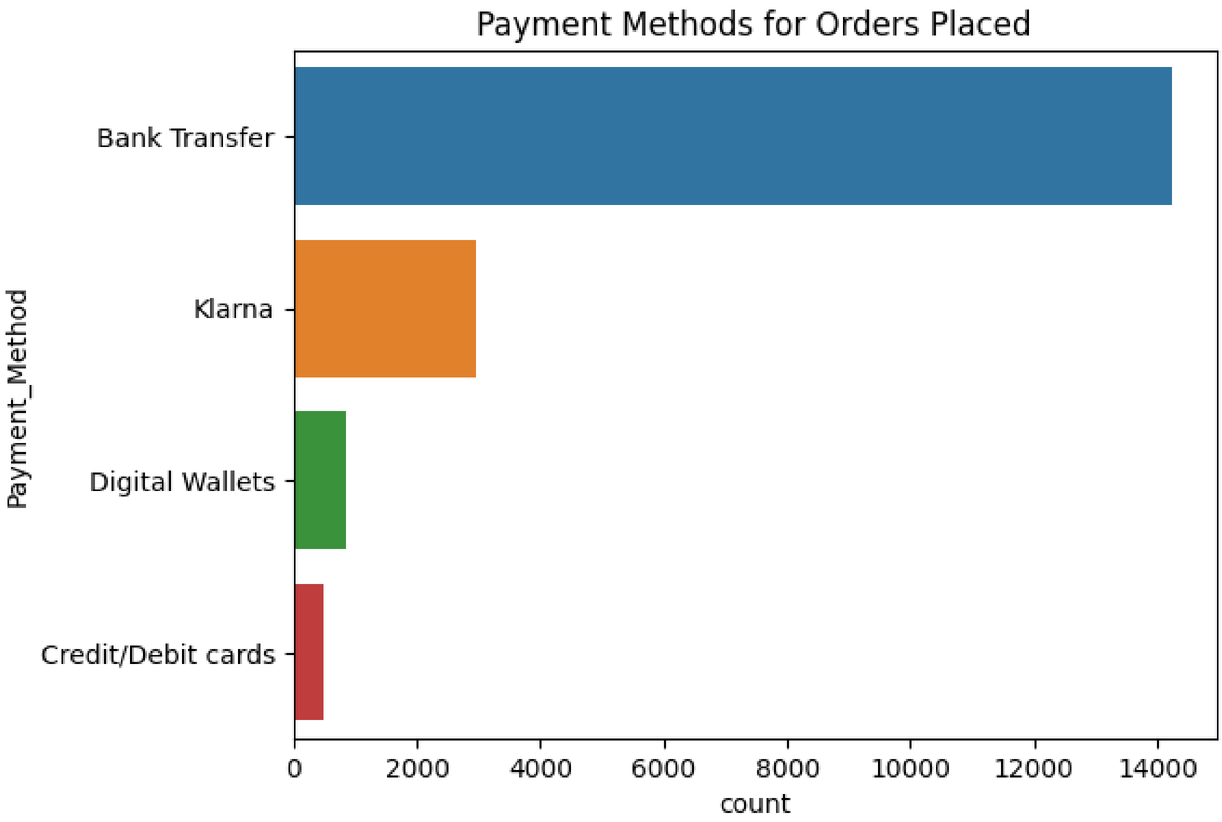
```
United Kingdom    16649
Germany             457
France              389
EIRE                260
Belgium              98
```

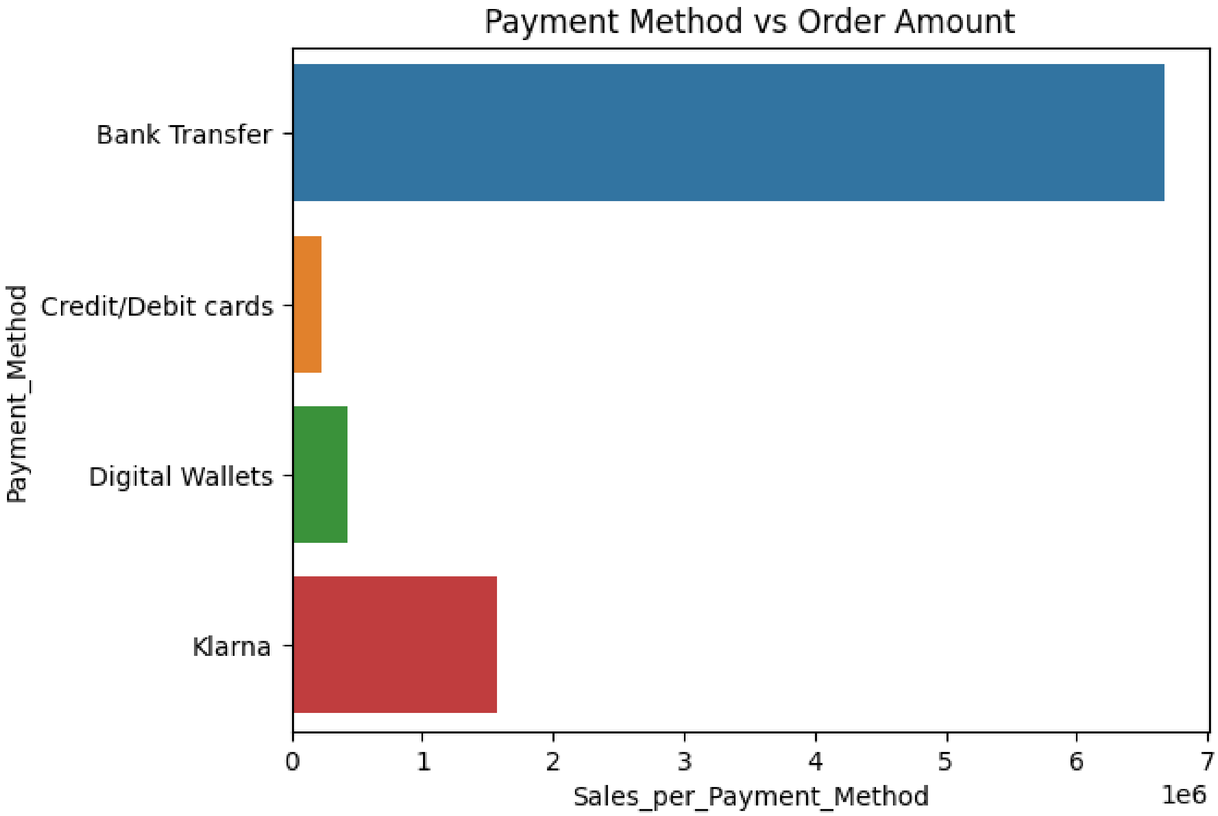
2) Is there a correlation between the country of the customer and the average order value?

```
Correlation between Country and Average Order Value: -0.11627391172320647
This value indicates a weak negative correlation
```

Payment Analysis:

1) What are the most common payment methods used by customers?


Payment Methods for Orders Placed
Payment_Method
Bank Transfer
Klarna
Digital Wallets
Credit/Debit cards
0
2000
4000
6000
8000
10000
12000
14000
count


Payment Method vs Order Amount
Payment_Method
Bank Transfer
Credit/Debit cards
Digital Wallets
Klarna
0
1
2
3
4
5
6
7
1e6
Sales_per_Payment_Method

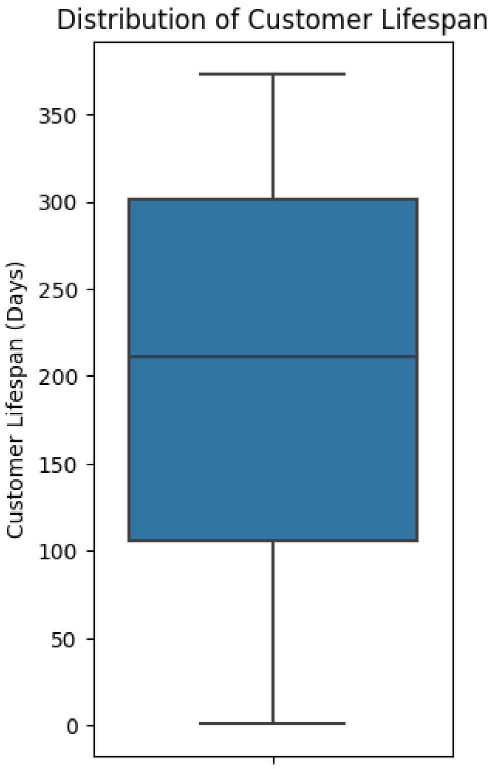
2) Is there a relationship between the payment method and the order amount?

```
Correlation between Payment Method and Order Amount: -0.6449052733694569
```

This strong negative correlation suggests that there is a significant relationship between the payment method and the order amount

Customer Behavior:

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



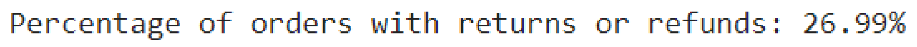
```
Average Customer Lifespan is 203.34 days
```

2) Are there any customer segments based on their purchase behavior?

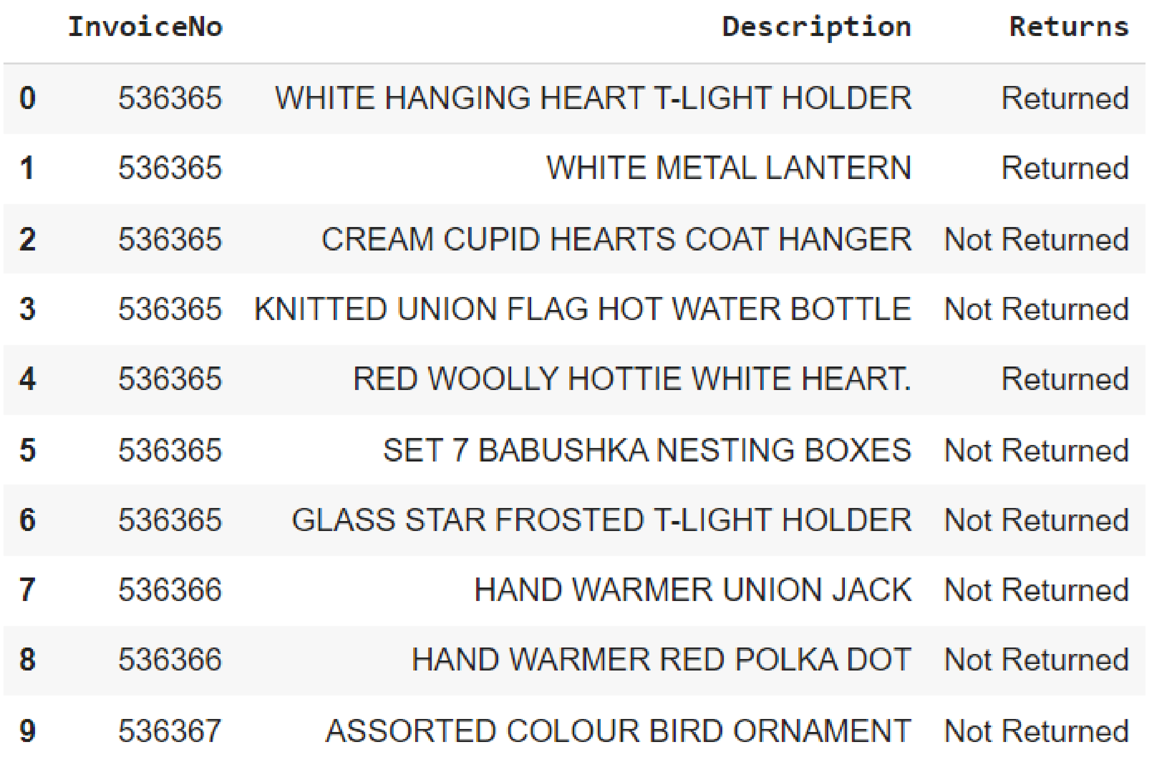
1. **Basic Segment:**
   - These customers have made purchases with moderate recency, frequency, and monetary values.
   - Recommendations: Engage with personalized promotions to encourage more frequent purchases and enhance loyalty.
2. **Frequent Segment:**
   - Customers in this segment exhibit high frequency and recency, indicating regular and recent purchases.
   - Recommendations: Reward loyalty with exclusive offers, loyalty programs, or early access to new products to maintain their engagement.
3. **Premium Segment:**
   - This segment comprises customers with high recency, frequency, and monetary values, suggesting significant spending.
   - Recommendations: Provide premium services, personalized recommendations, and exclusive perks to maximize their spending and enhance their overall experience.

Returns and Refunds:

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



```
Percentage of orders with returns or refunds: 26.99%
```



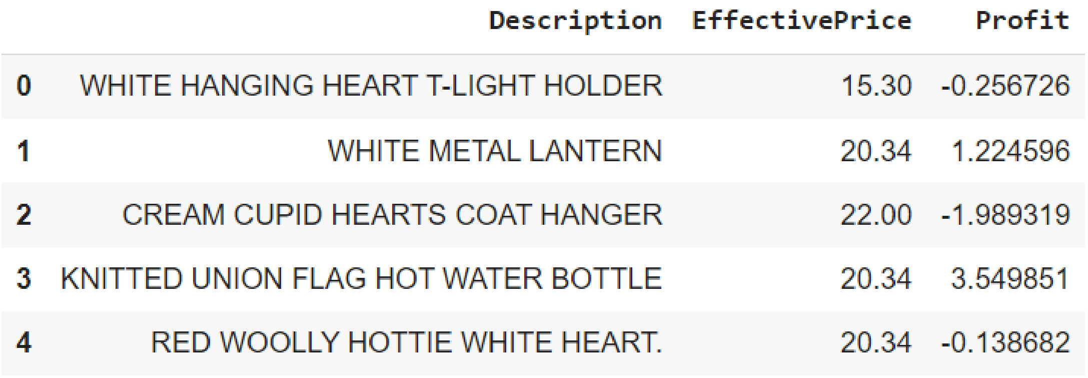
	InvoiceNo	Description	Returns
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	Returned
1	536365	WHITE METAL LANTERN	Returned
2	536365	CREAM CUPID HEARTS COAT HANGER	Not Returned
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	Not Returned
4	536365	RED WOOLLY HOTTIE WHITE HEART.	Returned
5	536365	SET 7 BABUSHKA NESTING BOXES	Not Returned
6	536365	GLASS STAR FROSTED T-LIGHT HOLDER	Not Returned
7	536366	HAND WARMER UNION JACK	Not Returned
8	536366	HAND WARMER RED POLKA DOT	Not Returned
9	536367	ASSORTED COLOUR BIRD ORNAMENT	Not Returned

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

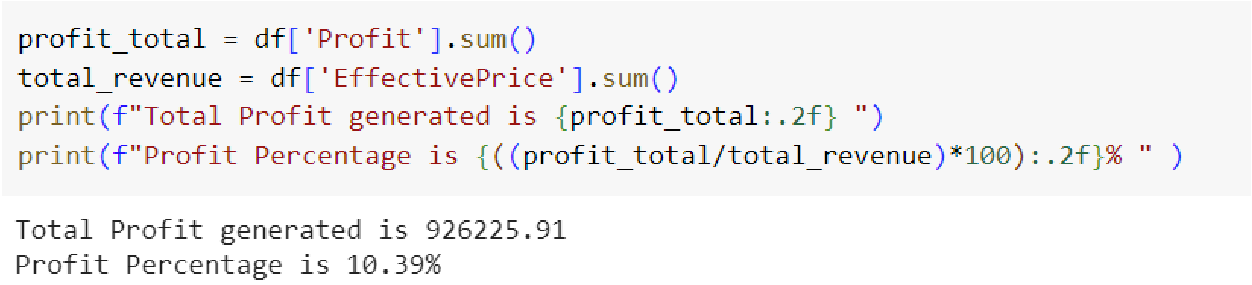
```
Chi-squared value: 103.3
P-value: 0.36
There is no significant correlation between product category and returns
```

Profitability Analysis:

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



	Description	EffectivePrice	Profit
0	WHITE HANGING HEART T-LIGHT HOLDER	15.30	-0.256726
1	WHITE METAL LANTERN	20.34	1.224596
2	CREAM CUPID HEARTS COAT HANGER	22.00	-1.989319
3	KNITTED UNION FLAG HOT WATER BOTTLE	20.34	3.549851
4	RED WOOLLY HOTTIE WHITE HEART.	20.34	-0.138682



```
profit_total = df['Profit'].sum()
total_revenue = df['EffectivePrice'].sum()
print(f"Total Profit generated is {profit_total:.2f} ")
print(f"Profit Percentage is {((profit_total/total_revenue)*100):.2f}% " )
```

```
Total Profit generated is 926225.91
Profit Percentage is 10.39%
```

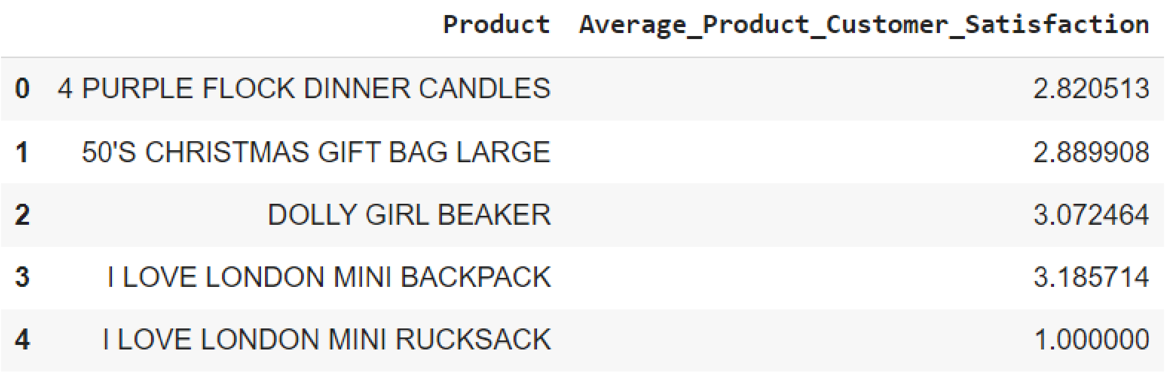
2) What are the top 5 products with the highest profit margins?

```
Top 5 products with the highest profit margins are
2466          PINK FLOCK PHOTO FRAME
2657                     RAIN PONCHO
3523      UNION JACK HOT WATER BOTTLE
3866             ZINC PLANT POT HOLDER
2512        PINK PAINTED KASHMIRI CHAIR
```

Customer Satisfaction:

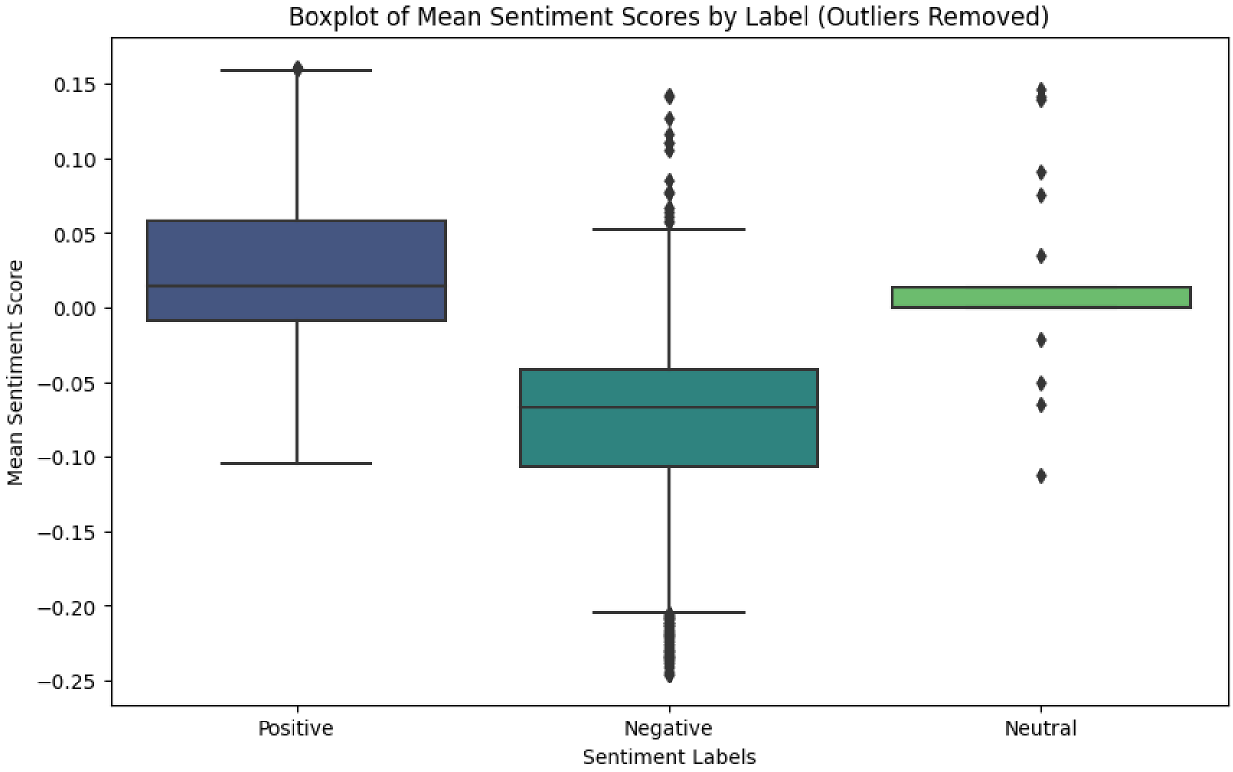
1) Is there any data available on customer feedback or ratings for products or services?

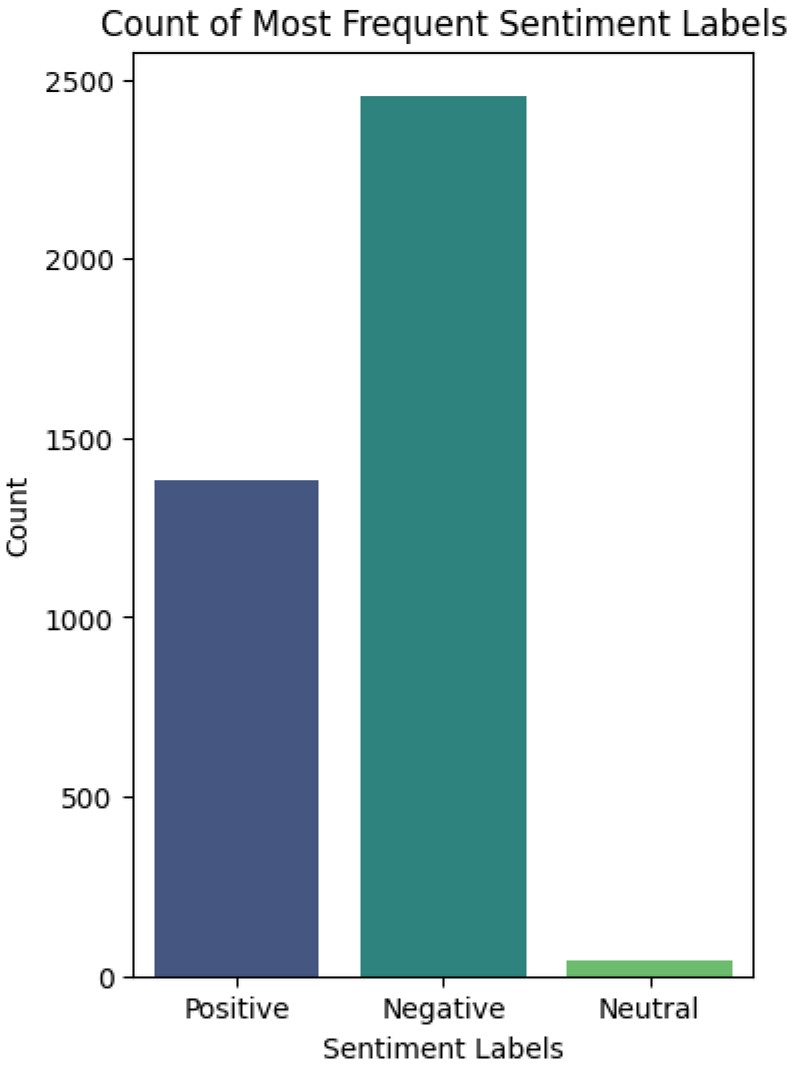
Data on Customer Satisfaction was generated using random customer feedback



	Product	Average_Product_Customer_Satisfaction
0	4 PURPLE FLOCK DINNER CANDLES	2.820513
1	50'S CHRISTMAS GIFT BAG LARGE	2.889908
2	DOLLY GIRL BEAKER	3.072464
3	I LOVE LONDON MINI BACKPACK	3.185714
4	I LOVE LONDON MINI RUCKSACK	1.000000

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




Count of Most Frequent Sentiment Labels
Count
2500
2000
1500
1000
500
0
Positive
Negative
Neutral
Sentiment Labels