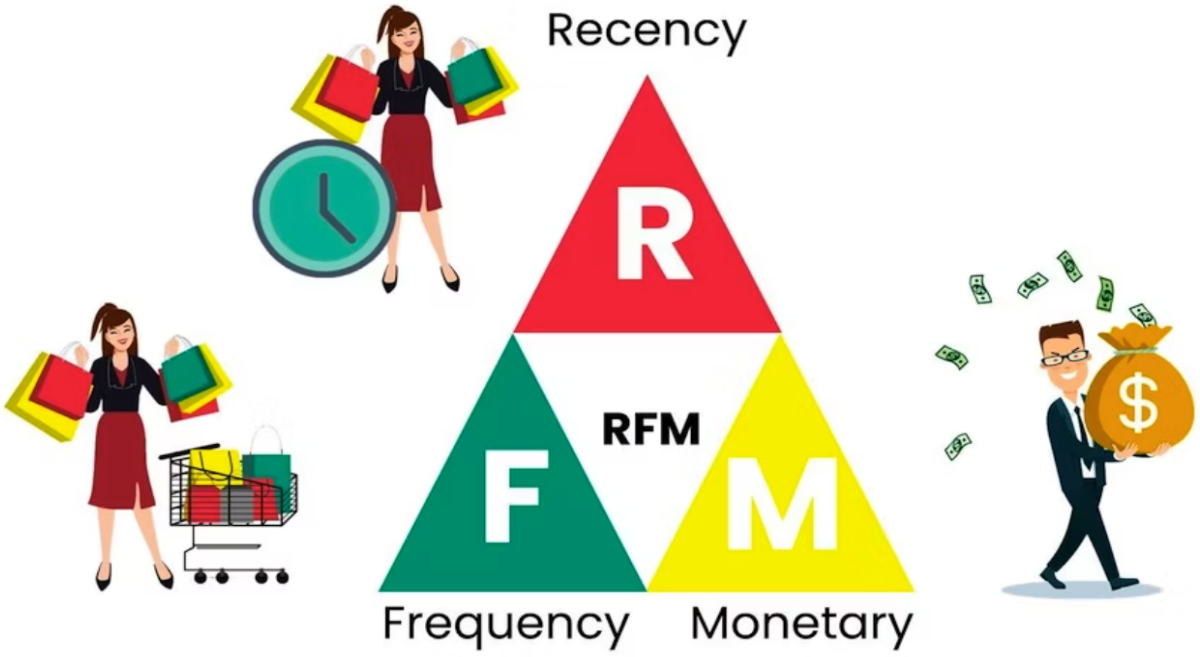


Retail Realms through RFM Analysis

IE6400 Foundations Data Analytics Engineering

Group Number 3

Jahnavi Mishra (002724552)

Sahita Bonthu (002823336)

Namrit Sheth (002244393)

Part 1: Introduction

In the dynamic landscape of the retail industry, this dataset serves as a valuable repository capturing key transactional details. From purchase quantities and unit prices to customer identifiers and corresponding countries, these records encapsulate essential insights into customer interactions within the online commerce realm. The application of RFM (Recency, Frequency, Monetary) analysis to this dataset presents a significant opportunity for unraveling actionable insights. By delving into the recency, frequency, and monetary dimensions of customer transactions, businesses can meticulously segment their customer base. This segmentation facilitates the formulation of targeted marketing strategies and personalized engagement initiatives. Effectively employing RFM analysis in this industry not only enhances customer retention but also provides a strategic advantage, empowering businesses to navigate the competitive online retail landscape with data-driven precision, fostering informed decision-making, and promoting sustainable growth.

Dataset Utilization: Leveraged a comprehensive e-commerce dataset from Kaggle as the foundation for the analysis.

RFM Analysis: Applied the RFM (Recency, Frequency, Monetary) model to dissect customer behavior.

Quantification of Customer Behavior: Quantified recency of purchases, frequency of transactions, and monetary value of transactions to understand customer engagement.

Data Preprocessing: Conducted meticulous data preprocessing, including cleaning, handling missing values, and converting data types for analysis readiness.

RFM Metric Calculation: Calculated Recency, Frequency, and Monetary metrics for each customer to create a robust customer segmentation model.

Customer Segmentation: Employed clustering techniques, such as K-Means clustering, to segment customers based on their RFM scores.

Optimal Cluster Determination: Experimented with different numbers of clusters to identify the optimal configuration for meaningful segmentation.

Segment Profiling: Analyzed and profiled each customer segment, describing their characteristics based on RFM scores and other relevant attributes.

Marketing Recommendations: Provided actionable marketing recommendations for each customer segment to enhance retention and maximize revenue.

Part 2: Data Sources

The main portion of data used in this analysis was sourced from catalog.data.gov.

The table contains various columns providing information related to the retail industry. The table contains the following :

Invoice Number (InvoiceNo):

Each transaction is uniquely identified by an Invoice Number. This identifier serves as a key reference point for tracking and analyzing individual transactions. The uniqueness of each InvoiceNo allows for precise monitoring of customer transactions and order details.

Stock Code (StockCode):

The Stock Code corresponds to a specific product or stock item involved in a transaction. This code aids in categorizing and organizing the diverse range of products available in the retail inventory. It plays a crucial role in understanding which products are popular, frequently purchased, or contribute significantly to overall sales.

Product Description (Description):

The Product Description column provides a concise yet informative overview of the products or stock items involved in each transaction. This information is valuable for understanding the nature of the products sold and aids in subsequent analysis of customer preferences and popular items.

Quantity Purchased (Quantity):

The Quantity column denotes the number of units of a particular product purchased in a given transaction. Tracking quantity data is essential for assessing demand patterns, identifying popular products, and managing inventory effectively.

Invoice Date (InvoiceDate):

The Invoice Date records the date and time when a transaction was initiated. This temporal information is crucial for conducting time-based analyses, such as identifying peak sales periods, seasonal trends, or assessing the impact of time on customer purchasing behavior.

Unit Price (UnitPrice):

Unit Price represents the cost of a single unit of the product involved in the transaction. This information is instrumental in calculating the total monetary value of each transaction and, consequently, understanding the financial aspects of customer interactions.

Customer ID:

Customer ID is a unique identifier assigned to each customer. This key enables the linkage of transactions to specific customers, facilitating the analysis of individual customer behavior, purchase history, and overall customer segmentation.

Country:

The Country column specifies the geographic location where each transaction occurred. Understanding the distribution of sales across different countries is essential for tailoring marketing strategies, recognizing global trends, and addressing any geographical variations in customer behavior.

Part 3: Results and Methods

In the data preprocessing workflow, we need to import the dataset into Python using Pandas and then explore its structure with functions like head(), info(), and describe(). Next, we handle missing values using techniques such as removal, and address data quality issues through cleaning steps like removing duplicates and handling outliers. Ensuring the correct data types for each column is crucial using functions like astype() for conversion. If there are numerical features with different scales, we apply feature scaling. Categorical variables are transformed into a numerical format, often through one-hot encoding or label encoding.This comprehensive approach ensures the data is cleaned, transformed, and ready for effective analysis or machine learning model training.The resulting dataset,referred to as “proj2”,now includes standardized numerical attributes and encoded categorical variables, rendering it well-prepared for the purposes of Customer Segmentation using RFM Analysis in our project.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

We calculate the Recency (R) metric for each customer in a dataset ‘proj2’. It first determines the maximum date of any purchase in the dataset (max_date). Then, it groups the data by customer ID, retrieves the latest purchase date for each customer, and calculates the number of days since their last purchase. The results are stored in a DataFrame (customer_recency). Finally, the DataFrame is sorted in descending order based on recency, providing a view of customers with the longest time since their last purchase. The code effectively computes and organizes the Recency metric for customer segmentation or analysis.

	CustomerID	InvoiceDate	Days_since_Lpurch
1046	13747	2010-12-01 10:37:00	373
3129	16583	2010-12-01 12:03:00	373
4096	17908	2010-12-01 11:45:00	373
1764	14729	2010-12-01 12:43:00	373
359	12791	2010-12-01 11:27:00	373
...	...	...	...
301	12713	2011-12-09 12:16:00	0
698	13263	2011-12-08 13:59:00	0
105	12476	2011-12-08 14:08:00	0
900	13536	2011-12-08 16:38:00	0
516	13013	2011-12-08 15:01:00	0

4372 rows × 3 columns

Now,We calculate the Frequency (F) metric for each customer in the dataset ‘proj2C’. It groups the data by customer ID, counts the unique number of invoices (representing orders) for each customer, and stores the result in a DataFrame (customer_frequency). The 'InvoiceNo' column is renamed to 'total_no_of_orders' for clarity, and the DataFrame is then sorted in descending order based on the total number of orders. This process effectively computes and organizes the Frequency metric, representing how often each customer makes a purchase. The resulting DataFrame provides insights into customer purchasing behavior, which can be useful for segmentation or further analysis.

	CustomerID	total_no_of_orders
1895	14911	248
330	12748	224
4042	17841	169
1674	14606	128
568	13089	118
...	...	...
1141	13877	1
2997	16400	1
1142	13878	1
1149	13886	1
990	13670	1

4372 rows × 2 columns

Next,we calculate the Monetary (M) metric for each customer in the dataset ‘proj2C’. It computes the total monetary value of a customer's purchases by multiplying the quantity and unit price for each item and summing these values for each customer. The resulting information is stored in a DataFrame (customer_monetary). The DataFrame includes customer IDs and their corresponding total monetary values, sorted in descending order based on monetary value. This Monetary metric is valuable for assessing the financial contribution of each customer and is commonly used in customer segmentation or analysis, providing insights into customer purchasing power.

	CustomerID	TotalPrice
1703	14646	279489.02
4233	18102	256438.49
3758	17450	187322.17
1895	14911	132458.73
55	12415	123725.45
...	...	...
125	12503	-1126.00
3870	17603	-1165.30
1384	14213	-1192.20
2236	15369	-1592.49
3756	17448	-4287.63

4372 rows × 2 columns

The merged three DataFrames (customer_recency, customer_frequency, and customer_monetary) based on customer ID to create an rfm_metrics DataFrame for RFM analysis.

	CustomerID	Days_since_Lpurch	total_no_of_orders	TotalPrice
1895	14911	0	248	132458.73
330	12748	0	224	28405.56
4042	17841	1	169	39869.05
1674	14606	0	128	11633.35
568	13089	2	118	57322.13
...	...	...	...	...
1141	13877	85	1	117.24
2997	16400	94	1	303.93
1142	13878	54	1	1271.57
1149	13886	70	1	243.56
990	13670	74	1	349.70

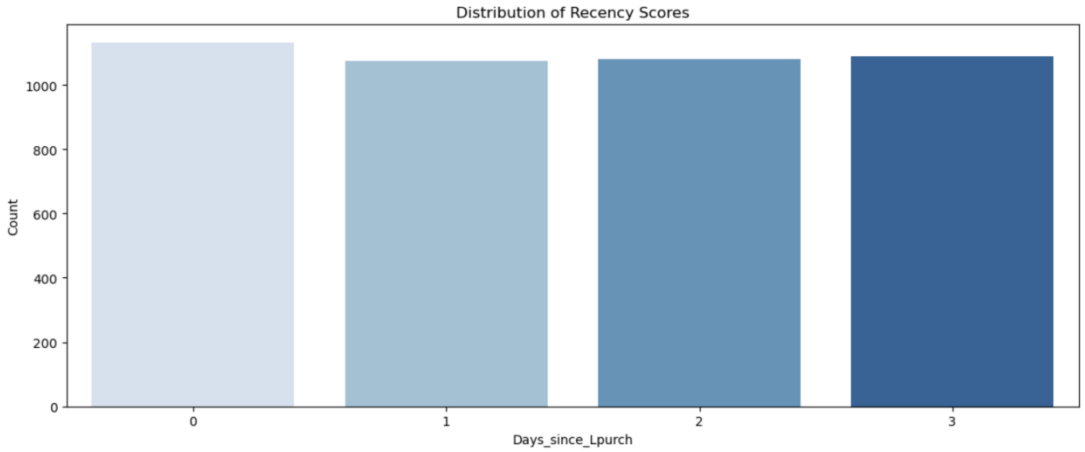
4372 rows × 4 columns

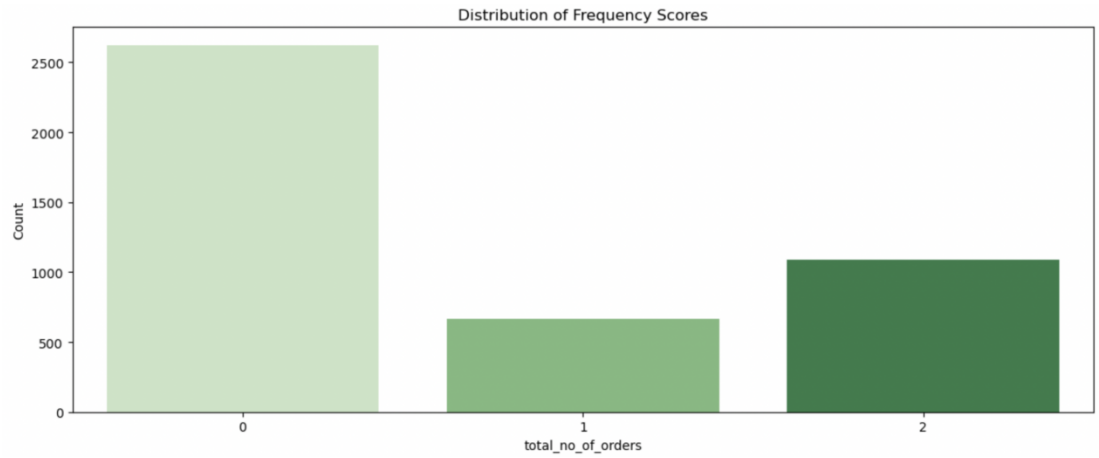
We assign RFM scores to each customer in the rfm_metrics DataFrame based on quartiles for Recency, Frequency, and Monetary values. Recency scores are determined by quartiles of the 'Days_since_Lpurch' column, Frequency scores by quartiles of the 'total_no_of_orders' column, and Monetary scores by quartiles of the 'TotalPrice' column. The resulting DataFrame includes these scores, allowing for detailed customer segmentation. The print statements display the

first few rows of the scored DataFrame and its shape, providing an initial overview of the RFM analysis results.

	CustomerID	Days_since_Lpurch	total_no_of_orders	TotalPrice	recency_score	frequency_score	monetary_score
0	12346	325	2	0.00	3	0	0
1	12347	1	7	4310.00	0	2	3
2	12348	74	4	1797.24	2	1	3
3	12349	18	1	1757.55	1	0	3
4	12350	309	1	334.40	3	0	1

We generate visualizations to depict the distribution of Recency, Frequency, and Monetary (RFM) scores for each customer in a 3x1 grid of subplots. Using Seaborn's countplot, it displays the count of customers within each score range for Recency, Frequency, and Monetary metrics. The color palettes ('Blues', 'Greens', 'Oranges') help distinguish the different components. The resulting visualizations offer insights into the distribution patterns of customers across the RFM score ranges, aiding in the interpretation of customer segmentation

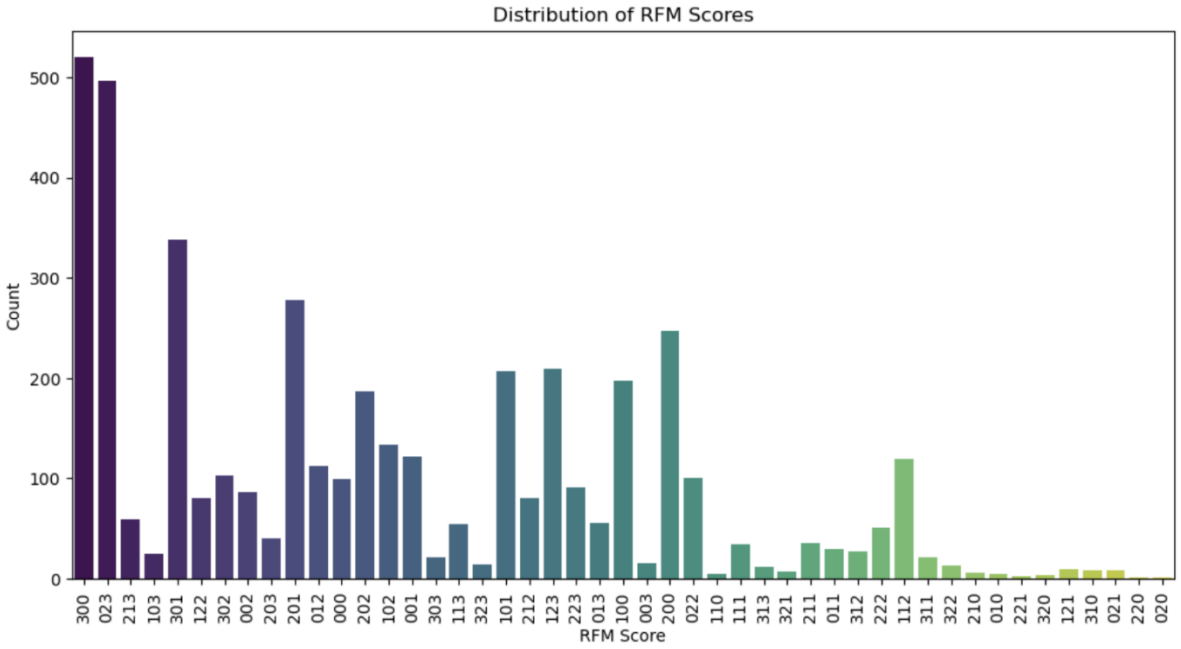




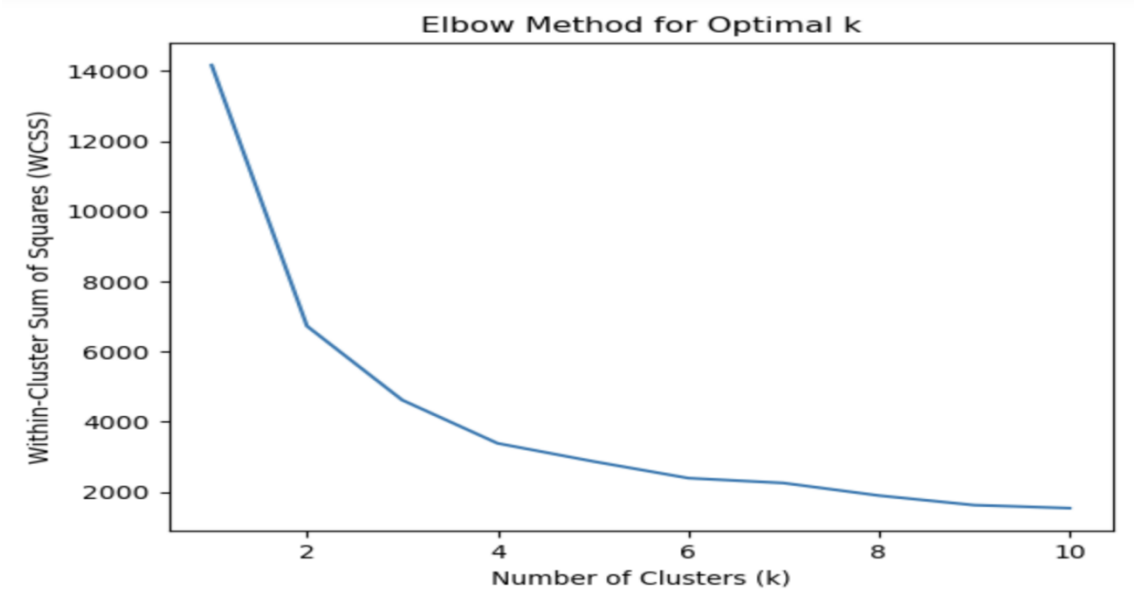
We combines individual Recency, Frequency, and Monetary (RFM) scores for each customer into a single 'rfm_score' column in the rfm_metrics DataFrame. This unified score, represented as a string, allows for a comprehensive assessment of each customer's overall RFM profile. The resulting DataFrame includes this composite score, providing a valuable basis for customer segmentation or analysis based on their combined recency, frequency, and monetary behavior.

	CustomerID	Days_since_Lpurch	total_no_of_orders	TotalPrice	recency_score	frequency_score	monetary_score	rfm_score
0	12346	325	2	0.00	3	0	0	300
1	12347	1	7	4310.00	0	2	3	023
2	12348	74	4	1797.24	2	1	3	213
3	12349	18	1	1757.55	1	0	3	103
4	12350	309	1	334.40	3	0	1	301

We now generate a countplot to visualize the distribution of combined RFM scores for customers in the rfm_metrics DataFrame. The plot displays the count of customers for each unique RFM score, offering insights into the distribution of customer segments based on their combined Recency, Frequency, and Monetary behavior. The plot is customized with a specified figure size, color palette ('viridis'), title ('Distribution of RFM Scores'), and rotated x-axis labels for improved readability. This visualization aids in understanding the composition of customer segments and can inform targeted marketing or engagement strategies based on RFM analysis



We use K-Means clustering to segment customers based on their RFM. It extracts the RFM scores from the rfm_metrics DataFrame and applies the Elbow Method to determine the optimal number of clusters (k). The Elbow Method graph is plotted to help visually inspect the point where the rate of decrease in Within-Cluster Sum of Squares (WCSS) slows down. The optimal number of clusters is then chosen . Finally, K-Means clustering is applied with the optimal number of clusters, and the resulting cluster assignments are added as a new 'Cluster' column to the original rfm_metrics DataFrame. This segmentation can be useful for targeted marketing or personalized strategies based on customer behavior.



We sort the rfm_metrics based on the 'Cluster' column in descending order, allowing for a quick examination of the distribution and characteristics of each identified customer segment.

	CustomerID	Days_since_Lpurch	total_no_of_orders	TotalPrice	recency_score	frequency_score	monetary_score	rfm_score	Cluster
610	13142	18	1	307.09	1	0	1	101	2
1468	14332	22	3	840.21	1	0	2	102	2
3356	16887	34	1	192.33	1	0	0	100	2
739	13318	0	3	640.76	0	0	1	001	2
3355	16885	14	3	450.41	0	0	1	001	2
...	...	...	...	...	...	...	...	...	...
2823	16163	224	3	441.60	3	0	1	301	0
2822	16162	251	1	37.40	3	0	0	300	0
1118	13848	91	3	1255.00	2	0	2	202	0
2819	16159	281	1	348.20	3	0	1	301	0
0	12346	325	2	0.00	3	0	0	300	0

4372 rows × 9 columns

We need to create a new DataFrame (segment_data) by merging the rfm_metrics , containing RFM scores, with a subset of columns ('CustomerID,' 'Country,' 'Description,' and 'UnitPrice') from the proj2C DataFrame. The merging is based on the common 'CustomerID' column. The resulting DataFrame is then sorted in descending order based on the 'Description' column. This process is likely undertaken to analyze and profile customer segments, with a focus on the types of products ('Description') purchased by customers within each segment.

	CustomerID	Days_since_Lpurch	total_no_of_orders	TotalPrice	recency_score	frequency_score	monetary_score	rfm_score	Cluster	Country	Description	UnitPrice
219558	15472	112	1	371.56	2	0	1	201	0	United Kingdom	ZINC WIRE SWEETHEART LETTER TRAY	3.75
332043	17290	3	2	513.01	0	0	1	001	2	United Kingdom	ZINC WIRE SWEETHEART LETTER TRAY	3.75
218330	15444	8	2	1411.21	0	0	2	002	2	United Kingdom	ZINC WIRE SWEETHEART LETTER TRAY	3.75
46935	13012	8	11	1296.62	0	2	2	022	1	United Kingdom	ZINC WIRE SWEETHEART LETTER TRAY	3.75
166024	14710	13	4	599.48	0	1	1	011	2	United Kingdom	ZINC WIRE SWEETHEART LETTER TRAY	3.75
...	...	...	...	...	...	...	...	...	...	...	...	...

											CANDLES
366288	17774	96	5	1203.78	2	1	2	212	1	United Kingdom	4 PURPLE FLOCK DINNER CANDLES	2.55
44210	12953	9	1	329.85	0	0	1	001	2	United Kingdom	4 PURPLE FLOCK DINNER CANDLES	2.55
388208	18055	6	11	6729.29	0	2	3	023	1	United Kingdom	4 PURPLE FLOCK DINNER CANDLES	2.55
167600	14725	24	1	655.58	1	0	2	102	2	United Kingdom	4 PURPLE FLOCK DINNER CANDLES	2.55

401604 rows × 12 columns

We need to create three separate DataFrames (segment_0_data, segment_1_data, and segment_2_data) to represent distinct customer segments identified through K-Means clustering. Each DataFrame is obtained by filtering the original segment_data DataFrame based on the 'Cluster' column, where values 0, 1, and 2 correspond to the different clusters. These segmented DataFrames allow for detailed analysis and profiling of customers within each identified cluster, facilitating insights into the distinct characteristics and behaviors of customers in different segments.

Here is the attributes from customers in Segment 0, including RFM scores, product-related details (unit price and description), and country information. This information is organized into a new DataFrame named segment_0_attributes for focused analysis and profiling of customers within this segment.

	recency_score	frequency_score	monetary_score	UnitPrice	Description	Country
0	3	0	0	1.04	MEDIUM CERAMIC TOP STORAGE JAR	United Kingdom
1	3	0	0	1.04	MEDIUM CERAMIC TOP STORAGE JAR	United Kingdom
288	3	0	1	2.10	CHOCOLATE THIS WAY METAL SIGN	Norway
289	3	0	1	2.10	METAL SIGN NEIGHBOURHOOD WITCH	Norway
290	3	0	1	0.85	RETRO MOD TRAY	Norway
...	...	...	...	...	...	...
400795	3	0	0	0.42	PENNY FARTHING BIRTHDAY CARD	United Kingdom
400796	3	0	0	16.95	SPACEBOY BABY GIFT SET	United Kingdom
400797	3	0	0	16.95	DOLLY GIRL BABY GIFT SET	United Kingdom
400798	3	0	0	1.65	LUNCH BAG DOILEY PATTERN	United Kingdom
400799	3	0	0	2.55	GUMBALL COAT RACK	United Kingdom

51958 rows × 6 columns

Here is the specific attributes from customers in Segment 1, including RFM scores, product-related details (unit price and description), and country information. This information is

organized into a new DataFrame named segment_1_attributes for focused analysis and profiling of customers within this segment.

	recency_score	frequency_score	monetary_score	UnitPrice	Description	Country
2	0	2	3	2.10	BLACK CANDELABRA T-LIGHT HOLDER	Iceland
3	0	2	3	4.25	AIRLINE BAG VINTAGE JET SET BROWN	Iceland
4	0	2	3	3.25	COLOUR GLASS. STAR T-LIGHT HOLDER	Iceland
5	0	2	3	0.65	MINI PAINT SET VINTAGE	Iceland
6	0	2	3	1.25	CLEAR DRAWER KNOB ACRYLIC EDWARDIAN	Iceland
...	...	...	...	...	...	...
401599	1	0	3	0.42	LIPSTICK PEN RED	United Kingdom
401600	1	0	3	2.10	HAND WARMER SCOTTY DOG DESIGN	United Kingdom
401601	1	0	3	1.25	SET OF 3 WOODEN SLEIGH DECORATIONS	United Kingdom
401602	1	0	3	0.39	PAINTED METAL STAR WITH HOLLY BELLS	United Kingdom
401603	1	0	3	0.29	SWISS CHALET TREE DECORATION	United Kingdom

309698 rows × 6 columns

Here is the specific attributes from customers in Segment 2, including RFM scores, product-related details (unit price and description), and country information. This information is organized into a new DataFrame named segment_2_attributes for focused analysis and profiling of customers within this segment.

	recency_score	frequency_score	monetary_score	UnitPrice	Description	Country
665	0	0	2	3.75	FAIRY CAKE DESIGN UMBRELLA	Austria
666	0	0	2	1.49	CERAMIC STRAWBERRY DESIGN MUG	Austria
667	0	0	2	9.95	CERAMIC CAKE STAND + HANGING CAKES	Austria
668	0	0	2	1.49	CERAMIC CAKE DESIGN SPOTTED PLATE	Austria
669	0	0	2	7.95	DOORMAT FAIRY CAKE	Austria
...	...	...	...	...	...	...
400808	0	0	0	12.75	REGENCY CAKESTAND 3 TIER	United Kingdom
400809	0	0	0	2.95	ROSES REGENCY TEACUP AND SAUCER	United Kingdom
400810	0	0	0	0.42	CARD CHRISTMAS VILLAGE	United Kingdom
400811	0	0	0	4.15	REGENCY SUGAR BOWL GREEN	United Kingdom
400812	0	0	0	3.25	REGENCY MILK JUG PINK	United Kingdom

39948 rows × 6 columns

Segment 0 attributes

IE6400 Foundations Data Analytics Engineering
Group 3

	recency_score	frequency_score	monetary_score	UnitPrice	Description	Country
count	51958.000000	51958.000000	51958.000000	51958.000000	51958	51958
unique	NaN	NaN	NaN	NaN	3332	31
top	NaN	NaN	NaN	NaN	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
freq	NaN	NaN	NaN	NaN	352	46861
mean	2.541765	0.165711	1.158243	4.312703	NaN	NaN
std	0.498257	0.461444	0.804037	172.675528	NaN	NaN
min	2.000000	0.000000	0.000000	0.000000	NaN	NaN
25%	2.000000	0.000000	1.000000	1.250000	NaN	NaN
50%	3.000000	0.000000	1.000000	1.950000	NaN	NaN
75%	3.000000	0.000000	2.000000	3.750000	NaN	NaN
max	3.000000	2.000000	3.000000	38970.000000	NaN	NaN

Segment 1 attributes

	recency_score	frequency_score	monetary_score	UnitPrice	Description	Country
count	309698.000000	309698.000000	309698.000000	309698.000000	309698	309698
unique	NaN	NaN	NaN	NaN	3787	27
top	NaN	NaN	NaN	NaN	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
freq	NaN	NaN	NaN	NaN	1578	272504
mean	0.478531	1.801549	2.835088	3.416502	NaN	NaN
std	0.724562	0.452003	0.380035	36.092727	NaN	NaN
min	0.000000	0.000000	1.000000	0.000000	NaN	NaN
25%	0.000000	2.000000	3.000000	1.250000	NaN	NaN
50%	0.000000	2.000000	3.000000	1.950000	NaN	NaN
75%	1.000000	2.000000	3.000000	3.750000	NaN	NaN
max	3.000000	2.000000	3.000000	8142.750000	NaN	NaN

Segment 2 attributes

	recency_score	frequency_score	monetary_score	UnitPrice	Description	Country
count	39948.000000	39948.000000	39948.000000	39948.000000	39948	39948
unique	NaN	NaN	NaN	NaN	3054	18
top	NaN	NaN	NaN	NaN	REX CASH+CARRY JUMBO SHOPPER	United Kingdom
freq	NaN	NaN	NaN	NaN	186	37363
mean	0.607790	0.103785	1.251202	2.829542	NaN	NaN
std	0.488249	0.308170	0.705772	6.941557	NaN	NaN
min	0.000000	0.000000	0.000000	0.000000	NaN	NaN
25%	0.000000	0.000000	1.000000	0.850000	NaN	NaN
50%	1.000000	0.000000	1.000000	1.650000	NaN	NaN
75%	1.000000	0.000000	2.000000	3.390000	NaN	NaN
max	1.000000	2.000000	2.000000	705.450000	NaN	NaN

Marketing Recommendations:

As we present our project recommendations, we advocate for a strategic focus on the top 5 customers known for their consistent high-value purchases. We propose the introduction of exclusive loyalty programs and personalized communication, including early access to sales and tailored discounts, to significantly enhance customer loyalty. Consistently seeking and incorporating customer feedback is a key element of our proposed strategy, ensuring a deep understanding of evolving preferences and alignment with customer expectations.

We emphasize the importance of understanding order patterns, particularly the notable peaks in order volumes on Thursdays and during lunchtime. To optimize resource utilization, we suggest seizing these patterns by implementing Thursday-specific promotions and encouraging early morning shopping on Fridays with enticing special offers.

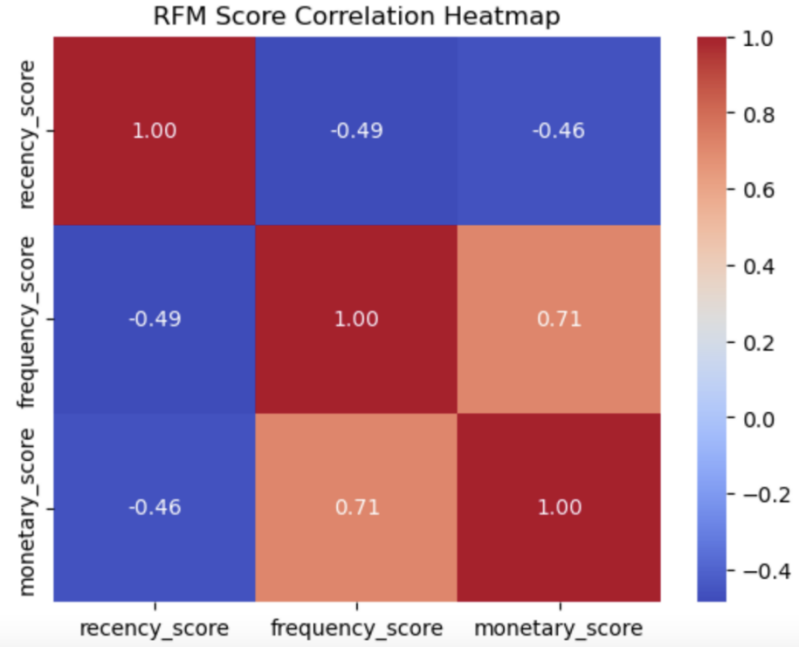
In the context of analyzing average order values by country, we recommend the implementation of targeted strategies. For countries with lower average order values, we propose the incorporation of upselling tactics and adjustments in pricing based on purchasing power to drive increased revenue. Conversely, for countries with higher average order values, our strategy involves the deployment of exclusive promotions to sustain engagement and maximize returns.

Addressing products with the lowest orders, we advise a comprehensive market analysis to reassess market demand and make necessary adjustments in marketing, pricing, or product positioning. Product improvements, such as enhancements or variations, are seen as vital, along with identifying opportunities for upselling to contribute to higher order values.

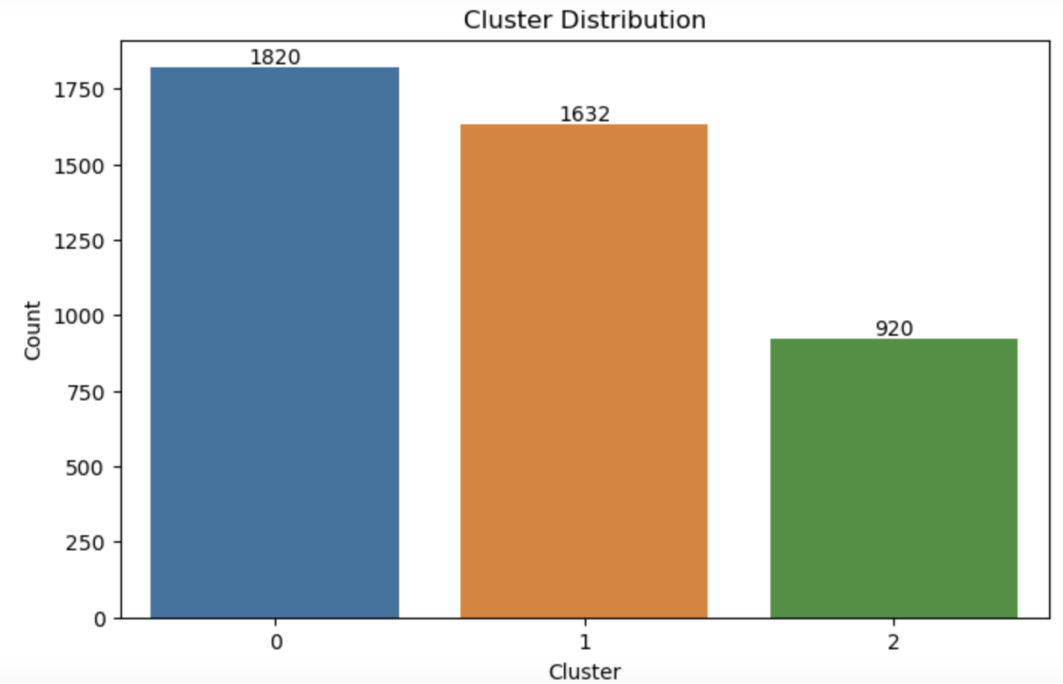
In regions with lower average order values, we suggest a nuanced approach, including localized pricing adjustments based on economic conditions, the initiation of special promotions, and exploration of market expansion opportunities or partnerships. These

strategies are anticipated to motivate larger purchases and foster enhanced brand engagement in these regions.

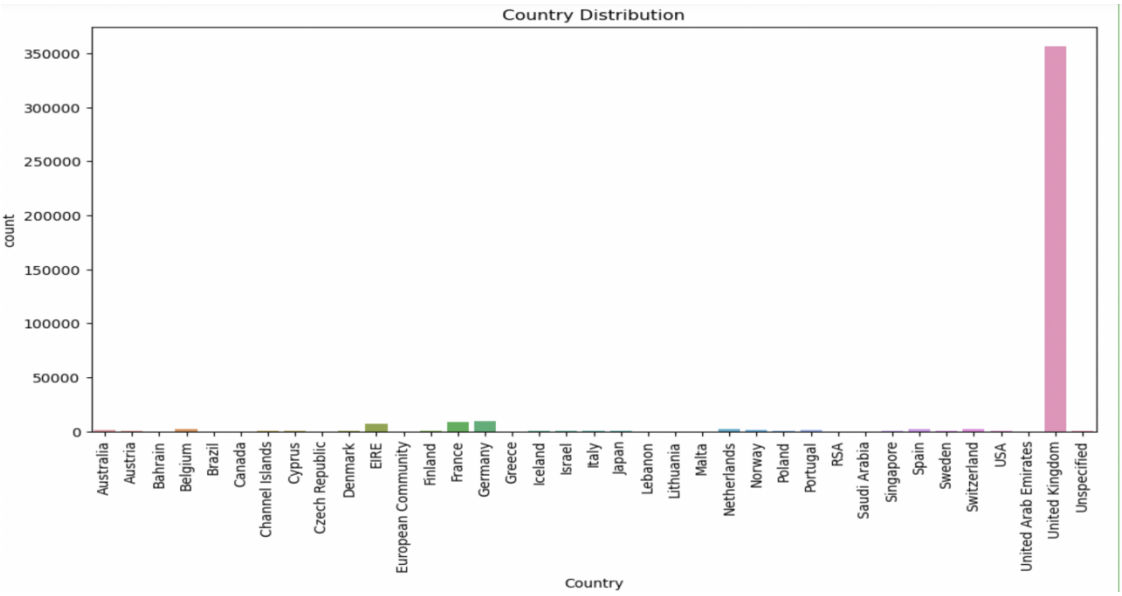
We create a heatmap visualizing the correlation between Recency, Frequency, and Monetary (RFM) scores in the rfm_metrics DataFrame. The heatmap displays correlation values, using the 'coolwarm' color map, providing insights into the relationships between these three components.



We create a bar chart to show the distribution of customers across different clusters. Each bar represents a cluster, and the height indicates the count of customers in that cluster. Annotations above the bars display the exact counts.



Here is a bar chart to visualize the distribution of customers across different countries. Each bar represents a country, and the height reflects the count of customers in that country. The x-axis labels are rotated for better readability.



Find the solutions to these questions:

1. Data Overview

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

The size of the dataset is 541909 rows × 8 columns

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

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

InvoiceNo - An exclusive reference number assigned to each transaction or invoice.
StockCode - A code used to identify the specific stock or product linked to the transaction.
Description - A written account detailing the stock or product involved in the transaction.
Quantity - The total count of products or items associated with the particular transaction.
InvoiceDate - The specific date and time when the invoice or transaction was created.
UnitPrice - The cost of a single unit of the respective product or item.
CustomerID - A distinct identifier assigned to the customer connected to the transaction.
Country - The nation where the customer involved in the transaction is situated or where the transaction took place.

- What is the time period covered by this dataset?

The range is between 12/1/2010,8:26 and 12/9/2011,12:50

2. Customer Analysis

- How many unique customers are there in the dataset?

There are 4339 unique customers

- 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: total_no_of_orders, dtype: float64
```

Here as observed the average number of order a customer places is approximately 4.

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

	CustomerID	Days_since_Lpurch	total_no_of_orders	TotalPrice
1895	14911	0	248	132458.73
330	12748	0	224	28405.56
4042	17841	1	169	39869.05
1674	14606	0	128	11633.35
568	13089	2	118	57322.13

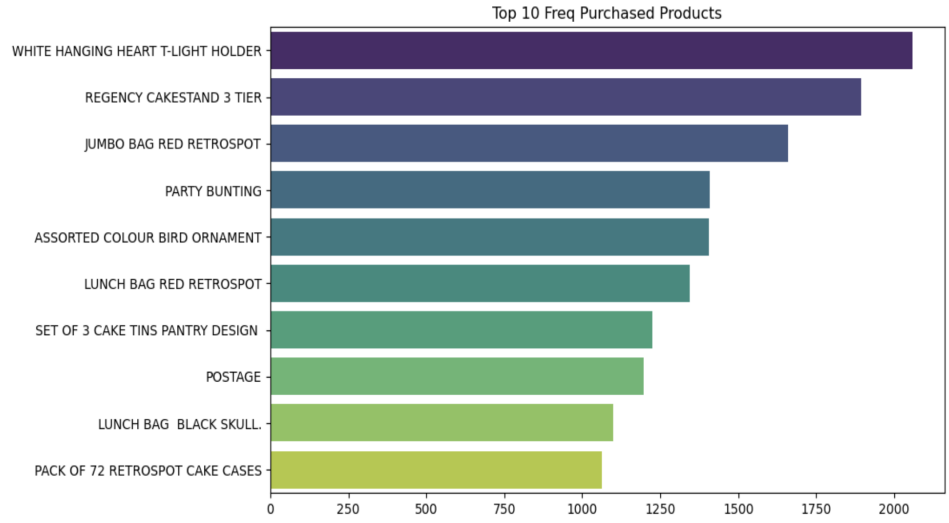
The Top 5 customers with most orders are customers with ID:

1. 12748
2. 14911
3. 17841
4. 13089
5. 14606

3. Product Analysis

- What are the top 10 most frequently purchased products?

```
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
```



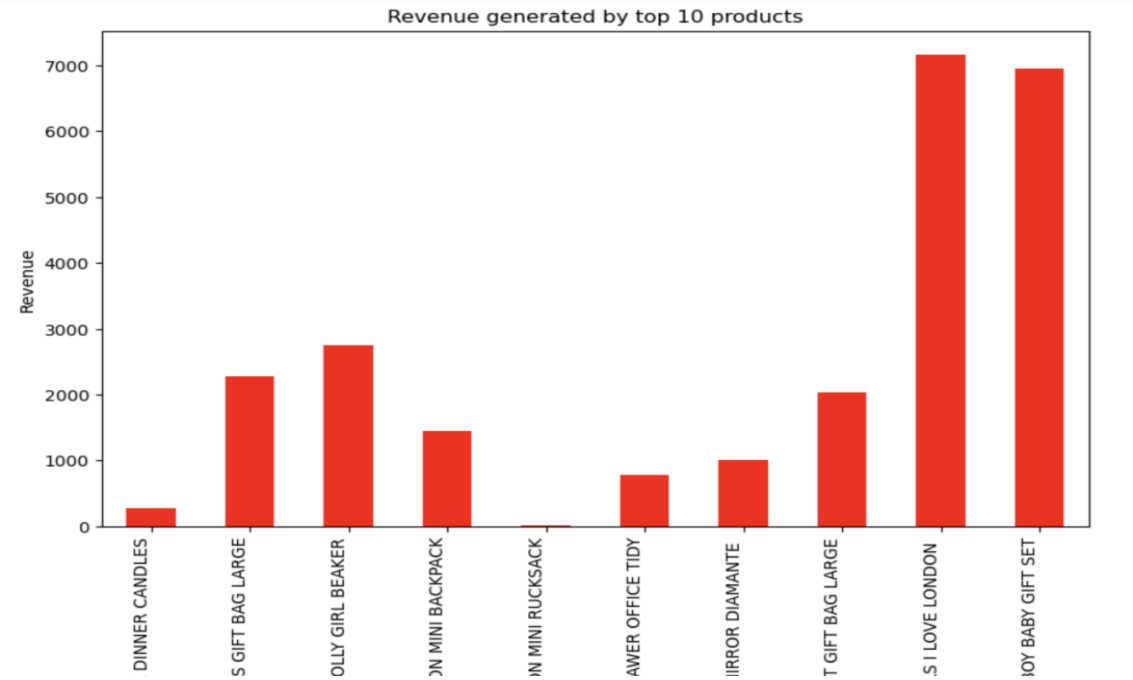
- What is the average price of products in the dataset?

We calculated the average price (UnitPrice) in the proj2 The variable avgprc holds the rounded average price value.

Hence,The average price of the products is: 3.47

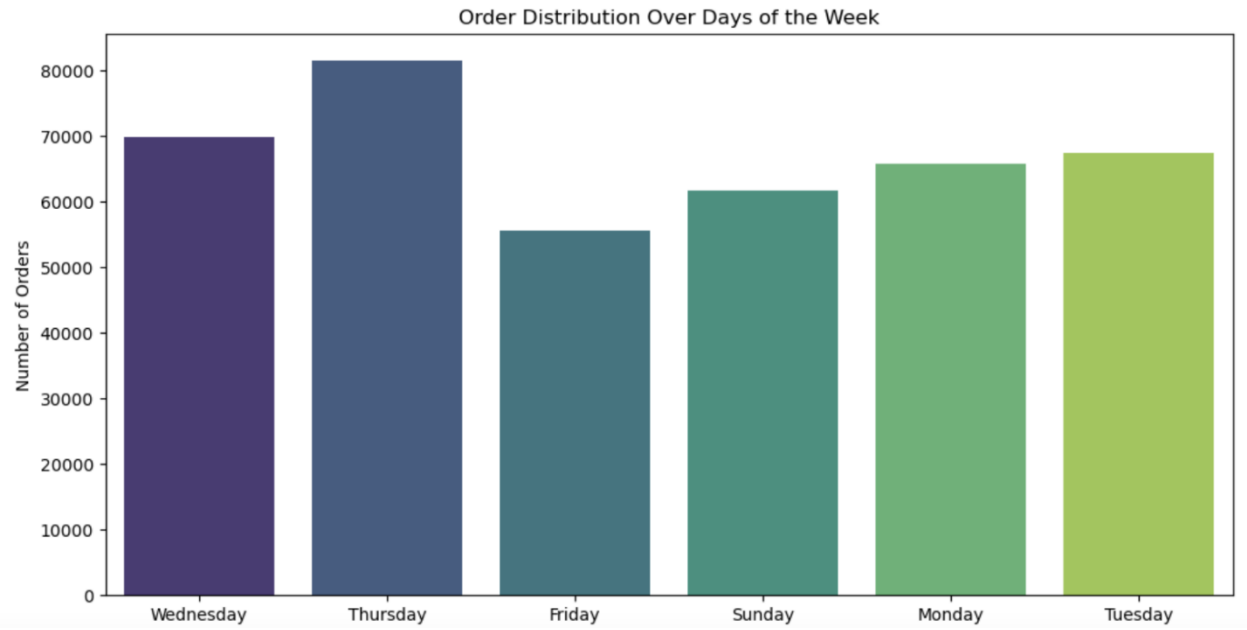
- Can you find out which product category generates the highest revenue?

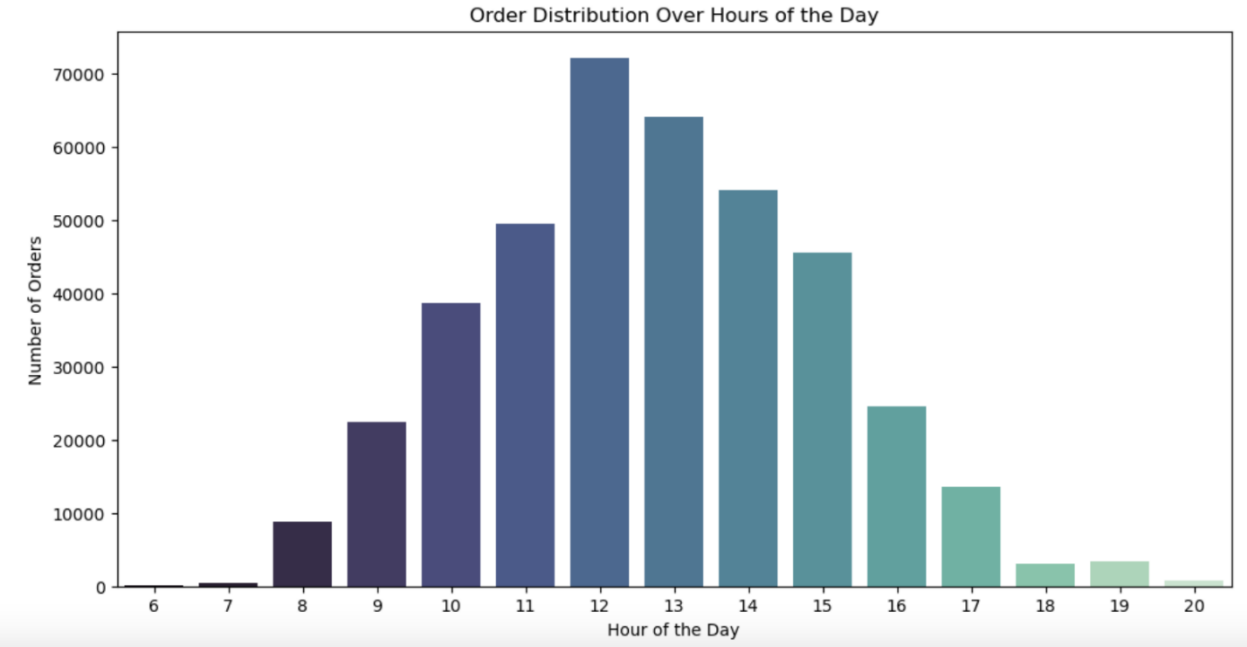
```
Description
REGENCY CAKESTAND 3 TIER               132567.70
WHITE HANGING HEART T-LIGHT HOLDER      93767.80
JUMBO BAG RED RETROSPOT                 83056.52
PARTY BUNTING                           67628.43
POSTAGE                                 66710.24
ASSORTED COLOUR BIRD ORNAMENT           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: TotalPrice, dtype: float64
```



4. Time Analysis

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





- What is the average order processing time?

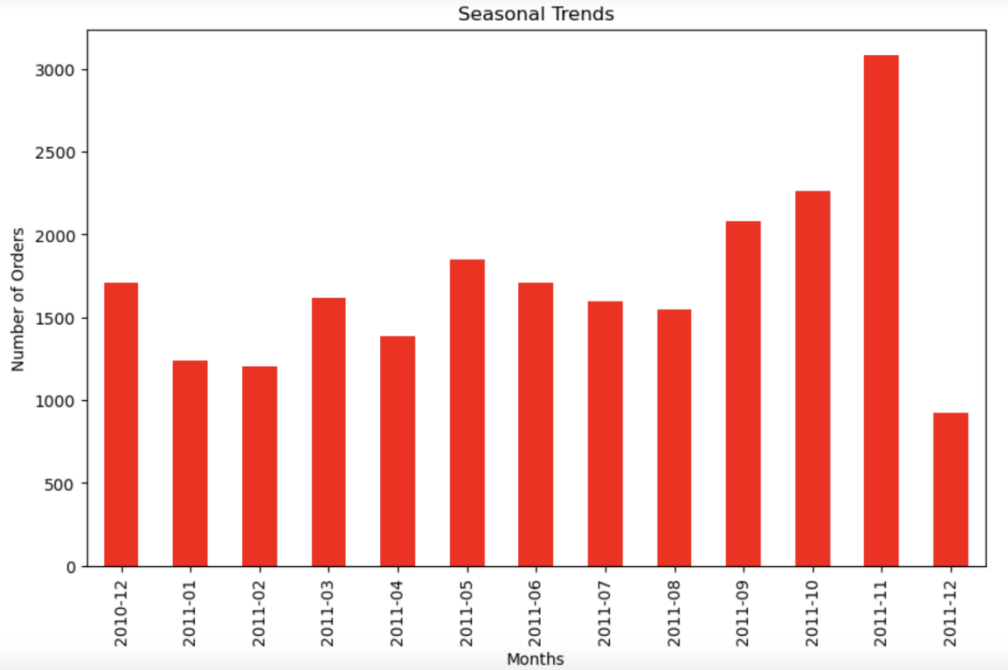
Regarding this query, insufficient data was available. Ordinarily, order processing time is determined as the time elapsed between the order date and the order fulfillment date. To conduct this analysis, we introduced a new column with random dates, incorporating a buffer period of one to ten days beyond the order dates. We considered InvoiceDate as the initial order processing date and the new column called final order fulfillment date.

Average Processing Time: 5 days 12:02:38.341052379

We took the difference between these 2 dates to get the processing time for each order.

- Are there any seasonal trends in the dataset?

```
InvoiceMonth
2010-12     1708
2011-01     1236
2011-02     1202
2011-03     1619
2011-04     1384
2011-05     1849
2011-06     1707
2011-07     1593
2011-08     1544
2011-09     2078
2011-10     2263
2011-11     3086
2011-12      921
Freq: M, Name: InvoiceNo, dtype: int64
```



5. Geographical Analysis

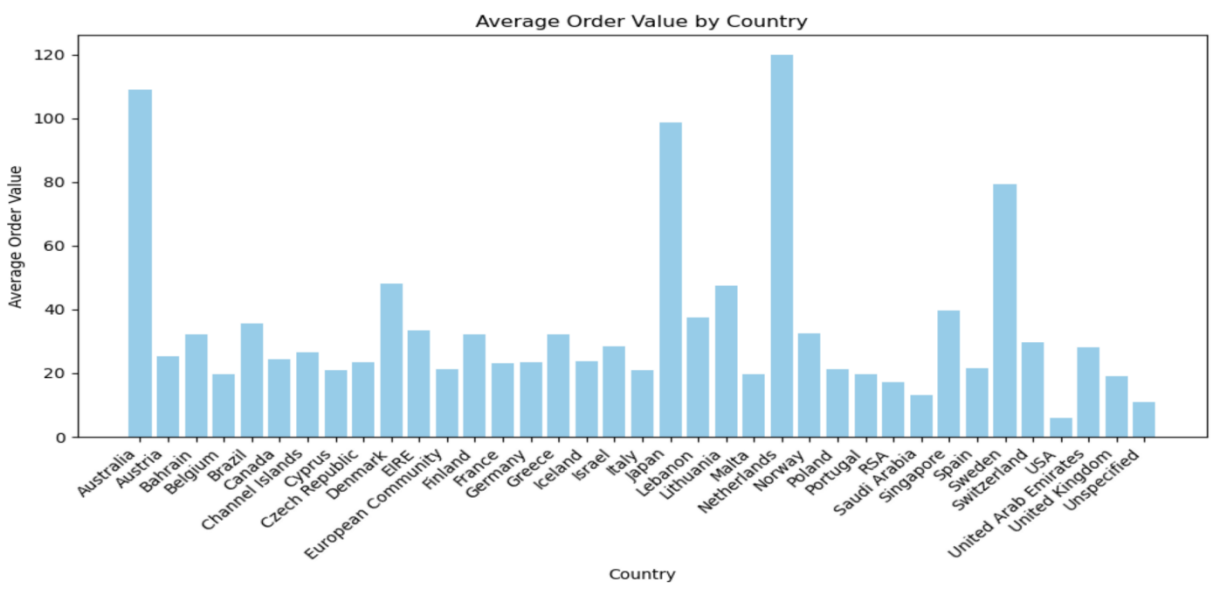
- Can you determine the top 5 countries with the highest number of orders?

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

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

	Country	AverageOrderValue
0	Australia	108.910787
1	Austria	25.322494
2	Bahrain	32.258824
3	Belgium	19.773301
4	Brazil	35.737500
5	Canada	24.280662
6	Channel Islands	26.520991
7	Cyprus	21.045434
8	Czech Republic	23.590667
9	Denmark	48.247147
10	EIRE	33.445054
11	European Community	21.176230
12	Finland	32.124806
13	France	23.200714
14	Germany	23.365978
15	Greece	32.263836
16	Iceland	23.681319
17	Israel	28.293117
18	Italy	21.034259
19	Japan	98.716816
20	Lebanon	37.641778
21	Lithuania	47.458857
22	Malta	19.728110

23	Netherlands	120.059696
24	Norway	32.378877
25	Poland	21.152903
26	Portugal	19.711598
27	RSA	17.281207
28	Saudi Arabia	13.117000
29	Singapore	39.827031
30	Spain	21.659822
31	Sweden	79.360976
32	Switzerland	29.696004
33	USA	5.948179
34	United Arab Emirates	27.974706
35	United Kingdom	18.914008
36	Unspecified	11.040539



6.Payment Analysis

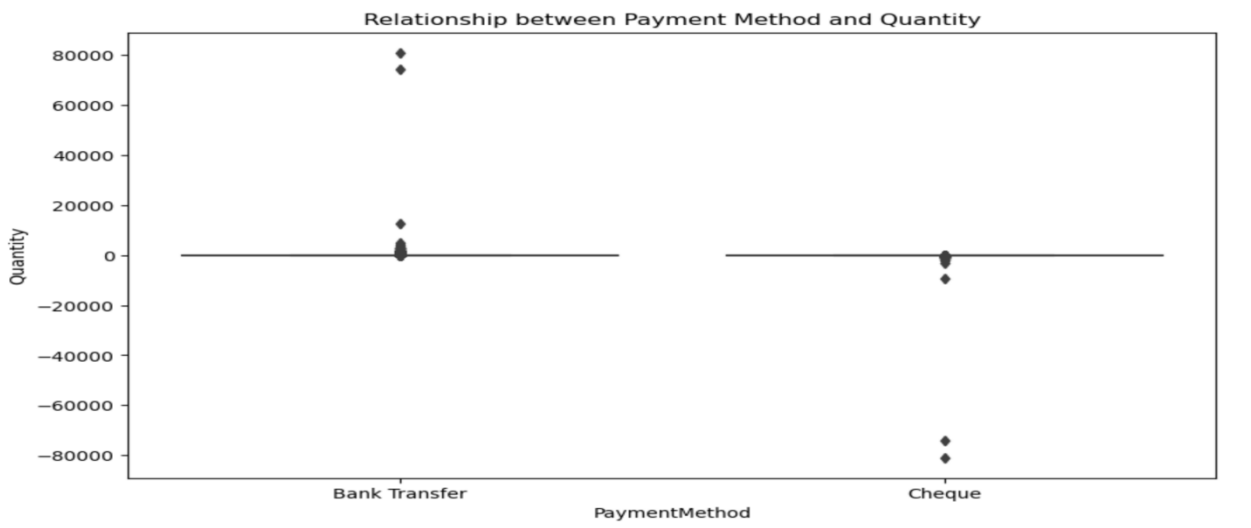
As we do not have payment column,we need to create a new column 'PaymentMethod' based on the presence of 'C' in 'InvoiceNo'.

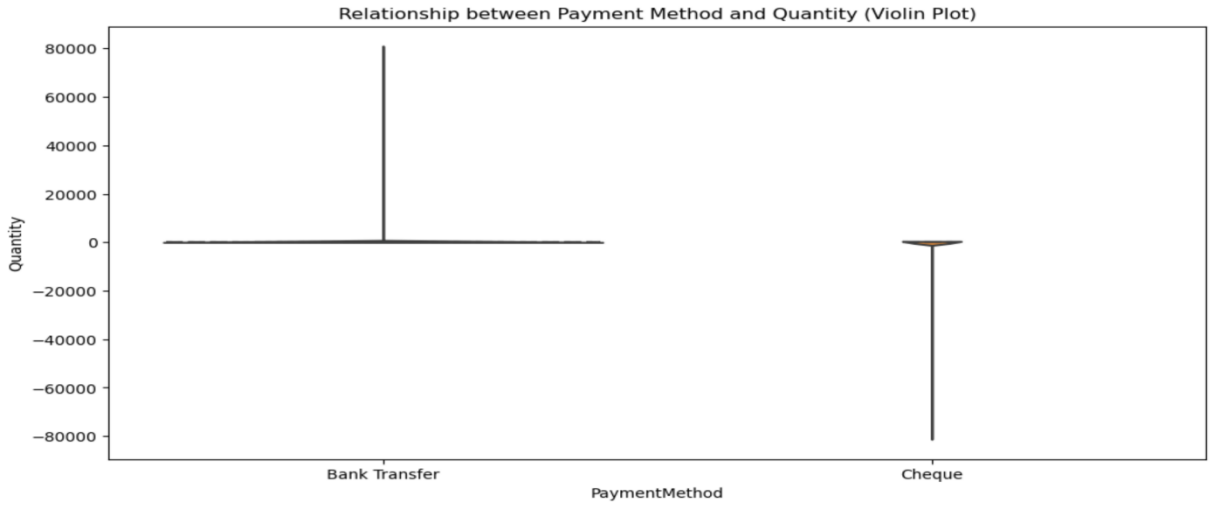
- What are the most common payment methods used by customers?

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice	DayOfWeek	HourOfDay	InvoiceMonth	PaymentMethod
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	15.30	Wednesday	8	2010-12	Bank Transfer
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	Wednesday	8	2010-12	Bank Transfer
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom	22.00	Wednesday	8	2010-12	Bank Transfer
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	Wednesday	8	2010-12	Bank Transfer
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	Wednesday	8	2010-12	Bank Transfer

```
Bank Transfer    392732
Cheque             8872
Name: PaymentMethod, dtype: int64
```

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





7. Customer Behavior

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

	CustomerID	Days_since_Lpurch	total_no_of_orders	TotalPrice	recency_score	frequency_score	monetary_score	rfm_score	Cluster
0	12346	325	2	0.00	3	0	0	300	0
1	12347	1	7	4310.00	0	2	3	023	1
2	12348	74	4	1797.24	2	1	3	213	1
3	12349	18	1	1757.55	1	0	3	103	1
4	12350	309	1	334.40	3	0	1	301	0

Result :91.04711802378775
Hence on an average customers have a difference of approx 92 days before their first and last purchase.

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

No,here are no customer segments based on their purchase behavior. The clustering algorithm has grouped customers into distinct segments, and the 'Cluster' column in the rfm_metrics DataFrame represents these segments. Each segment is characterized by specific patterns in recency, frequency, and monetary value, providing insights into different customer behaviors and preferences.

8. Returns and Refunds

As we do not have returns column, we created a new binary column named 'Return' in the proj2 DataFrame to indicate whether an order had a return.

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

The percentage of orders with returns or refunds is: 2.18%

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

```
Description
 4 PURPLE FLOCK DINNER CANDLES      0.000000
 50'S CHRISTMAS GIFT BAG LARGE      0.009091
 DOLLY GIRL BEAKER                  0.014599
 I LOVE LONDON MINI BACKPACK        0.000000
 I LOVE LONDON MINI RUCKSACK        0.000000
                                       ...
ZINC T-LIGHT HOLDER STARS SMALL     0.012448
ZINC TOP  2 DOOR WOODEN SHELF       0.181818
ZINC WILLIE WINKIE  CANDLE STICK    0.005208
ZINC WIRE KITCHEN ORGANISER         0.000000
ZINC WIRE SWEETHEART LETTER TRAY    0.000000
Name: Return, Length: 3896, dtype: float64

Chi-square statistic: 20949.239430416943
P-value: 0.0
```

9. Profitability Analysis

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

The total profit generated by the company is: $8278519.42

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

```
Top 5 products with the highest profit margins:
Description
 4 PURPLE FLOCK DINNER CANDLES    100.0
 I LOVE LONDON MINI BACKPACK      100.0
 I LOVE LONDON MINI RUCKSACK      100.0
 NINE DRAWER OFFICE TIDY          100.0
 RED SPOT GIFT BAG LARGE          100.0
Name: ProfitMargin, dtype: float64
```

10. **Customer Satisfaction**

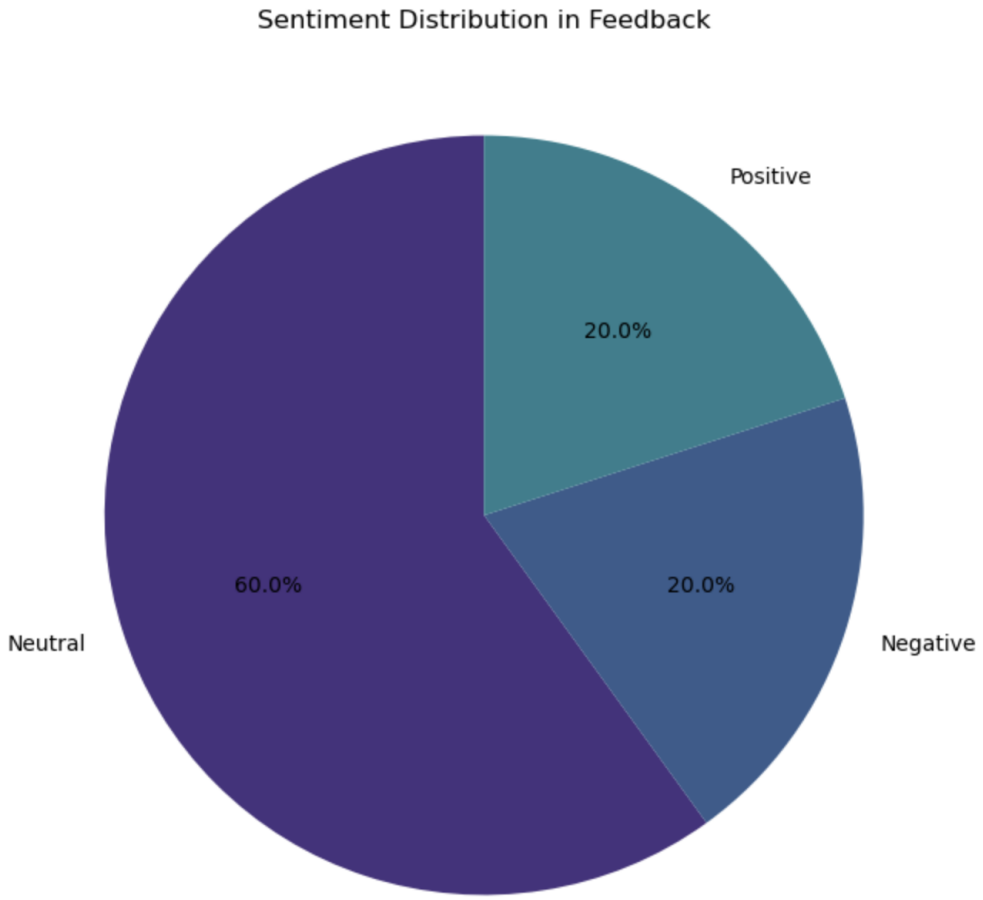
As we do not have feedback column,we add a new column named 'Feedback' to the proj2 DataFrame. This column contains random values selected from the feedback_list for each corresponding row in the dataset.

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

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice	DayOfWeek	HourOfDay	InvoiceMonth
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	15.30	Wednesday	8	2010-12
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	Wednesday	8	2010-12
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom	22.00	Wednesday	8	2010-12
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	Wednesday	8	2010-12
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	Wednesday	8	2010-12

PaymentMethod	Return	Profit	Revenue	ProfitMargin	Feedback
Bank Transfer	False	15.30	15.30	100.0	bring new colours
Bank Transfer	False	20.34	20.34	100.0	Average
Bank Transfer	False	22.00	22.00	100.0	Good
Bank Transfer	False	20.34	20.34	100.0	stop the production
Bank Transfer	False	20.34	20.34	100.0	Average

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


Sentiment Distribution in Feedback
Positive
20.0%
60.0%
20.0%
Neutral
Negative