Customer Segmentation using RFM Analysis

IE6400 – Foundations of data analytics Engineering Project 2

Final Report Group Number 27

Vishal Reddy Vookanti (002810767)

Hrudhvik Nangineni (002801706)

Siva Abhishek Sirivella (002204549)

Sacchit Shah (002820122)

Yaswanth Reddy Nalamalapu (002842412)

Introduction:

In this project, we delve into the world of E-Commerce, armed with a dataset brimming with transactional details from a UK-based online retail store specializing in unique all-occasion gifts. Our mission is to unravel insights that can guide targeted marketing and customer retention strategies using the powerful RFM (Recency, Frequency, Monetary) analysis method.

The dataset, spanning from December 1, 2010, to September 12, 2011, captures the pulse of a registered non-store online retailer frequented by both individual customers and wholesalers. With a focus on customer segmentation, we aim to understand and categorize customers based on their recent purchasing behavior, purchase frequency, and monetary value.

Our journey begins with essential data preprocessing steps, ensuring the dataset is clean, handling missing values, and fine-tuning data types if necessary. We then move on to the core of our analysis—RFM calculation. Recency reflects how recently a customer made a purchase, Frequency quantifies how often they buy, and Monetary encapsulates the total value of their purchases.

Segmentation comes next, as we assign RFM scores to customers based on quartiles or custom-defined bins. These scores amalgamate to form a comprehensive RFM score for each customer. Clustering techniques, such as K-Means clustering, will be deployed to create meaningful segments.

As we unravel the story of our customers, we'll dive into specific questions. How many unique customers grace our dataset? What does the distribution of orders per customer look like? Who are the top 5 customers by order count? We'll traverse through product analysis, time trends, geographical nuances, payment methods, customer behavior, returns and refunds, profitability, and even peek into the realm of customer satisfaction.

With each step, we aim to not only answer specific questions but to weave a narrative that sheds light on the dynamics of this E-Commerce landscape. With statistical analysis and compelling visualizations, our goal is to provide actionable marketing recommendations and present the untold stories hidden within the dataset.

Data Acquisition and handling:

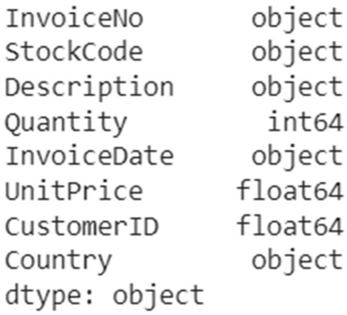
Data is obtained from the from the Kaggle data sets named E-Commerce dataset provided by The UCI Machine Learning Repository which has transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail which mainly sells unique gift items. It has wholesalers as its main customers.

Per the UCI Machine Learning Repository, this data was made available by Dr. Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

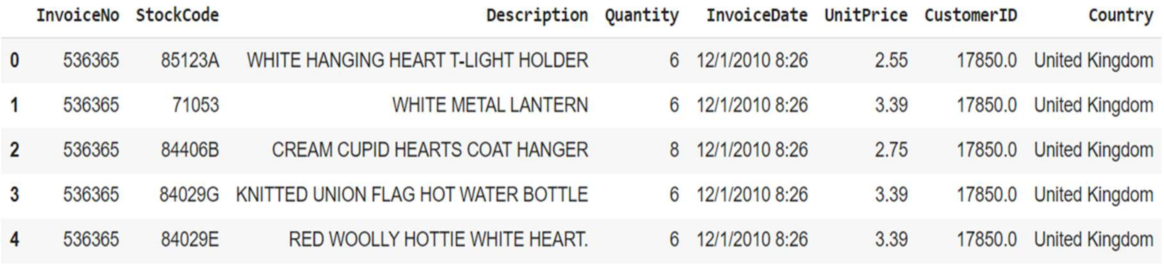
This data is then uploaded into Jupyter Notebook.

Data Inspection:

The Initial inspection has shown that the data set contains541909 rows with a total of 8 columns which are Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country with datatypes ranging from Objects, int and float. Here Upon Closer Inspection we can see that some of the rows in Customer ID and Description are not filled which are to addressed in a way by which we can maintain data integrity and Quality.



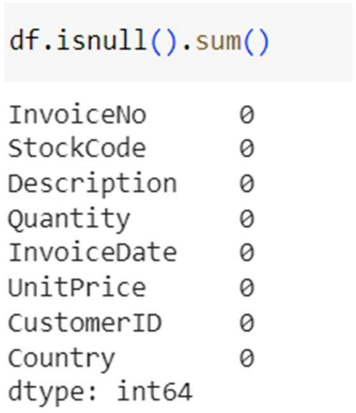
```
InvoiceNo       object
StockCode       object
Description     object
Quantity         int64
InvoiceDate     object
UnitPrice      float64
CustomerID     float64
Country         object
dtype: object
```



	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

Data Cleaning:

Customer Id is a such a field that is required for identifying the wholesaler/ person who bought a product and used in grouping the products bought by certain customers so for the missing values in this field we cannot fill some random or most occurring values as that may disturb the segmentation of other customer as the number of missing rows are in a significant number.



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

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

In the column of Description, we have a total of 1454 missing values which will not hinder our analysis so to maintain the data Quality and not skew the results for some products we have dropped rows containing the missing values.

For the duplicate data which amounted to 10062 rows we have removed the duplicate entries which has brought the data to be analyzed to 401604 rows.

Now after eliminating the outliers, we made the data consistent by Changing Customer ID to integer data to segment the customers accordingly and changed Invoice Date in to a standard date by making that row in date time data type.

Extra Data calculated based on the present data:

We have added a column maned Total Price which stores the total price for a particular order which is obtained by multiplying quantity and unit price column of the particular row.

By all these processing we have transformed the data into 401564 rows with a total of 9 columns.

RFM Analysis:

RFM Analysis is a form of analysis used to analyses some data sets which can be used to categorize customers to form marketing Strategies.

Here RFM means:

Recency (R): How recently a customer made a purchase. Calculate the number of

days since the customer's last purchase.

Frequency (F): How often a customer makes a purchase. Calculate the total number

of orders for each customer.

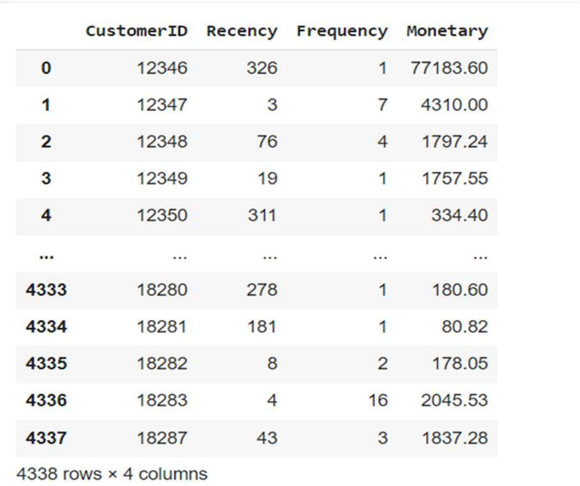
Monetary (M): The total monetary value of a customer's purchases. Calculate the

sum of the total price for each customer

RFM Calculation:

Calculated the Recency (R) using the quantity and invoice date to find the customers last purchase; Frequency (F) using customer Id and Invoice Date to find how often a customer is purchasing; and Monetary (M) using the total price created before to get the sum of total purchases done by the customer.

Using all these calculations we have created a RFM data frame.

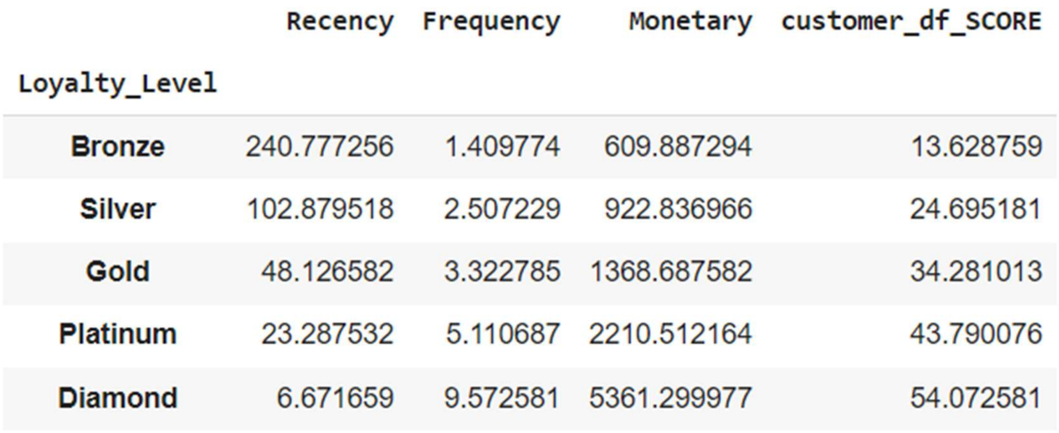


	CustomerID	Recency	Frequency	Monetary
0	12346	326	1	77183.60
1	12347	3	7	4310.00
2	12348	76	4	1797.24
3	12349	19	1	1757.55
4	12350	311	1	334.40
...	...	...	...	...
4333	18280	278	1	180.60
4334	18281	181	1	80.82
4335	18282	8	2	178.05
4336	18283	4	16	2045.53
4337	18287	43	3	1837.28

4338 rows × 4 columns

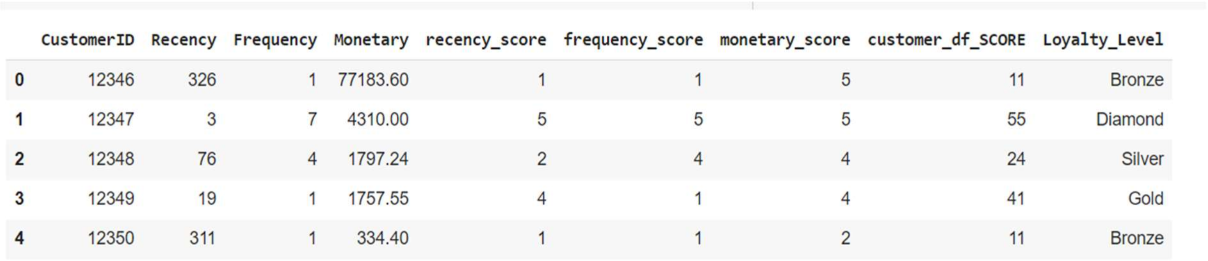
RFM segmentation:

Assigned PFM scores to each customers using a custom – defined bins using the loyalty system with five segmentations named as Bronze, Silver, Gold, Platinum, Diamond.



	Recency	Frequency	Monetary	customer_df_SCORE
Loyalty_Level				
Bronze	240.777256	1.409774	609.887294	13.628759
Silver	102.879518	2.507229	922.836966	24.695181
Gold	48.126582	3.322785	1368.687582	34.281013
Platinum	23.287532	5.110687	2210.512164	43.790076
Diamond	6.671659	9.572581	5361.299977	54.072581

Then combined the RFM scores to create a single RFM score for each Customer:

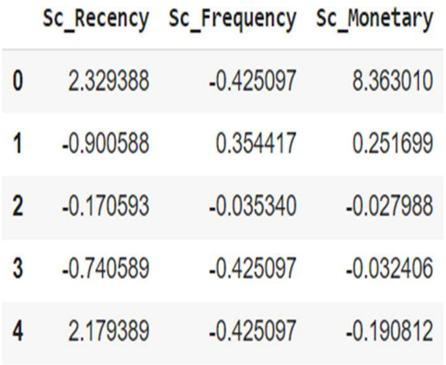


	CustomerID	Recency	Frequency	Monetary	recency_score	frequency_score	monetary_score	customer_df_SCORE	Loyalty_Level
0	12346	326	1	77183.60	1	1	5	11	Bronze
1	12347	3	7	4310.00	5	5	5	55	Diamond
2	12348	76	4	1797.24	2	4	4	24	Silver
3	12349	19	1	1757.55	4	1	4	41	Gold
4	12350	311	1	334.40	1	1	2	11	Bronze

Customer Segmentation:

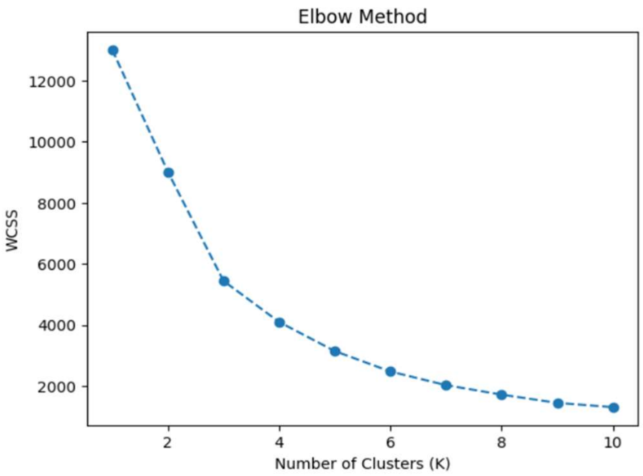
In this customer segmentation analysis, we employed the RFM (Recency, Frequency, Monetary) method to categorize customers based on their purchasing behavior. We started by preprocessing the data, including importing the dataset, handling missing values, and converting data types. The RFM metrics, representing Recency, Frequency, and Monetary aspects, were then calculated for each customer.

To ensure uniform scaling of the RFM metrics, we applied standardization using the StandardScaler. This step transformed the RFM values into a common scale for further analysis.

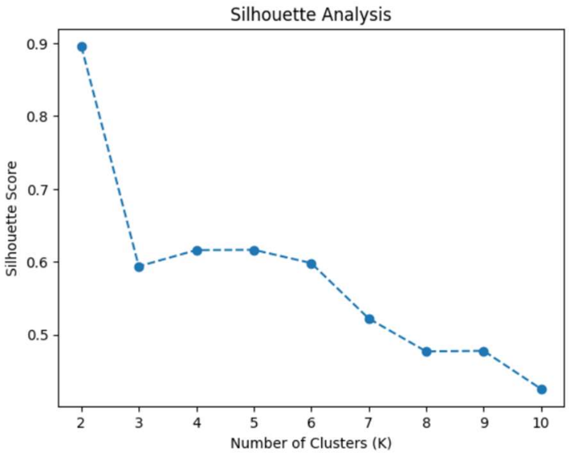


	Sc_Recency	Sc_Frequency	Sc_Monetary
0	2.329388	-0.425097	8.363010
1	-0.900588	0.354417	0.251699
2	-0.170593	-0.035340	-0.027988
3	-0.740589	-0.425097	-0.032406
4	2.179389	-0.425097	-0.190812

We have used The Elbow Method to identify the optimal number of clusters. By plotting the Within-Cluster-Sum-of-Squares (WCSS) against the number of clusters, we observed an "elbow" point to determine the suitable cluster count. Based on the Elbow Method, we chose an optimal number of clusters (e.g., 4 clusters). The K-Means algorithm was then applied to assign customers to these clusters, and additional columns were added to the dataset to represent the cluster assignments.



Also used Silhouette Analysis to validate the quality of the clusters. This analysis provides insights into how well-separated the clusters are and aids in assessing the appropriateness of the chosen cluster count.



By Doing this we have a dataset which includes cluster assignments and standardized RFM values. Each cluster represents a group of customers with similar purchasing behaviors. The analysis can be interpreted based on the characteristics of each cluster, providing valuable insights for marketing strategies.

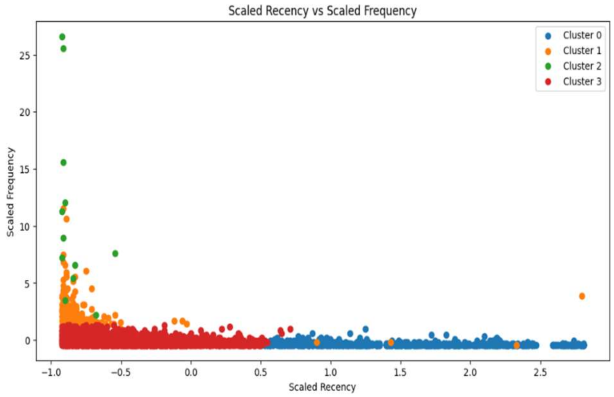
The K-Means clustering analysis successfully grouped customers into distinct segments and the Silhouette Score reinforced the quality of the clustering, indicating well-defined clusters.

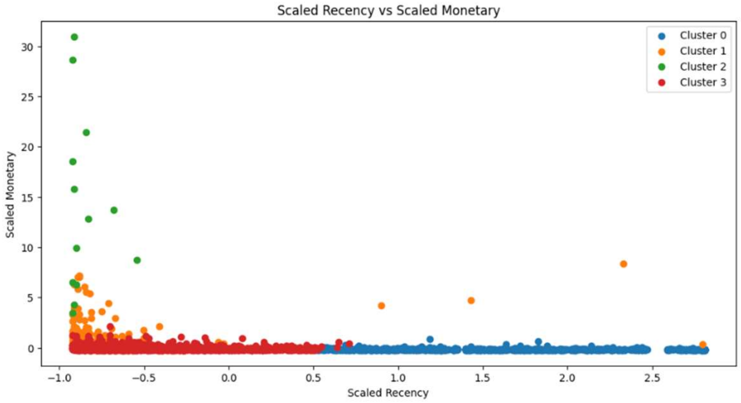
We have visualized:

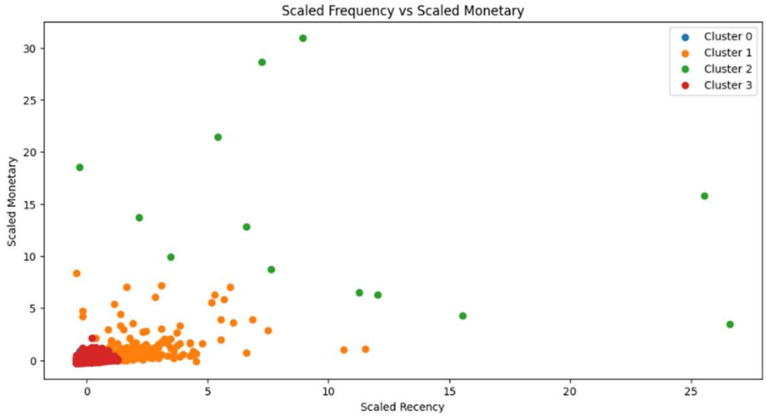
The 3D plot and visualizations of scaled features provided a comprehensive understanding of the clusters' characteristics.

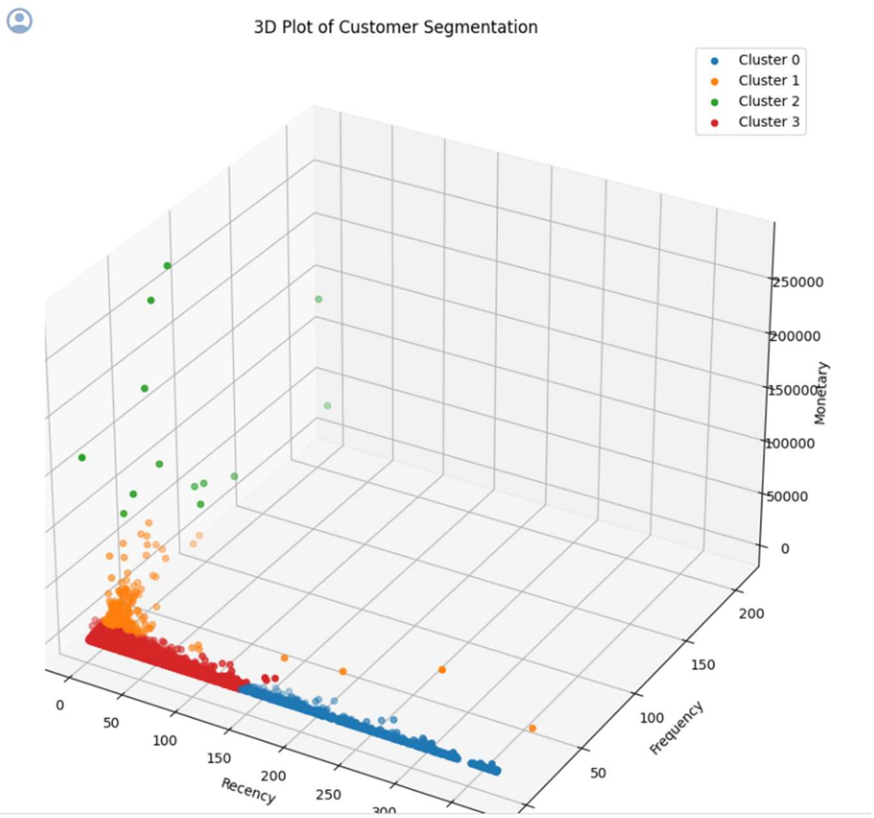
The distribution of customers across clusters was visualized through a pie chart, facilitating a concise representation of the segmentation results.

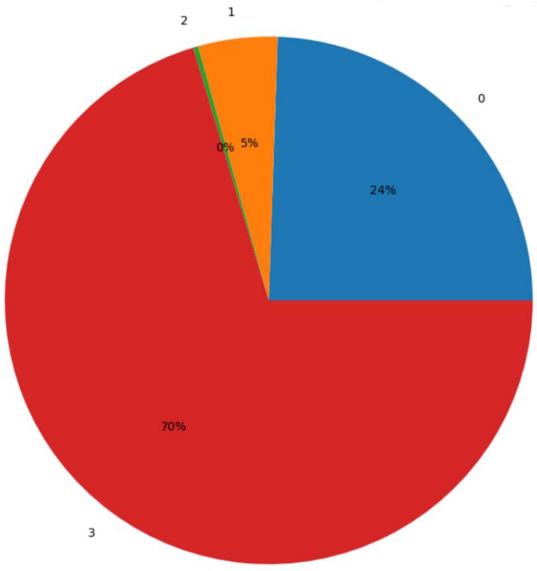
These findings serve as a foundation for tailored marketing strategies and enhanced customer engagement.








3D Plot of Customer Segmentation
Cluster 0
Cluster 1
Cluster 2
Cluster 3
250000
200000
150000
100000
50000
0
Monetary
200
150
100
50
Frequency
0
50
100
150
200
250
300
Recency


2
1
0
0%
5%
24%
70%
3

Meaning of the clusters used is:

Cluster 0: "Hibernating customers" - Those are the customers that buys at the lowest frequency, the lowest recently and that spend the least money.

Cluster 1: " Recent customers" - Those are customers that have been active quite recently that might be interesting to keep stimulated.

Cluster 2: "Exceptional customers" - Those are the customers that we want to keep, that buys at the highest frequency, the most recently and that spend the most money.

Cluster 3: "Punctual customers" - Those are the customers that buys more punctually items on the website.

Segment Profiling for Each customer:

Diamond: People in this group are more frequent buyers with average days since the last purchase is 7 and average number of times they have transacted in the platform is about 10 times in the last 1 year. Also, their average sales value is 5.3K pounds. These are your most loyal customers, who bought recently, most often, and are heavy spenders.

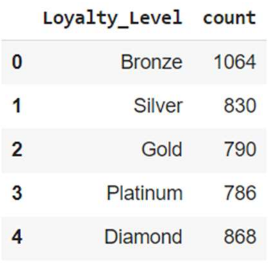
Platinum: This group has an average frequency of 5 times and recency of 23 days. Also, their average sales value is 2.2K pounds. These are your loyal customers with a good frequency and who spent a good amount.

Gold: This group has an average frequency of 3-4 times and recency of 48 days. This group is also high spenders with average sales of about 1.3K pounds. These are your recent customers with an average frequency and who spent a good amount.

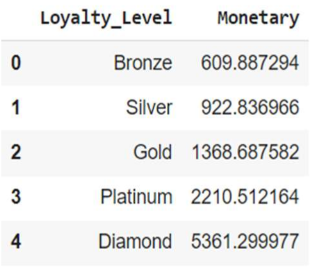
Silver: People in this group have made a transaction on the platform about 102 days ago. Their frequency and monetary values are 3 times and 922 pounds respectively. These are your customers who purchased a decent number of times and spent good amounts, but haven't purchased recently.

Bronze: This is the dormant group with average days since their last purchase is 240. They have transacted around 2 times in the platform with average sales of 609 pounds. These are customers who used to visit and purchase in your platform, but haven't been visiting recently.

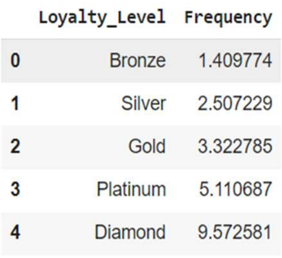
We have also calculated Average amount spent and Average Frequency for different segments.



	Loyalty_Level	count
0	Bronze	1064
1	Silver	830
2	Gold	790
3	Platinum	786
4	Diamond	868



	Loyalty_Level	Monetary
0	Bronze	609.887294
1	Silver	922.836966
2	Gold	1368.687582
3	Platinum	2210.512164
4	Diamond	5361.299977



	Loyalty_Level	Frequency
0	Bronze	1.409774
1	Silver	2.507229
2	Gold	3.322785
3	Platinum	5.110687
4	Diamond	9.572581

Marketing Recommendations:

Now Based on all the Analyses the following Recommendations are made for marketing:

Diamond: Reward these customers so that they can become an early adopter for your future products and help to promote your brand.

Platinum: Offer loyalty programs and reward these customers and give special discounts and help them become your Diamond members.

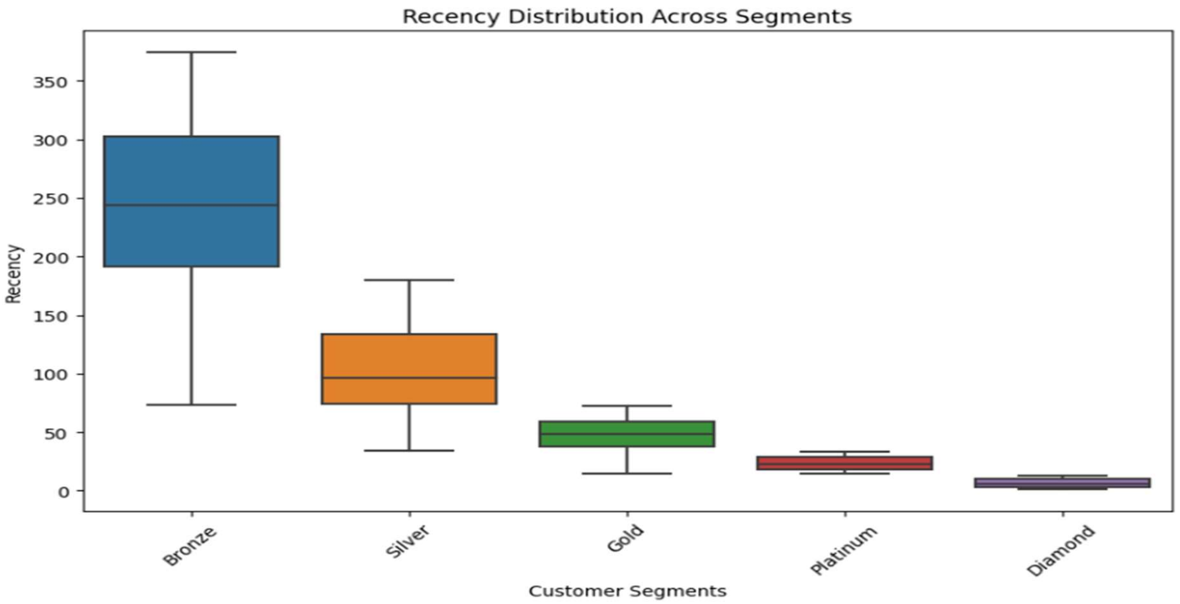
Gold: Offer memberships and recommend related products to upsell them and help them become your Platinum members.

Silver: Sending them personalized campaigns, offers, and product recommendations will help to reconnect with them.

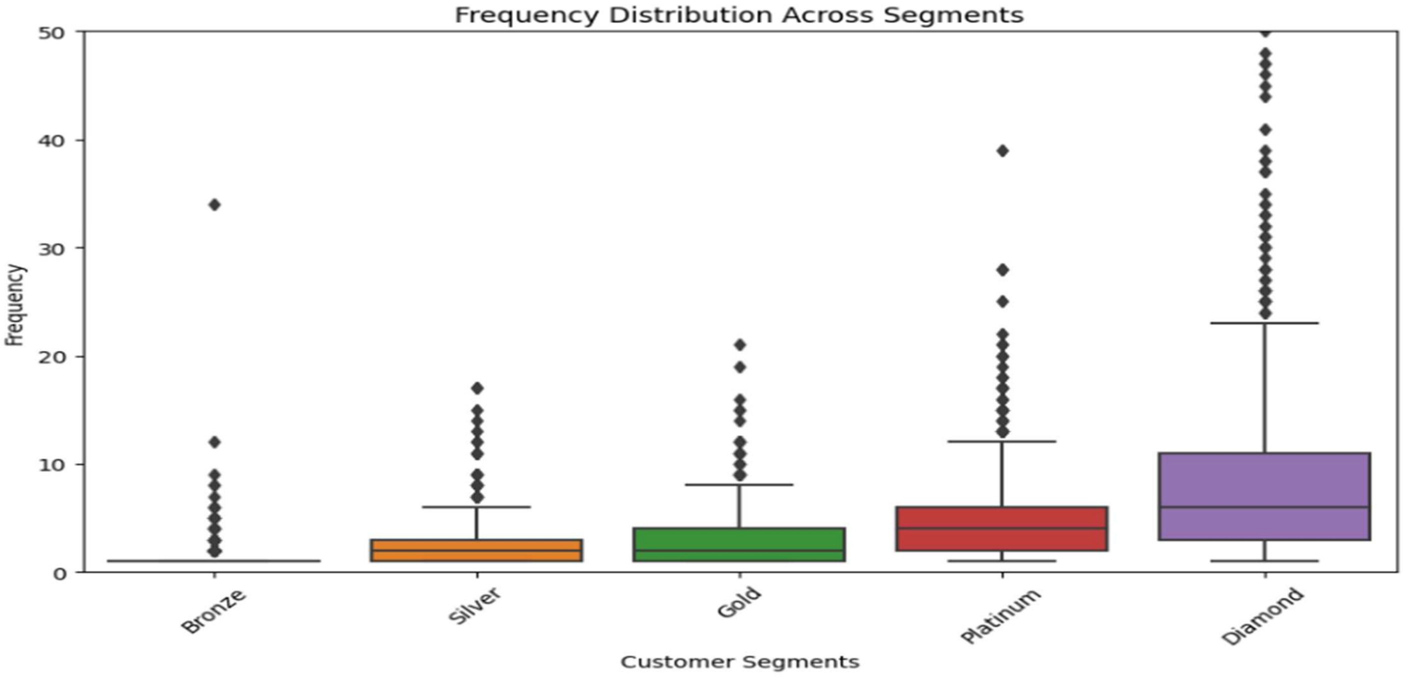
Bronze: Bring them back with relevant promotions, and run surveys to find out what went wrong and avoid losing them to a competitor.

Visualizations:

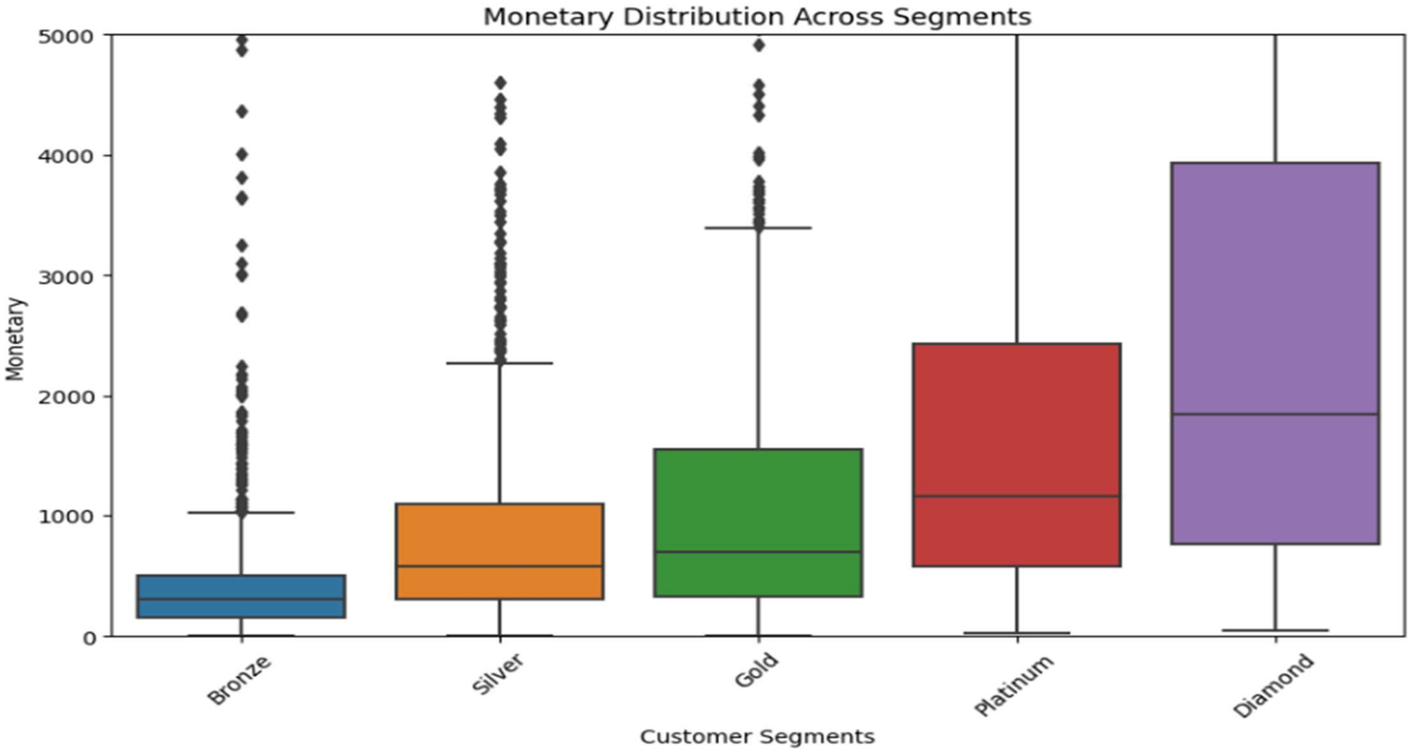
Visualization using Box plot for **Recency Distribution across the different Segments Developed:**



Visualization using Box plot for **Frequency Distribution across the different Segments Developed:**



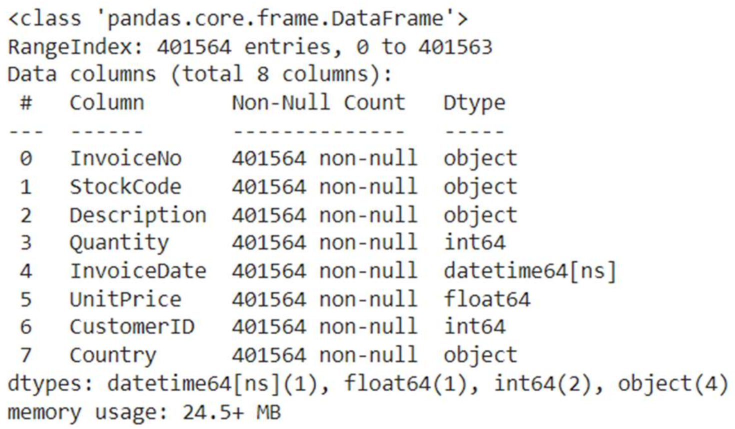
Visualization using Box plot for **Monetary Distribution across the different Segments Developed:**



Data Overview:

Size of the dataset after performing cleaning operations is:

401564 Rows and 8 Columns.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 401564 entries, 0 to 401563
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    401564 non-null  object
 1   StockCode    401564 non-null  object
 2   Description  401564 non-null  object
 3   Quantity     401564 non-null  int64
 4   InvoiceDate  401564 non-null  datetime64[ns]
 5   UnitPrice    401564 non-null  float64
 6   CustomerID   401564 non-null  int64
 7   Country      401564 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(2), object(4)
memory usage: 24.5+ MB
```

Description of each column in the dataset

*InvoiceNo: Invoice number that consists 6 digits. If this code starts with letter 'c', it indicates a cancellation.

*StockCode: Product code that consists 5 digits.

*Description: Product name.

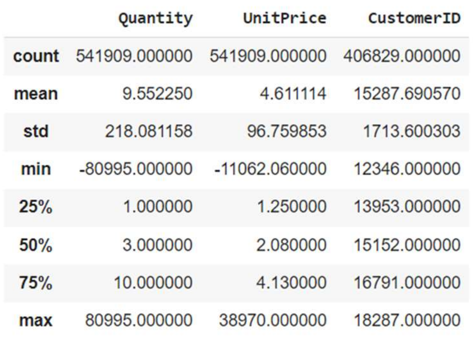
*Quantity: The quantities of each product per transaction.

*InvoiceDate: Represents the day and time when each transaction was generated.

*UnitPrice: Product price per unit.

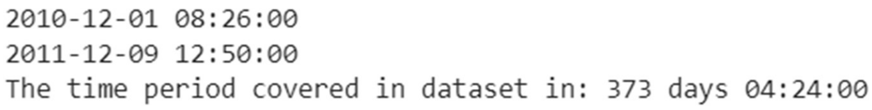
*CustomerID: Customer number that consists 5 digits. Each customer has a unique customer ID.

*Country: Name of the country where each customer resides.



	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

The Time period covered in this dataset is also obtained which is:

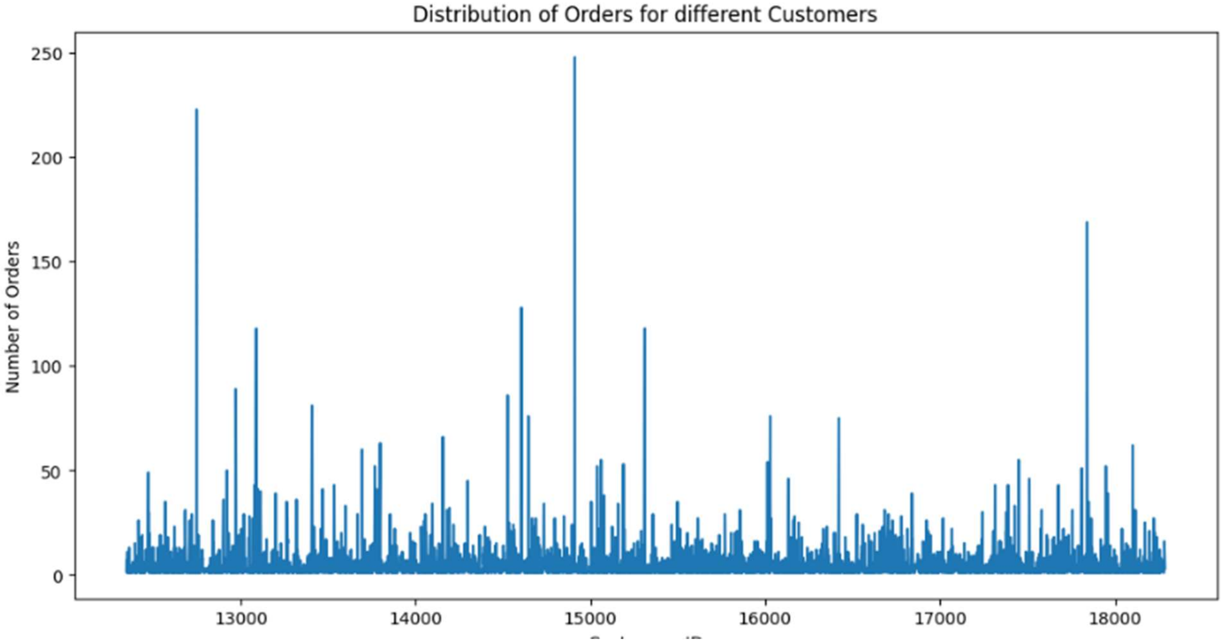


```
2010-12-01 08:26:00
2011-12-09 12:50:00
The time period covered in dataset in: 373 days 04:24:00
```

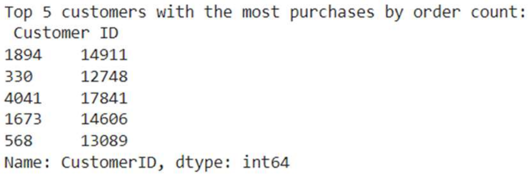
Customer Analysis:

We initiated the analysis by determining the total number of unique customers in the dataset using the 'CustomerID' column, which gives us the count of total 4371 unique customers (The count of unique customers serves as a foundational metric to comprehend the customer base and its diversity).

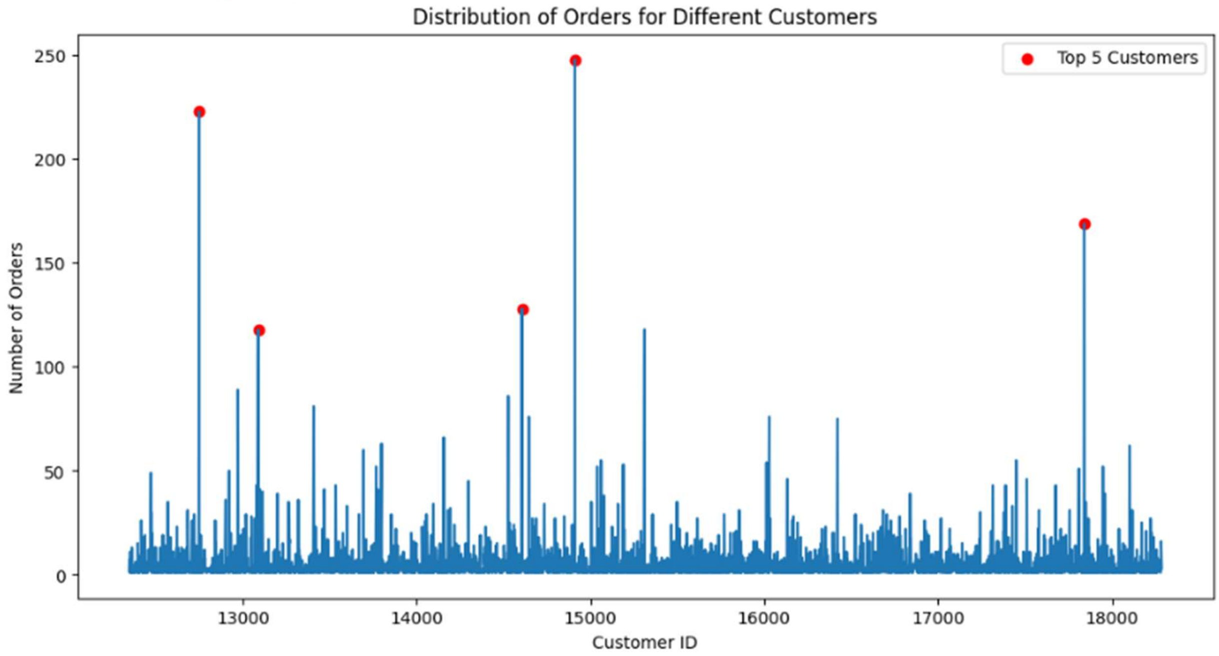
Next, we focused on the distribution of orders across different customers. By grouping the data based on 'CustomerID' and counting the unique 'InvoiceNo' for each customer, we created a visual representation of the order distribution.



To identify the top-performing customers, we sorted the customer data based on the number of orders in descending order. The top 5 customers with the highest order counts were then presented.

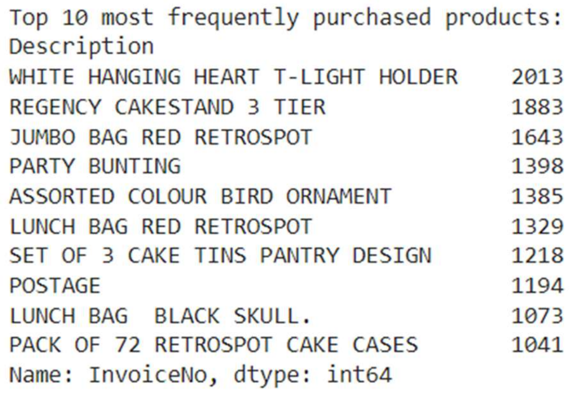


```
Top 5 customers with the most purchases by order count:
 Customer ID
1894    14911
330     12748
4041    17841
1673    14606
568     13089
Name: CustomerID, dtype: int64
```



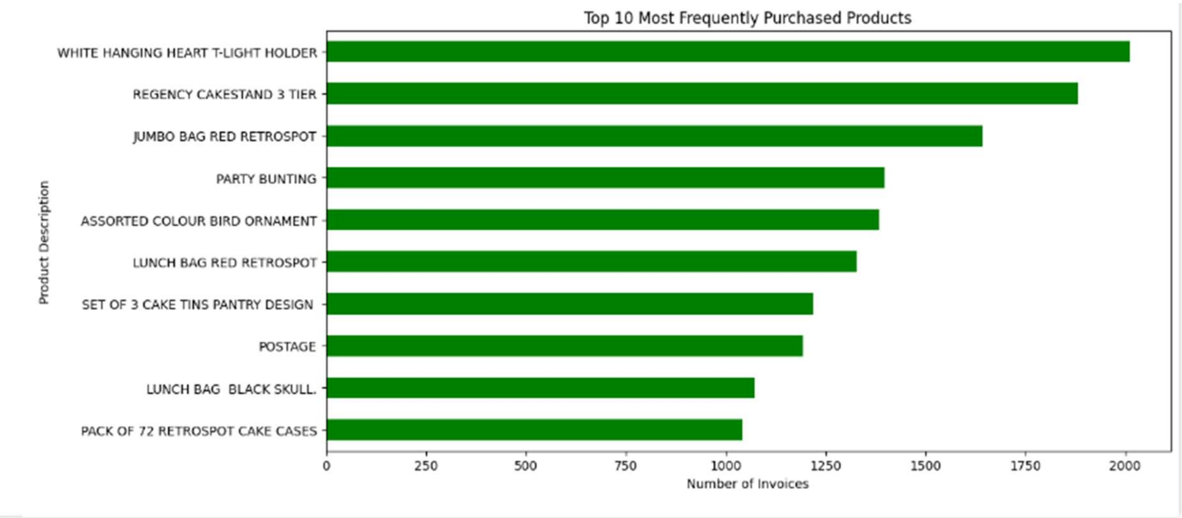
Product Analysis:

Here we have focused on identifying the most frequently purchased products to gain insights into customer preferences and popular inventory items.



```
Top 10 most frequently purchased products:
Description
WHITE HANGING HEART T-LIGHT HOLDER    2013
REGENCY CAKESTAND 3 TIER              1883
JUMBO BAG RED RETROSPOT               1643
PARTY BUNTING                         1398
ASSORTED COLOUR BIRD ORNAMENT         1385
LUNCH BAG RED RETROSPOT               1329
SET OF 3 CAKE TINS PANTRY DESIGN      1218
POSTAGE                               1194
LUNCH BAG  BLACK SKULL.               1073
PACK OF 72 RETROSPOT CAKE CASES       1041
Name: InvoiceNo, dtype: int64
```

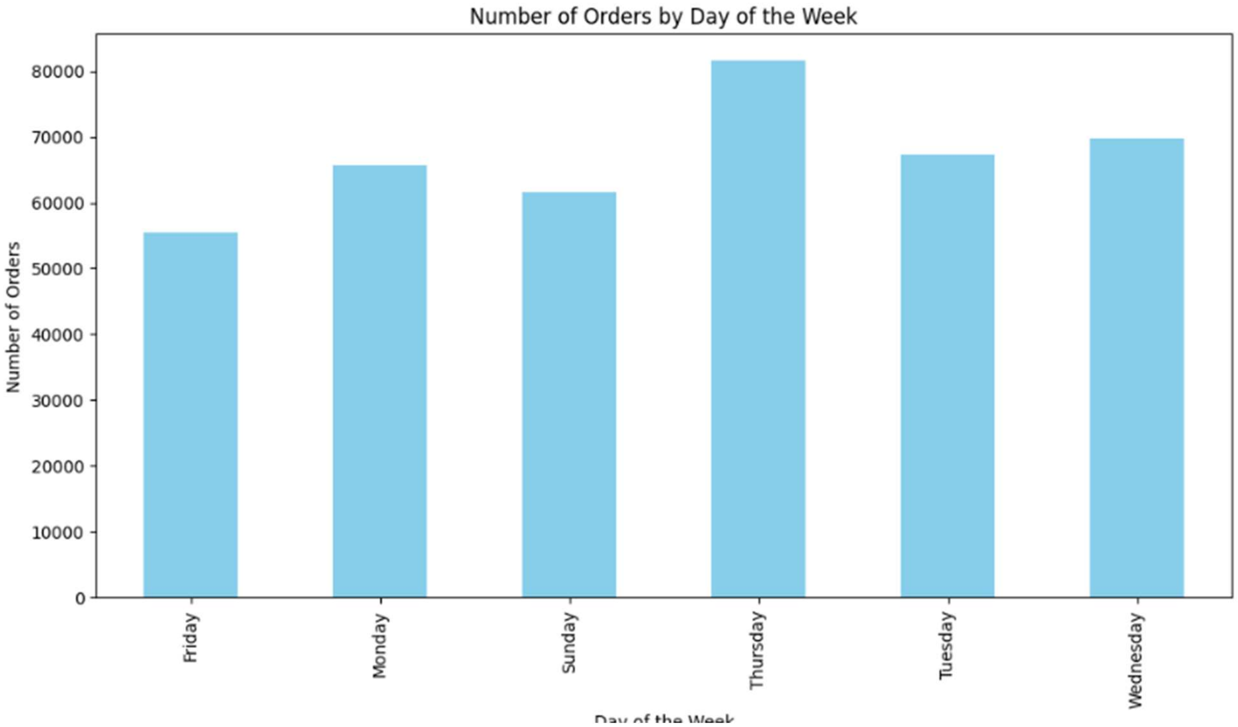
A horizontal bar chart was generated to represent the top 10 most frequently purchased products. The chart provides a clear overview of product popularity, with each bar corresponding to the number of invoices for a specific product.

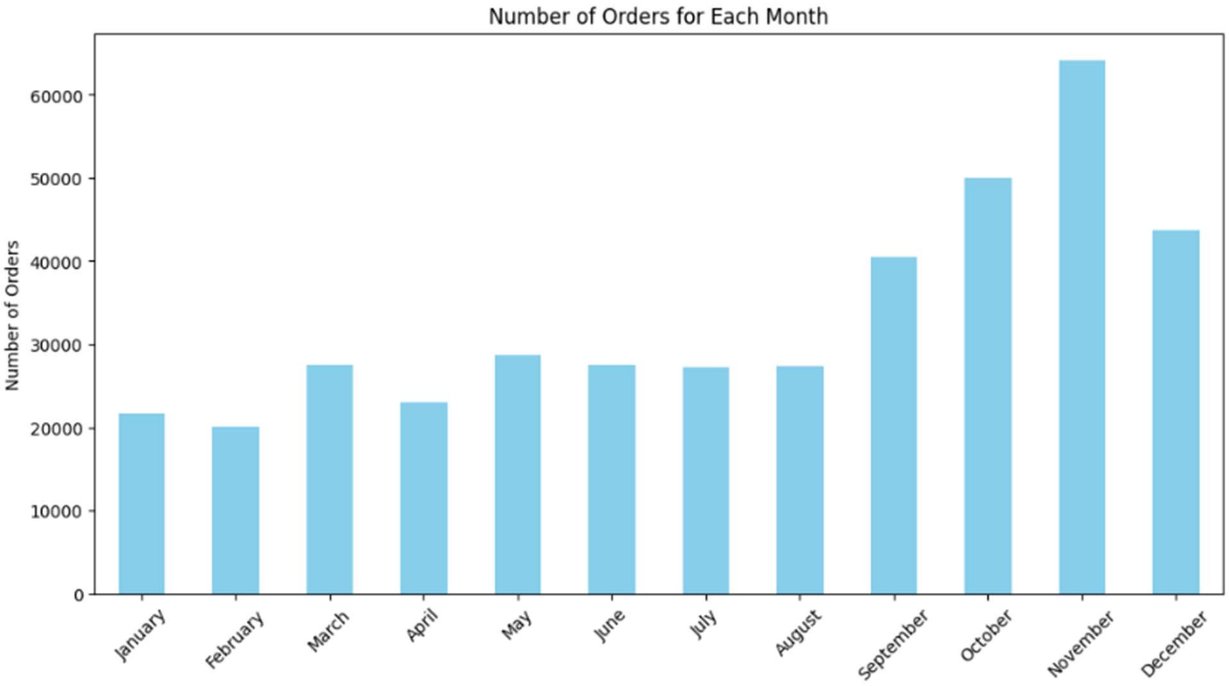


Then we have found that the average price of products to be **1.73** and identified the product category generating highest revenue to **be Paper Craft, Little Birdie** with a total Revenue amount of **168469.00.**

Time Analysis:

By visualizing using bar graph we have identified that Thursday is the day when most orders are placed with at least a 10,000 orders difference between the second highest.

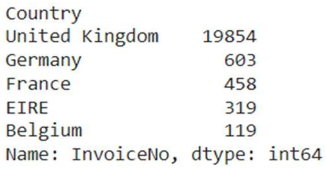




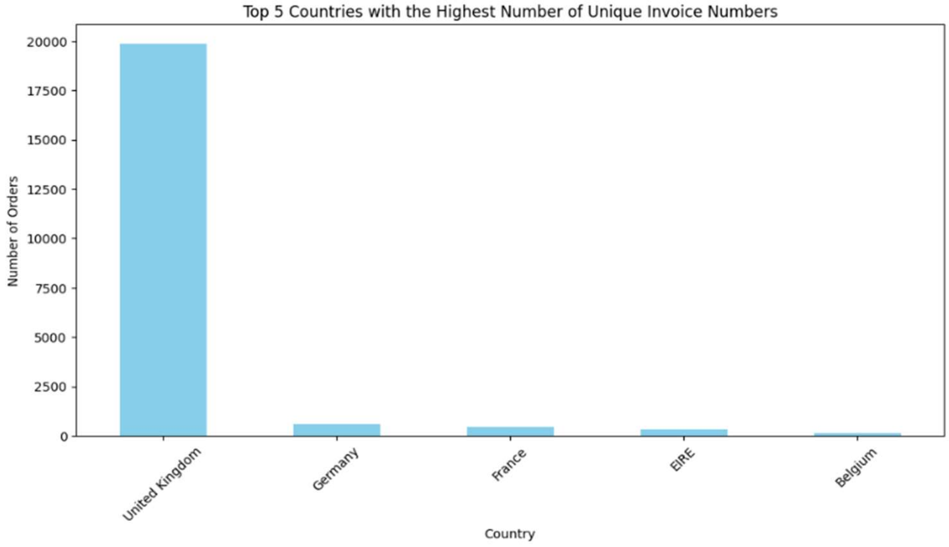
By seeing the above graph, we can say that the months of October to December experiences a surge in number of orders placed which ranges from around 40,000 to 60,000 orders; where as in the remaining months its around 20,000 to 30,000.

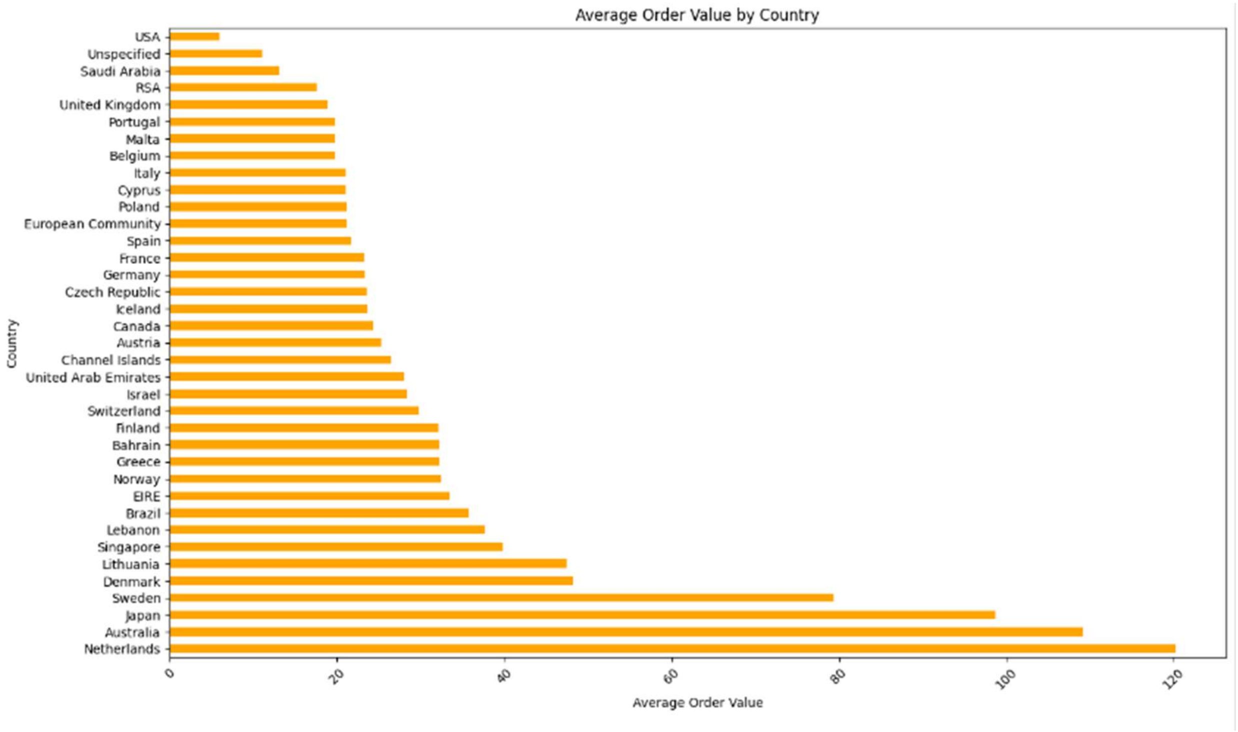
Geographical Analysis:

From the dataset we were able to get that Uk, Germany, France, Eire, Belgium are the top five countries with the highest order volumes.



```
Country
United Kingdom    19854
Germany             603
France              458
EIRE                319
Belgium             119
Name: InvoiceNo, dtype: int64
```



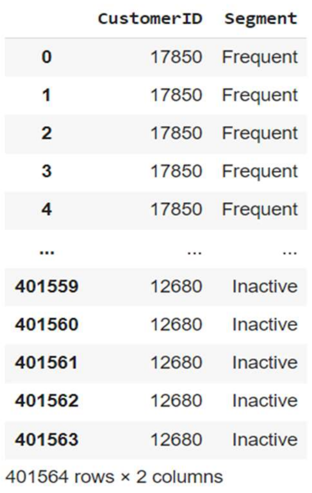


But we were not able to draw correlations between the customer country and the order country because of insufficient data of customer country.

Customer Behavior:

By using the customer Id and invoice date we can observe that average customer activity comes around 243 days on an average. Which says that the wholesalers who are the main customers buy products on an average once or twice a year.

We have also segment customers by their buying frequency into three categories of Inactive, Infrequent, and frequent using quartile method.

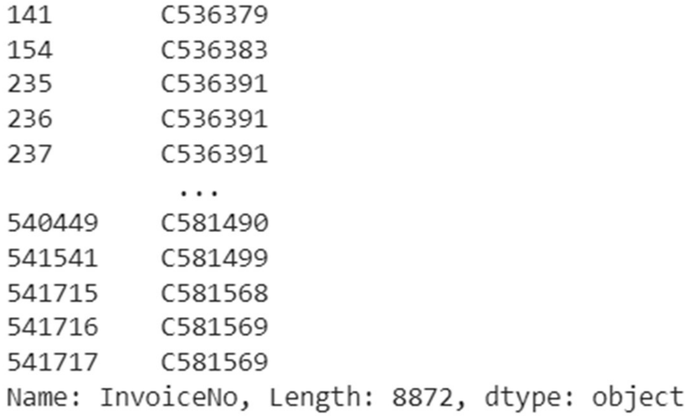


	CustomerID	Segment
0	17850	Frequent
1	17850	Frequent
2	17850	Frequent
3	17850	Frequent
4	17850	Frequent
...	...	...
401559	12680	Inactive
401560	12680	Inactive
401561	12680	Inactive
401562	12680	Inactive
401563	12680	Inactive

401564 rows × 2 columns

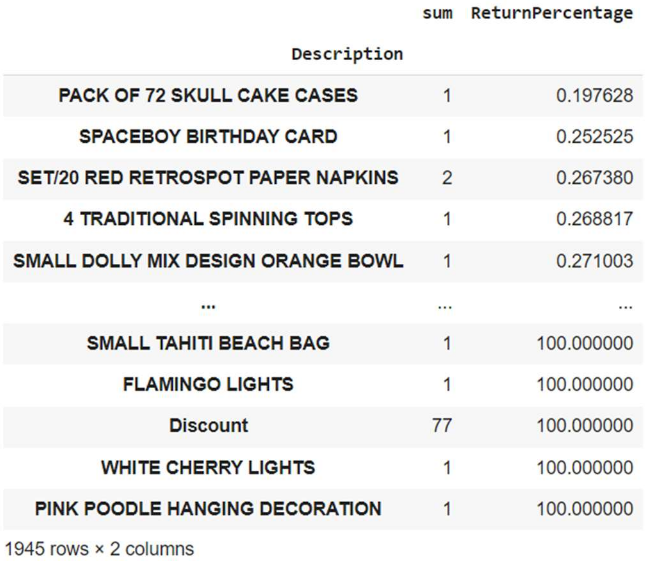
Return and refunds:

By checking the cancelled transactions by checking if the quantity is negative integer in the dataset we were able to get the them which comes to a number of 8872 transactions which is around 2.21%. It can be said that only a short amount of orders are returned and there are no issues with the Quality of the Products.



```
141       C536379
154       C536383
235       C536391
236       C536391
237       C536391
            ...
540449    C581490
541541    C581499
541715    C581568
541716    C581569
541717    C581569
Name: InvoiceNo, Length: 8872, dtype: object
```

We could not form a solid return correlation in association with the product category because of the various products which are returned were ordered in small number of times or only once. But we can see that the 100% of the products with discount are returned which may say that the Quality of the products on discount are not up to par with other products of same category (This is just an assumption based on the data).



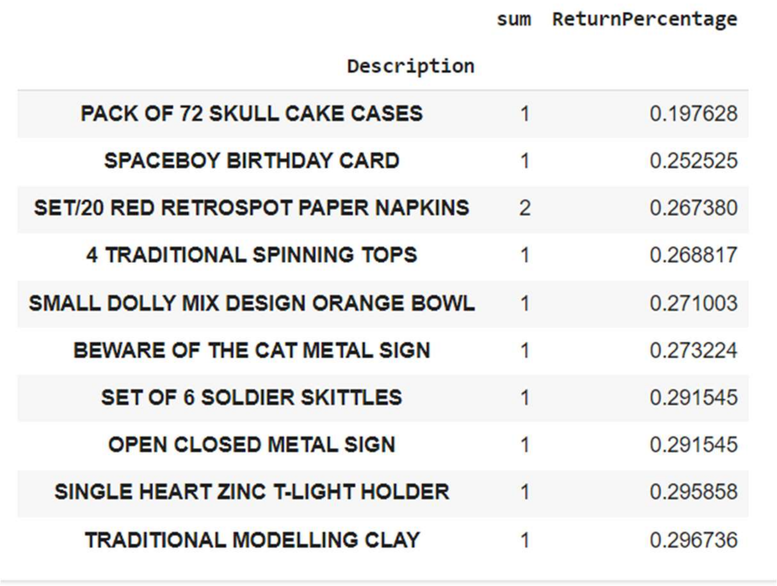
	sum	ReturnPercentage
Description		
PACK OF 72 SKULL CAKE CASES	1	0.197628
SPACEBOY BIRTHDAY CARD	1	0.252525
SET/20 RED RETROSPOT PAPER NAPKINS	2	0.267380
4 TRADITIONAL SPINNING TOPS	1	0.268817
SMALL DOLLY MIX DESIGN ORANGE BOWL	1	0.271003
...	...	...
SMALL TAHITI BEACH BAG	1	100.000000
FLAMINGO LIGHTS	1	100.000000
Discount	77	100.000000
WHITE CHERRY LIGHTS	1	100.000000
PINK POODLE HANGING DECORATION	1	100.000000

1945 rows × 2 columns

Customer Satisfaction:

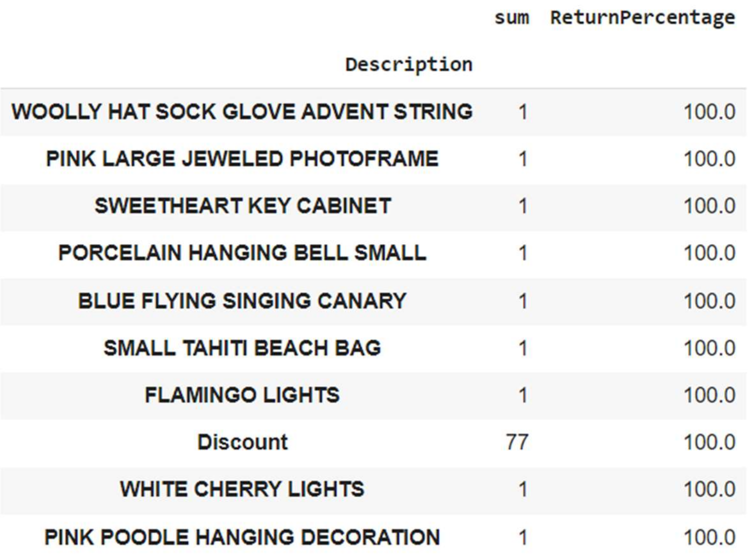
There are no now reviews or ratings for the products purchased so we cannot gleam the exact sentiment of the customer but based on the return percentage metrics we can analyze weather some products are able to satisfy the customer or not by getting the top 10 products which are returned and top 10 products that have less returned based on their return percentages.

Top 10 products with least return percentage:



	sum	ReturnPercentage
Description		
PACK OF 72 SKULL CAKE CASES	1	0.197628
SPACEBOY BIRTHDAY CARD	1	0.252525
SET/20 RED RETROSPOT PAPER NAPKINS	2	0.267380
4 TRADITIONAL SPINNING TOPS	1	0.268817
SMALL DOLLY MIX DESIGN ORANGE BOWL	1	0.271003
BEWARE OF THE CAT METAL SIGN	1	0.273224
SET OF 6 SOLDIER SKITTLES	1	0.291545
OPEN CLOSED METAL SIGN	1	0.291545
SINGLE HEART ZINC T-LIGHT HOLDER	1	0.295858
TRADITIONAL MODELLING CLAY	1	0.296736

Top 10 Products with Most Return Percentages:



	sum	ReturnPercentage
Description		
WOOLLY HAT SOCK GLOVE ADVENT STRING	1	100.0
PINK LARGE JEWELED PHOTOFRAME	1	100.0
SWEETHEART KEY CABINET	1	100.0
PORCELAIN HANGING BELL SMALL	1	100.0
BLUE FLYING SINGING CANARY	1	100.0
SMALL TAHITI BEACH BAG	1	100.0
FLAMINGO LIGHTS	1	100.0
Discount	77	100.0
WHITE CHERRY LIGHTS	1	100.0
PINK POODLE HANGING DECORATION	1	100.0

Limitations and Future Possibilities:

From the data set available to us we were able to analyses from many point of views but there are some types of analysis we were not able perform and they are:

The order processing time for the products, which if available we can see what is the average time taking for the product to reach the customer. This can be done if the shipped date is also available in the dataset and Order processing Time can now be obtained by calculating the days between the shipped date and invoice date.

Due to non-availability of the Customer's Country we were not able to gather what percent of products are ordered by the customers from a different country from their own and whether it is because of the product shortage or non-availability of a particular product in that country.

We were not able perform analyses on the preferred Payment method of the customers when they purchase a product and if its available we can perform Payment analyses to draw conclusions between payment methods and the order amount. This can be achieved if there is data for payment method and this can then be statistically analyzed using t-test.

As we do not have the cost price of the product at which the company has acquired the product we cannot Calculate the profit margins for the product and cannot see which products generate most profits which can be used in marketing the products and which products are suitable for a discount in some seasons.

If there is a category for recording the Reviews, we can see the customer satisfaction and improve upon them to reduce the number of returns and avoid loses for some products and even generate profits by improving the aspects based on the suggestions. Now this can be achieved using NLP Techniques using libraries in Python like NLTK or spaCy to obtain sentiment Analysis.