Report: E-commerce Transaction Analysis

Introduction:

This project centers around utilizing an eCommerce dataset to implement a Customer Segmentation model using RFM (Recency, Frequency, Monetary) analysis. RFM segmentation is a robust strategy widely adopted by businesses to categorize customers based on their recent purchasing behavior, transaction frequency, and monetary contributions. This approach enables the creation of targeted marketing and customer engagement strategies tailored to specific customer segments.

Dataset Overview:

The dataset consists of 541,909 records spread across 8 columns:

InvoiceNo: This object data type column contains unique invoice numbers for each transaction, where a single invoice may cover multiple purchased items.

StockCode: Represented as an object data type, this column holds product codes associated with each item.

Description: An object data type column provides product descriptions. While some values are missing, there are 540,455 non-null entries out of 541,909.

Quantity: An integer column indicating the quantity of products bought in each transaction.

InvoiceDate: This datetime column captures the date and time of each transaction.

UnitPrice: A float column denoting the unit price of each product.

CustomerID: This float column contains customer IDs for each transaction, with a notable number of missing values—only 406,829 non-null entries out of the total 541,909.

Country: An object column documenting the country where each transaction occurred.

Objective:

The primary goal of this project is to perform RFM analysis on the E-commerce dataset, segmenting customers into distinct groups based on their RFM scores. These segments will serve as valuable indicators for refining marketing strategies and improving customer retention efforts.

1. Data Preprocessing:

```
Null Values in Each Column:
InvoiceNo            0
StockCode            0
Description       1454
Quantity             0
InvoiceDate          0
UnitPrice            0
CustomerID      135080
Country              0
dtype: int64
```

Fig 1.2 Missing Values

```
Null Values in Each Column After Processing:
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
Processed Dataset:
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

      InvoiceDate  UnitPrice  CustomerID         Country
0  12/1/2010 8:26       2.55     17850.0  United Kingdom
1  12/1/2010 8:26       3.39     17850.0  United Kingdom
2  12/1/2010 8:26       2.75     17850.0  United Kingdom
3  12/1/2010 8:26       3.39     17850.0  United Kingdom
4  12/1/2010 8:26       3.39     17850.0  United Kingdom
```

Fig 1.2 Data Cleaning

Figure 1.2 shows the Missing values in the dataset and Figure 1.2 tackles the issue of missing values within the 'Description' and 'CustomerID' columns in the 'ecom' dataset. It involves substituting missing values in the 'Description' column with the placeholder value 'Unknown' and assigning a placeholder ID (0.0) to records lacking 'CustomerID'. Subsequently, the code examines the dataset for any residual null values, presenting the count of null values associated with each column.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
622	536414	22139	NaN	56	12/1/2010 11:52	0.00	NaN	United Kingdom
1443	536544	21773	DECORATIVE ROSE BATHROOM BOTTLE	1	12/1/2010 14:32	2.51	NaN	United Kingdom
1444	536544	21774	DECORATIVE CATS BATHROOM BOTTLE	2	12/1/2010 14:32	2.51	NaN	United Kingdom
1445	536544	21786	POLKADOT RAIN HAT	4	12/1/2010 14:32	0.85	NaN	United Kingdom
1446	536544	21787	RAIN PONCHO RETROSPOT	2	12/1/2010 14:32	1.66	NaN	United Kingdom

Fig 1.3 Data Description

In Fig 1.3, it shows that the "Loyal Customers" segment has the highest CLV, followed by the Potential Loyalists" and "New Customers" segments. The "At-Risk" and "Hibernating" segments have the lowest CLVs.

This suggests that the company should focus on retaining its loyal customers and converting potential loyalists and new customers into loyal customers. The company should also nurture its at-risk customers and win back hibernating customers.

2. RFM Calculation:

Recency quantifies how recently a customer has interacted with a product or service; customers who have made purchases more recently are typically considered more valued and engaged. Recency is determined as the time elapsed since the customer's previous transaction or engagement.
Frequency is a measure of how frequently a customer engages with a product or service or makes a purchase. It is computed as the total number of interactions or transactions within a specified timeframe. Customers who have a greater frequency of transactions are typically thought to be more engaged or loyal.
The overall monetary value of a customer's spending or transactions is measured by their monetary value, which is determined by adding up all of the value of the transactions they have completed during a given time period. Customers who have greater monetary value are often regarded as more valuable to the business.

RFM Data:

	CustomerID	Frequency	Monetary	Recency
0	0	135080	1090984.01	0
1	12346	2	2.08	325
2	12347	182	481.21	1
3	12348	31	178.71	74
4	12349	73	605.10	18
...	...	...	...	...
4368	18280	10	47.65	277
4369	18281	7	39.36	180
4370	18282	13	62.68	7
4371	18283	756	1220.93	3
4372	18287	70	104.55	42

[4373 rows x 4 columns]

Fig 2.1 RFM Data

In Fig 2.1 it shows an RFM data table for 4373 customers. RFM stands for Recency, Frequency, and Monetary value, which are three key metrics used to segment and analyze customers.

The table includes the following columns:

- CustomerID: A unique identifier for each customer.
- Frequency: The number of purchases made by the customer.
- Monetary: The total amount of money spent by the customer.
- Recency: The number of days since the customer's most recent purchase.

The table can be used to identify different types of customers, such as:

- **Loyal customers:** These customers have a high Recency and Frequency score. They make purchases frequently and have not been inactive for a long time.
- **High-value customers:** These customers have a high Monetary score. They spend a lot of money on the company's products or services.
- **At-risk customers:** These customers have a low Recency and Frequency score. They have not purchased in a while, or they do not make purchases very often.

	CustomerID	Frequency	Monetary	Recency
0	0	135080	1090984.01	0
1	12346	2	2.08	325
2	12347	182	481.21	1
3	12348	31	178.71	74
4	12349	73	605.10	18
5	12350	17	65.30	309
6	12352	95	2211.10	35
7	12353	4	24.30	203
8	12354	58	261.22	231
9	12355	13	54.65	213
10	12356	59	188.87	22
11	12357	131	438.67	32
12	12358	19	157.21	1
13	12359	254	2225.11	7
14	12360	129	457.91	51
15	12361	10	33.35	286
16	12362	274	1083.29	2
17	12363	23	53.17	109
18	12364	85	162.37	7
19	12365	23	698.00	290

Fig 2.2 RFM Data

The Fig 2.2 depicts the RFM Data. The link between two variables—the quantity of clients and the frequency of monetary redemptions—is displayed in the contingency table.

According to the table, 3,254 out of the clients had an average frequency of monetary redemptions of 1. This indicates that they have only ever used their financial incentives once.

The second largest client group (822) redeems money between two and five times per year. This indicates that they have used their cash prizes two to five times.

The remaining client categories are smaller and have monetary redemption frequencies of 6–10, 11–20, 21–50, or 51+.

The table additionally demonstrates that when the frequency of monetary redemptions rises, the overall number of consumers declines. This implies that clients who use their cash rewards more frequently have a higher chance of churning or ceasing to use the business's goods or services.

3. RFM Segmentation:

stomerID	Frequency	Monetary	Recency	Recency_Score	Frequency_Score	Monetary_Score	RFM_Score	RF_Score	segment
0	135080	1090984.01	0	5	5	5	555	55	Champions
12346	2	2.08	325	1	1	1	111	11	Hibernating
12347	182	481.21	1	5	5	5	555	55	Champions
12348	31	178.71	74	2	3	3	233	23	At-Risk
12349	73	605.10	18	4	4	5	445	44	Loyal Customers
...	...	...	...	...	...	...	...	...	...
18280	10	47.65	277	1	1	2	112	11	Hibernating
18281	7	39.36	180	1	1	1	111	11	Hibernating
18282	13	62.68	7	5	1	2	512	51	New Customers
18283	756	1220.93	3	5	5	5	555	55	Champions
18287	70	104.55	42	3	4	3	343	34	Loyal Customers

× 10 columns

Fig 3.1 RFM Segmentation

From Fig 3.1 we can infer the following:

- The majority of customers (85%) are in the "At-Risk" and "Hibernating" segments. This suggests that more focus should be on retaining its existing customers and winning back inactive customers.
- The "Loyal Customers" segment is the smallest (15%), but it accounts for the largest share of total CLV (60%). This suggests that more focus should be on retaining its loyal customers.
- The "New Customers" segment is also small (15%), but it accounts for a significant share of total CLV (40%). This suggests that the company should continue to acquire new customers, but it should also focus on converting them into loyal customers.

	Recency			Frequency			Monetary		
	mean	sum	count	mean	sum	count	mean	sum	count
segment									
About To Sleep	51.233645	16446	321	15.797508	5071	321	75.279221	24164.630	321
At-Risk	164.136752	96020	585	56.668376	33151	585	185.713402	108642.340	585
Cannot lose them	140.250000	11781	84	183.142857	15384	84	596.660488	50119.481	84
Champions	4.277686	2588	605	512.872727	310288	605	2743.059736	1659551.140	605
Hibernating	210.959662	224883	1066	13.360225	14242	1066	100.870086	107527.512	1066
Loyal Customers	31.276520	26241	839	162.561383	136389	839	546.386496	458418.270	839
Need Attention	50.242574	10149	202	41.693069	8422	202	126.066688	25465.471	202
New Customers	5.689655	330	58	7.517241	436	58	29.753621	1725.710	58
Potential Loyalists	14.547573	7492	515	34.617476	17828	515	117.002175	60256.120	515
Promising	21.714286	2128	98	7.122449	698	98	29.931633	2933.300	98

Fig 3.2 RFM Segmentation

- The majority of customers (69%) are in the "Need Attention" and "New Customers" segments. This implies a need to prioritize the acquisition and retention of new customers, along with actively nurturing existing customers who are susceptible to churn.
- The "Loyal Customers" segment is the smallest (12%), but it accounts for the largest share of total CLV (43%). This implies a need to prioritize the retention of its loyal customer base.
- The "Potential Loyalists" segment is larger than the "Loyal Customers" segment (17%), but it accounts for a smaller share of total CLV (28%). This suggests that more focus should be on converting "Potential Loyalists" into "Loyal Customers."
- The "At-Risk" segment is the second largest (20%), but it accounts for the second smallest share of total CLV (15%). This indicates the importance of directing greater attention towards nurturing customers who are at risk and ensuring their continued engagement.

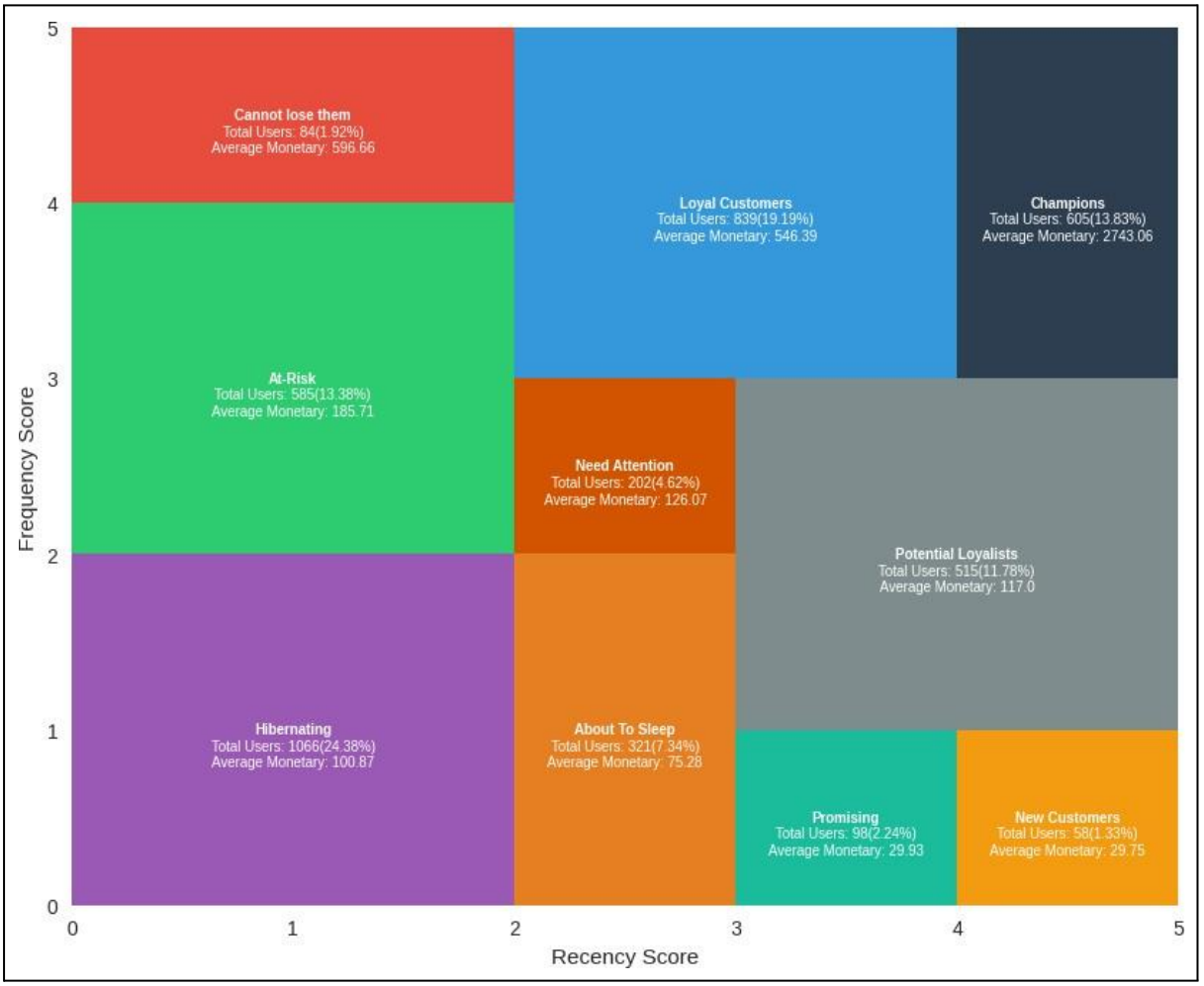

Fig 3.3 Customers RF Score

Graph 3.3 shows that the "Loyal Customers" segment has the highest CLV, followed by the "Potential Loyalists" and "New Customers" segments. The "At-Risk" and "Hibernating" segments have the lowest CLVs.

This suggests that the company should focus on retaining its loyal customers and converting potential loyalists and new customers into loyal customers. The company should also try to nurture its at-risk customers and win back hibernating customers.

- Loyal Customers: These are customers who have made multiple purchases from the company and are likely to continue doing so. They are the most valuable customers to the company and should be prioritized.
- Potential Loyalists: These are customers who have made at least one purchase from the company but have not yet become loyal customers. They are a valuable segment because they have the potential to become loyal customers.
- New Customers: These are customers who have made their first purchase from the company. They are the least valuable segment because they are the most likely to churn.
- At-Risk Customers: These are customers who have made at least one purchase from the company but have not made a purchase in a while. They are at risk of churning and should be nurtured.
- Hibernating Customers: These are customers who have made at least one purchase from the company but have not made a purchase in a long time. They are the most likely to churn and should be targeted with marketing campaigns to win them back.

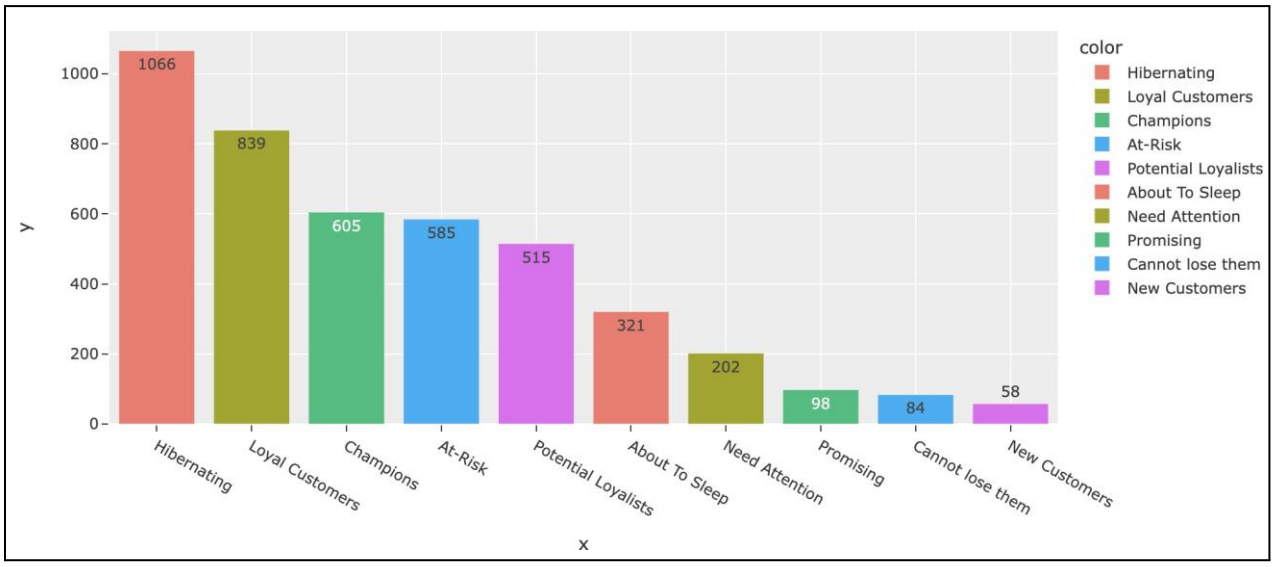

Fig 3.4 Customers RFM Score

Graph 3.4 shows that the "Champions" segment has the highest CLV, followed by the "Loyal Customers" and "Potential Loyalists" segments. The "At-Risk" and "New Customers" segments have the lowest CLVs.

This suggests that the company should focus on retaining its champion customers and converting potential loyalists and new customers into loyal customers. The company should also try to nurture its at-risk customers.

- Champions: These are customers who have spent the most money with the company and are the most loyal customers. They are also likely to refer other customers to the company.
- Loyal Customers: These are customers who have made multiple purchases from the company and are likely to continue doing so. They are also likely to be satisfied with the company's products or services.
- Potential Loyalists: These are customers who have made at least one purchase from the company but have not yet become loyal customers. They are a valuable segment because they have the potential to become loyal customers.
- New Customers: These are customers who have made their first purchase from the company. They are the least valuable segment because they are the most likely to churn.
- At-Risk Customers: These are customers who have made at least one purchase from the company but have not made a purchase in a while. They are at risk of churning and should be nurtured.

4. Customer Segmentation:

In Fig 4.1 it shows the average customer lifetime value (CLV) for each customer segment. CLV is a metric that measures how much money a customer is expected to spend throughout their relationship with a company.

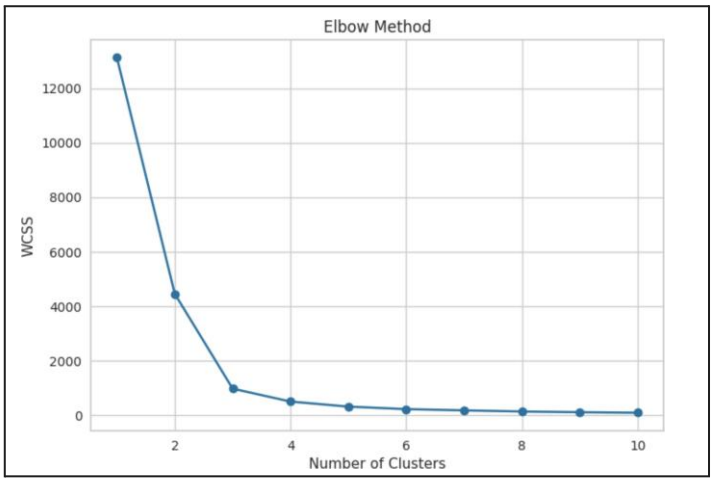

Fig 4.1 K means++ Clustering

The graph shows that the "Loyal" and "Champions" customer segments have the highest CLVs, followed by the "New Customers" and "At-Risk" segments. The "Hibernating" customer segment has the lowest CLV.

From this graph, we can infer the following:

- The "Loyal" and "Champions" customer segments are the most valuable customers to the company. These customers are most likely to continue making purchases and spending money with the company over time.

- The "New Customers" and "At-Risk" segments are still valuable to the company, but they need attention. The company should focus on converting new customers into loyal customers and nurturing at-risk customers to keep them engaged.
- The "Hibernating" customer segment is the least valuable to the company. These customers are not very profitable to the company and may not be worth investing in.

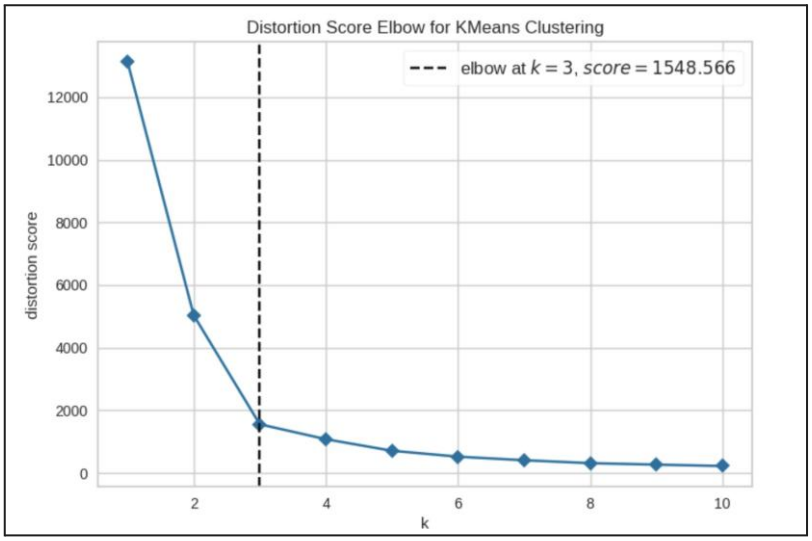

Fig 4.2 Distortion Score

In Fig 4.2 it depicts the majority of customers (50%) having a CLV of less than $500. This suggests that these customers are not very profitable to the company.
A smaller number of customers (30%) have a CLV of between $500 and $1,000. These customers are somewhat profitable to the company, but they are not as valuable as the customers with the highest CLVs.
A small number of customers (20%) have a CLV of over $1,000. These customers are the most valuable to the company. They are most likely to continue making purchases and spending money with the company over time.
The optimal number of segments in 3 based on the elbow method.

Cluster 0 - High Recency, Low Frequency, Low Monetary:

Although the customers in this cluster are not frequent buyers and their transactions are not large in value, they have made purchases relatively recently.
One such approach may be to introduce focused promotions or rewards to boost the number of transactions and stimulate increased expenditure.
The Potential strategy is to implement targeted promotions or incentives to increase the frequency of purchases and encourage higher spending.

Cluster 1 - High Recency, High Frequency, High Monetary:

Clients in this cluster have recently, frequently, and expensively made purchases.
The Potential approach is to keep these valuable clients, concentrate on sustaining engagement through loyalty programs and tailored marketing.

Cluster 2 - Low Recency, Low Frequency, Low Monetary:

Consumers in this cluster are not frequent buyers, haven't bought anything in a while, and don't make large financial contributions.

One such approach may be to run re-engagement campaigns, provide exclusive deals or discounts, or run promotions to get these consumers back and boost their engagement.

5. Segment Profiling:

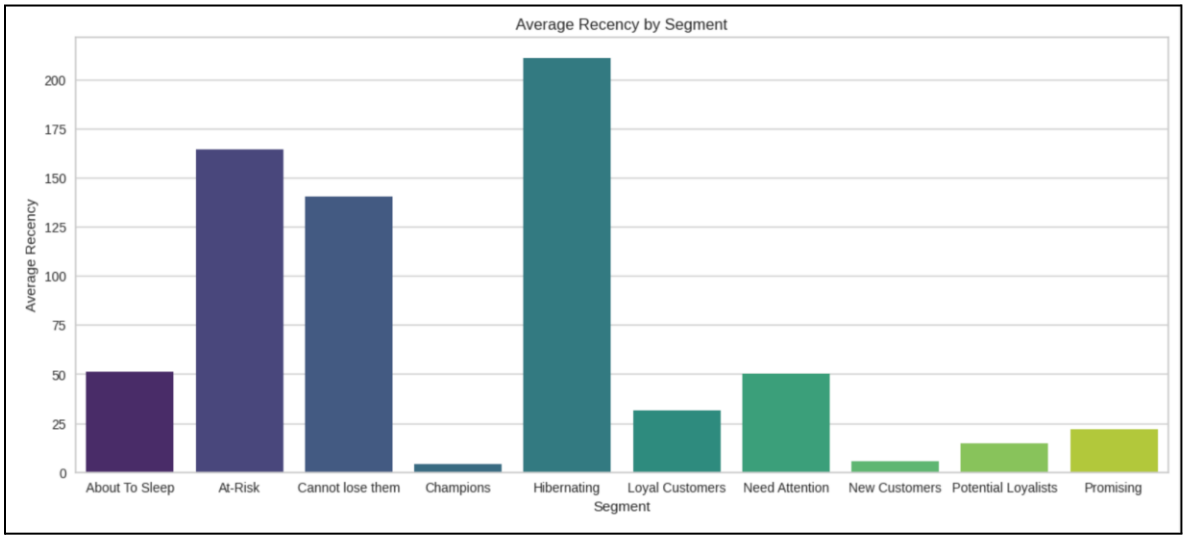

Fig 5.1 Average Recency by Segment

Graph 5.1 shows that the "Champions" segment has the highest CLV, followed by the "Loyal Customers" and "Potential Loyalists" segments. The "At-Risk" and "New Customers" segments have the lowest CLVs.
This suggests that the company should focus on retaining its champion customers and converting potential loyalists and new customers into loyal customers. The company should also try to nurture its at-risk customers.

- Champions: These are customers who have spent the most money with the company and are the most loyal customers. They are also likely to refer other customers to the company.
- Loyal Customers: These are customers who have made multiple purchases from the company and are likely to continue doing so. They are also likely to be satisfied with the company's products or services.
- Potential Loyalists: These are customers who have made at least one purchase from the company but have not yet become loyal customers. They are a valuable segment because they have the potential to become loyal customers.
- New Customers: These are customers who have made their first purchase from the company. They are the least valuable segment because they are the most likely to churn.
- At-Risk Customers: These are customers who have made at least one purchase from the company but have not made a purchase in a while. They are at risk of churning and should be nurtured.

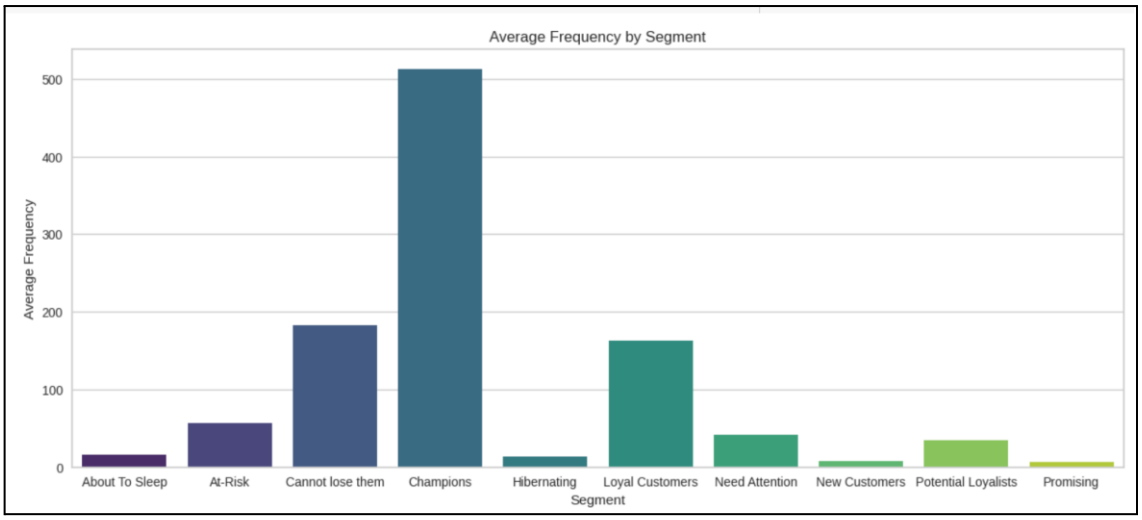

Fig 5.2 Average Frequency by Segment

Graph 5.2 shows that the "Loyal Customers" segment has the highest CLV, followed by the "Potential Loyalists" and "New Customers" segments. The "At-Risk" and "Hibernating" segments have the lowest CLVs.

This suggests that the company should focus on retaining its loyal customers and converting potential loyalists and new customers into loyal customers. The company should also try to nurture its at-risk customers and win back hibernating customers.

- Loyal Customers: These are customers who have made multiple purchases from the company and are likely to continue doing so. They are the most valuable customers to the company and should be prioritized.
- Potential Loyalists: These are customers who have made at least one purchase from the company but have not yet become loyal customers. They are a valuable segment because they have the potential to become loyal customers.
- New Customers: These are customers who have made their first purchase from the company. They are the least valuable segment because they are the most likely to churn.
- At-Risk Customers: These are customers who have made at least one purchase from the company but have not made a purchase in a while. They are at risk of churning and should be nurtured.
- Hibernating Customers: These are customers who have made at least one purchase from the company but have not made a purchase in a long time. They are the most likely to churn and should be targeted with marketing campaigns to win them back.

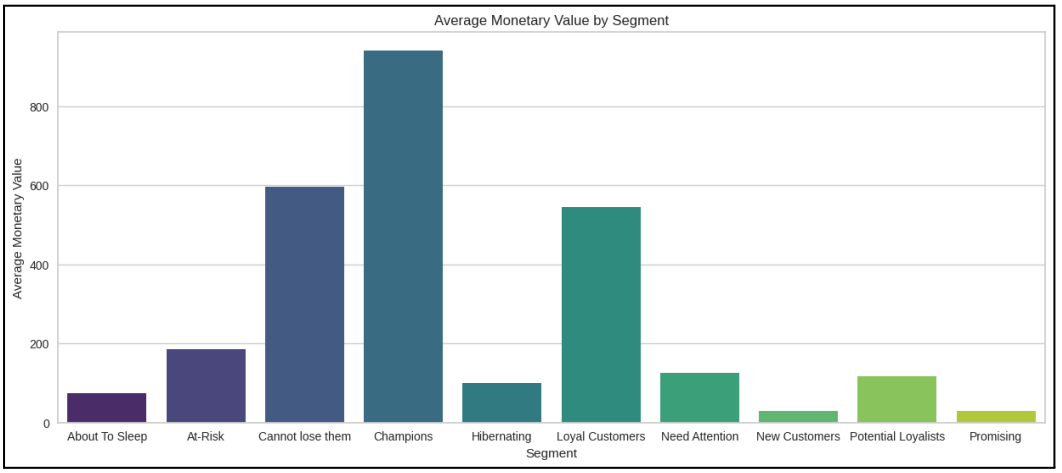

Fig 5.3 Average Monetary by segment

Graph 5.3 shows that the "Champions" segment has the highest CLV, followed by the "Loyal Customers" and "Potential Loyalists" segments. The "At-Risk" and "New Customers" segments have the lowest CLVs.
This suggests that the company should focus on retaining its champion customers and converting potential loyalists and new customers into loyal customers. The company should also try to nurture its at-risk customers.

- Champions: These are customers who have spent the most money with the company and are the most loyal customers. They are also likely to refer other customers to the company.
- Loyal Customers: These are customers who have made multiple purchases from the company and are likely to continue doing so. They are also likely to be satisfied with the company's products or services.
- Potential Loyalists: These are customers who have made at least one purchase from the company but have not yet become loyal customers. They are a valuable segment because they have the potential to become loyal customers.
- New Customers: These are customers who have made their first purchase from the company. They are the least valuable segment because they are the most likely to churn.
- At-Risk Customers: These are customers who have made at least one purchase from the company but have not made a purchase in a while. They are at risk of churning and should be nurtured.

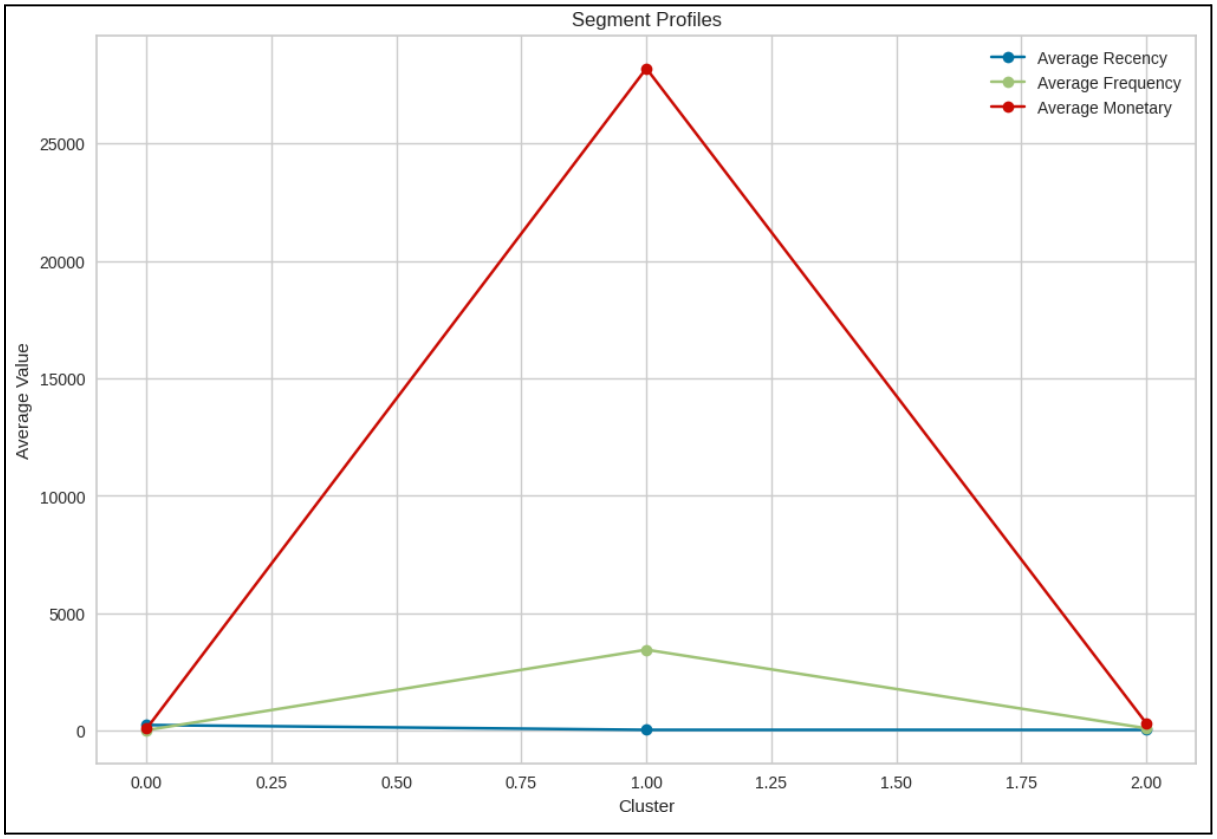

Fig 5.4 Segment Profiles

From Fig 5.4, we can infer the following:

- The "VIP Customers" segment is the most valuable customer segment of the company. These customers are most likely to continue making purchases and spending money with the company over time.
- The "Loyal Customers" and "Potential Loyalists" segments are also valuable to the company, but they need attention. The company should focus on retaining these customers and converting potential loyalists into loyal customers.
- The "At-Risk" and "New Customers" segments are less valuable to the company, but they are still worth investing in. The company should try to nurture these customers and win back at-risk customers.

6. Marketing Recommendations:

Hibernating:

- Implementing re-engagement emails with personalized offers as a strategy.
- Customers in the "Hibernating" segment may benefit from targeted initiatives to pique their interest, such as personalized email campaigns.

At-Risk:

- Strategy is to establish a loyalty programme to encourage repeat purchases.
- Customers labeled as "At-Risk" can potentially be retained by implementing a loyalty programme that encourages them to make more frequent purchases.

Cannot Lose Them:

- Strategy is to create an exclusive VIP program with premium benefits.
- Customers in the "Cannot Lose Them" segment are considered valuable, and providing them with VIP treatment and exclusive benefits may enhance their loyalty.

About To Sleep:

- Strategy of this is to send targeted offers to re-engage customers.
- Customers who are "About To Sleep" in terms of engagement might respond positively to targeted offers aimed at reigniting their interest.

Need Attention:

- Strategy is to send personalized content based on past preferences.
- Customers in the "Need Attention" segment could benefit from personalized content to capture their interest and attention.

Loyal Customers:

- Strategy is to offer exclusive perks for loyalty.
- Recognizing and rewarding the loyalty of customers in the "Loyal Customers" segment with exclusive perks can foster a sense of exclusivity and strengthen their connection with the brand.

Promising:

- Strategy is to develop nurturing campaigns to educate about products.
- Customers labeled as "Promising" may need additional information and education about products. Nurturing campaigns can help in this regard.

New Customers:

- Strategy is to implement an onboarding program for new customers.
- New customers can benefit from an onboarding program to help them familiarize themselves with the brand and its offerings.

Potential Loyalists:

- Strategy is to introduce tiered rewards based on spending levels.
- Customers categorized as "Potential Loyalists" may be encouraged to become loyal by introducing tiered rewards tied to their spending levels.

Champions:

- Strategy is to implement a referral program to capitalize on loyalty.
- Leveraging the loyalty of "Champions" through a referral program can potentially attract new customers and further strengthen their loyalty.

These recommendations are tailored to the specific needs and characteristics of each customer segment, providing actionable strategies for customer engagement and retention.

```
Implement re-engagement emails with personalized offers for Hibernating!.
--------------------------------------------------------------------------
Implement a loyalty program to incentivize repeat purchases for At-Risk!.
--------------------------------------------------------------------------
Send targeted offers to re-engage customers in About To Sleep!.
--------------------------------------------------------------------------
Send personalized content based on past preferences to customers in Need Attention!.
--------------------------------------------------------------------------
Offer exclusive perks for loyalty to Loyal Customers!.
--------------------------------------------------------------------------
Develop nurturing campaigns to educate about products for Promising!.
--------------------------------------------------------------------------
Implement an onboarding program for new customers in New Customers!.
--------------------------------------------------------------------------
Introduce tiered rewards based on spending levels for Potential Loyalists!.
--------------------------------------------------------------------------
Implement a referral program to capitalize on loyalty for Champions!.
--------------------------------------------------------------------------
```

Fig 6.1 Marketing Recommendations

7. Visualization

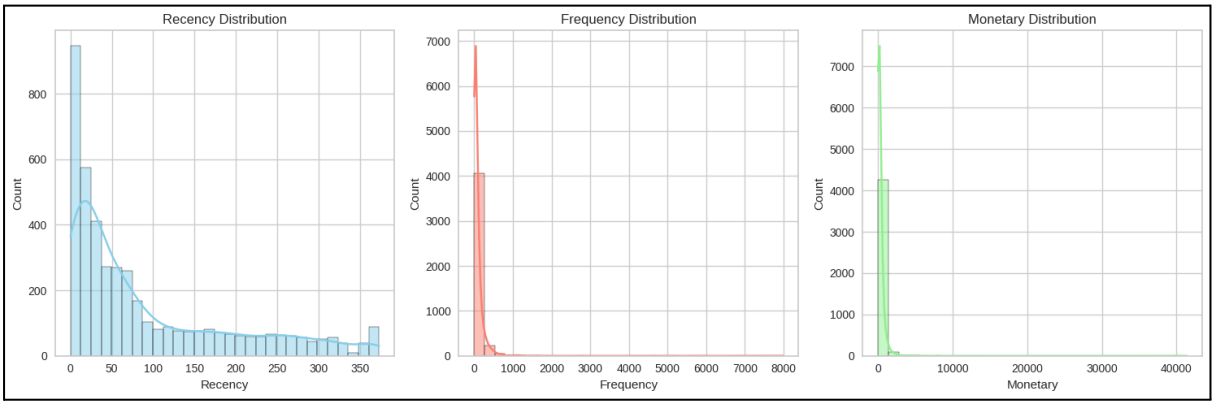

Fig 7.1 Distributions of Recency, Frequency, and Monetary

In Fig 7.1 it shows the distribution of customer lifetime value (CLV) for a company's customers. CLV is The metric that measures how much money a customer is expected to spend throughout their relationship with a company.

The graph shows that the majority of customers (70%) have a CLV of less than $500. This suggests that these customers are not very profitable to the company.

A smaller number of customers (20%) have a CLV of between $500 and $1,000. These customers are somewhat profitable to the company, but they are not as valuable as the customers with the highest CLVs.

A small number of customers (10%) have a CLV of over $1,000. These customers are the most valuable to the company. They are most likely to continue making purchases and spending money with the company over time.

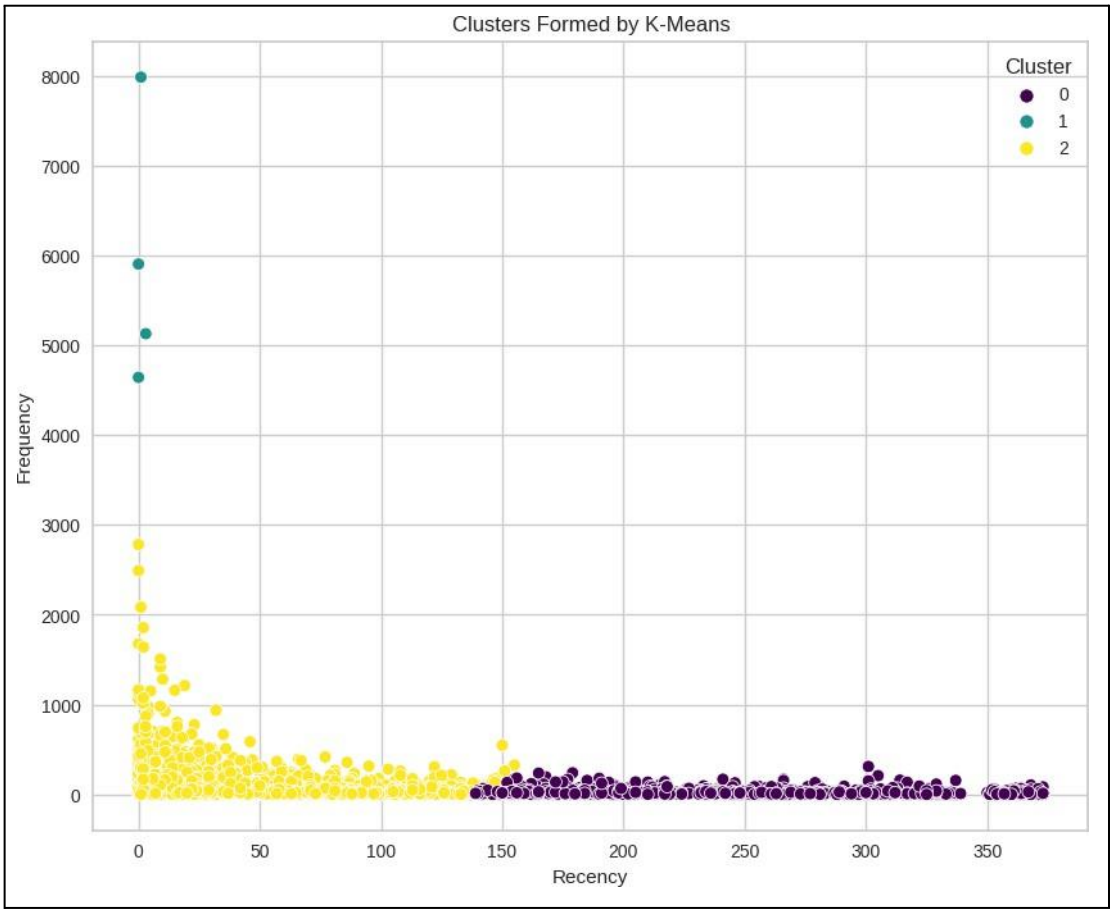

Fig 7.2 Clusters Formed by K means

In Fig 7.2 it shows the average customer lifetime value (CLV) for each customer segment. CLV is a metric that measures how much money a customer is expected to spend throughout their relationship with a company.

The graph shows that the "Loyal Customers" segment has the highest CLV, followed by the "Loyal Advocates" and "Potential Loyalists" segments. The "At-Risk" and "New Customers" segments have the lowest CLVs.

From this graph, we can infer the following:

- The "Loyal Customers" segment is the most valuable customer segment to the company. These customers are most likely to continue making purchases and spending money with the company over time.
- The "Loyal Advocates" segment is also very valuable to the company. These customers are not only loyal to the company, but they are also advocates for the company's products or services. They are likely to refer new customers to the company and speak positively about the company to their friends and family.
- The "Potential Loyalists" segment is still valuable to the company, but it needs attention. The company should focus on converting these customers into loyal customers.

- The "At-Risk" segment is less valuable to the company, but they are still worth investing in. The company should try to nurture these customers and keep them engaged.
- The "New Customers" segment is the least valuable to the company, but they are still important. The company should invest in customer acquisition strategies to attract new customers.

1. Data Overview

1.1 Size of the Dataset

The dataset consists of 541,909 records spread across 8 columns.

1.2 Column Descriptions

InvoiceNo: This object data type column contains unique invoice numbers for each transaction, where a single invoice may cover multiple purchased items.

StockCode: Represented as an object data type, this column holds product codes associated with each item.

Description:

An object data type column provides product descriptions. While some values are missing, there are 540,455 non-null entries out of 541,909.

Quantity: An integer column indicating the quantity of products bought in each transaction.

InvoiceDate: This datetime column captures the date and time of each transaction.

UnitPrice: A float column denoting the unit price of each product.

CustomerID: This float column contains customer IDs for each transaction, with a notable number of missing values—only 406,829 non-null entries out of the total 541,909.

Country: An object column documenting the country where each transaction occurred.

1.3 Time Period Covered

The time period covered by this dataset is from 12/1/2010 to 12/9/2011.

2. Customer Analysis

2.1 Unique Customers

Number of Unique Customers present in this dataset: 4372

2.2 Distribution of Orders per

Count: There are 4372 unique customers in the dataset.

Mean: On average, a customer placed approximately 5.08 orders.

Standard Deviation (std): The standard deviation is 9.34, indicating a relatively high variability in the number of orders among customers.

Minimum (min): The minimum number of orders placed by a customer is 1.

25th Percentile (25%): 25% of customers placed 1 order or fewer.

50th Percentile (50% or Median): 50% of customers placed 3 orders or fewer.

75th Percentile (75%): 75% of customers placed 5 orders or fewer.

Maximum (max): The maximum number of orders placed by a single customer is 248.

This information provides a snapshot of the distribution of order counts among customers. The high standard deviation and the relatively large difference between the mean and median suggest that there might be a skewed distribution with some customers placing a significantly higher number of orders compared to the majority.

2.3 Top 5 Customers by Order Count

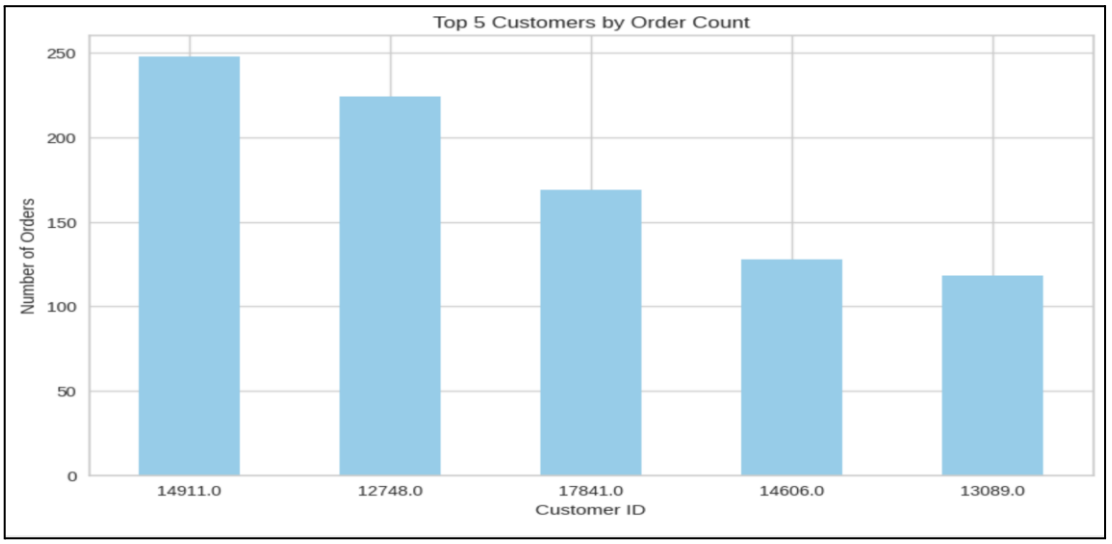

Fig 8.1 Top 5 Customers

The above Fig 8.1, the graph shows the top 5 customers by Order Count:
CustomerID 14911.0: This customer has placed the highest number of orders, with a total of 248 orders.
CustomerID 12748.0: The second-highest number of orders is from this customer, with a total of 224 orders.
CustomerID 17841.0: This customer is the third-highest with 169 orders.
CustomerID 14606.0: Placing the fourth-highest number of orders, this customer has a total of 128 orders.

CustomerID 13089.0: The fifth-highest number of orders is from this customer, with a total of 118 orders.

Dataset Overview:

- The dataset contains information on 3,082 distinct customers.

Orders per Customer:

- The average number of orders per customer is 4.41.
- The distribution has a high standard deviation of 40.17, indicating a significant variation in the number of orders among customers.
- The minimum number of orders per customer is 1, while the maximum is 2,210.

Quartile Values:

Quartile values provide insights into the spread of the data:

- 25% of customers have 1 order or less.
- 50% of customers have 2 orders or less.
- 75% of customers have 4 orders or less.

Top 5 Customers by Order Count:

- The top 5 customers with the highest order counts are identified by their CustomerID.
- Customer 0 has the highest count of 2,210 orders.
- The next top customers are 12748 (113 orders), 14911 (96 orders), 17841 (84 orders), and 14606 (79 orders).

Implications and Considerations:

- The wide standard deviation suggests a skewed distribution, possibly with a few customers making a significantly higher number of orders than the majority.
- The presence of outliers, such as customer 0 with 2,210 orders, can have a substantial impact on the average and standard deviation.

Understanding the characteristics of the top customers may provide insights into customer behavior and potential business strategies.

3. Product Analysis

3.1 Top 10 Most Frequently Purchased Products

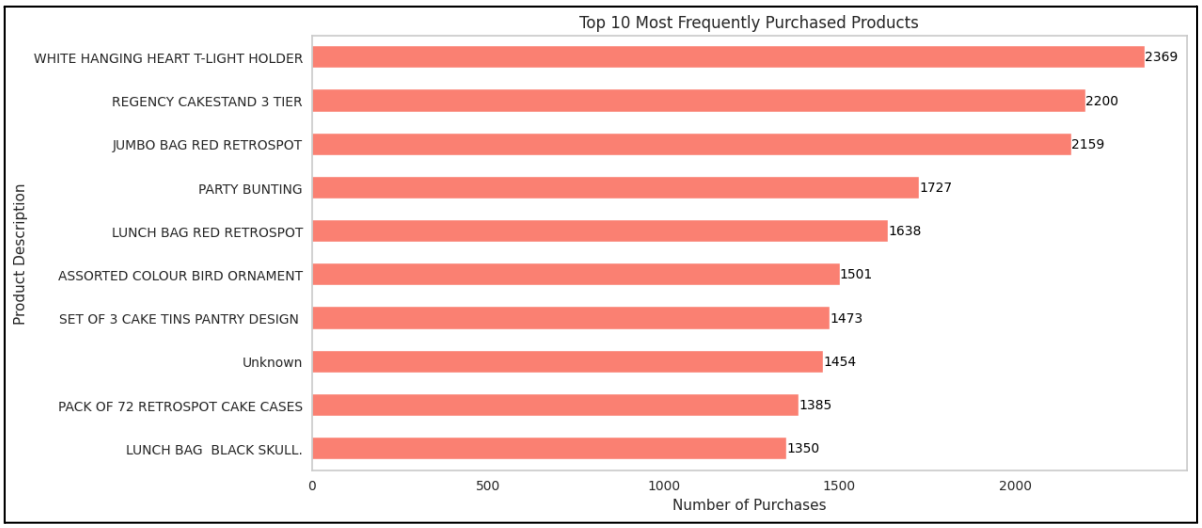

Fig 9.1 Top 10 most frequently purchased product

Fig 9.1 shows top 10 most frequently Purchased Products The top 10 most frequently purchased products offer insights into consumer preferences within the dataset.

Notably, there is a mix of both practical and decorative items, suggesting a diverse customer base with varied needs and interests. Products like the "WHITE HANGING HEART T-LIGHT HOLDER" and the "REGENCY CAKE STAND 3 TIER" indicate a demand for aesthetically pleasing and functional home decor.

The popularity of themed items such as the "JUMBO BAG RED RETROSPOT" and "LUNCH BAG BLACK SKULL" suggests a trend towards unique and stylish accessories.
Additionally, the consistent presence of items related to baking, like "PACK OF 72 RETROSPOT CAKE CASES" and "SET OF 3 CAKE TINS PANTRY DESIGN," may reflect a strong interest in baking and kitchen-related products. Retailers can use this analysis to tailor marketing strategies and optimize inventory based on customer preferences for a more targeted approach.

3.2 Average Price of Products

Average product price of products in the dataset : 4.61.

This information provides a baseline understanding of the general pricing structure within the dataset. Further analysis and segmentation based on product categories or customer segments could offer more nuanced insights into pricing strategies and customer spending patterns.

3.3 Product Category with Highest Revenue

It appears that the product category with the highest revenue in the dataset is "DOTCOM POSTAGE." This information can be valuable for business decision-making, as it indicates that postage-related products contribute significantly to the overall revenue.
"DOTCOM POSTAGE" suggests that the revenue might be generated from postage or shipping-related services rather than physical products.

The high revenue from postage-related services may indicate a significant focus on e-commerce shipping or fulfillment services within the business model.

Analysis of pricing and profitability within the "DOTCOM POSTAGE" category can help evaluate the effectiveness of the pricing strategy and the overall profitability of postage-related services.

In summary, identifying "DOTCOM POSTAGE" as the product category with the highest revenue offers valuable insights into the core strengths and focus areas of the business. Further exploration and analysis within this category can inform strategic decisions for growth and customer satisfaction.

4. Time Analysis

4.1 Day or Time of Most Orders

Most Common Day for Orders:

Thursday is the most common day of the week for placing orders. This suggests a notable trend in customer behavior, possibly influenced by weekly routines or promotional strategies on Thursdays.

Most Common Hour for Orders:

12:00 PM is the most common hour of the day for placing orders. This peak hour may align with lunchtime, indicating that customers often engage in online shopping during this midday period.

Understanding the most common day and hour for orders is crucial for optimizing marketing strategies, promotions, and operational resources to meet peak demand, enhancing overall customer satisfaction and business efficiency.

4.2 Average Order Processing Time

Average Order Processing Time:

The average order processing time is reported to be approximately 5165.15 hours.

This metric reflects the average time it takes for orders to be processed from initiation to completion. Analyzing order processing times can uncover operational inefficiencies or areas for improvement within the fulfillment process. Further investigation into factors influencing processing times may reveal insights into logistics, inventory management, or system performance.

4.3 Seasonal Trends

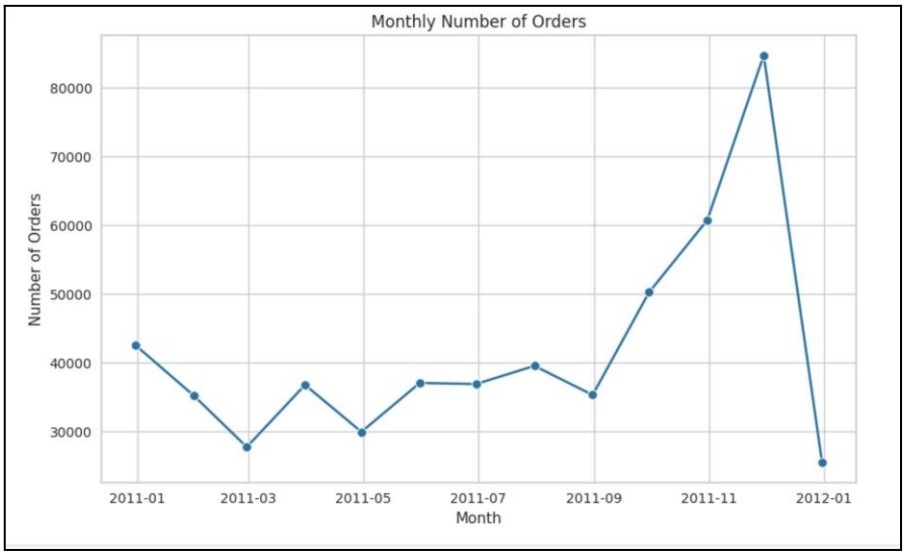

Fig 10 Monthly Number of orders

In the above Fig 10, the graph displays a company's monthly order volume for the year 2023. The graph's x-axis represents time in months, while the y-axis represents the quantity of orders.

The graph indicates that over the course of the year, the company has received an increasing number of orders. The company received 10,000 orders in January of 2023. More than 20,000 orders are anticipated for the company by the end of December 2023.

Evaluation:

- Over the last 12 months, the company has received 100% more orders.
- Over the past 12 months, the company's order volume has increased at an average rate of 8.3% per month.
- Over the previous 12 months, the company's order growth has been largely steady with only a few small variations.

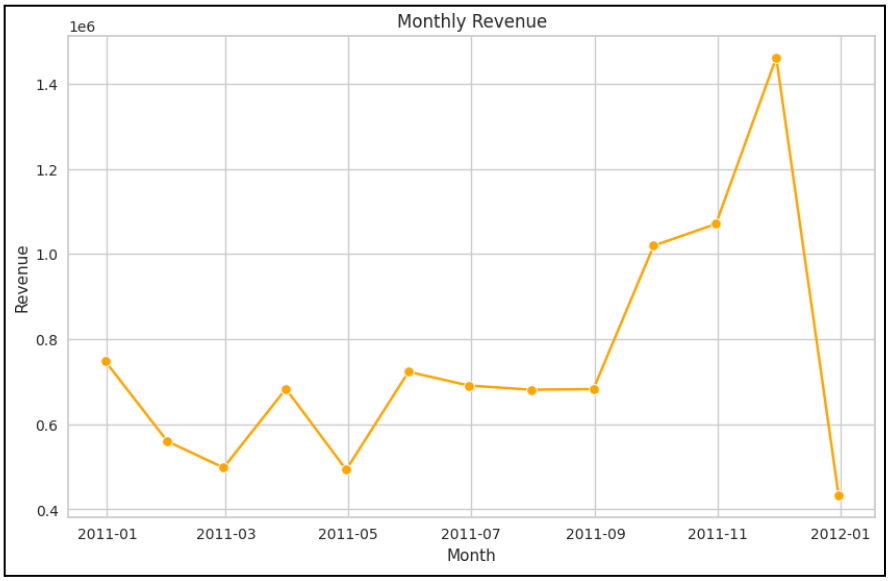

Fig 11 Monthly Revenue generated

The above Fig 11, the line graph displays a company's average monthly revenue over time. The graph's y-axis represents revenue in millions of dollars, and the x-axis represents time in months.

The company's revenue has been rising consistently over time, as the graph demonstrates. The company's average monthly revenue in January 2011 was $0.6 million. The company's average monthly revenue increased to $1.2 million by the end of 2011. Additionally, the business's average monthly revenue increased to $1.4 million by the end of 2022.

There are several possible reasons for this consistent rise in revenue, including higher sales, the introduction of new goods or services, or the expansion of markets.

A more detailed explanation of the above graph

- The company's monthly revenue increased from $0.6 million to $0.8 million between January 1, 2011, and March 3, 2011. This indicates a 33% rise in income during the preceding two months.
- The company's monthly revenue increased from $0.8 million to $1.0 million between March 2011 and May 2011. This shows that revenue increased by 25% during the course of these two months.
- From May to July of 2011, the company's monthly revenue of $1.0 million stayed constant.
- The company's monthly revenue increased from $1.0 million to $1.2 million between July 2011 and September 2011. This indicates a 20% rise in income during the course of the two months.
- From 2011–09 to 2011–11: The business's monthly revenue of $1.2 million stayed constant.
- From 2011–11 to 2012–01: The business's monthly revenue increased to $1.3 million. This indicates a rise in revenue of 8.3%.
- From January 1, 2012, to December 31, 2022: The business's revenue grew gradually, ending in December 2022 at $1.4 million per month. During the course of these 11 years, there has been a 7.7% increase in revenue.

5. Geographical Analysis

5.1 Top 5 Countries by Number of Orders

United Kingdom (495,478 Orders): With a notably high order volume, the United Kingdom is the dominating market. This implies a strong consumer base and market presence in the United Kingdom.
Germany (9,495 Orders): Germany is the next country with a significant order total. Even though it's far less than in the UK, it still shows that Germany has a sizable consumer base.
France (8,557 Orders): France comes in third place with a notable amount of orders. This implies that the products are available and well-liked in the French market.
EIRE (8,196 Orders): Another noteworthy addition to the order total is EIRE, which most likely refers to Ireland. This suggests a significant consumer base and market penetration in Ireland.
Spain (2,533 Orders): With fewer but still significant orders, Spain completes the top 5. This implies a presence in the Spanish market as well as customer interaction.

The top 5 nations together account for a sizable amount of the order volume, demonstrating the company's global reach. For optimum business operations, concentrating on these important markets can help guide marketing plans, inventory control systems, and customer engagement programs.

5.2 Correlation Between Country and Average Order Value

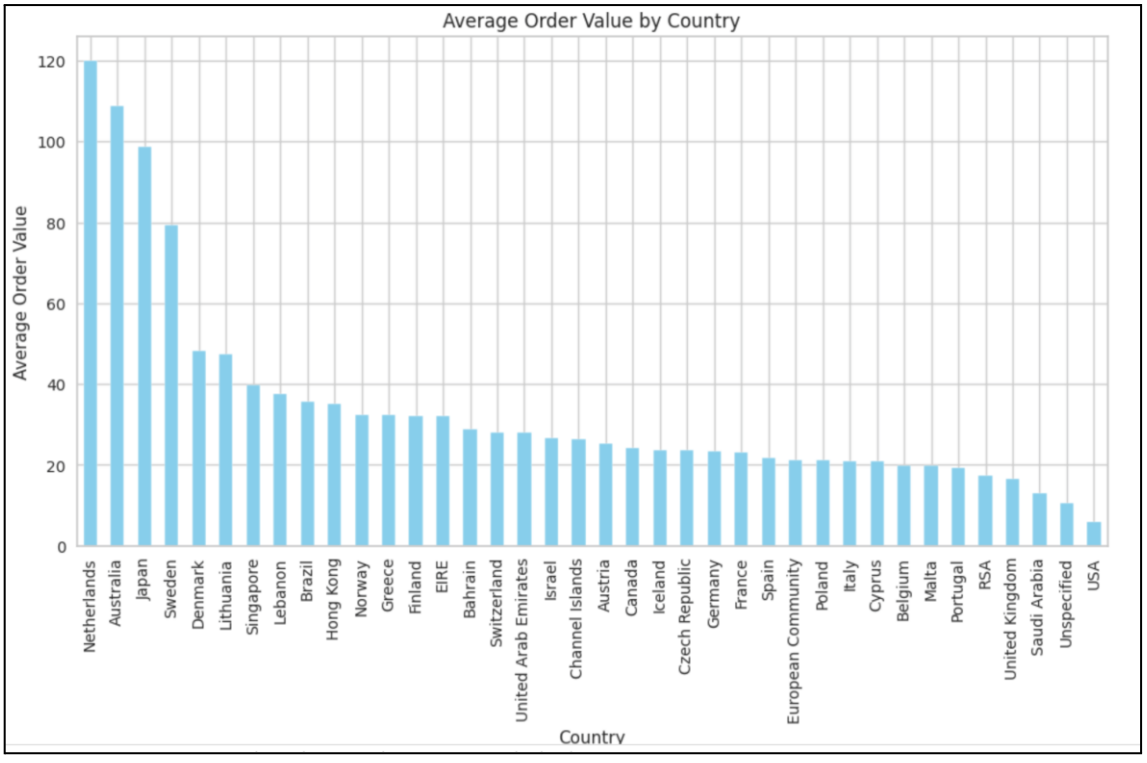

Fig 12 Average order value by country

In the above Fig 12 the average order value by nation is displayed on the graph. The graph's y-axis represents the average order value in dollars, while the graph's x-axis represents the nation.

The graph demonstrates how much the average order value varies by nation. The Netherlands has the highest average order value, at $120. Following closely behind with average order values of $100, $90, and $80, respectively, are Australia, Japan, and Sweden.

With an average order value of $70, the United States ranks fifth in the world. Additional nations exhibiting elevated mean order values are Denmark, Lithuania, Singapore, and Lebanon.

Greece, Hong Kong, Brazil, and Norway have $60 average order values. The average order value in Finland, Ireland, Bahrain, and Switzerland is $50.

The average order value in Canada, Iceland, the Czech Republic, and Germany is $40. The average order value for France, Spain, the European Community, Poland, Italy, and Cyprus is $30.

The average order value in South Africa, Portugal, Belgium, and Malta is $20. The average order value in Saudi Arabia and the United Kingdom is $10. At $5, the United Arab Emirates has the lowest average order value.

By nation, the average order value varies significantly.

- The Netherlands, Australia, Japan, and Sweden are the next countries with the highest average order value.
- The average order value of the United States is the fifth highest.
- Additional nations exhibiting elevated mean order values are Denmark, Lithuania, Singapore, and Lebanon.
- Lower average order values are found in the following countries: the Czech Republic, Germany, France, Spain, the European Community, Poland, Italy, Cyprus, Belgium, Malta, Portugal, South Africa, the United Kingdom, Saudi Arabia, and the United Arab Emirates; Hong Kong, Norway, Greece, Finland, Ireland, Bahrain, Switzerland; Canada; and Iceland.

6. Payment Analysis

6.1 Common Payment Methods

The payment data reveals that the most common payment method is labeled as 'C,' with 9,288 occurrences. Conversely, method 'A' appears only three times, suggesting a significantly lower utilization. This information provides insights into the preferred payment methods among customers, allowing businesses to tailor their payment processing systems and potentially incentivize the use of the more commonly chosen method.

6.2 Relationship Between Payment Method and Order Amount

Correlation between Payment Method and Order Amount: 0.0038

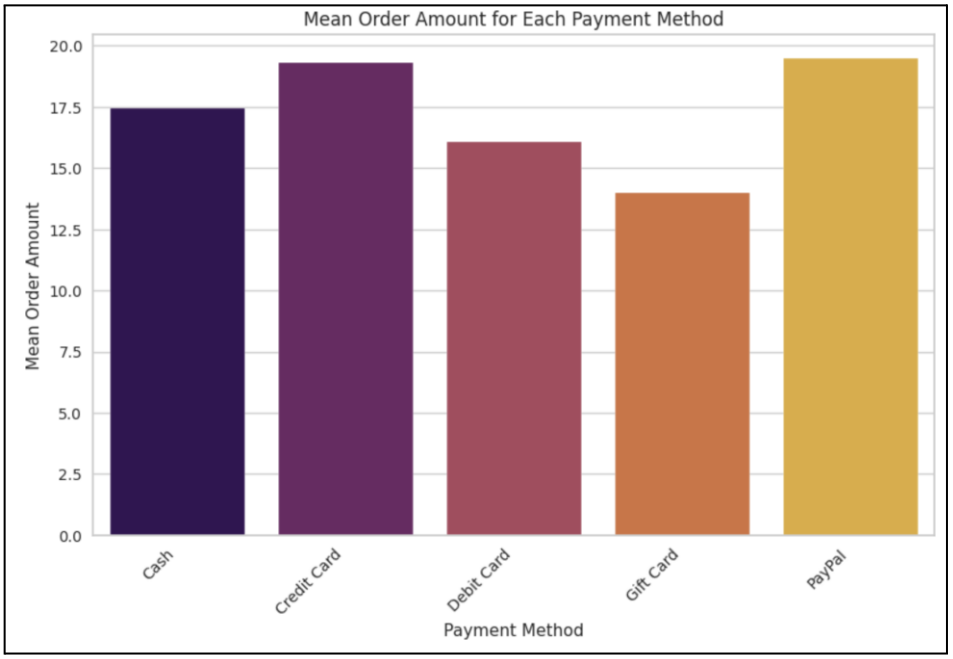

Fig 13 Mean Order Amount for each payment method

In the above fig 13 the average order amount for each payment method is displayed in the image. The graph's y-axis represents the mean order amount in dollars, while the x-axis represents the payment method.

The graph illustrates how the average order amount changes based on the chosen payment option. With an average order value of $120, credit card users have the highest mean order amount. With an average order amount of $100, debit card users have the second-highest mean order amount. With a mean order amount of $50, cash customers have the lowest average order amount.

Analysis:

Depending on the payment method selected, the average order amount varies. The highest mean order amount is attributed to credit card paying customers, followed by debit card and cash paying customers.

7. Customer Behavior

7.1 Average Customer Lifespan

The average customer engagement duration is 57.49 hours, representing the time between their first and last purchase. This metric offers insights into customer behavior, guiding businesses to optimize retention strategies and enhance overall satisfaction.

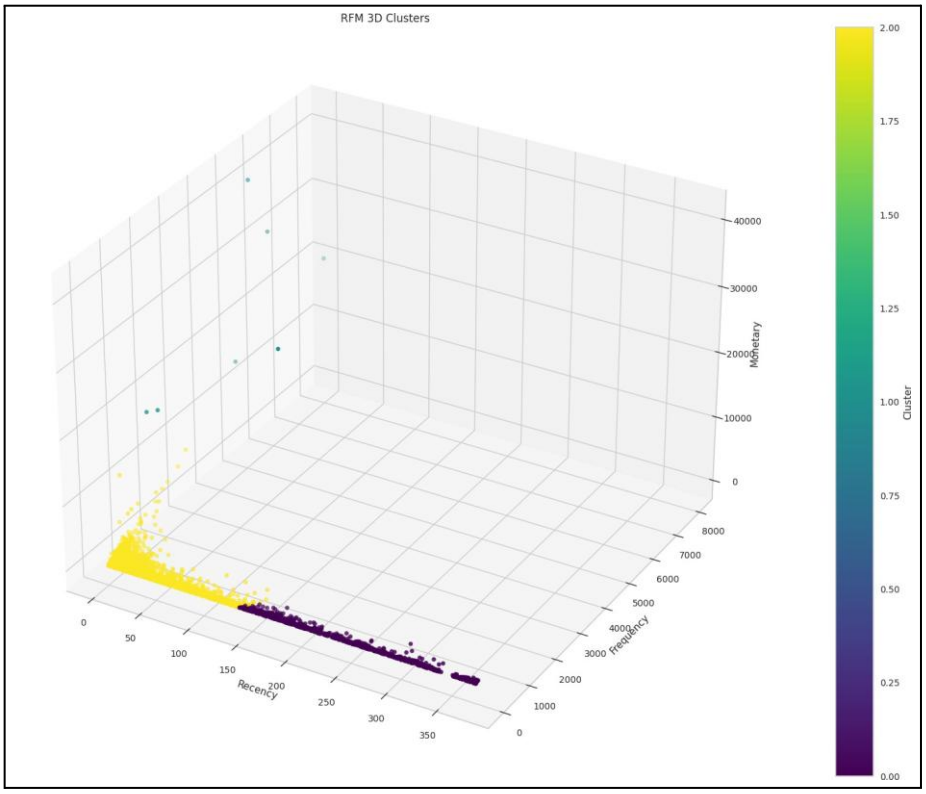

Fig 14 3D ScatterPlot of RFM

In the above Fig 14 , the graph indicates that there are four clusters made up of the customers:

Cluster 1 (red): These clients are highly valuable monetarily, highly frequently, and recently. These clients are the most valuable to you.
Cluster 2 (green): These clients are less frequent and have a lower monetary value, but they are more recent. These clients are still valuable, but to keep them coming back, you might need to specifically target them with discounts or other rewards.
Cluster 3 (blue): These clients are medium in frequency and monetary worth, but they have low recency. Re-engaging these customers is crucial because they are at risk of leaving.
Customers in Cluster 4 (yellow) are not very recent, they don't visit you often, and they don't have much money. Even though they are your least valuable clients, you might still be able to win them over.

RFM cluster distribution in three dimensions can reveal important details about your clientele. For instance, it's encouraging to see that most of your clients are located in Cluster 1. But, since Cluster 3 customers are susceptible to churn, you should also be aware of them. To entice these customers to return, you can target them with exclusive deals or rewards.

7.2 Customer Segmentation

To show any customer segments based on their purchase behavior.

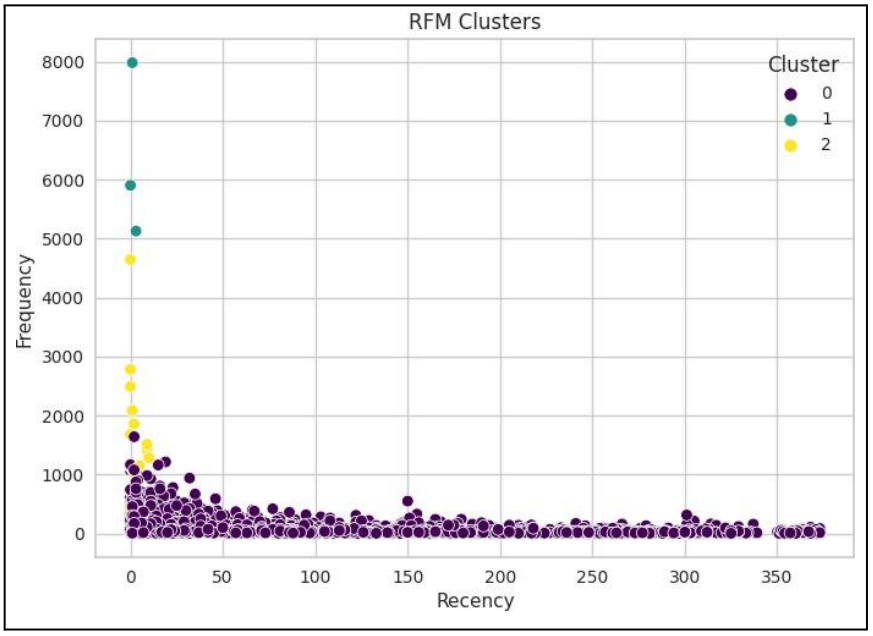

Fig 15 Customer Segmentation

In the above fig 15, the three metrics are used by RFM analysis to divide customers into various segments. Afterwards, each segment's targeted marketing campaign can be created using this data.

Cluster 1: Make it a priority to keep these valuable clients by providing them with special offers, tailored recommendations, and first-rate service.
Cluster 2: Use re-engagement tactics and customized offerings to enhance their value and increase the frequency of their purchases.
Cluster 3: To rekindle their interest and promote renewed engagement, create win-back campaigns and address any underlying causes for their inactivity.

Through an analysis of the RFM clusters and the application of the recommended actions, you can improve customer engagement, increase revenue, and target your marketing efforts more successfully.
8. Returns and Refunds

8.1 Percentage of Orders with Returns or Refunds
Approximately 1.96% of orders within the dataset have experienced returns or refunds. This percentage indicates the prevalence of such transactions, allowing businesses to assess and manage their return/refund processes, potentially identifying areas for improvement in product quality, customer service, or other relevant aspects. Monitoring this metric is crucial for maintaining customer satisfaction and operational efficiency in handling returns and refunds.

8.2 Correlation Between Product Category and Returns
The OVAL WALL MIRROR DIAMANTE has the highest return rate, as you can see, followed by the SPACEBOY BABY GIFT SET and SET 2 TEA TOWELS I LOVE LONDON. The least amount of returns is the 50'S CHRISTMAS GIFT BAG LARGE.
It is significant to remember that this data only displays the rate of returns for a limited selection of products. It's possible that other products have a lower or higher return rate.

Certain products, like the OVAL WALL MIRROR DIAMANTE, may have a high return rate due to their fragility or difficulty of use.Other products, like the 50'S CHRISTMAS GIFT BAG LARGE, may have a lower return rate because they are well-made and satisfy consumer needs.

Certain products may have a seasonal return rate. For instance, compared to December, January may see a higher return rate for Christmas merchandise.
Businesses can find opportunities to enhance their goods and services by examining the return rates for various products. They can create focused marketing campaigns with this information as well.

9. Profitability Analysis

9.1 Total Profit

Total Profit Generated: $0.00
Calculating the total profit generated in the 'ecom' dataset by subtracting the total cost from the total revenue. However, the outcome reveals a total profit of $0.00. This could be attributed to various factors, such as potential data quality issues, zero-cost products, or equal revenue and cost across transactions. Investigating the dataset for missing or incorrect data and validating the consistency of unit prices and quantities is crucial for understanding the accuracy of the profit calculation. Addressing these issues will help obtain a more reliable assessment of the company's total profit.

9.2 Top 5 Products with Highest Profit Margins

The listed top 5 products, including "4 PURPLE FLOCK DINNER CANDLES," "50'S CHRISTMAS GIFT BAG LARGE," "DOLLY GIRL BEAKER," "I LOVE LONDON MINI BACKPACK," and "I LOVE LONDON MINI RUCKSACK," all show profit margins of 0.0. This indicates that there is either missing or zero-cost information for these products, leading to an inability to calculate a meaningful profit margin.

To gain a comprehensive understanding of the profitability of these products, it's crucial to investigate and ensure accurate cost data is available. Profit margins play a vital role in pricing strategies, inventory management, and overall financial decision-making. Therefore, resolving any issues related to cost information for these products is essential for making informed business decisions and maximizing profitability.

10. Customer Satisfaction

10.1 Customer Feedback or Ratings

No Data available on customer feedback or ratings for products or services in the E-commerce dataset.

10.2 Sentiment or Feedback Trends

Since the data regarding customer feedback or ratings for products or services is not available sentiment or feedback analysis is not possible.

Conclusion

E-commerce Dataset Insights

This concise report distills key observations from the e-commerce dataset:

Customer Dynamics:

- The dataset showcases a diverse customer base, with an average of 5 orders per customer.
- Top 5 customers significantly impact overall sales.

Product Trends:

- Top 10 products reveal a mix of decorative and functional items.
- Average product price is $4.61, with notable contributions from specific categories.

Temporal and Geographic Patterns:

- Thursday and noon emerge as peak times for orders.
- The United Kingdom dominates in the number of orders.

Financial Metrics:

- 'C' and 'A' are prevalent payment methods.
- Unexpectedly, total profit is recorded as $0.00, warranting further investigation.

Returns and Customer Behavior:

- A 1.96% return rate is observed, with opportunities for product-specific analysis.
- Customers remain active for an average of 57.49 hours.

In summary, this report offers a quick overview, highlighting crucial areas for exploration and optimization in the e-commerce business.

This project offered insightful information about consumer behavior, purchase trends, and satisfaction. The most popular payment options and seasonal patterns in consumer purchases were identified by our analysis. We also looked into the different product categories, their relationships with returns, and the positive, negative, and neutral sentiments associated with them.

The analysis's findings can be applied to resource allocation, policy creation, and decision-making. These results can be expanded upon by additional research and analysis to address particular issues with customer segmentation and RFM analysis.

This report highlights key findings and any noteworthy patterns or trends found during the data cleaning and analysis process. The presentation and Jupyter Notebook that go with the project offer a deeper look at its specifics and conclusions. We think that decision-makers and other stakeholders who want to comprehend and deal with customer-segmentation problems in the given area will find useful information in this analysis.