Comprehensive Data Analysis on Sale Data

Ayush Kumar Mishra

September 24, 2024

Dashboard

For seeing all the code live interactively,

Visit Dashboard

https://ayusheda-assign.streamlit.app/

Contents

1. Introduction

In this document, I will formulate how i did analysis on the data.
The data contains information about the orders, customers, products, and sales.
The goal of this analysis is to provide insights into customer behavior, sales trends, SKU performance, and other key metrics.
The analysis will be performed using Python and various data analysis libraries such as pandas, NumPy, and Matplotlib. The analysis will cover the following key areas:

- Customer behavior analysis
- Sales trends analysis
- SKU performance analysis
- Order analysis
- Cohort analysis
- Geographic analysis
- Time-based analysis
- Customer lifetime value (CLV) analysis
- Basket analysis
- Price sensitivity analysis
- And more...

1. Data Preparation and Overview

Loading and Inspecting the Dataset

- Load the dataset and check its structure.

	Unnamed: 0	user_id	order_date	order_id	sku_id	warehouse_name	quantity	placed_gmv
0	0	0868733	2022-09-16	262052	2567941	USA	1.0	331.60
1	1	0868733	2022-09-16	262052	434572f	USA	1.0	416.52
2	2	0868733	2022-09-16	262052	8ae8fa4	USA	2.0	45.00
3	3	0868733	2022-09-16	262052	c9932dc	USA	3.0	67.50
4	4	0868733	2022-09-16	262052	35c7c3b	USA	1.0	340.71

- Inspect data for missing values, duplicates, and correct data types. - Their are no missing values and duplicates in the dataset.

```
missing_values = df.isnull().sum()
print("Missing values in each column:\n", missing_values)
```

✓ 0.0s

```
Missing values in each column:
 Unnamed: 0        0
user_id           0
order_date        0
order_id          0
sku_id            0
warehouse_name    0
quantity          0
placed_gmv        0
dtype: int64
```

Statistical Summary

```
summary_stats = df.describe()
summary_stats
```

✓ 0.0s

	Unnamed: 0	order_id	quantity	placed_gmv
count	130000.000000	1.300000e+05	130000.000000	130000.000000
mean	64999.500000	6.822964e+05	1.591008	1336.445672
std	37527.911835	3.202138e+05	1.854480	2735.577056
min	0.000000	2.387230e+05	1.000000	4.200000
25%	32499.750000	3.236010e+05	1.000000	371.500000
50%	64999.500000	8.655470e+05	1.000000	591.900000
75%	97499.250000	9.787400e+05	2.000000	1310.490000
max	129999.000000	1.064487e+06	137.000000	216814.080000

Answer

One thing we can observe from summary is that Quantity and Placed GMV are skewed and have outliers.
As 75 percentile is 2 and 50 percentile is 1 for Quantity and 75 percentile is 1310.49 and 50 percentile is 591.90
for Placed GMV.
Whereas their Max values are 137 and 216814 which is much higher than 75 percentile.

Date Formatting

This step is essential because the date column is in string format. We need to convert it to a datetime format for further analysis.

```
df['order_date'] = pd.to_datetime(df['order_date'], errors='coerce')

print(df.dtypes)
```

✓ 0.0s

```
Unnamed: 0                int64
user_id                  object
order_date       datetime64[ns]
order_id                  int64
sku_id                   object
warehouse_name           object
quantity                float64
placed_gmv              float64
dtype: object
```

2. Customer Behavior Analysis

Customer Purchase Frequency

Let's look at the distribution of frequency by which customers are placing orders .

```
purchase_frequency['order_count'].describe()
```

✓ 0.0s

```
count    3660.000000
mean       35.519126
std        52.486606
min         1.000000
25%         7.000000
50%        17.000000
75%        43.000000
max       833.000000
Name: order_count, dtype: float64
```

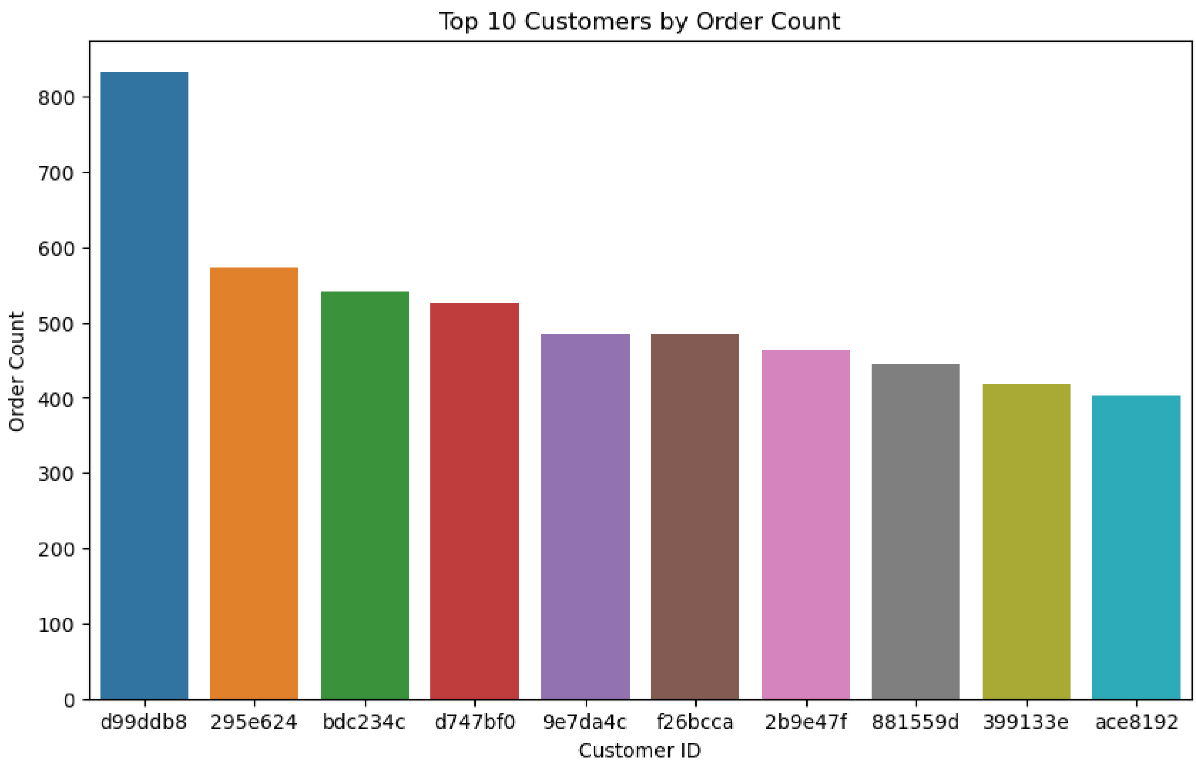
Insights

Insights:

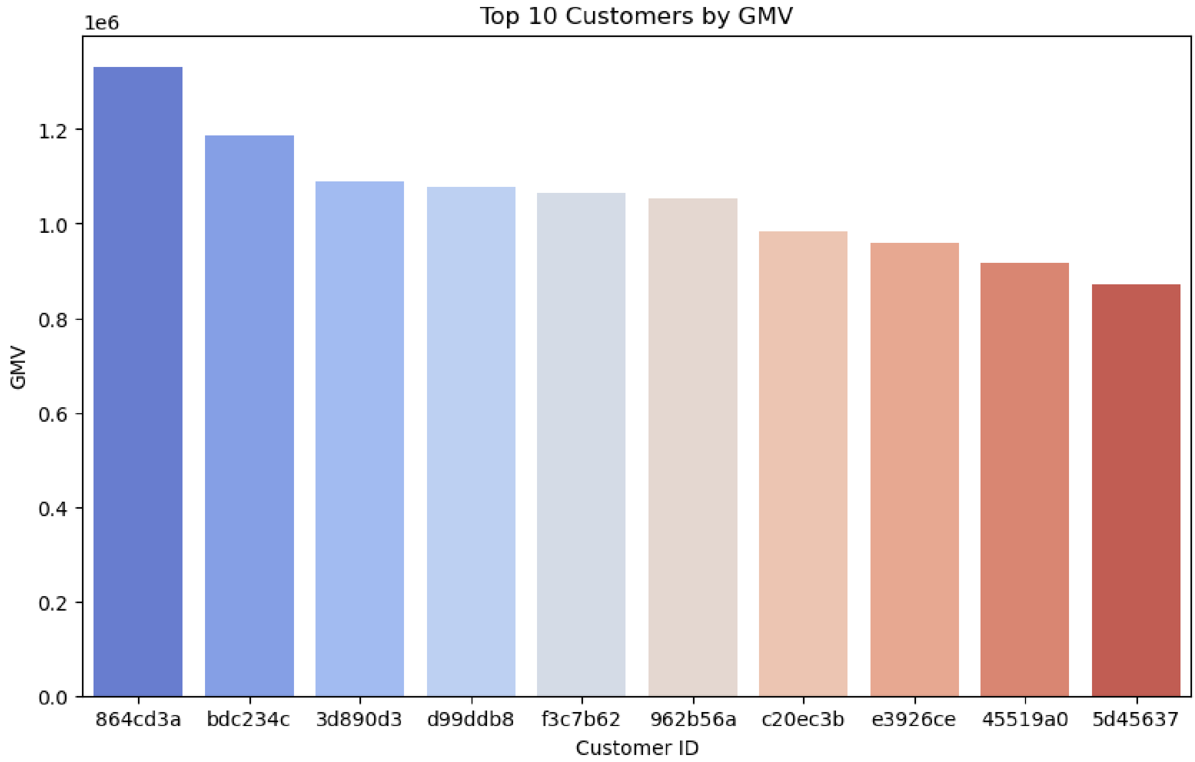
- More than 50% of customers have placed orders less than 17 times which is almost half than means
. meaning few people are buying a lot.
- And 75% of customers have placed orders less than 43 times.
- Just **293 people** out of 130000 have placed orders more than 100 times.

Top Customers

- Based on Order frequency, I am identifying the top customers.



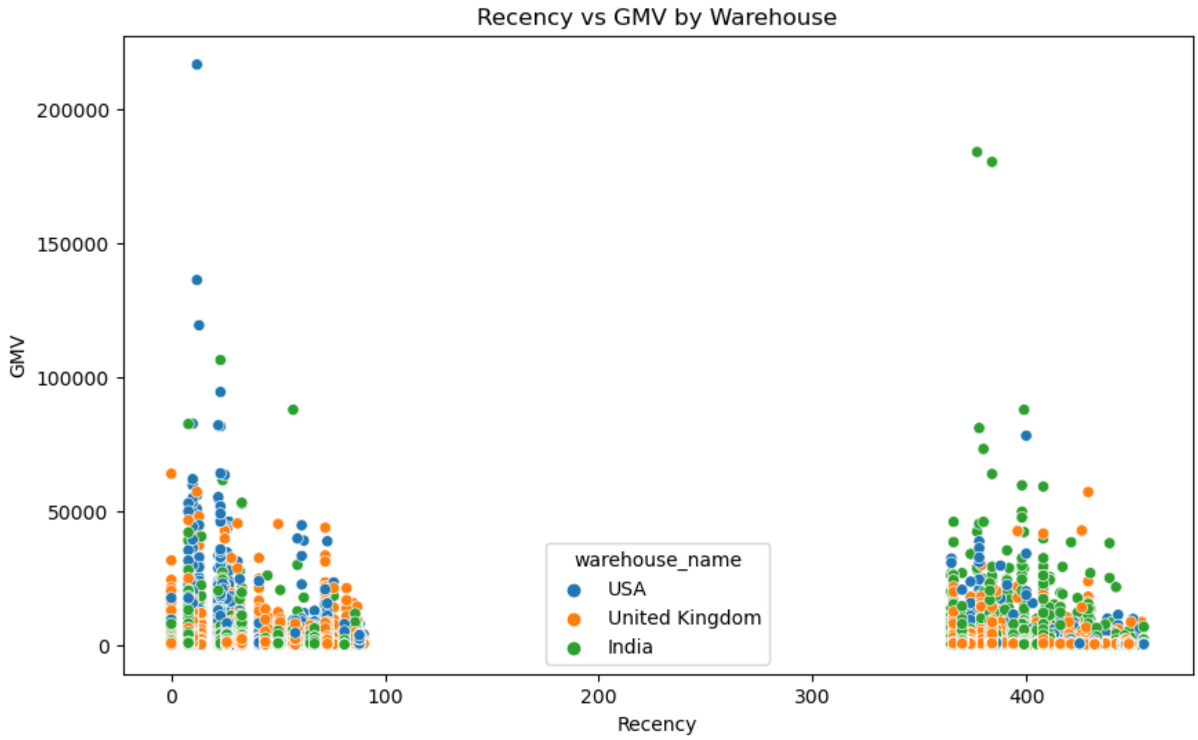
- Based on GMV, I am identifying the top customers.



RFM Analysis

RFM analysis is a powerful way to segment customers based on their behavior.

- Recency: When the customer last made a purchase. Here i am calculating the recency of the customers.

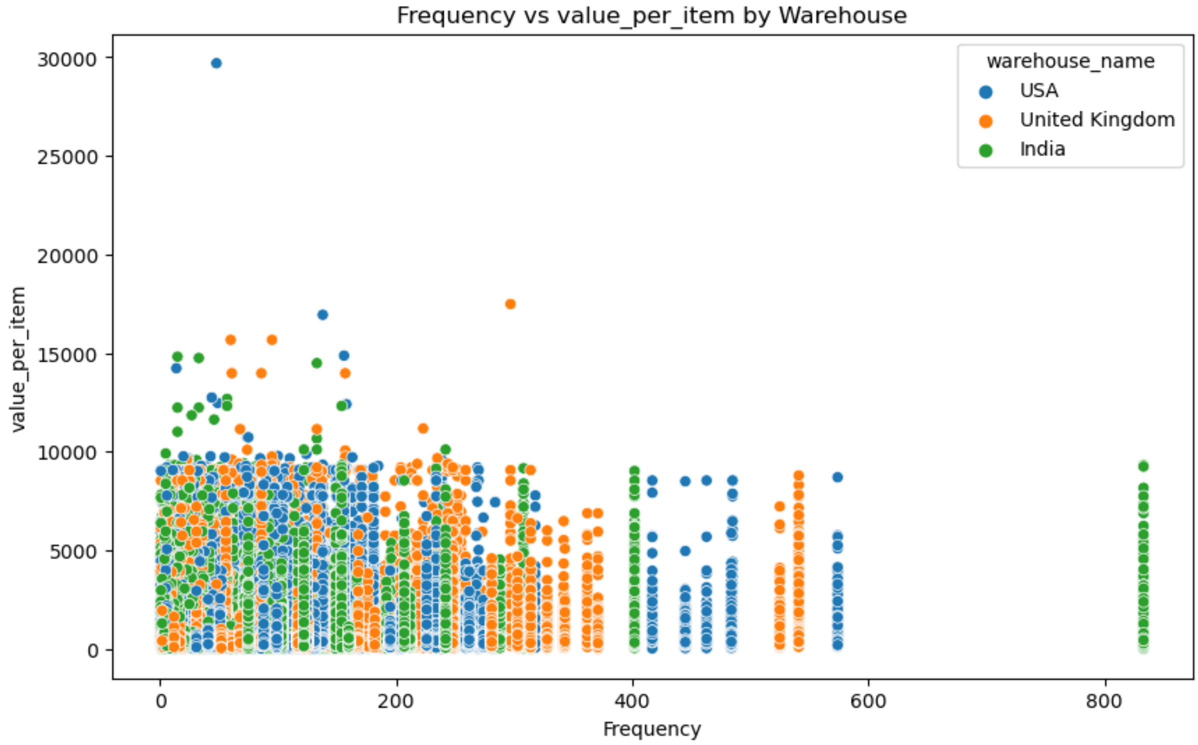


Insights

From the above graph, There are two types of custumers:-

* One who are frequent buyers and have bought recently less than 100 days.
* One who are seasonal buyers and have come to buy only after a year.

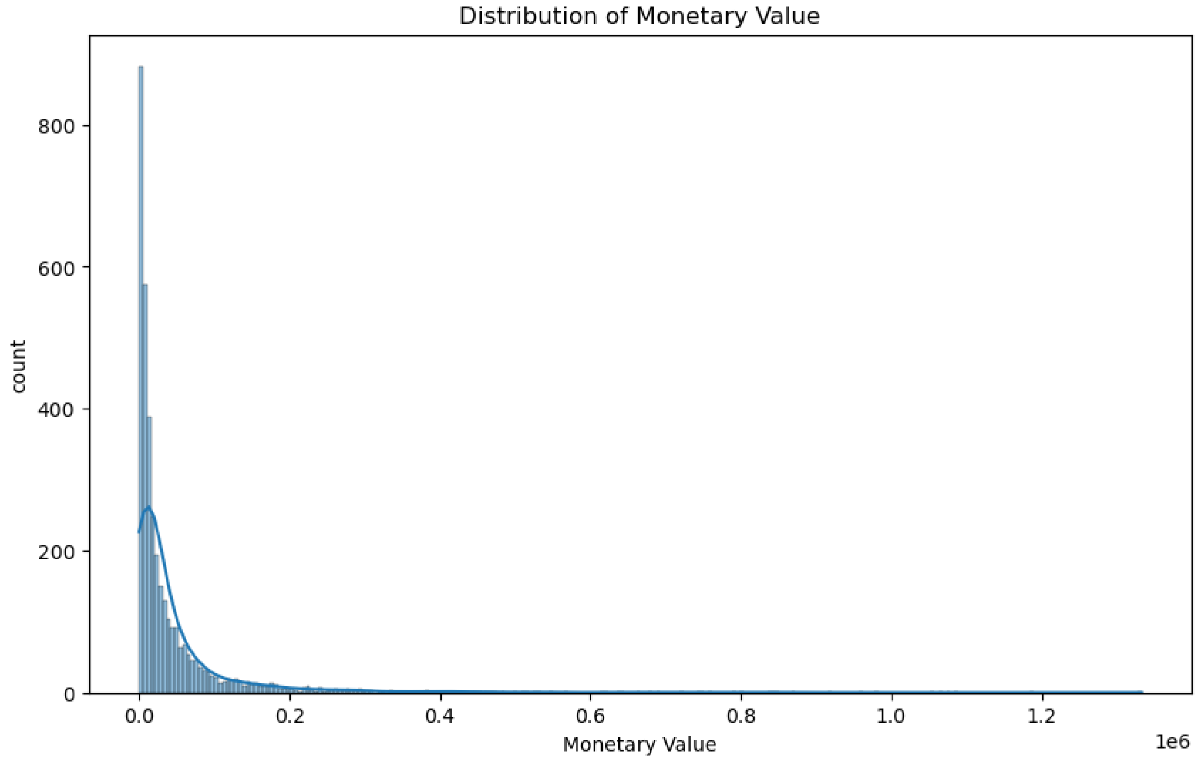
- Frequency: How often the customer made purchases. Here i am calculating the how frequent customers have come to place orders.



Insights

From the above graph, One observations is that low frequent buyers have more value_per_item than high frequent buyers.

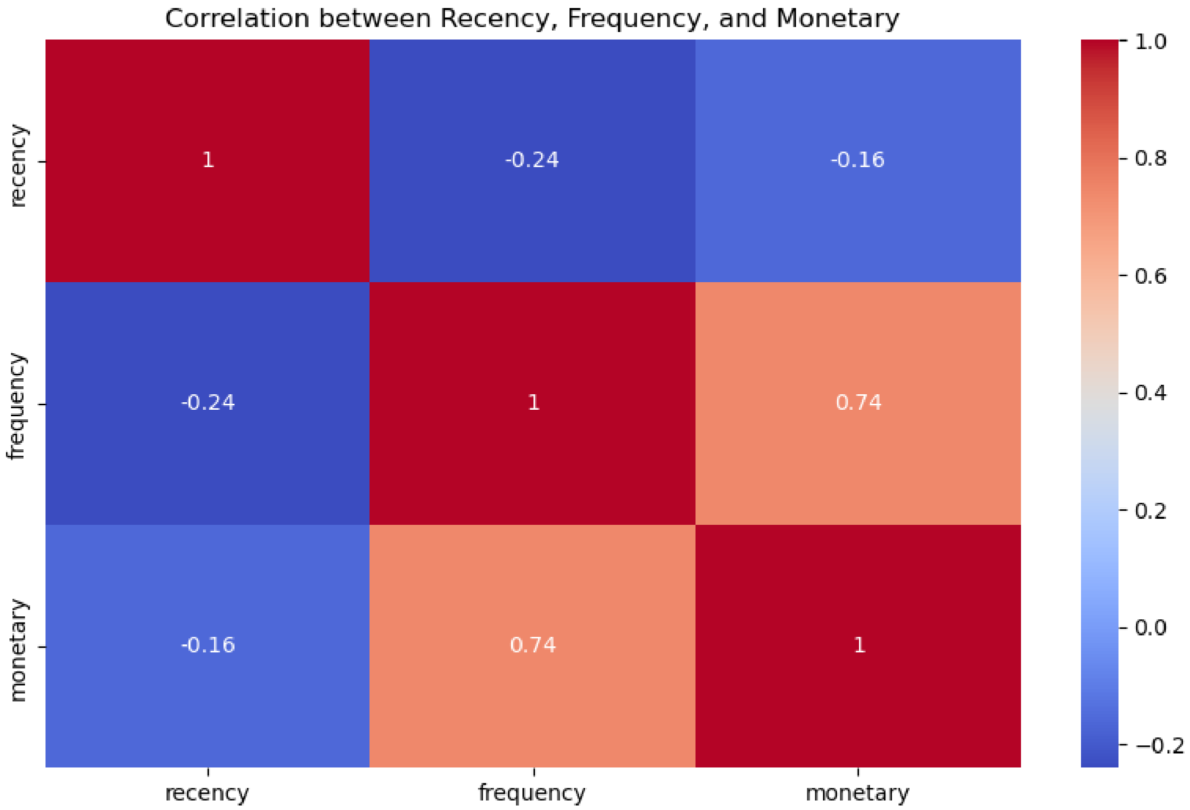
– Monetary: How much money the customer has spent. Here i am calculating the how much money customers have spent.



Insights

Majority of the people have spend less than 0.2 * 10^6.

Now let's see the relationship between recency, frequency, and monetary values.



Insights

- From the above graph, we can see that there is a positive correlation between frequency and monetary value.

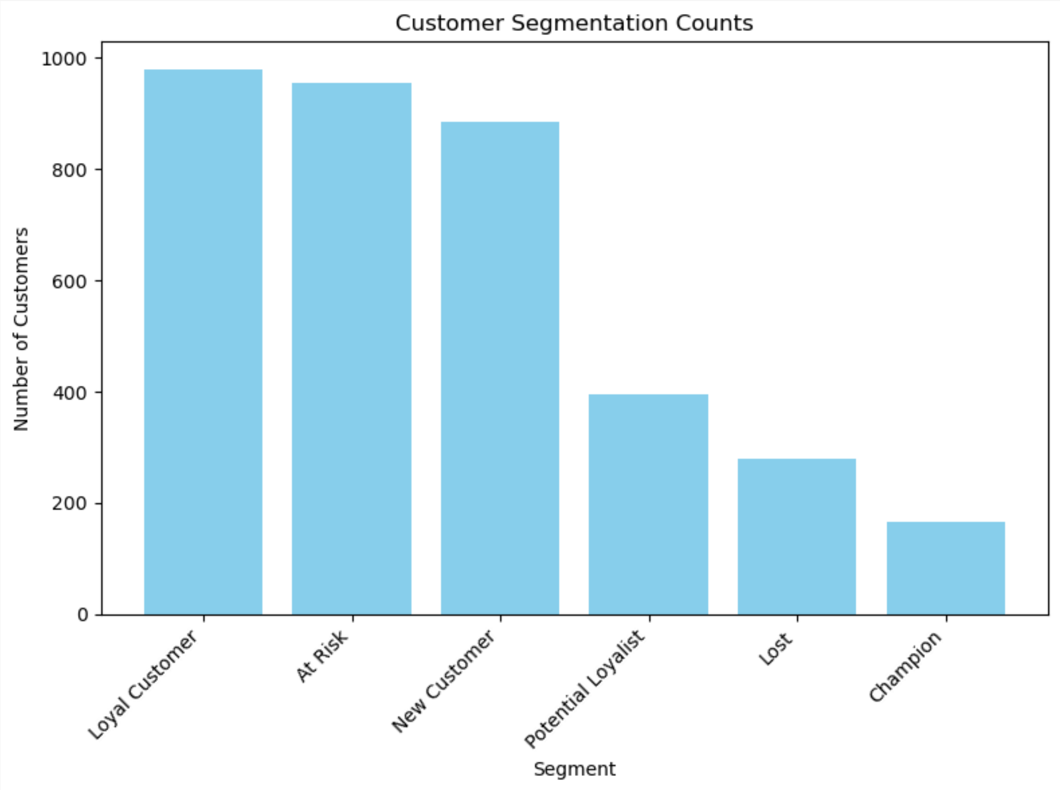
- But there is a negative correlation between recency and frequency and monetary value.

Score based on all three recency, frequency, and monetary values.

	user_id	recency	frequency	monetary	recency_score	frequency_score	monetary_score	RFM_score
0	0000e88	67	3	9491.60	2	1	2	212
1	000159a	13	98	84908.69	4	5	5	455
2	000c1b2	23	3	5304.84	4	1	2	412
3	0039abd	12	3	2098.24	4	1	1	411
4	003b0e5	76	9	2525.84	2	2	1	221

Answer

Based on this score, i segmented customers into different categories such as:

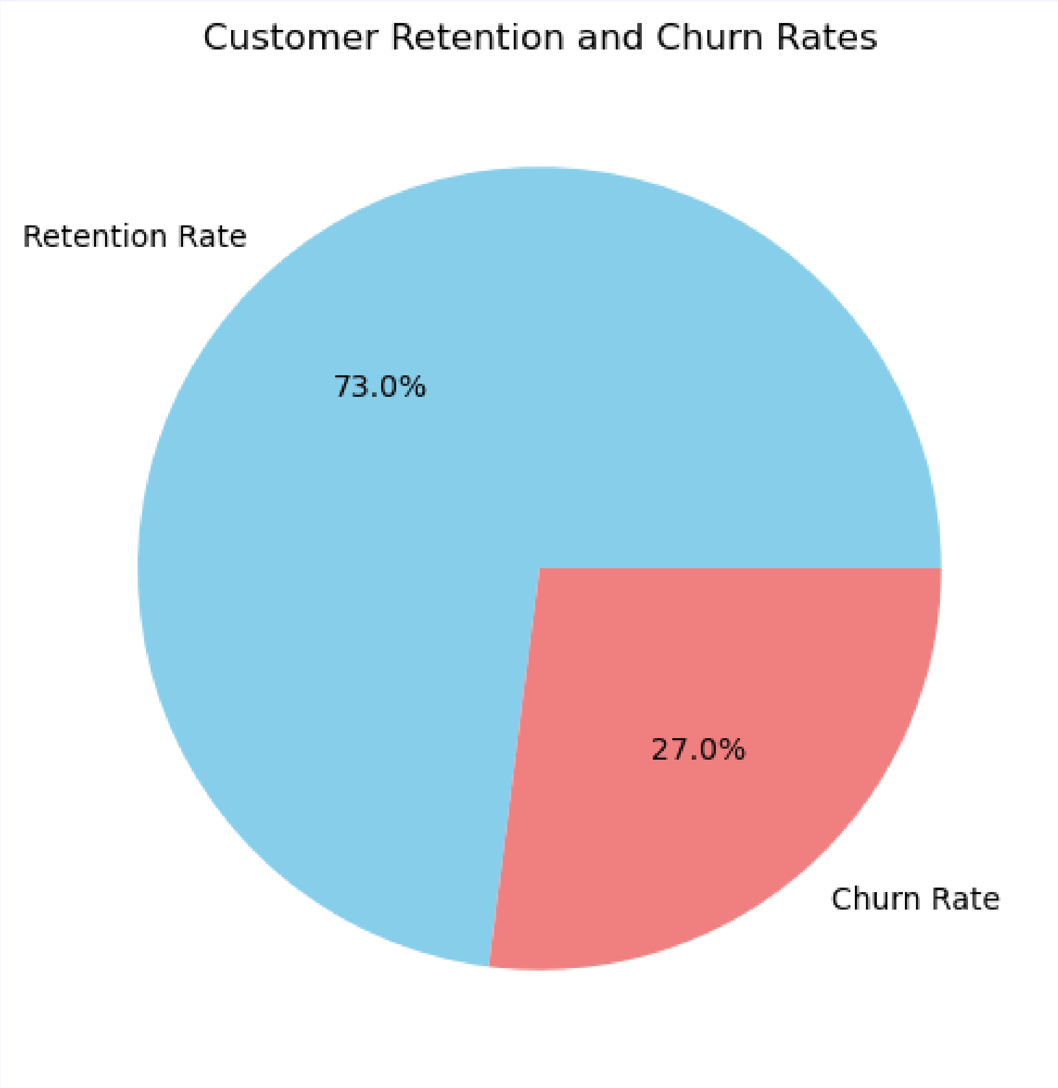


- **Champions:** Customers with high recency, frequency, and monetary scores (R = 4-5, F = 4-5, M = 4-5).
- **Loyal Customers:** Customers with high frequency and monetary scores but may have slightly lower recency (R = 3-5, F = 4-5, M = 4-5).
- **Potential Loyalists:** Customers with high recency and frequency but lower monetary value (R = 4-5, F = 3-5, M = 2-3).
- **New Customers:** High recency.
- **At Risk:** Low recency, frequency, and monetary value.
- **Lost:** Low recency, frequency, and monetary value.

Customer Retention and Churn

- Analyze customer retention rates and identify potential churn risks.

Customer Retention and Churn Rate

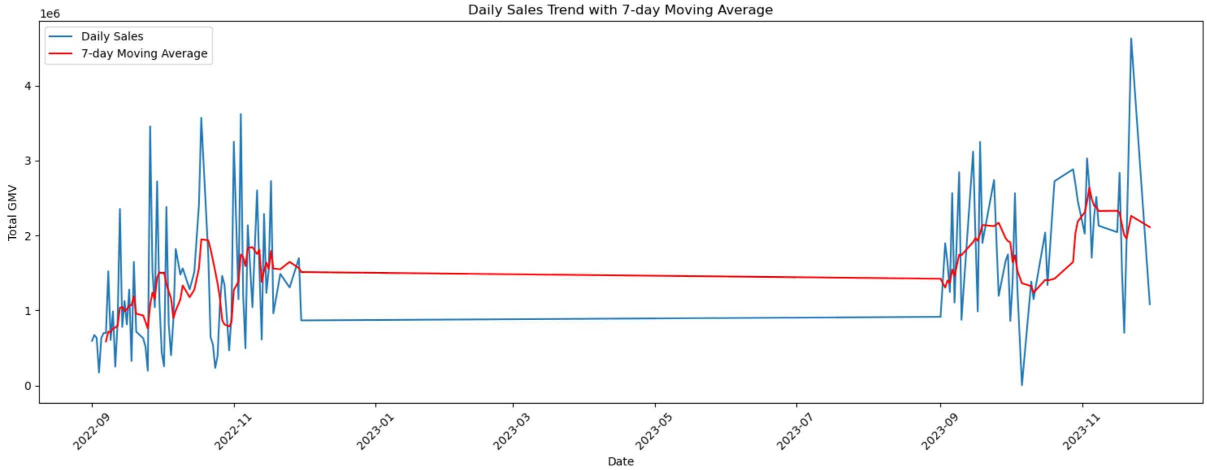


Clearly from the pie chart we can see that 73% of the customers are retained and 23% are churned.

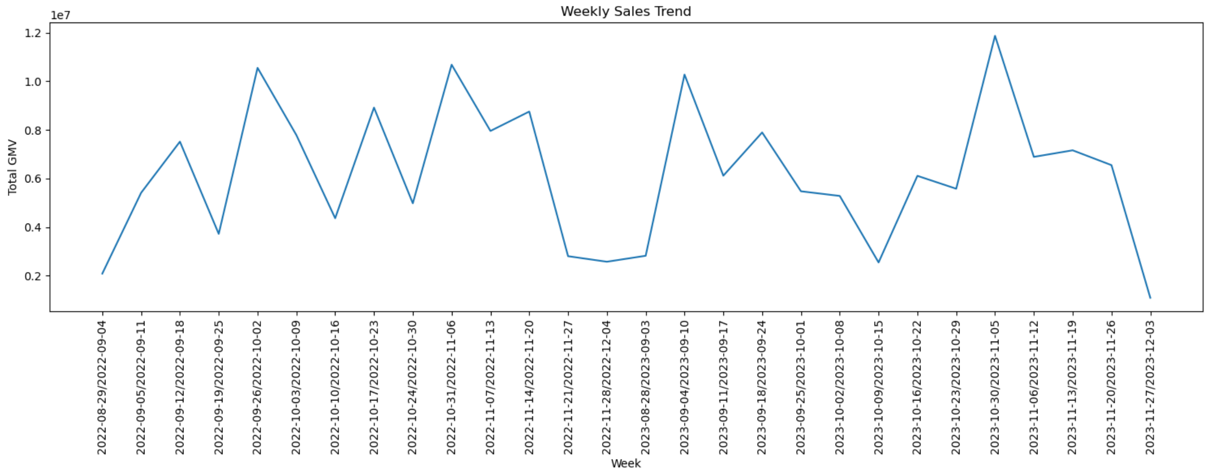
3. Sales Trends Analysis

Time-based Trends

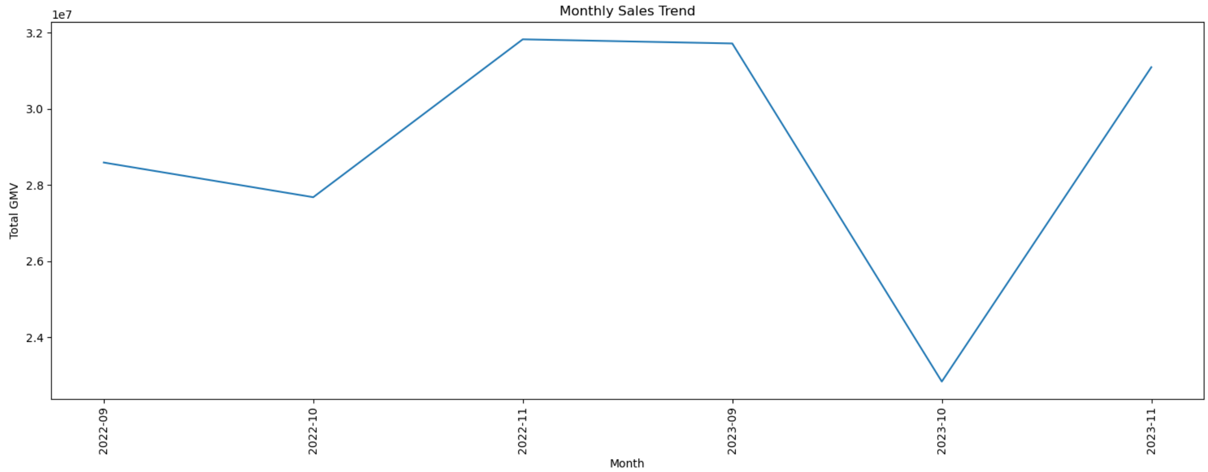
– Analyze daily, weekly, and monthly sales trends.



Daily sales trends with 7-days moving average.



Weekly sales trends.



Monthly sales trends.

Time-based sales trends.

Peak Sales Periods

Identify peak sales periods and any seasonality trends.

Peak Sales Periods by Days

Order Date	Placed GMV
2023-11-22	4,629,566.36
2022-11-04	3,623,119.11
2022-10-18	3,572,387.45
2022-09-26	3,457,606.77
2022-11-01	3,252,048.05
2023-09-18	3,250,288.29
2023-09-15	3,120,405.10
2023-11-03	3,030,412.93
2023-10-28	2,885,263.97
2023-09-09	2,846,904.68

Top 10 Peak Sales Days

Peak Sales Periods by weeks and months

Top 10 Peak Sales Weeks

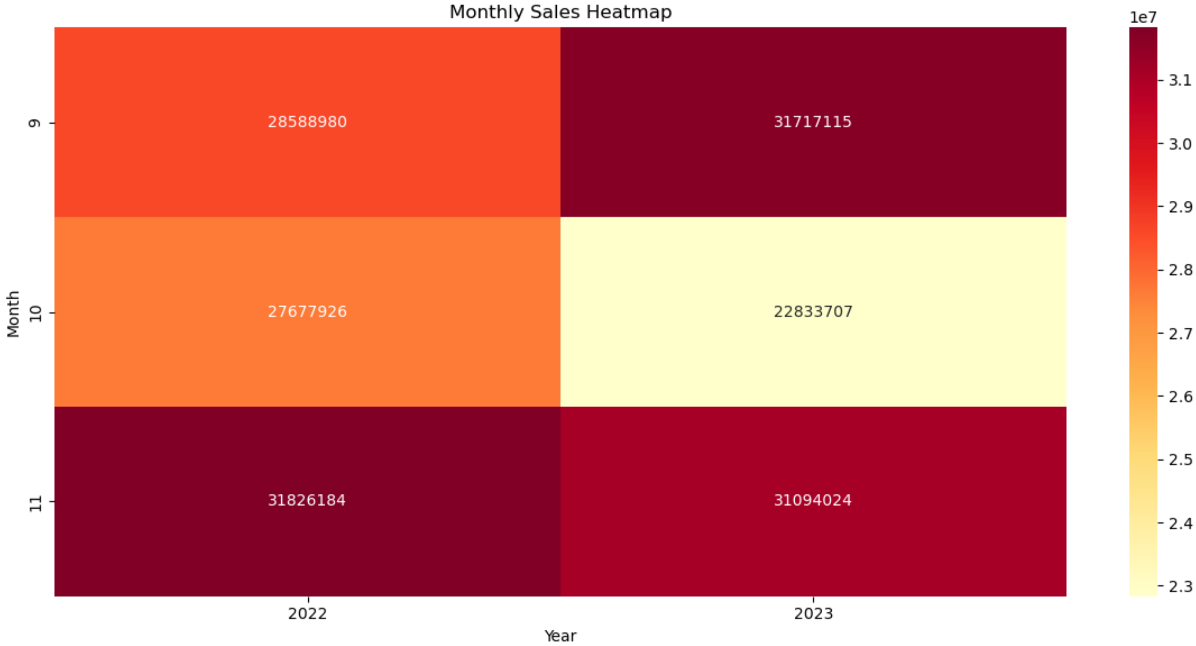
Week	Placed GMV
2023-10-30/2023-11-05	11,877,473.08
2022-10-31/2022-11-06	10,685,698.73
2022-09-26/2022-10-02	10,557,566.61
2023-09-04/2023-09-10	10,280,051.70
2022-10-17/2022-10-23	8,919,020.51
2022-11-14/2022-11-20	8,756,086.30
2022-11-07/2022-11-13	7,956,940.37
2023-09-18/2023-09-24	7,895,670.74
2022-10-03/2022-10-09	7,787,362.60
2022-09-12/2022-09-18	7,517,063.73

Top 10 Peak Sales Months

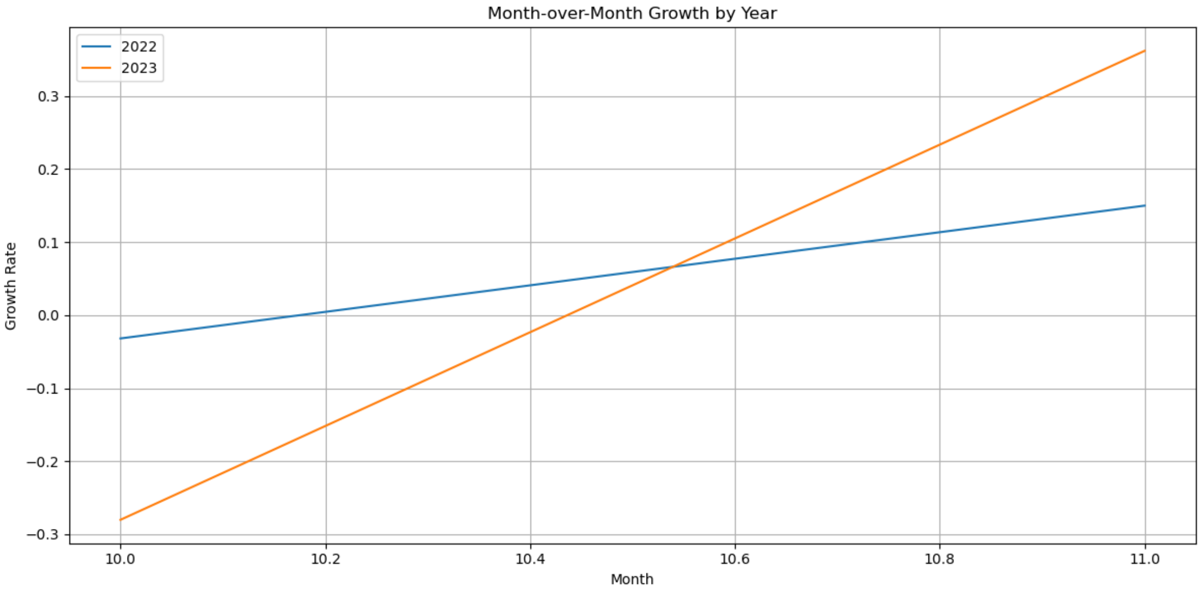
Month	Placed GMV
2022-11	31,826,184.02
2023-09	31,717,114.58
2023-11	31,094,024.46
2022-09	28,588,980.49
2022-10	27,677,926.44
2023-10	22,833,707.32

From above tables, following observations can be made:

- The top 10 peak sales days have placed GMV ranging from 2.8M to 4.6M.
- Most of the peak sales days are in the months of September, October, and November.

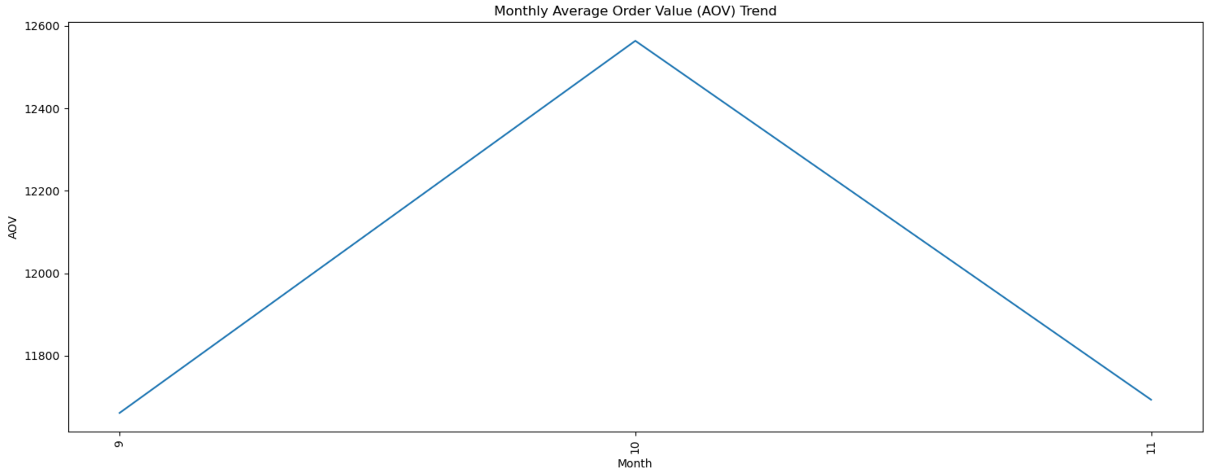


- Also we can see find month by month over year sales.



Average Order Value (AOV)

- Analyze trends in average order value over time.



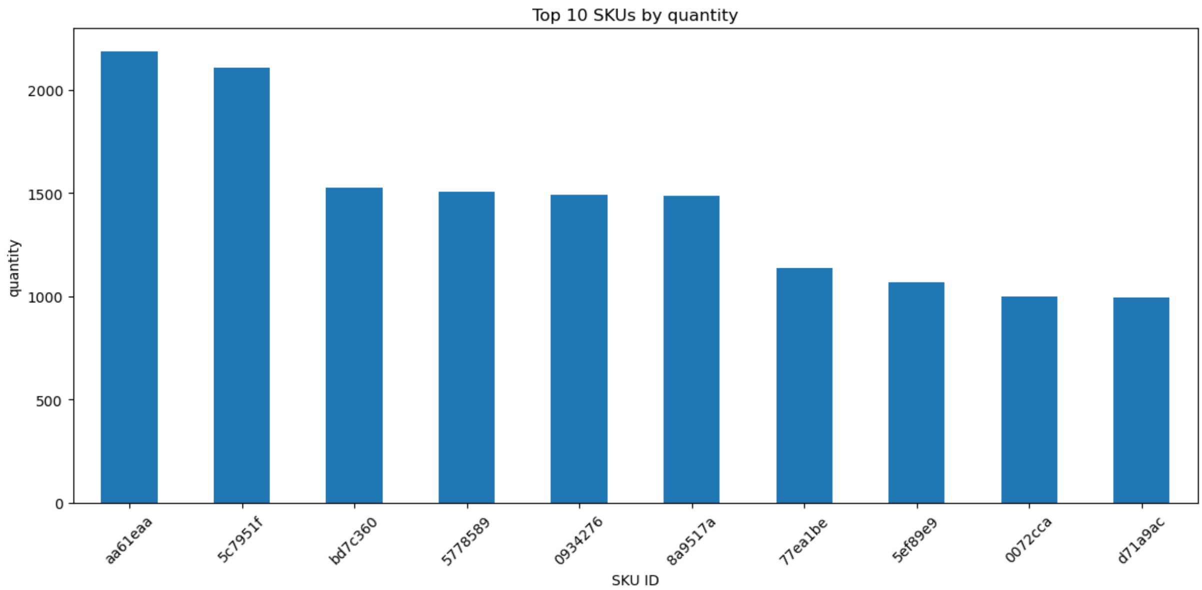
Insights:

* The aov is clearly showing an increase in sales upto october and then a decrease from there.
* It means months aroung October has higher sales and demand than other months beacuse of festivals season.

4. SKU Performance Analysis

Top-Selling SKUs

First lets see which are SKUs that are top selling.



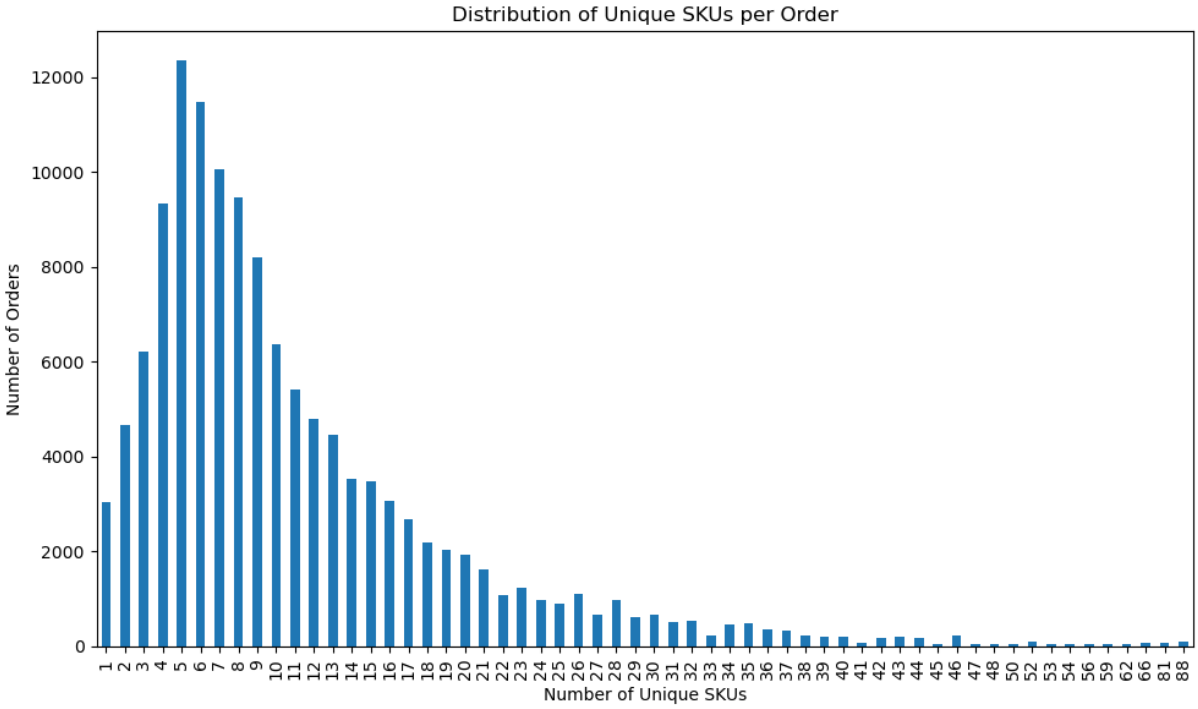
Top 10 Selling SKUs by Quantity. Also lets see the top selling SKUs by GMV.

```
Top 10 SKUs by placed_gmv:

sku_id
aa61eaa    3358729.54
bd7c360    3033729.76
5778589    2004186.77
0072cca    1818349.65
8a9517a    1770669.35
0934276    1691484.70
5c7951f    1674269.10
ee90f3e    1285239.63
be170aa    1270098.64
6323dad    1269127.86
Name: placed_gmv, dtype: float64
```

SKU Diversity

Analyze the diversity of SKUs in customer orders. What this means is that how many unique SKUS are their in any order$_I D$.



Insights

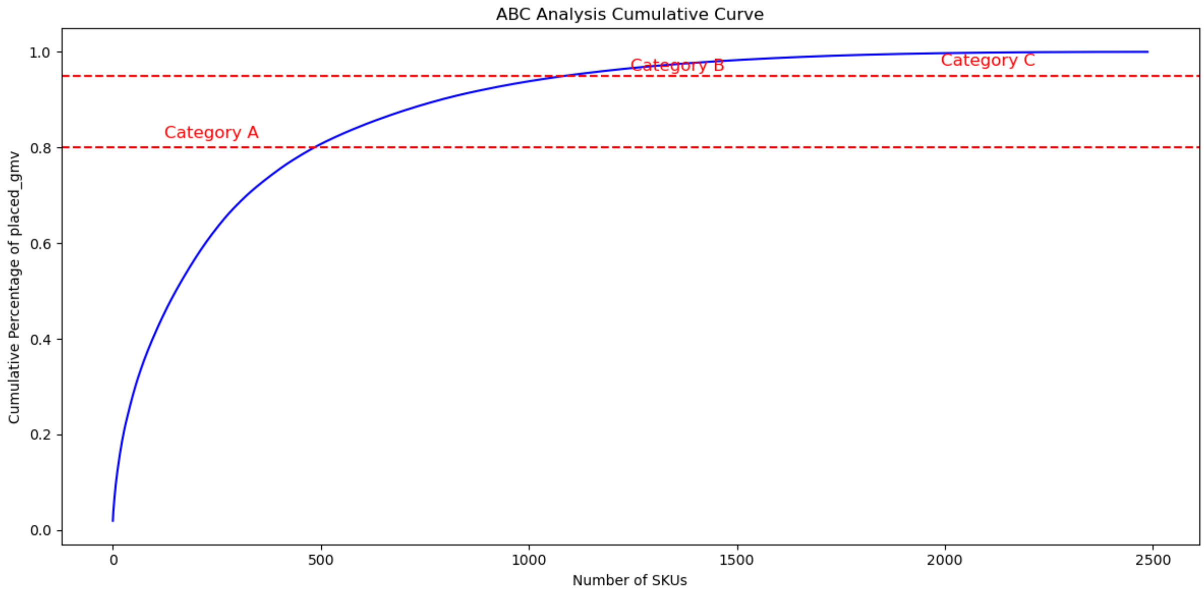
Insights:

- Majority of the orders contain 1-14 unique SKUs, meaning customers tend to buy a variety of products in a single order.
- But customers buying more than 14-15 unique SKUs are very less.
- 7 is the most common number of unique SKUs in an order.

ABC Analysis

Perform ABC analysis to categorize SKUs based on sales contribution.

- Category A SKUs (up to 80% of GMV) are the most critical for driving sales and revenue.
- Category B SKUs contribute moderately (the next 15% of GMV).
- Category C SKUs (final 5%) are the least significant for overall revenue.



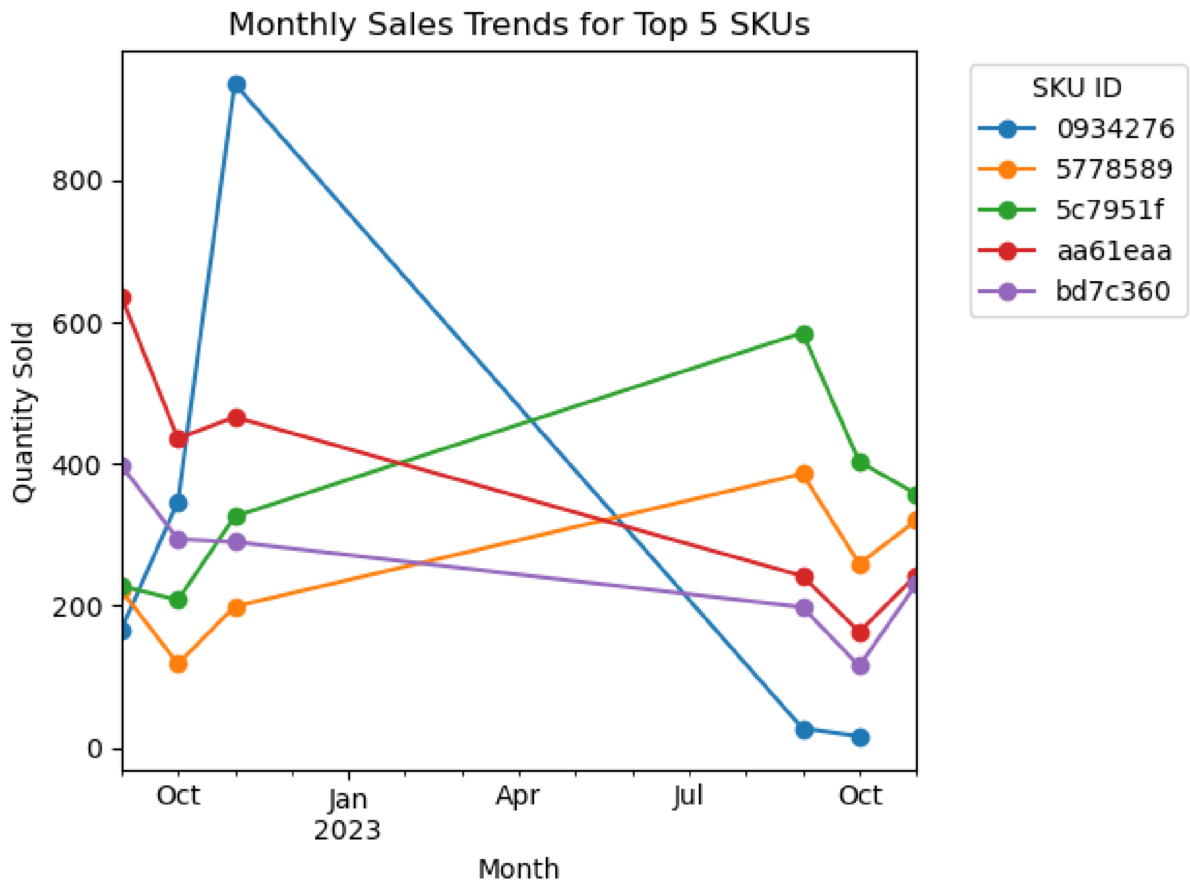
One key observations that can be noticed is that **Only less than 500 unique SKU's are contributing to 80% of the GMV. Extensive results is as follows:**

```
ABC Analysis Results:
placed_gmv
A      484
B      605
C     1398
Name: count, dtype: int64

Percentage of SKUs in each category:
placed_gmv
A     19.453376
B     24.316720
C     56.189711
Name: count, dtype: float64
```

Purchase Patterns

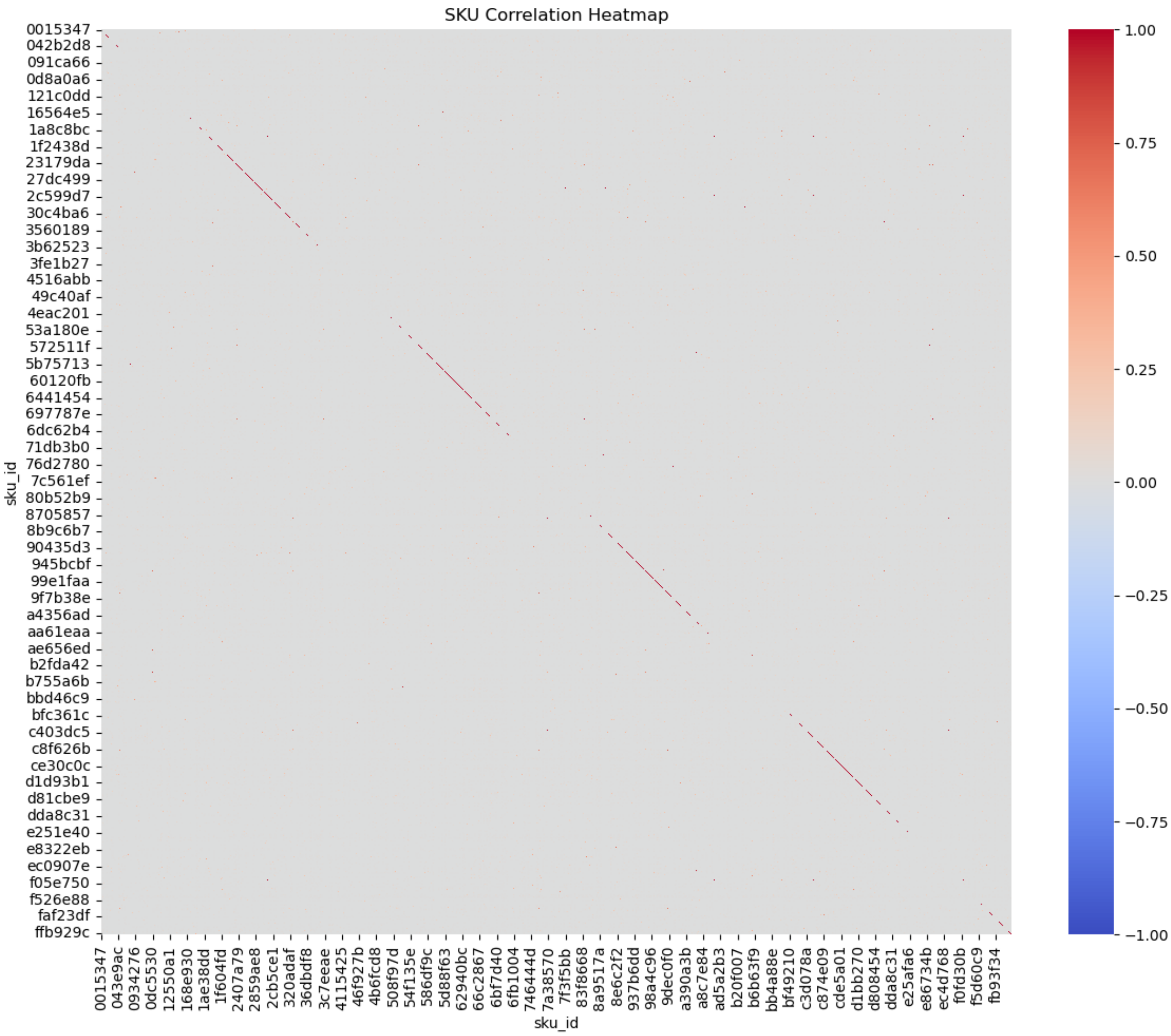
Purchase patterns for top-selling SKUs.



As already seen, the top selling SKUs are sold more in the month of October and November.

Correlated SKU Pairs:

Correlated SKU pairs that are frequently purchased together.

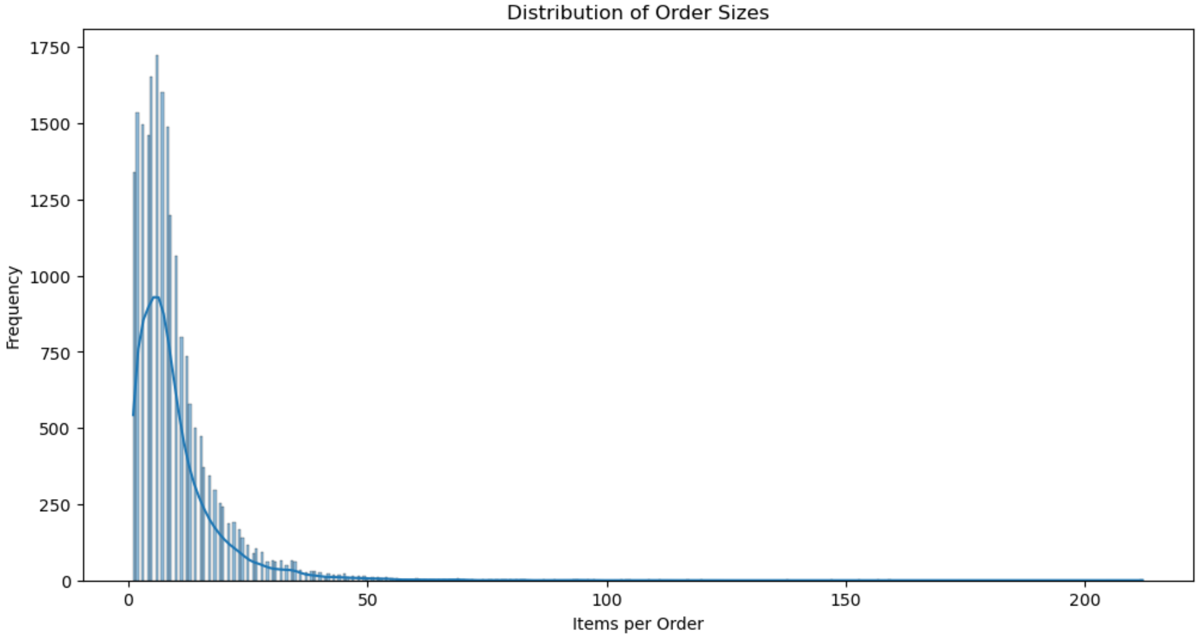



```
Top 10 Correlated SKU Pairs:
```

sku_id	sku_id	sku_corr
320adaf	a6ae1cb	0.991596
a6ae1cb	320adaf	0.991596
320adaf	581e5e8	0.991561
581e5e8	320adaf	0.991561
	a6ae1cb	0.987324
a6ae1cb	581e5e8	0.987324
a2633d6	af6266b	0.979263
af6266b	a2633d6	0.979263
320adaf	ed18f1c	0.962180
ed18f1c	320adaf	0.962180

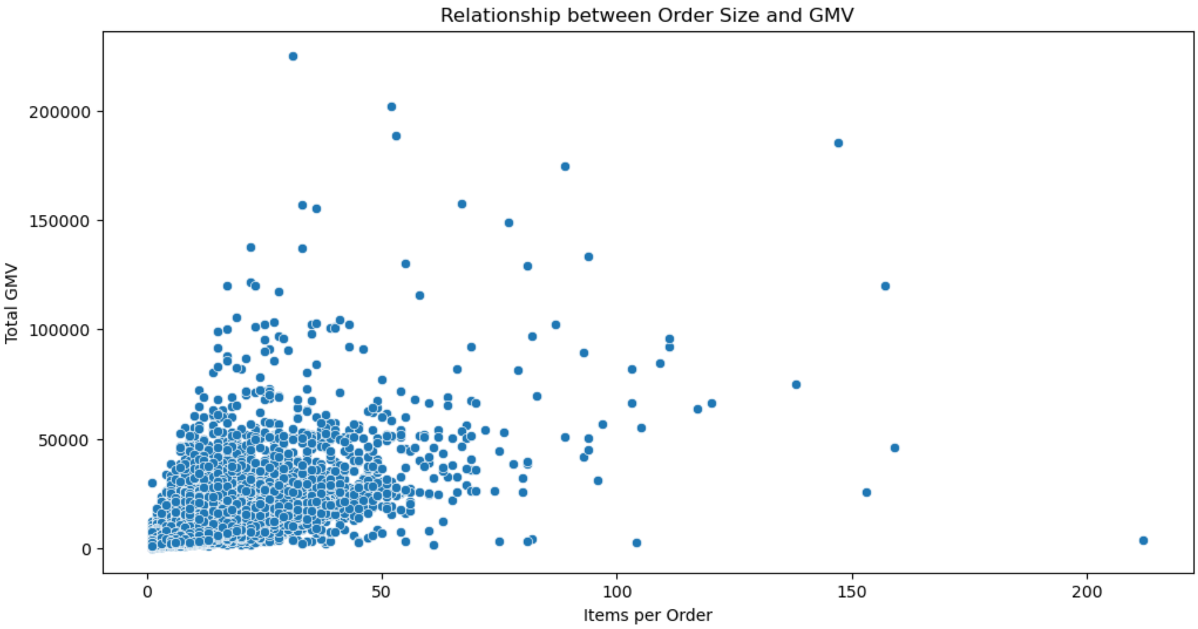
5. Order Analysis

Order Sizes



Insights: Majority of the items per order is less than 30 only, indicating that most customers tend to buy a small number of items in a single order.

Relationship Between Order Size and GMV

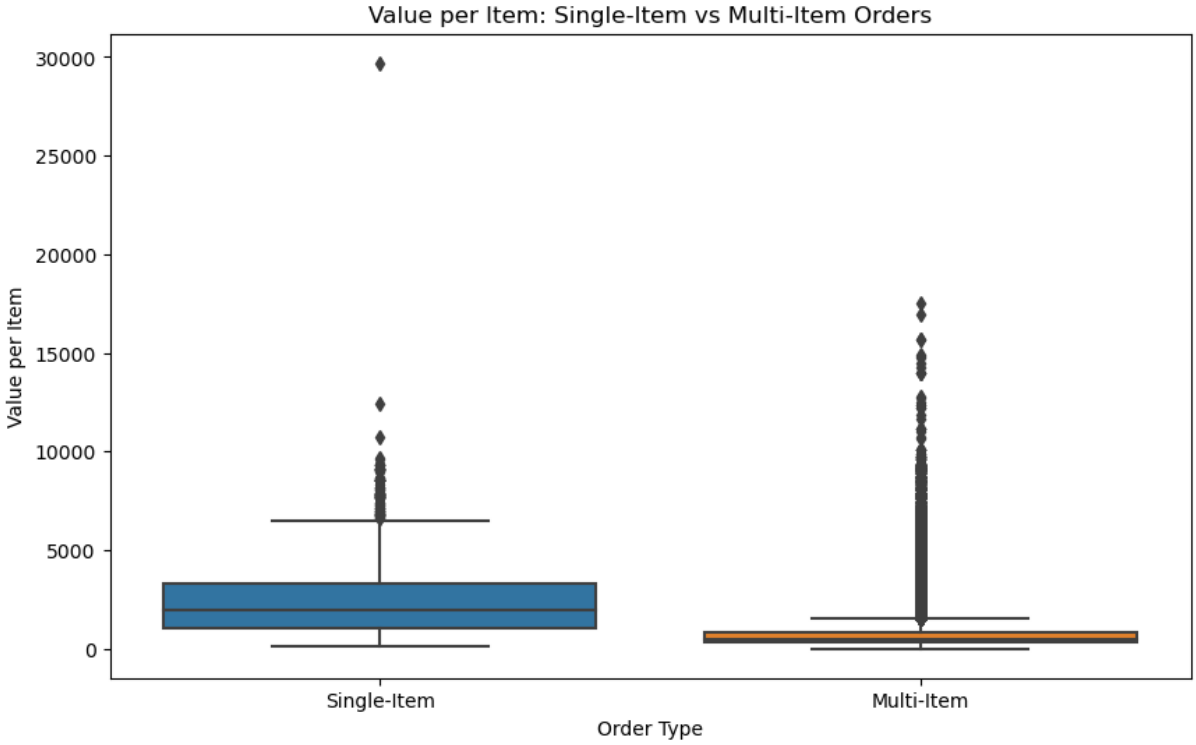


```
Average GMV by order size:
    items_per_order      total_gmv
0               1.0    3376.119590
1               2.0    3969.668248
2               3.0    4018.977017
3               4.0    4210.006826
4               5.0    5036.383487
..              ...            ...
95            147.0  185110.060000
96            153.0   25552.700000
97            157.0  119976.360000
98            159.0   46079.400000
99            212.0    3579.440000
```

Multi-item Orders

```
Comparison of single-item vs multi-item orders:
                   avg_gmv  order_count
is_multi_item
False          3376.119590         1340
True           8546.590083        19799

From this we can see that multi-item orders have higher average GMV than single
Seasonal patterns in orders:
   month   quantity   placed_gmv
0      9   9.416080  7858.495579
1     10  10.319842  8699.902473
2     11   9.747356  8215.198914
```



**Insights: If a customer is buying more items in a single order,
then the value per item is low, meaning they are buying cheaper items if buying multiple items.
But if they are buying single item, then the value of that item is high.**

6. Cohort Analysis

Customer Cohorts

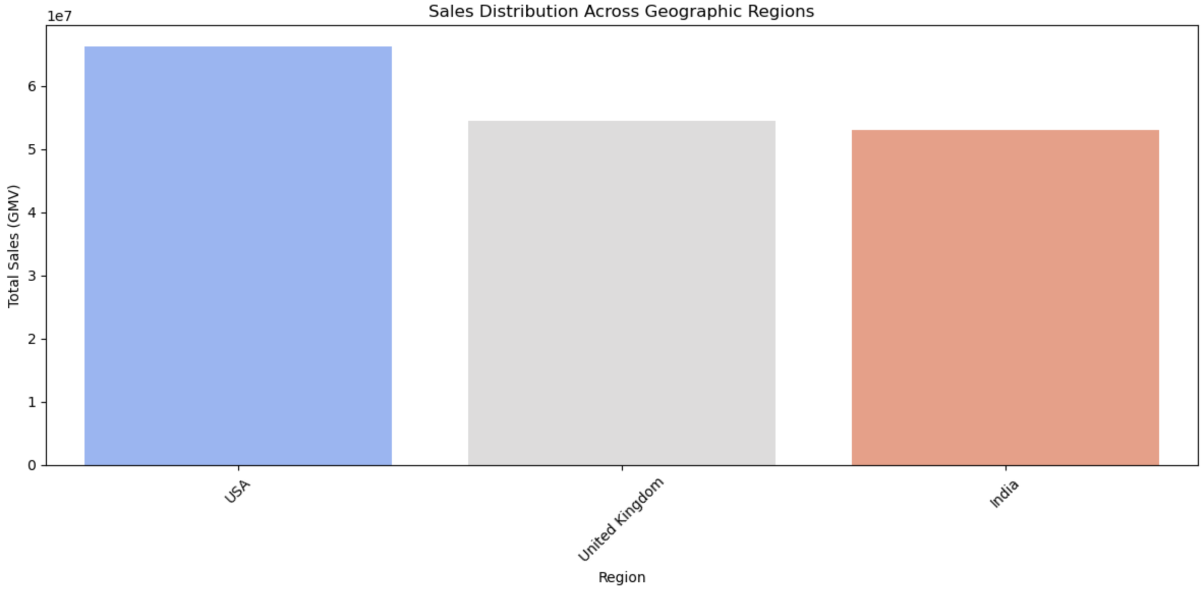
- Create cohorts based on the first purchase date of customers.

Cohort Retention

- **High Initial Retention:** All cohorts begin with a 100% retention rate, as expected at the start of the analysis period.
- **Declining Retention:** Retention rates drop consistently over time for all cohorts, showing typical user drop-off after acquisition.
- **Best Performance by 2022-09 Cohort:** The 2022-09 cohort maintains relatively high retention rates across longer periods, with 47.47% after Period 4 and 43.07% after Period 6.
- **Improved Retention for Recent Cohorts:** The 2023-09 and 2023-10 cohorts show improved retention compared to earlier cohorts, with the 2023-09 cohort peaking at 63.47% in Period 2.
- **Sharp Decline for 2022-10 Cohort:** The 2022-10 cohort experiences a sharp drop from 100% to 57.03% after the first period and continues to decline significantly afterward.
- **Retention Stabilization:** Some cohorts, like the 2022-09, show retention stabilization around 30-40% over extended periods.

7. Geographic Analysis

- Which of the Region has highest average sales.



Insights: From the plot, USA has the highest average sales while India has lowest.

- Identify high-performing and underperforming areas.

```
High-performing areas by region:
warehouse_name
USA    66248696.45

Underperforming areas:
warehouse_name
United Kingdom    54423087.40
India             53066153.46
```

Promotion Opportunities

On which days, promotions should be given to increase sales.

```
Days of the week with below-average sales:
day_of_week
2             19264706.77
3             19741434.06
```

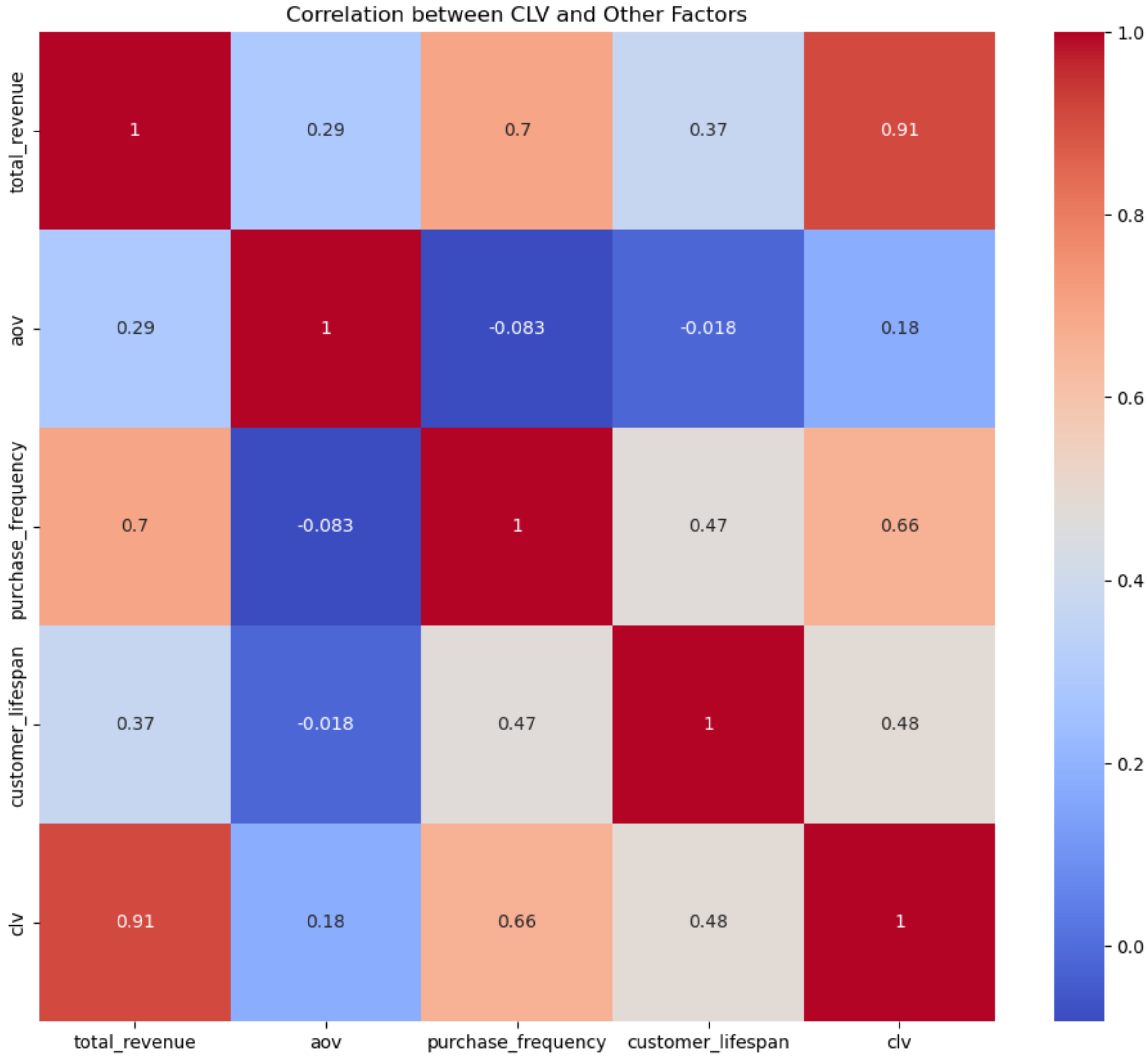
```
5               22500186.65
6               17285881.25
```

8. Customer Lifetime Value (CLV) Analysis

CLV Calculation

- CLV is calculated by multiplying the total revenue generated by a customer with their estimated lifespan, assuming yearly CLV.
- Metrics such as total revenue, average order value (AOV), purchase frequency, and customer lifespan are included in the calculation.

CLV Influencing Factors



- **Total Revenue:** Strong positive correlation with CLV (0.91). Customers generating higher revenue have significantly higher CLV.
- **Purchase Frequency:** Moderate positive correlation with CLV (0.66). Frequent buyers tend to have higher CLV.
- **Customer Lifespan:** Weak to moderate positive correlation with CLV (0.48). The longer a customer stays active, the higher their CLV, but this impact is smaller compared to revenue and frequency.
- **Average Order Value (AOV):** Weak positive correlation with CLV (0.18). AOV has a relatively small impact, indicating that frequent purchases are more important than high-value purchases.

Insights on revenue, AOV, and purchase frequency

- **Total Revenue and Purchase Frequency:** Strong positive correlation (0.70). Customers who purchase more frequently tend to generate higher total revenue.
- **AOV and Purchase Frequency:** Slight negative correlation (-0.083). Customers making larger purchases tend to buy less frequently, suggesting a trade-off between order size and frequency.

Customer Segmentation Based on CLV

- **High-Value Customers:** Identified as those in the top 25% of CLV. These customers contribute a significant portion of revenue and tend to have higher purchase frequency.
- **Low-Value Customers:** Identified as those in the bottom 25% of CLV. These customers have lower revenue, fewer purchases, and shorter lifespans.

CLV Distribution

- The distribution of CLV shows a skewed pattern, with a smaller number of high-value customers contributing a disproportionate amount of revenue, while a larger segment consists of low-value customers.

9. Basket Analysis

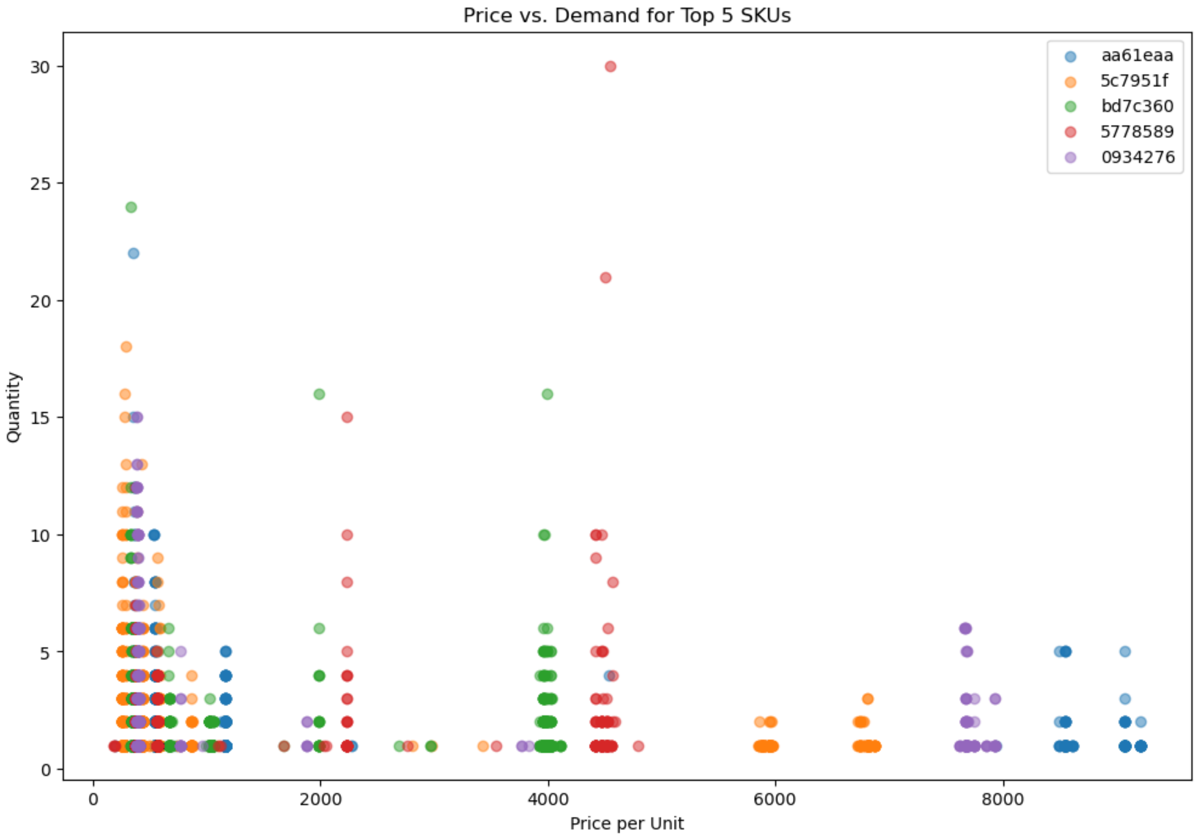
Market Basket Analysis

In this section, I will find which pairs are most likely to bought together.

```
Most freq. product combinations in multi-item orders:          Count  SKU_ID
(aa61eaa, bd7c360)                                              183    12
(0eeddec, 8705857)                                              159    11
(d0990b0, f4575a8)                                              143     8
(941d30b, cb91396)                                              137     8
(92e3cb7, d0990b0)                                              115     6
(8ae2033, 92e3cb7)                                              114     6
(0eeddec, 77ea1be)                                              110     6
(0934276, 56f9240)                                              104     6
(380b808, 5ef89e9)                                              103     5
(0085272, 385c311)                                               98     5
```

10. Price Sensitivity Analysis

Price vs. Demand



- The demand for SKUs is generally higher at lower prices, as seen by the dense clustering of points on the left side of the plot, where the price per unit is less than 2000.
- The SKU labeled '0934276' (purple) shows a broad distribution of prices, ranging from very low to as high as 8000, with a consistent but low quantity sold across this range.
- The SKU labeled 'aa61eaa' (blue) demonstrates significant demand even at higher prices, peaking around 8000, suggesting a premium product with stable demand.
- There's a notable drop in demand as the price increases beyond 2000 for most SKUs, indicating price sensitivity among customers for these products.
- Most of the data points for all SKUs are concentrated around quantities less than 10, indicating that bulk purchases are less common, regardless of price.

- The SKU '5778589' (red) has instances of high demand (quantities around 15-20) at lower price points, suggesting it is a popular choice when priced competitively.

Thank You