Comprehensive Data Analysis on Sale Data

Ayush Kumar Mishra

September 23, 2024

Dashboard

For seeing all the code live interactively,

Visit Dashboard

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

Contents

1. Introduction

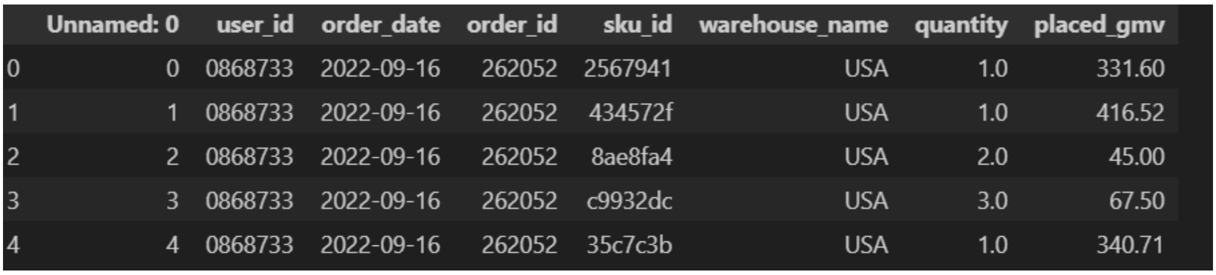
In this document, I will formulate how i did analysis on the data.
The data contains information about the orders, customers, products, and sales.
The goal of this analysis is to provide insights into customer behavior, sales trends, SKU performance, and other key metrics.
The analysis will be performed 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...

Data Preparation and Overview

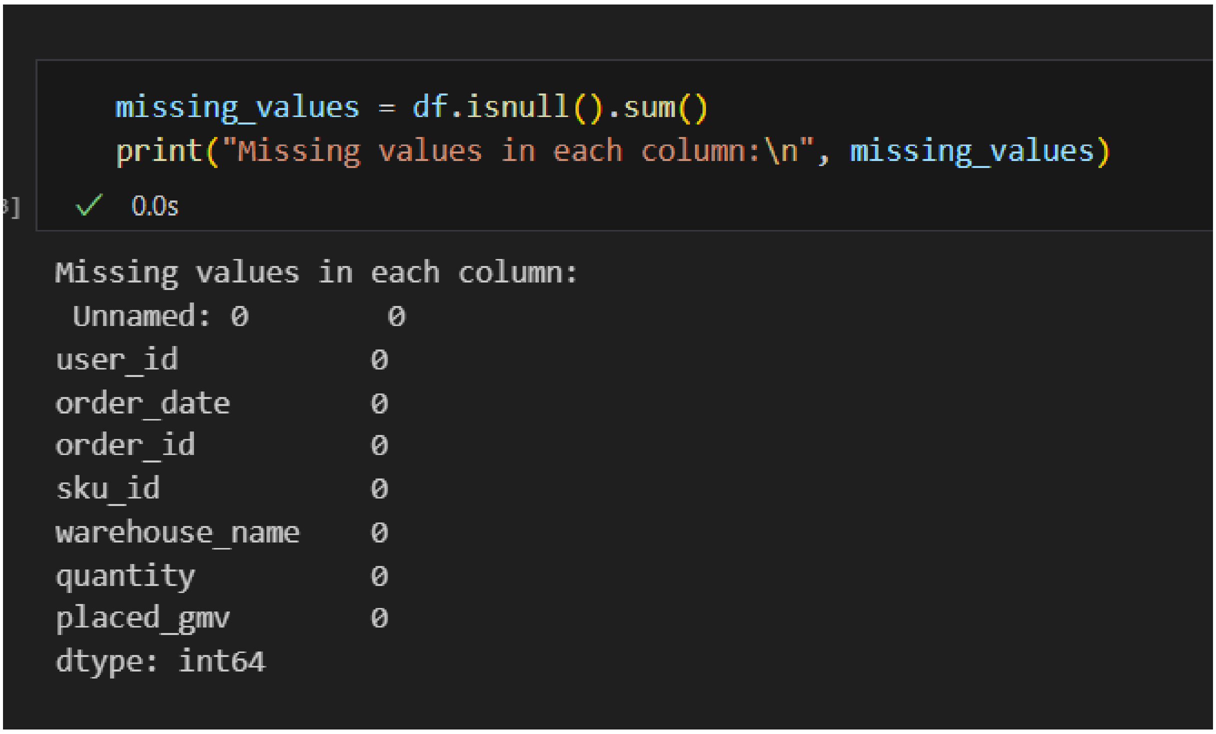
Loading and Inspecting the Dataset

- 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.

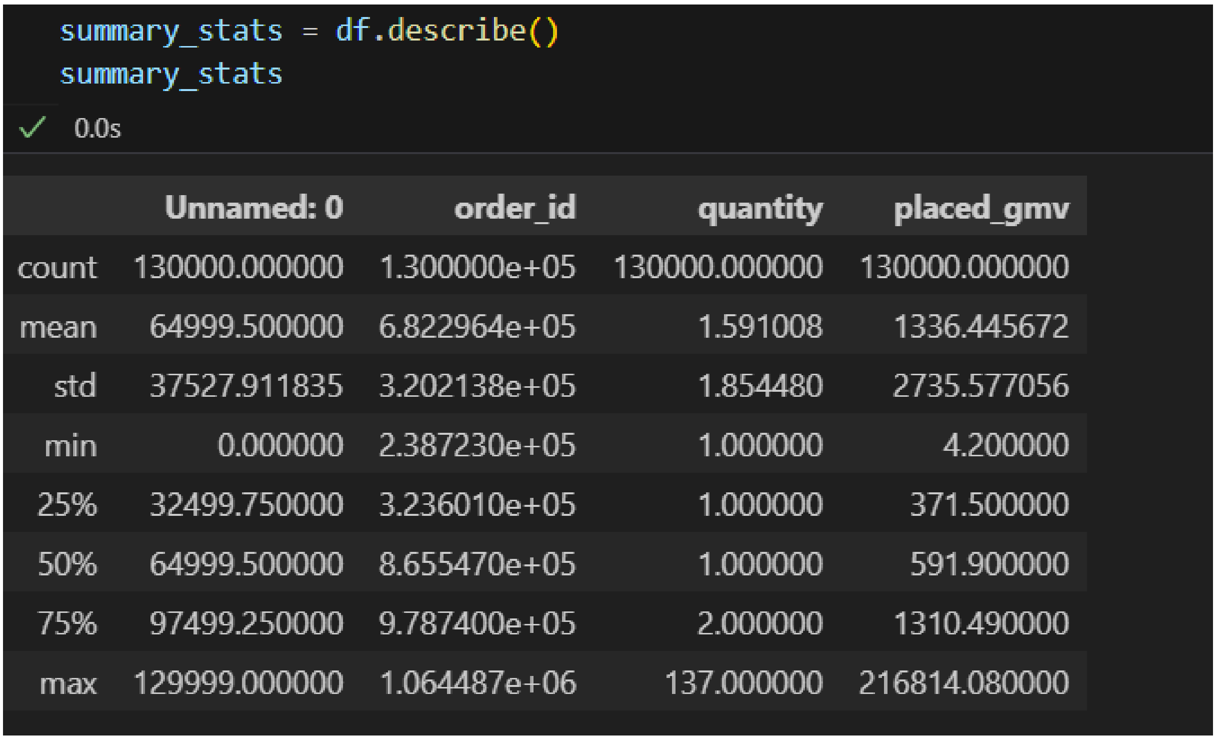


```
missing_values = df.isnull().sum()
print("Missing values in each column:\n", missing_values)
```

✓ 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



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

✓ 0.0s

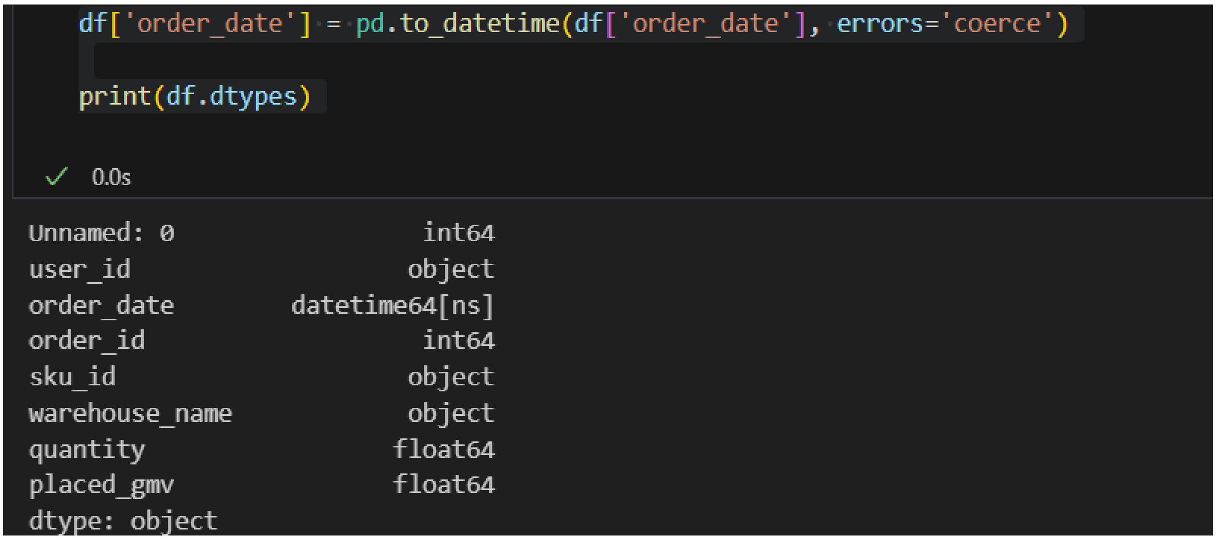
	Unnamed: 0	order_id	quantity	placed_gmv
count	130000.000000	1.300000e+05	130000.000000	130000.000000
mean	64999.500000	6.822964e+05	1.591008	1336.445672
std	37527.911835	3.202138e+05	1.854480	2735.577056
min	0.000000	2.387230e+05	1.000000	4.200000
25%	32499.750000	3.236010e+05	1.000000	371.500000
50%	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)
```

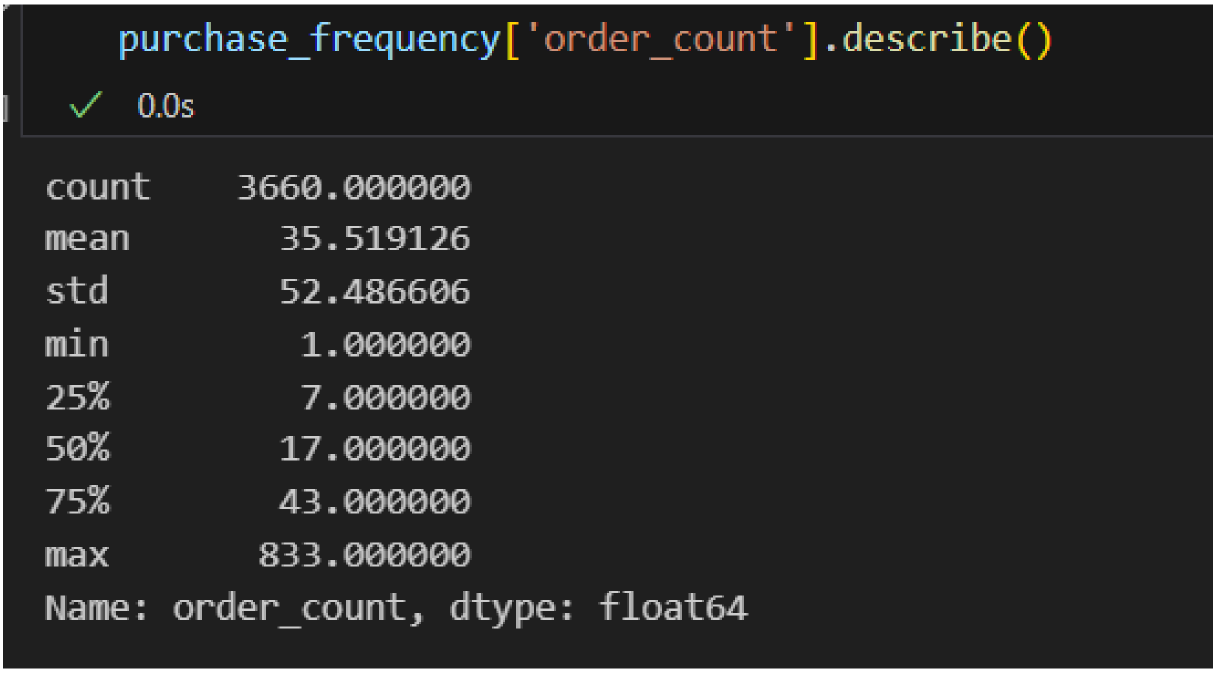
✓ 0.0s

```
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
```

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
```

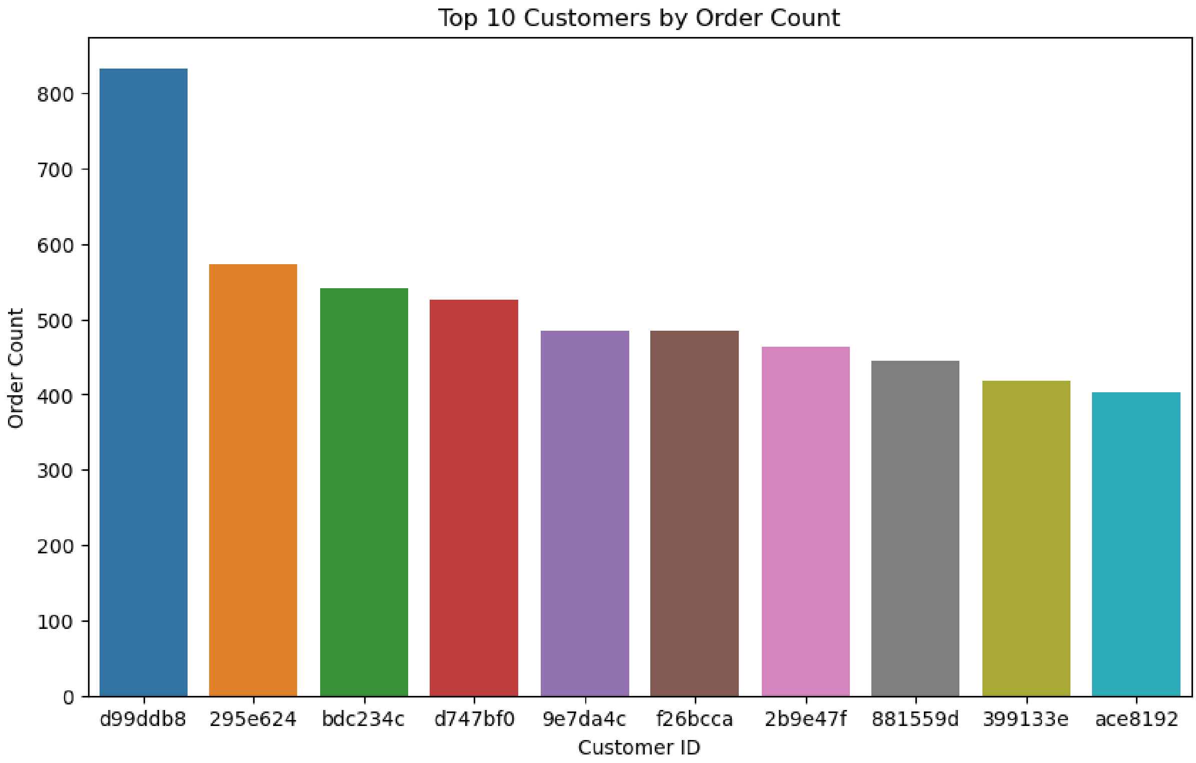
Insights

Insights:

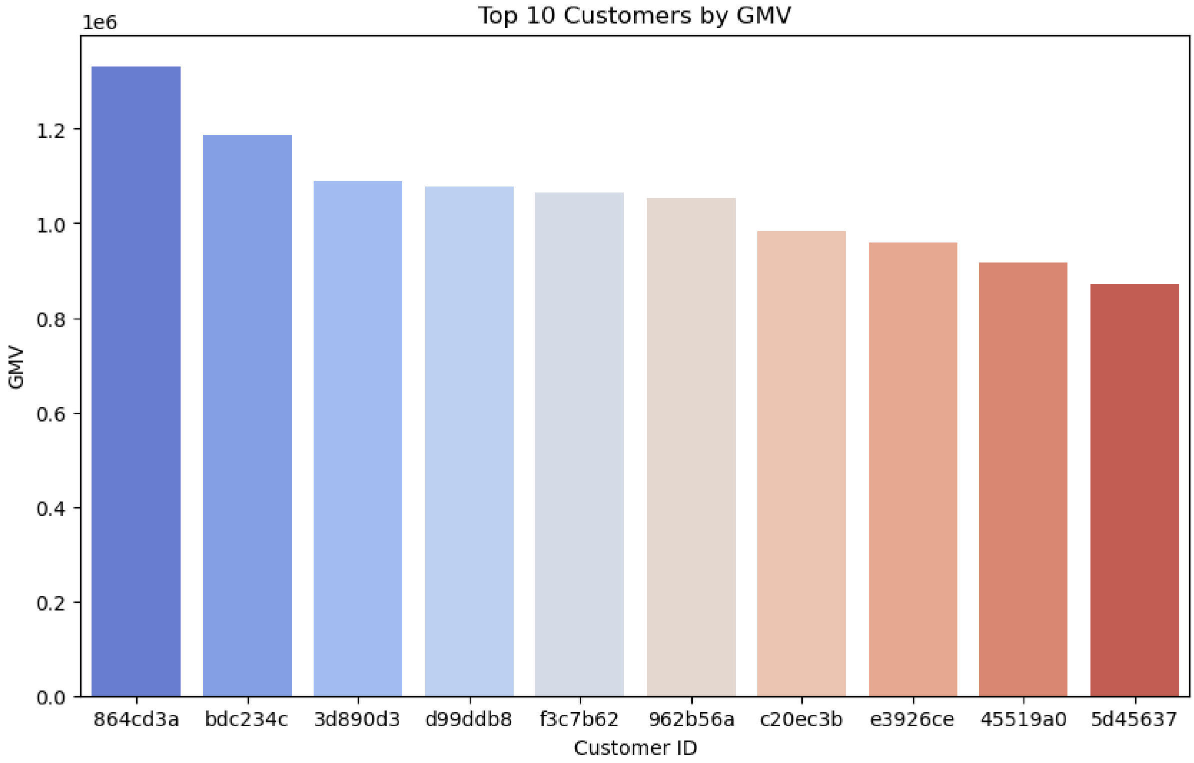
- More than 50% of customers have placed orders less than 17 times which is almost half than means
  . meaning few people are buying a lot.
- And 75% of customers have placed orders less than 43 times.
- Just **293 people** out of 130000 have placed orders more than 100 times.

Top Customers

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



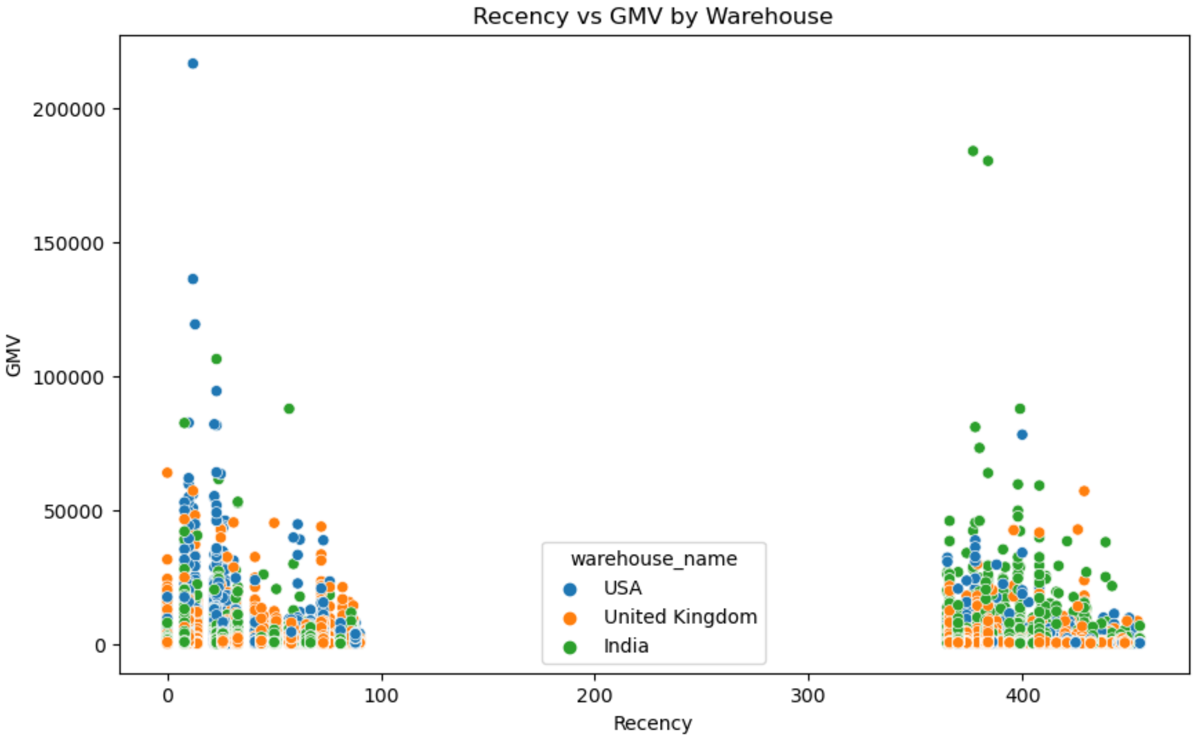
- Based on GMV, I am identifying the top customers.



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.

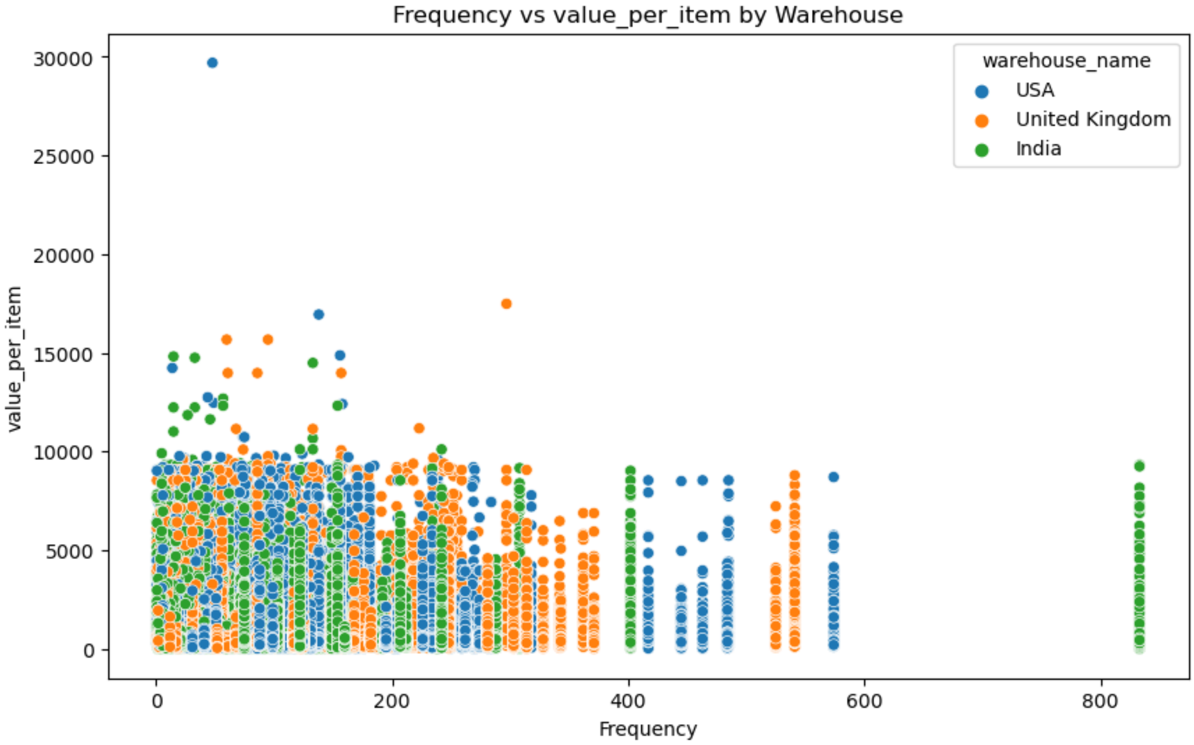


Insights

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.

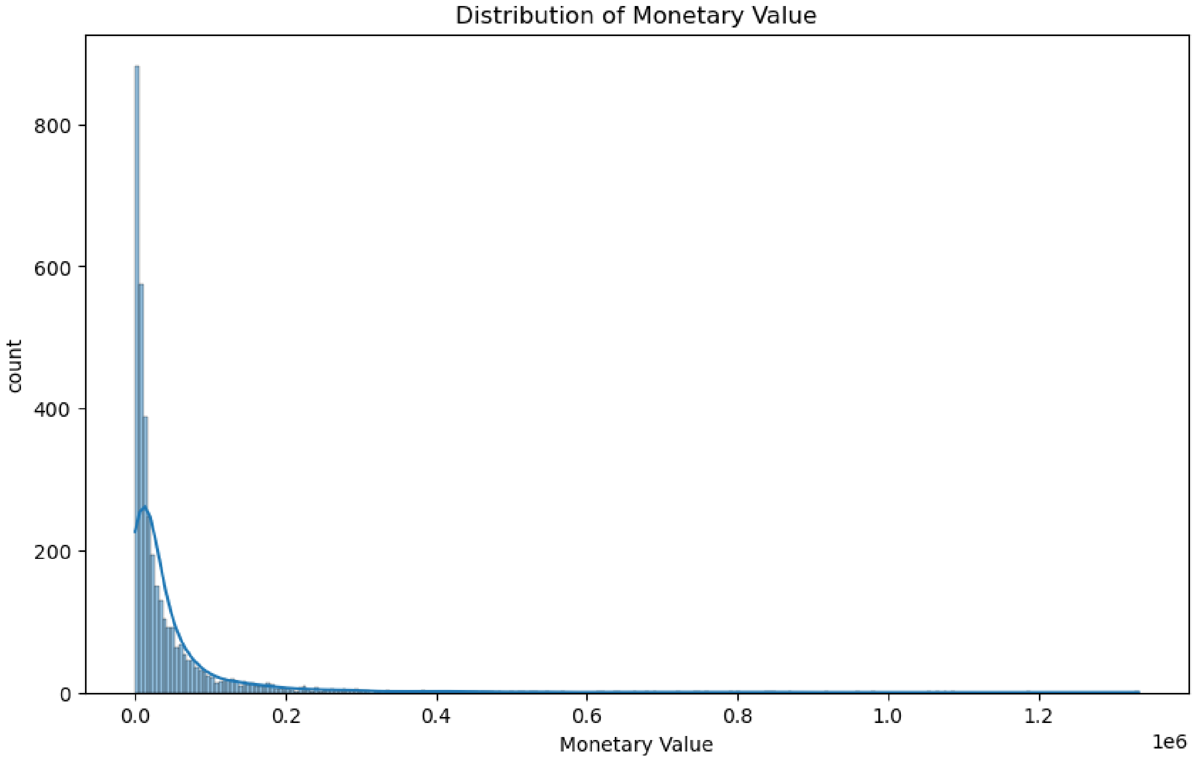
- Frequency: How often the customer made purchases. Here i am calculating the how frequent customers have come to place orders.



Insights

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

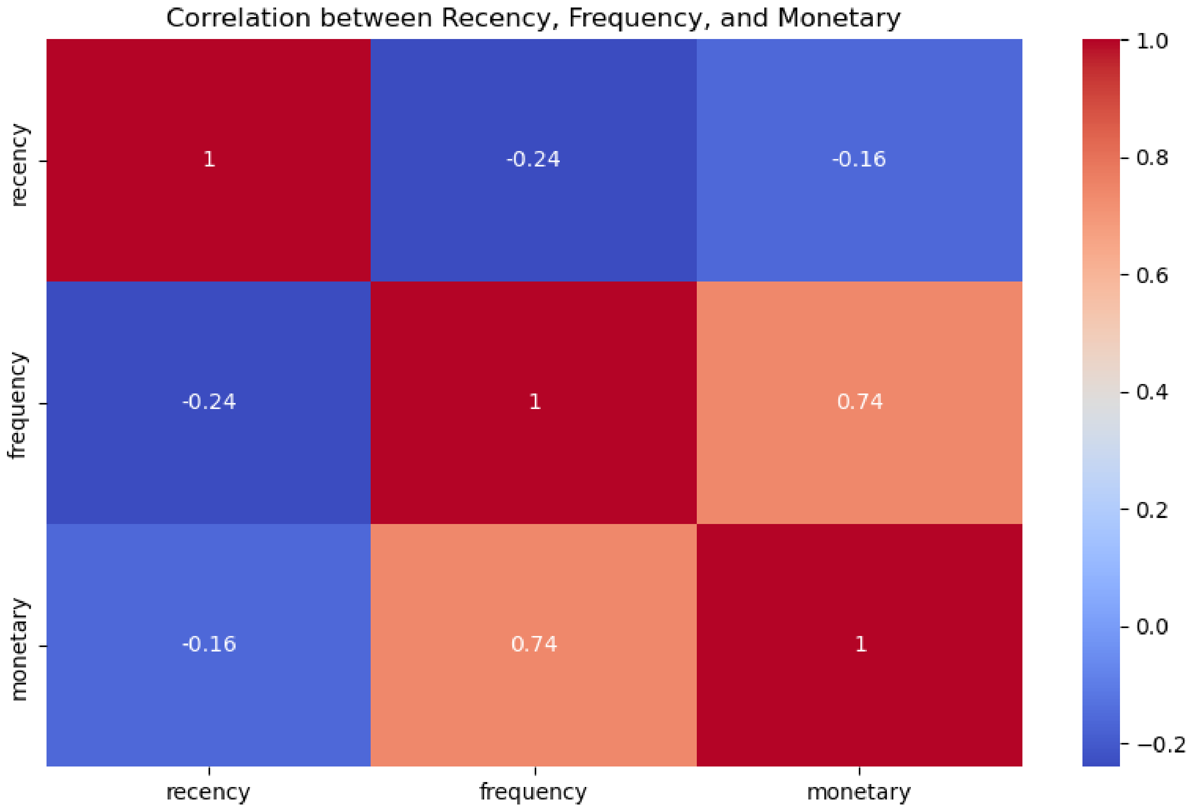
- Monetary: How much money the customer has spent. Here i am calculating the how much money customers have spent.



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.

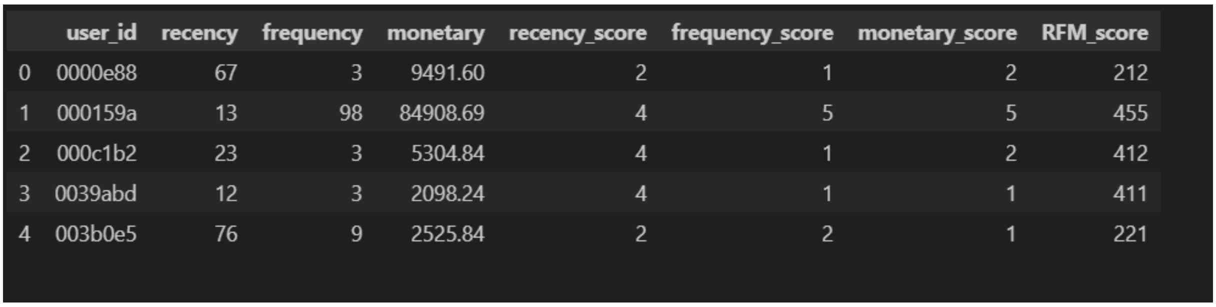


Insights

- 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.

Sales Trends Analysis

Time-based Trends

- Analyze daily, weekly, and monthly sales trends.

Peak Sales Periods

- Identify peak sales periods and any seasonality trends.

Year-over-Year Growth

- Calculate year-over-year growth in sales.

Average Order Value (AOV)

- Analyze trends in average order value over time.

SKU Performance Analysis

Top-Selling SKUs

- Identify the top-selling SKUs based on quantity sold and GMV.

SKU Diversity

- Analyze the diversity of SKUs in customer orders.

ABC Analysis

- Perform ABC analysis to categorize SKUs based on sales contribution.

Purchase Patterns

- Examine SKU purchase patterns and correlations between items.

Order Analysis

Order Sizes

- Analyze the number of items per order.

Relationship Between Order Size and GMV

- Examine the relationship between order size and GMV.

Multi-item Orders

- Identify patterns in orders containing multiple items.

Cohort Analysis

Customer Cohorts

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

Cohort Retention

- Analyze retention rates and purchasing behavior over time for each cohort.

Geographic Analysis

- Analyze sales distribution across different geographic regions.
- Identify high-performing and underperforming areas.

Time-based Analysis

Order Patterns by Day and Time

- Analyze patterns in order timing by day of the week and time of day.

Promotion Opportunities

- Identify potential opportunities for targeted promotions based on time-based analysis.

Customer Lifetime Value (CLV) Analysis

CLV Calculation

- Calculate customer lifetime value (CLV) for various customer segments.

CLV Influencing Factors

- Identify factors that influence CLV.

Basket Analysis

Market Basket Analysis

- Perform market basket analysis to identify frequently co-purchased items.

Product Recommendations

- Generate product recommendations based on customer purchase patterns.

Price Sensitivity Analysis

Price vs. Demand

- Analyze the relationship between price changes and demand for different SKUs.

Price Optimization Opportunities

- Identify opportunities for optimizing pricing strategies.

Visualization and Reporting

- Create informative visualizations to present key insights from the data analysis.
- Prepare a comprehensive report summarizing findings and recommendations.

Advanced Analytics (Optional)

Predictive Modeling

- Develop predictive models for future sales and customer behavior.

Customer Segmentation via Clustering

- Perform clustering analysis to identify distinct customer segments.

Action Plan and Recommendations

- Based on the insights, develop actionable recommendations to improve business performance.