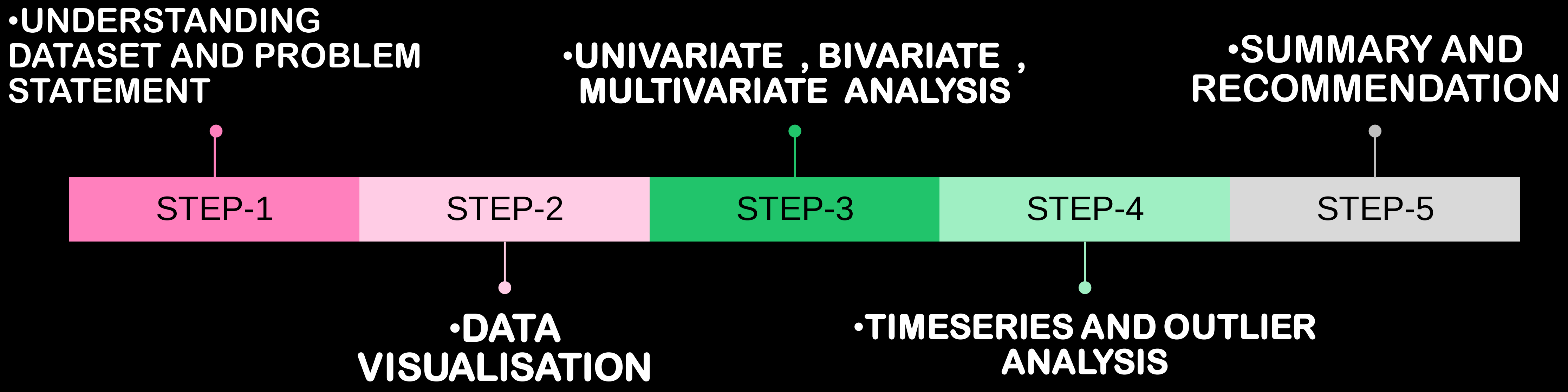
DATA SCIENCE OMNI-MART RETAIL DATASET ANALYSIS.

Presented by : Team DS Enthusiasts

Timeline



STEP-1

•UNDERSTANDING DATASET AND PROBLEM STATEMENT

STEP-2

•DATA VISUALISATION

STEP-3

•UNIVARIATE , BIVARIATE , MULTIVARIATE ANALYSIS

STEP-4

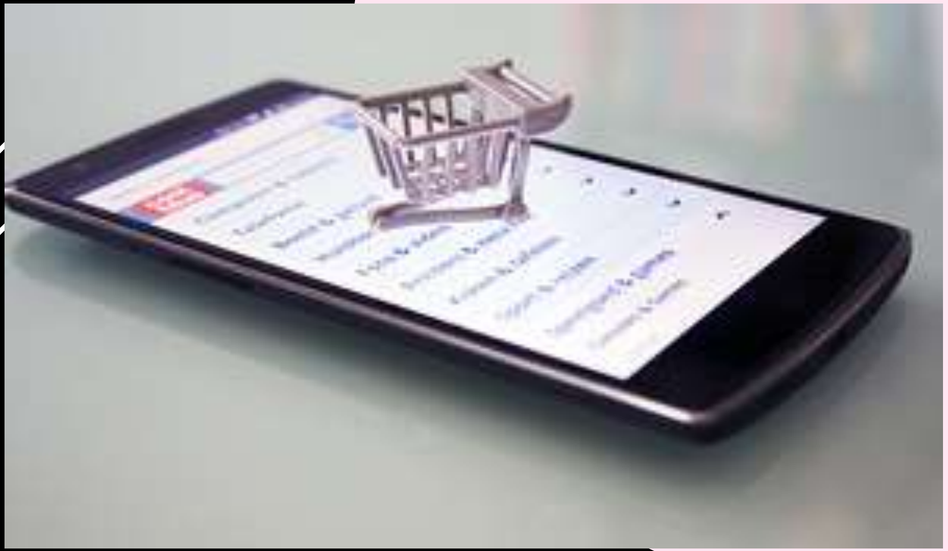
•TIMESERIES AND OUTLIER ANALYSIS

STEP-5

•SUMMARY AND RECOMMENDATION

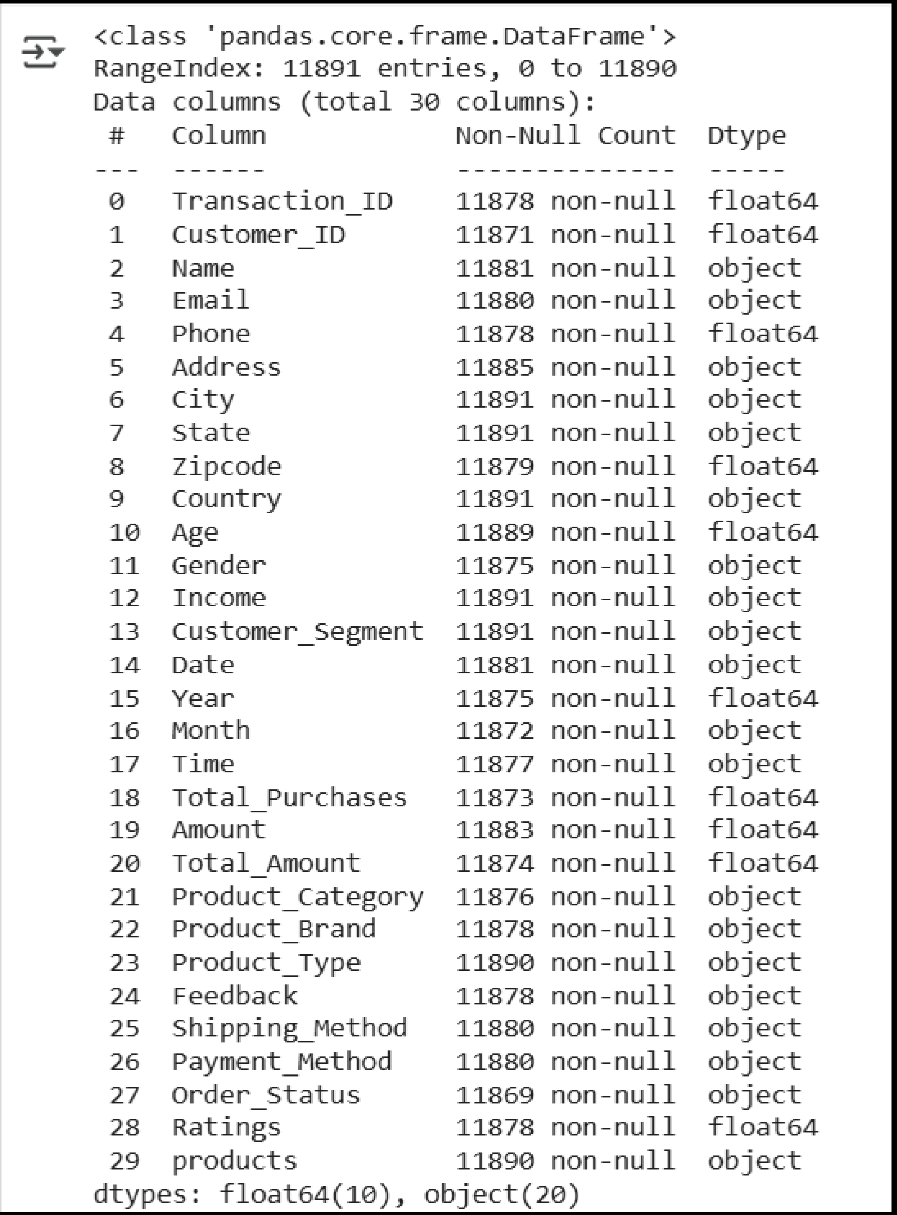
Problem statement

The Retail Challenge: Optimizing Customer Experience and Sales at Omni-Mart Retailers

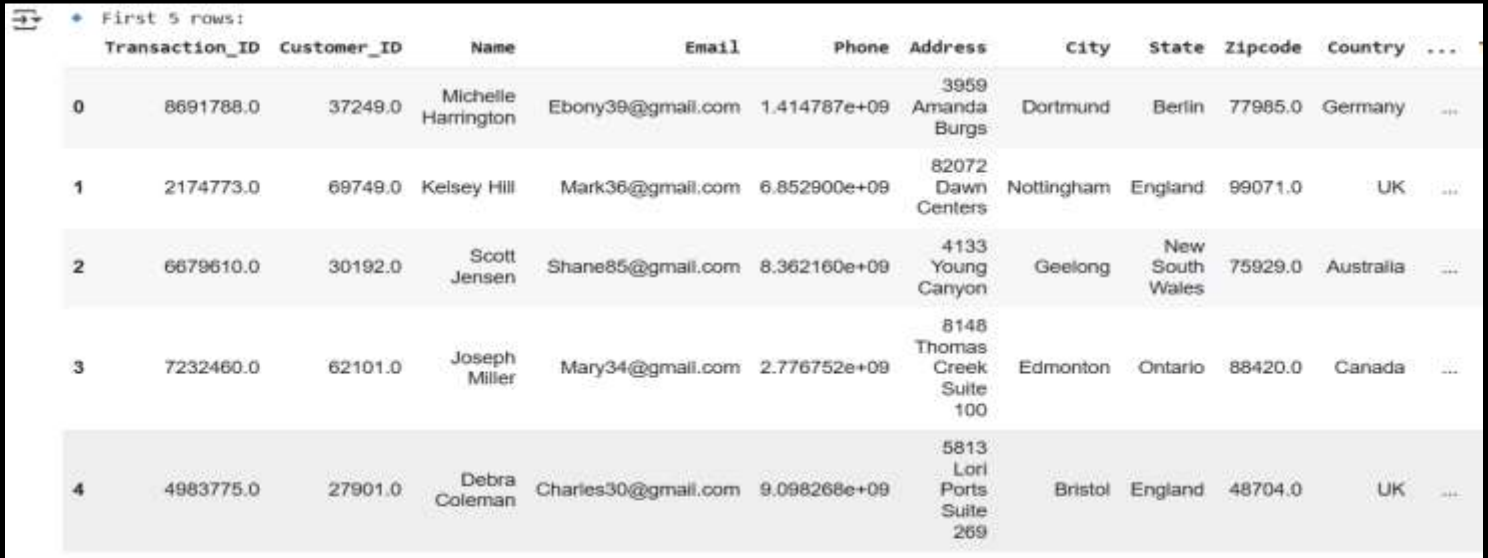


Omni-Mart Retailers is a multinational company with a vast database of customer transactions and feedback. The company's goal is to gain a deeper understanding its customer base to improve sales, increase customer retention, and optimize its marketing strategies of Your challenge is to act as a data analyst for Omni-Mart. Using the provided dataset, your team must perform a comprehensive Exploratory Data Analysis (EDA) to uncover actionable insights. Your analysis should focus on answering key business questions and identifying opportunities for growth.

DATA DESCRIPTION

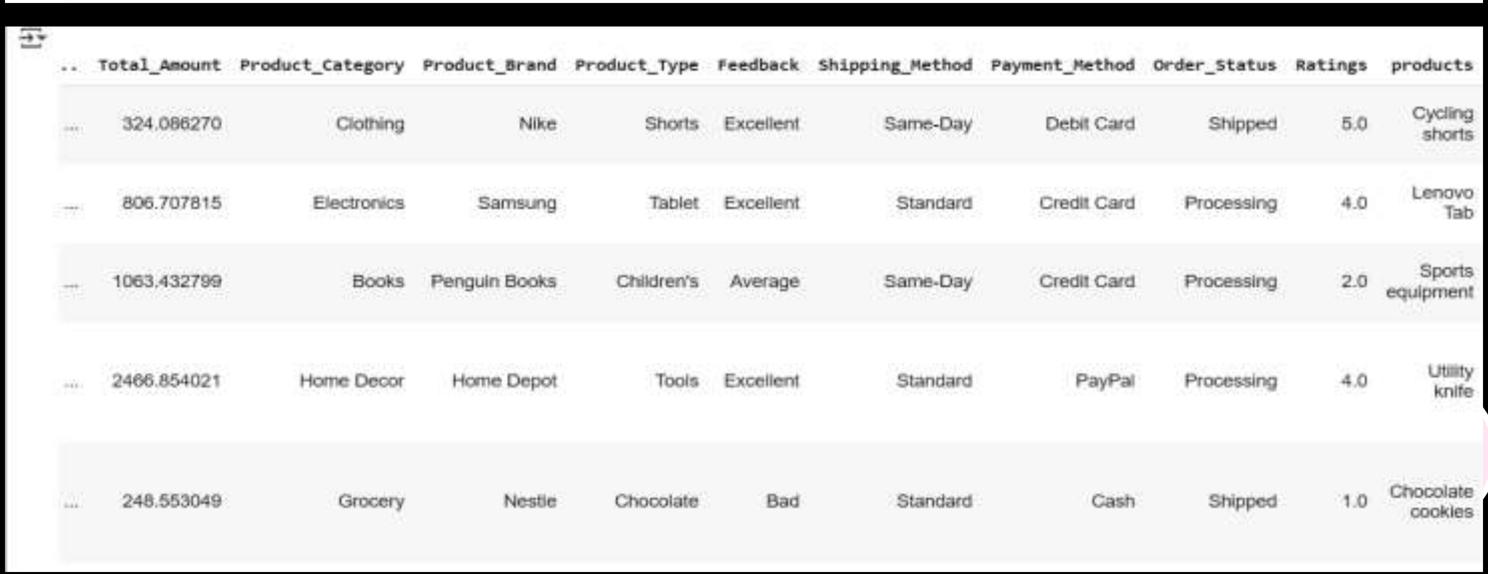


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11891 entries, 0 to 11890
Data columns (total 30 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Transaction_ID    11878 non-null  float64
 1   Customer_ID       11871 non-null  float64
 2   Name              11881 non-null  object
 3   Email             11880 non-null  object
 4   Phone             11878 non-null  float64
 5   Address           11885 non-null  object
 6   City              11891 non-null  object
 7   State             11891 non-null  object
 8   Zipcode           11879 non-null  float64
 9   Country           11891 non-null  object
 10  Age               11889 non-null  float64
 11  Gender            11875 non-null  object
 12  Income            11891 non-null  object
 13  Customer_Segment  11891 non-null  object
 14  Date              11881 non-null  object
 15  Year              11875 non-null  float64
 16  Month             11872 non-null  object
 17  Time              11877 non-null  object
 18  Total_Purchases   11873 non-null  float64
 19  Amount            11883 non-null  float64
 20  Total_Amount      11874 non-null  float64
 21  Product_Category  11876 non-null  object
 22  Product_Brand     11878 non-null  object
 23  Product_Type      11890 non-null  object
 24  Feedback          11878 non-null  object
 25  Shipping_Method   11880 non-null  object
 26  Payment_Method    11880 non-null  object
 27  Order_Status      11869 non-null  object
 28  Ratings           11878 non-null  float64
 29  products          11890 non-null  object
dtypes: float64(10), object(20)
```



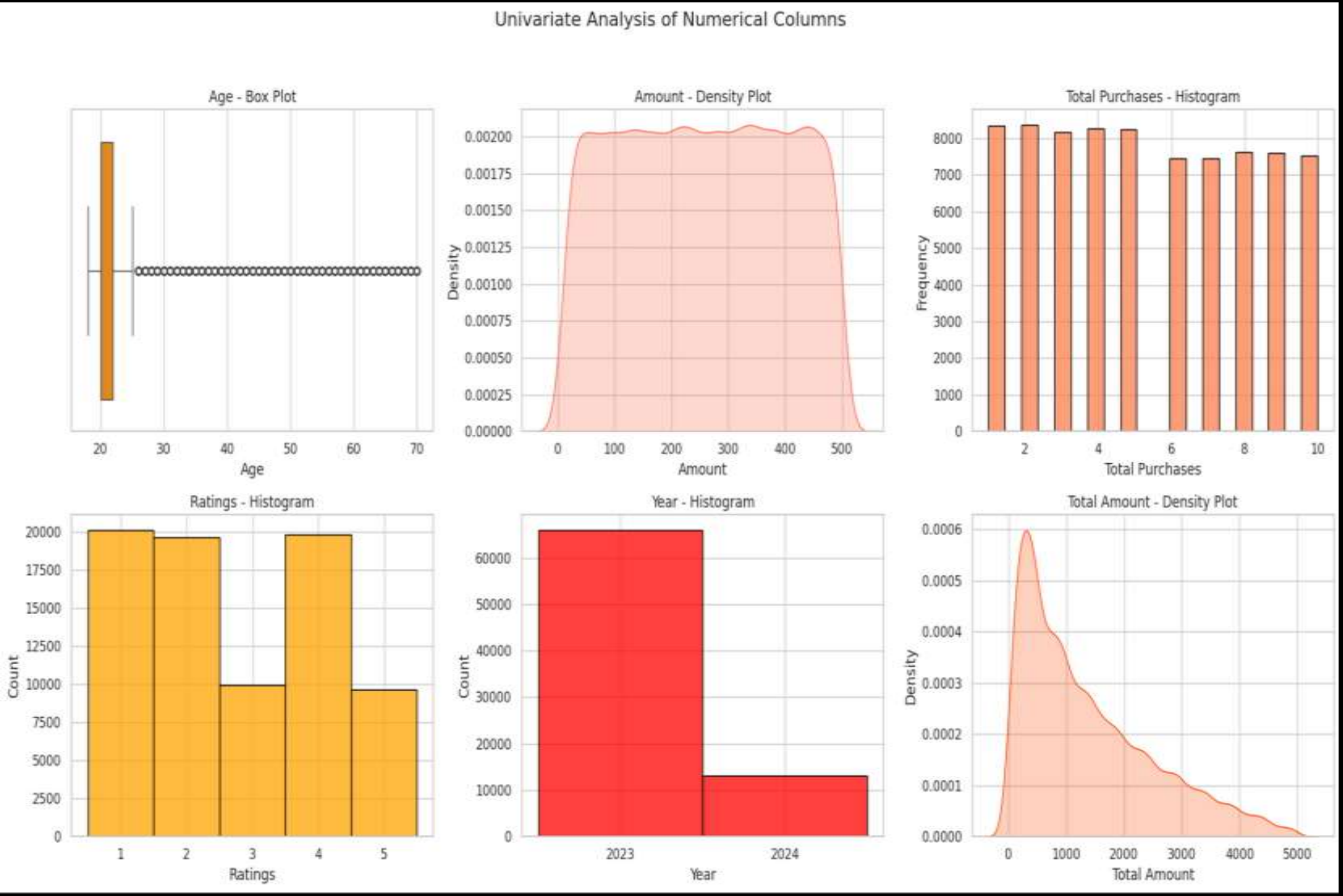
First 5 rows:

	Transaction_ID	Customer_ID	Name	Email	Phone	Address	City	State	Zipcode	Country	...
0	8691788.0	37249.0	Michelle Harrington	Ebony39@gmail.com	1.414787e+09	3959 Amanda Burgs	Dortmund	Berlin	77985.0	Germany	...
1	2174773.0	69749.0	Kelsey Hill	Mark36@gmail.com	6.852900e+09	82072 Dawn Centers	Nottingham	England	99071.0	UK	...
2	6679610.0	30192.0	Scott Jensen	Shane85@gmail.com	8.362160e+09	4133 Young Canyon	Geelong	New South Wales	75929.0	Australia	...
3	7232460.0	62101.0	Joseph Miller	Mary34@gmail.com	2.776752e+09	8148 Thomas Creek Suite 100	Edmonton	Ontario	88420.0	Canada	...
4	4983775.0	27901.0	Debra Coleman	Charles30@gmail.com	9.098268e+09	5813 Lori Ports Suite 269	Bristol	England	48704.0	UK	...



..	Total_Amount	Product_Category	Product_Brand	Product_Type	Feedback	Shipping_Method	Payment_Method	Order_Status	Ratings	products
...	324.086270	Clothing	Nike	Shorts	Excellent	Same-Day	Debit Card	Shipped	5.0	Cycling shorts
...	806.707815	Electronics	Samsung	Tablet	Excellent	Standard	Credit Card	Processing	4.0	Lenovo Tab
...	1063.432799	Books	Penguin Books	Children's	Average	Same-Day	Credit Card	Processing	2.0	Sports equipment
...	2466.854021	Home Decor	Home Depot	Tools	Excellent	Standard	PayPal	Processing	4.0	Utility knife
...	248.553049	Grocery	Nestle	Chocolate	Bad	Standard	Cash	Shipped	1.0	Chocolate cookies

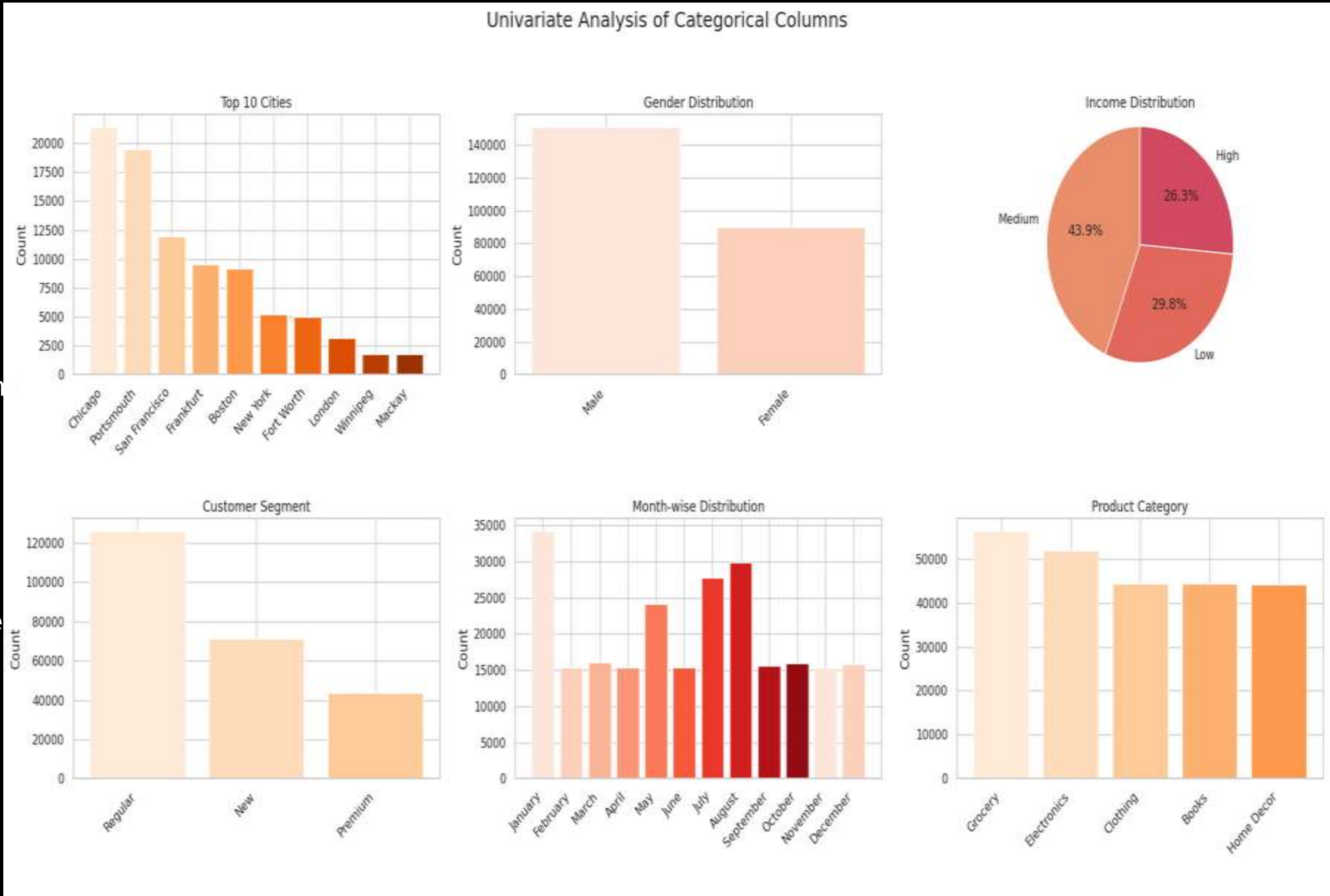
UNIVARIATE ANALYSIS - Numerical Columns



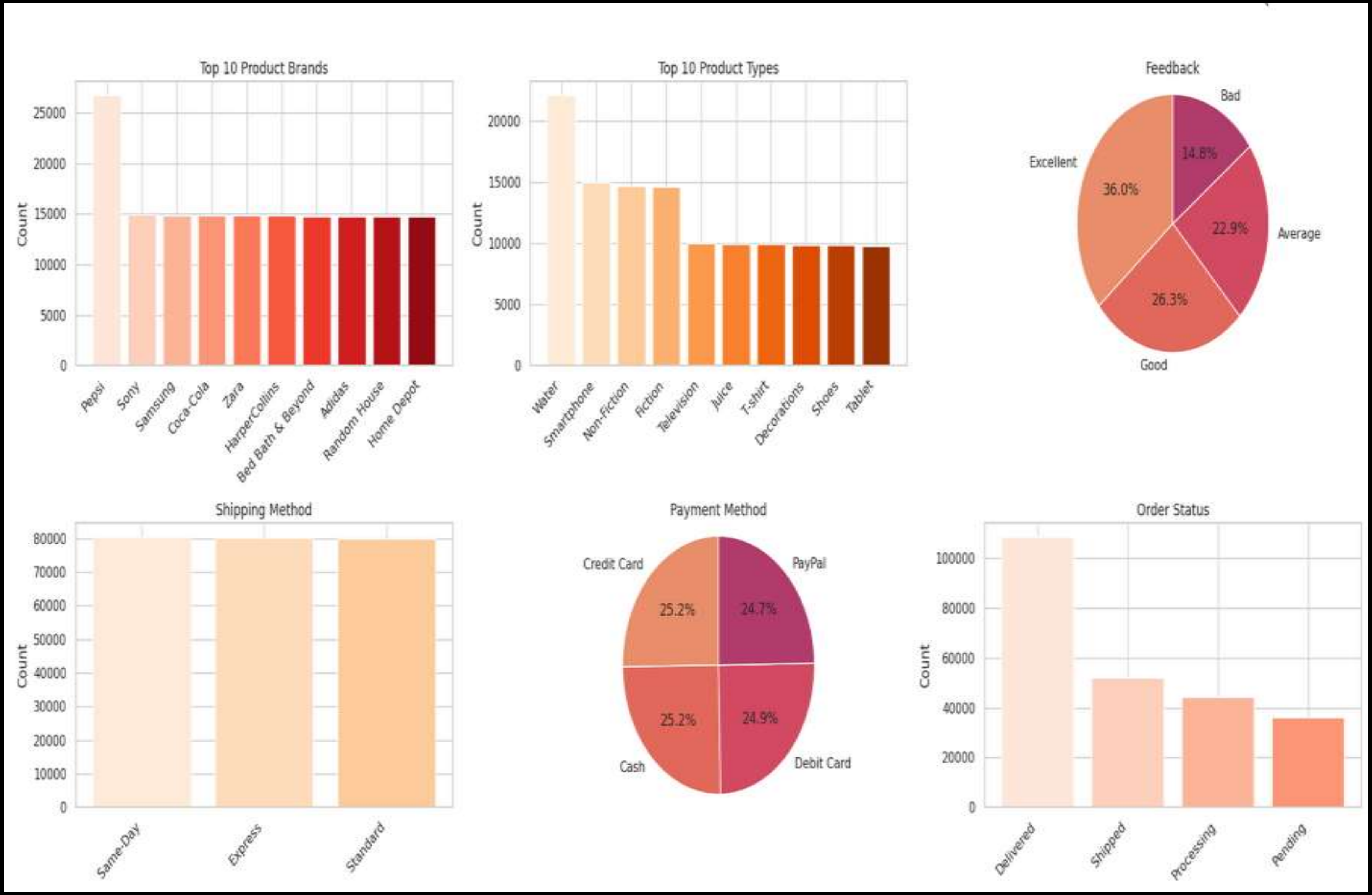
- Age is concentrated around 18–25 with outliers up to 70, suggesting a mostly young customer base.
- Amount shows a nearly uniform distribution, indicating consistent spending patterns.
- Total Purchases are fairly balanced, with most customers buying between 2–10 times.
- Ratings are mostly low to mid-range, hinting at potential dissatisfaction or critical customers.
- Year distribution shows most data comes from 2023, with a drop in 2024.
- Total Amount Spent is right-skewed — a few customers contribute to high-value purchases.

UNIVARIATE ANALYSIS - Categorical columns

- Top Cities: Chicago leads, with most customers than any other city.
- Gender: Majority of the customer base is male.
- Income: It is dominated by the medium-income group, followed by low-income an then few high-income customers.
- Customer Segment: Mostly regular buyers, with few new or premium customers.
- Month-wise Sales: Steady throughout the year, with peaks in April and December.
- Product Category: Grocery and clothing are the most purchased categories.



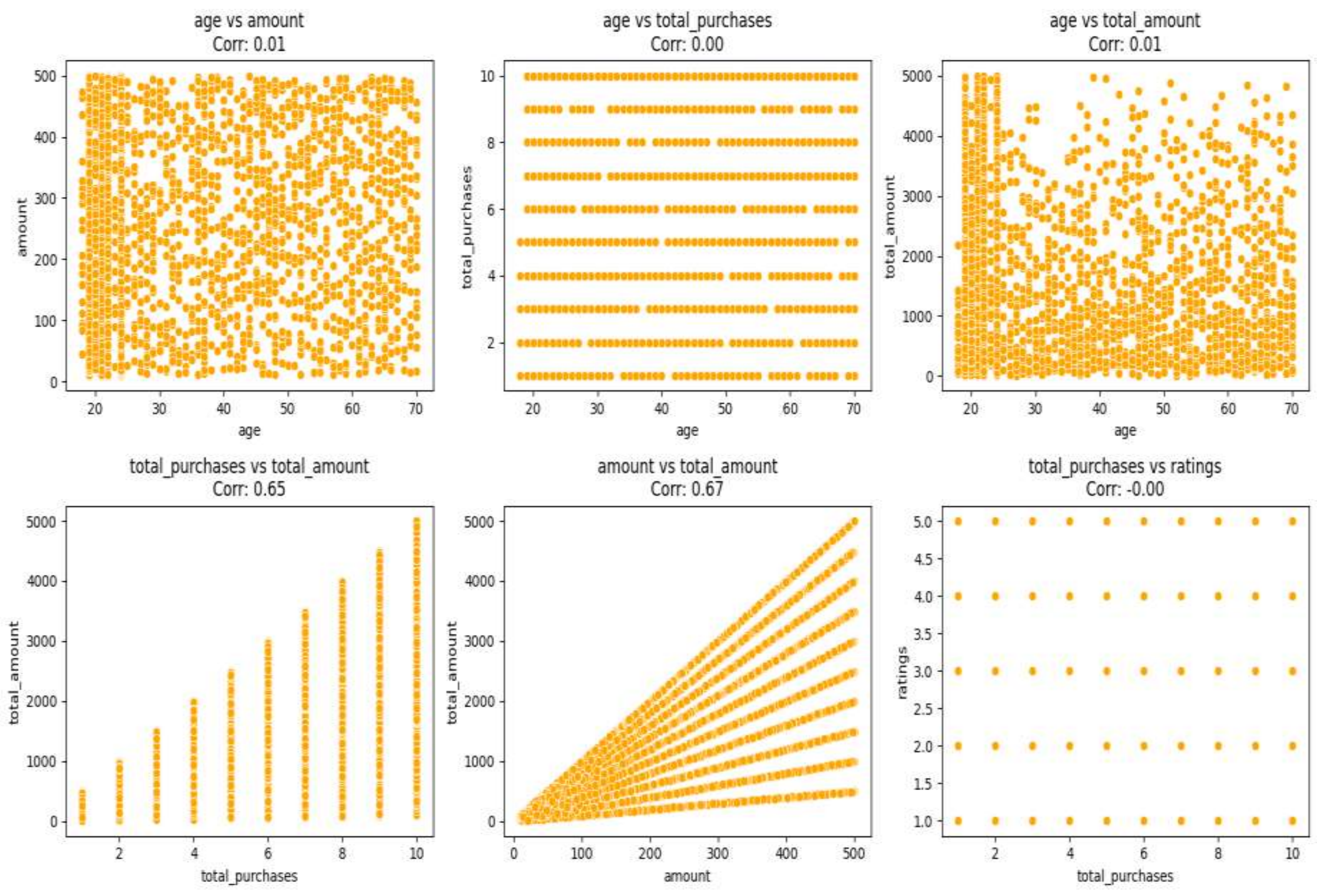
UNIVARIATE ANALYSIS - Categorical columns



- Top Product Brands: Pepsi dominates the sales volume followed by Sony and Samsung.
- Top Product Types: Water and Smart phone lead in sales.
- Feedback: Most people have excellent followed by good review, and negative reviews are quite less.
- Shipping Method: Mostly same-day and express with standard being slightly less used.
- Payment Method: Dominated by debit and credit cards, with cash and PayPal used less.
- Order Status: Majority of the products have been delivered and amount of pending orders are significantly less

Bivariate Analysis- Numerical vs Numerical

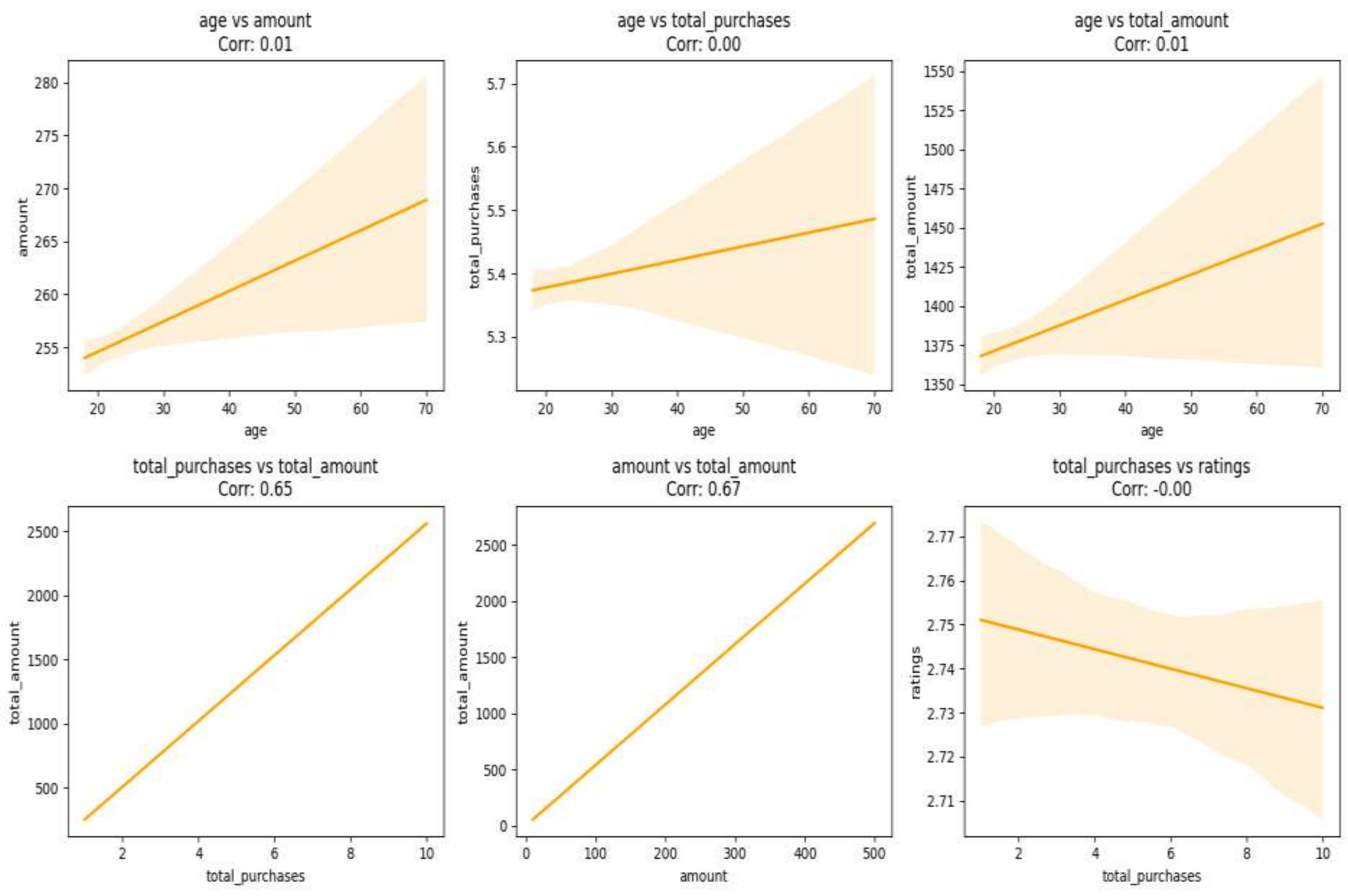
1.SCATTERPLOT



Observations:

- Age vs Amount / Total_Purchases / Total_Amount – Middle-age groups often cluster at moderate–high spending and purchase levels.
- Total_Purchases vs Total_Amount – Strong positive correlation; total spend rises mainly with purchase frequency.
- Amount vs Total_Amount – Strong correlation; larger single purchases contribute to higher overall spend.
- Total_Purchases vs Ratings – Frequent buyers tend to give slightly higher ratings.
- Outliers – A few customers with very high amounts or purchases stand apart; could indicate VIPs or data anomalies.

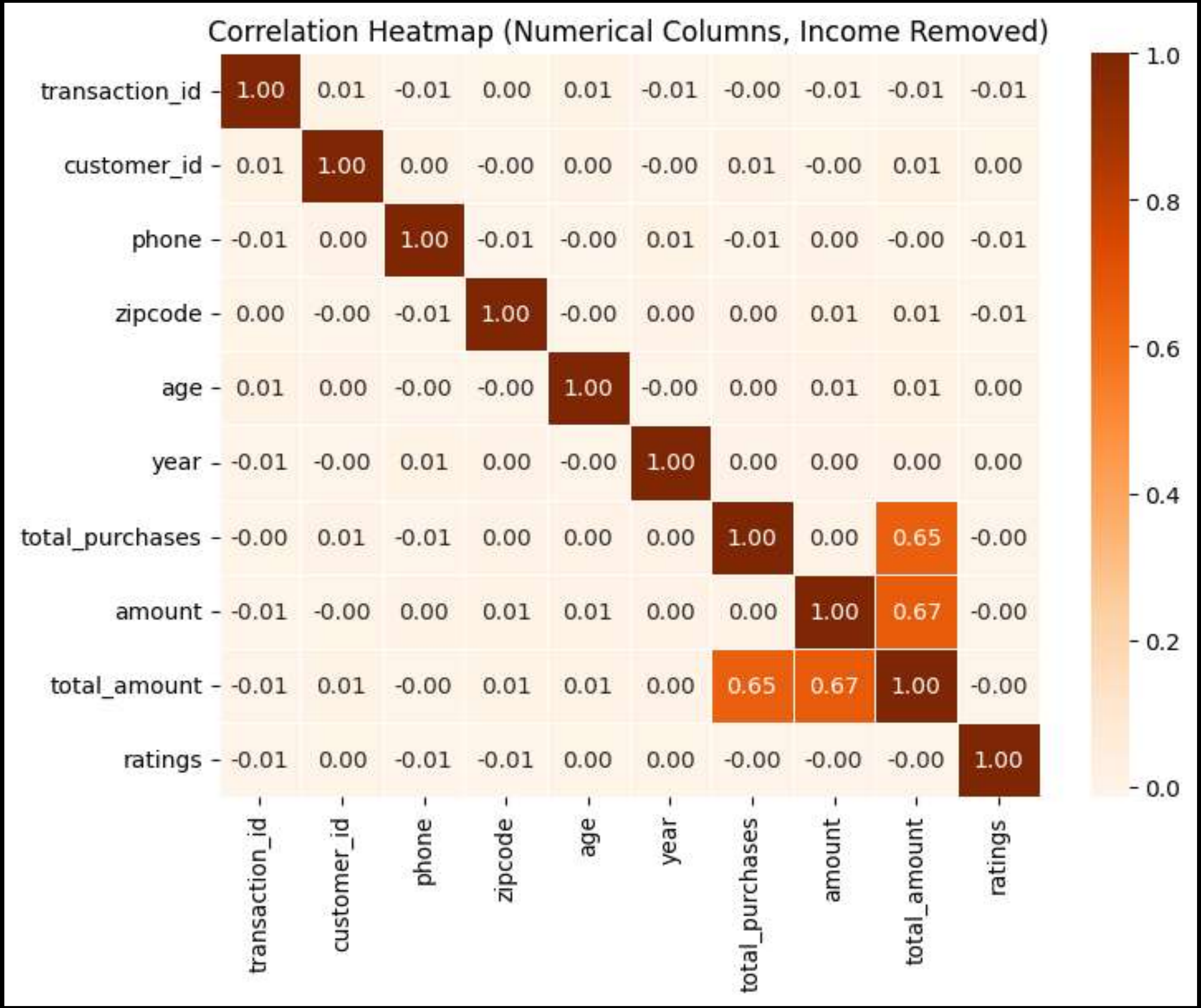
2. Regression Line



Observations:

1. Age vs Amount – Middle-aged customers generally spend more per purchase; slight upward trend.
2. Age vs Total_Purchases – Purchase frequency increases moderately with age, peaking at middle age.
3. Age vs Total_Amount – Total spend rises with age initially, then flattens for older groups.
4. Total_Purchases vs Total_Amount – Strong positive trend; more purchases → higher total spend.
5. Amount vs Total_Amount – Larger single purchases contribute to higher overall revenue.
6. Total_Purchases vs Ratings – Frequent buyers show slightly higher ratings; trend is mild.

3.Correlation Heatmap



Correlation Heatmap (Numerical Columns, Income Removed)

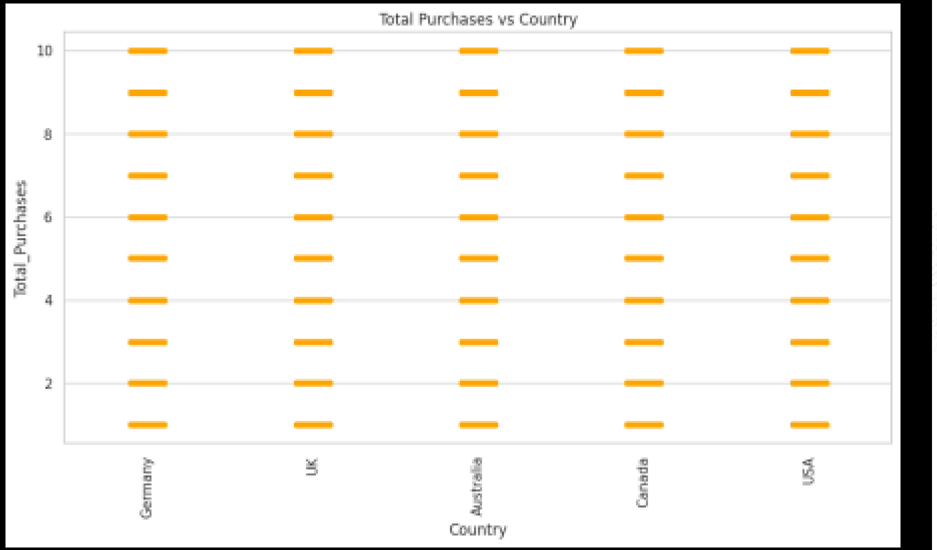
	transaction_id	customer_id	phone	zipcode	age	year	total_purchases	amount	total_amount	ratings
transaction_id	1.00	0.01	-0.01	0.00	0.01	-0.01	-0.00	-0.01	-0.01	-0.01
customer_id	0.01	1.00	0.00	-0.00	0.00	-0.00	0.01	-0.00	0.01	0.00
phone	-0.01	0.00	1.00	-0.01	-0.00	0.01	-0.01	0.00	-0.00	-0.01
zipcode	0.00	-0.00	-0.01	1.00	-0.00	0.00	0.00	0.01	0.01	-0.01
age	0.01	0.00	-0.00	-0.00	1.00	-0.00	0.00	0.01	0.01	0.00
year	-0.01	-0.00	0.01	0.00	-0.00	1.00	0.00	0.00	0.00	0.00
total_purchases	-0.00	0.01	-0.01	0.00	0.00	0.00	1.00	0.00	0.65	-0.00
amount	-0.01	-0.00	0.00	0.01	0.01	0.00	0.00	1.00	0.67	-0.00
total_amount	-0.01	0.01	-0.00	0.01	0.01	0.00	0.65	0.67	1.00	-0.00
ratings	-0.01	0.00	-0.01	-0.01	0.00	0.00	-0.00	-0.00	-0.00	1.00

Observations:

1. Total_Purchases vs Total_Amount – Strong positive correlation; more purchases → higher total spend.
2. Amount vs Total_Amount – High positive correlation; bigger single purchases contribute to higher total revenue.
3. Age vs Amount / Total_Purchases / Total_Amount – Weak to moderate correlation; age has some influence on spending and purchase behavior.
4. Total_Purchases vs Ratings – Mild positive correlation; frequent buyers slightly tend to give higher ratings.
5. Ratings vs other KPIs – Mostly weak correlations; customer satisfaction is not strongly tied to purchase amount or frequency.

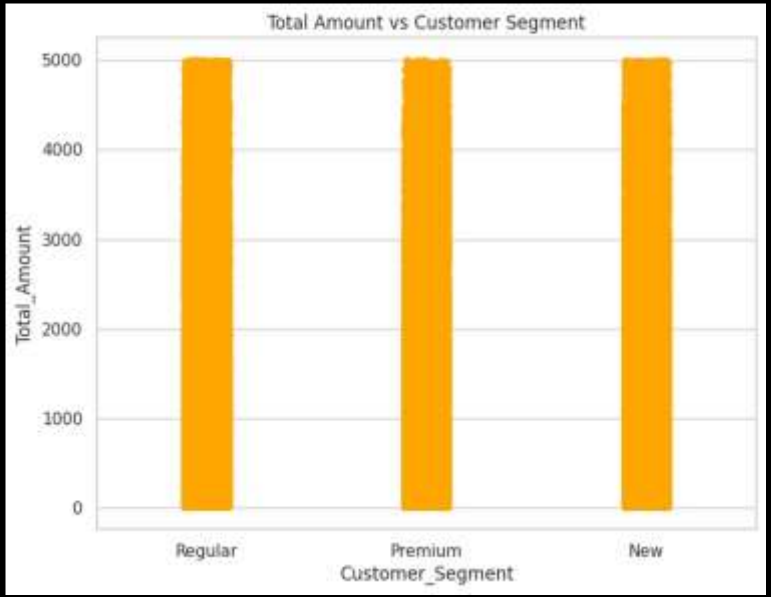
Bivariate Analysis- Numerical vs Categorical

1.Strip Plot



1. Total Purchases vs Country

- The plot shows that **all countries (Germany, UK, Australia, Canada, USA)** have customers distributed across the full purchase range (from 1 to 10).
- There is **no clear dominance** of one country in terms of purchase volume; the distribution looks fairly even.
- This suggests that purchase behavior is **consistent across countries**, with no strong regional skew.

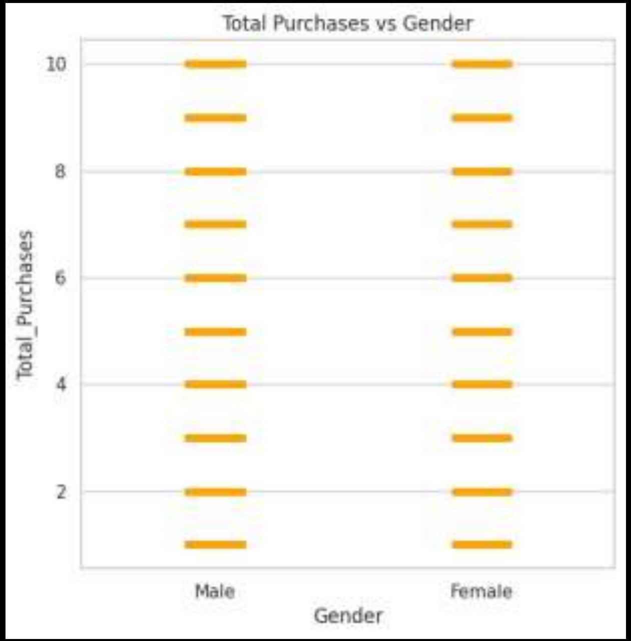


2. Total Amount vs Customer Segment

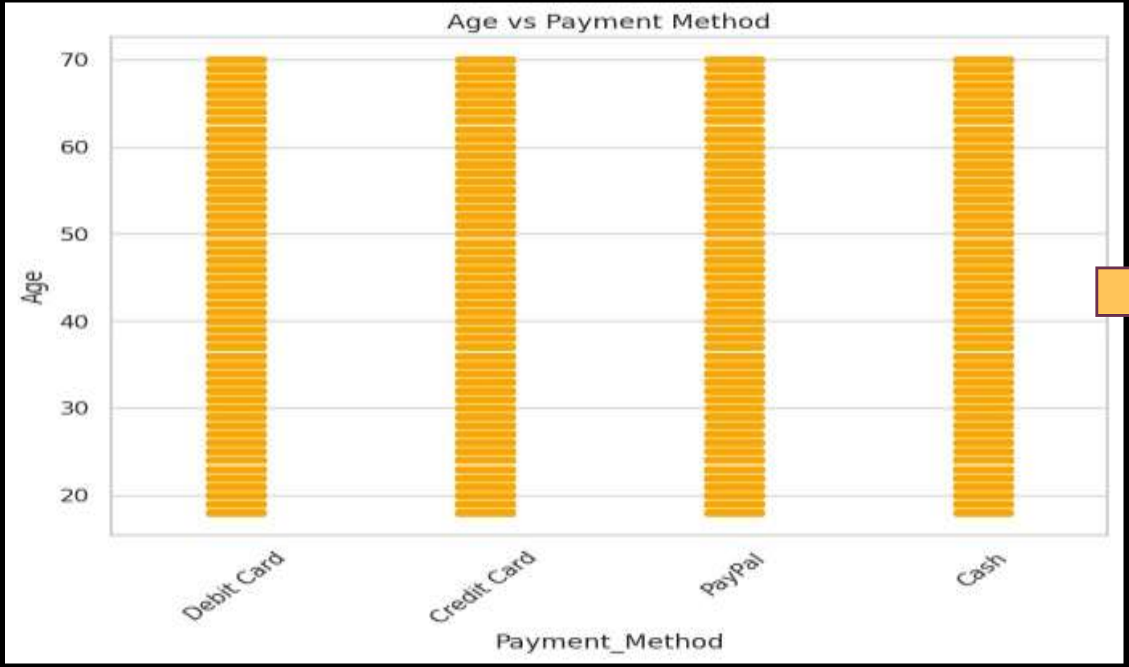
- All three customer segments (**Regular, Premium, and New**) show amounts ranging from **low to very high values (up to around 5000)**.
- There doesn't appear to be a major difference in **spending range** between the segments.
- However, **Premium and Regular customers might be expected to cluster at higher amounts**, while New customers may have more spread, but here all three appear quite similar.

Total Purchases vs Gender:

- The distribution of total purchases appears **similar for both males and females**.
- There is no clear gender bias in purchasing behavior, as both genders have purchases spread across the full range from 1 to 10**.**

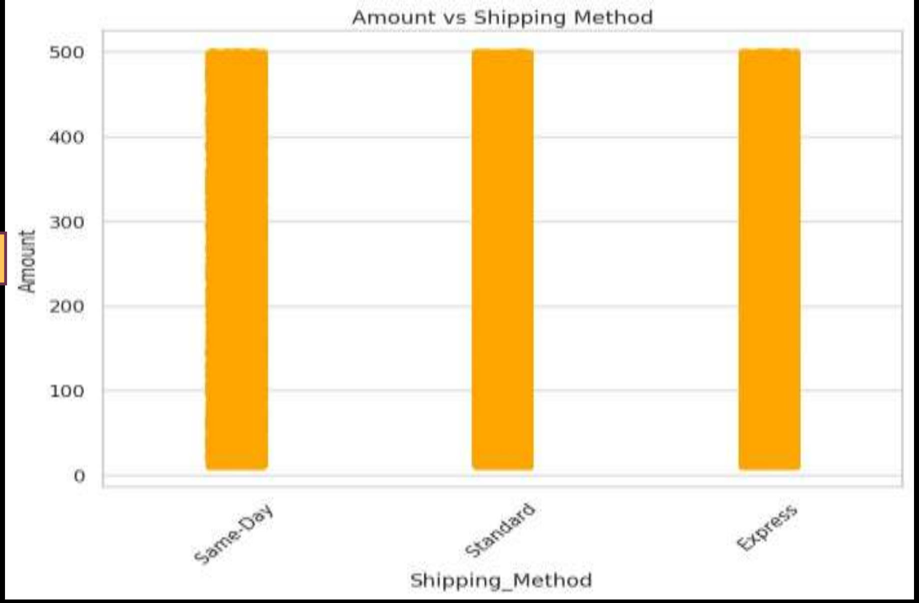


Strip plot



Age vs Payment Method:

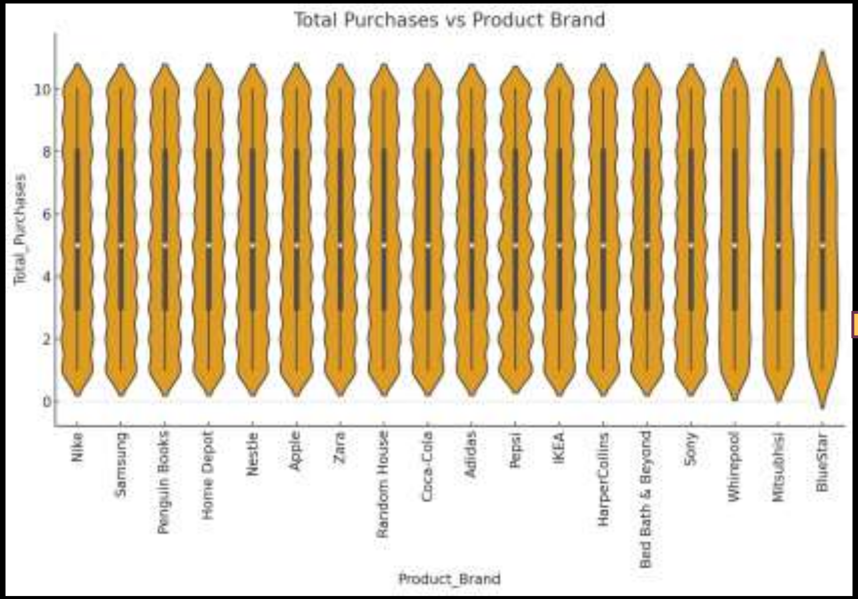
- The age distribution for all payment methods (Debit Card, Credit Card, PayPal, Cash) looks **quite uniform and consistent**.
- No specific age group seems to favor any particular payment method over others.



Amount vs Shipping Method:

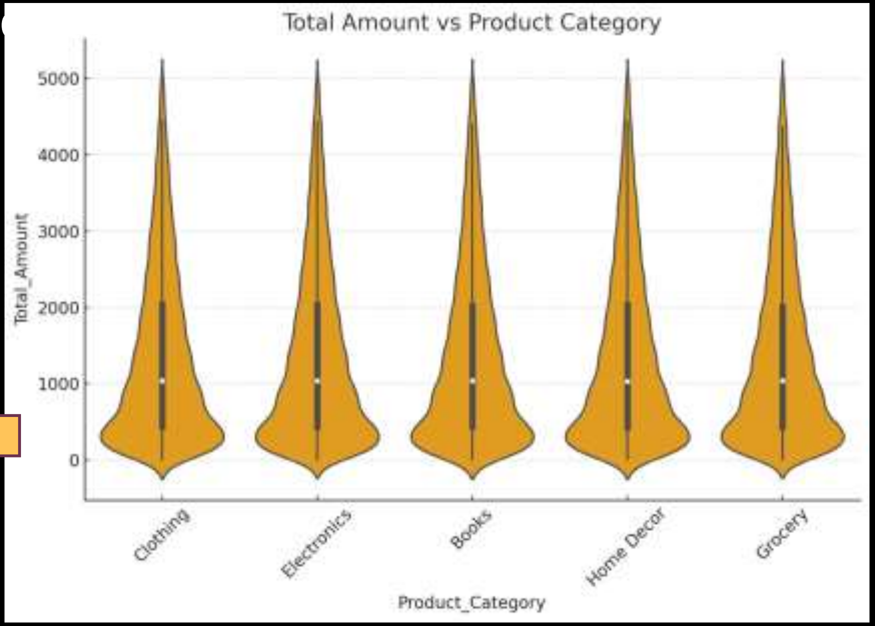
- The purchase amount distribution looks **very consiste across all shipping methods (Same-Day, Standard, Express)**.
- There is no significant difference in the amount based on shipping method.

Violin Plots



2. Total Purchases vs Product Brand

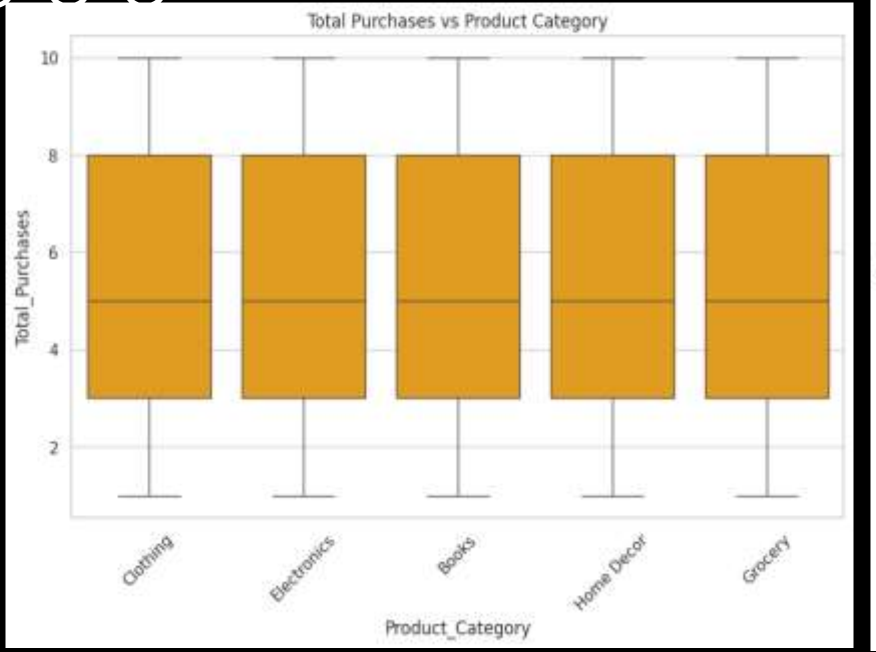
- For all brands (**Nike, Samsung, Apple, Zara, Pepsi, etc.**) the distribution of **Total Purchases** is very similar.
- Purchases mostly cluster around **3–6 items per brand**.
- Extreme values (higher number of purchases per brand) are rare, showing **consistent buying behavior** across brands.
- No brand stands out as having a significantly higher or lower spread compar



1. Total Amount vs Product Category

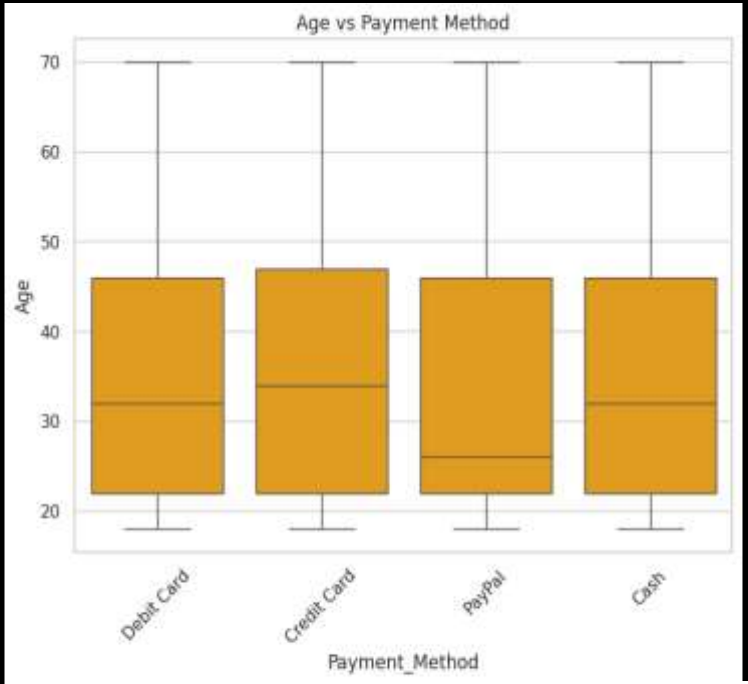
- The distribution of **Total Amount** is quite similar across categories like **Clothing, Electronics, Books, Home Decor, and Grocery**.
- Most transactions are clustered at lower amounts (close to 0–1000), but a few high-value purchases go up to around **5000**.
- The spread is widest at lower values, suggesting **more small purchases than large ones** across all categories.

2. BOX Plots



Total Purchases vs Product Category (Box Plot):

- The distribution of total purchases is **quite uniform across product categories** (Clothing, Electronics, Books, Home Decor, Grocery).
- Median total purchases seem to hover around 5 purchases for all categories.
- There are no strong outliers or significant differences in purchase behavior between categories

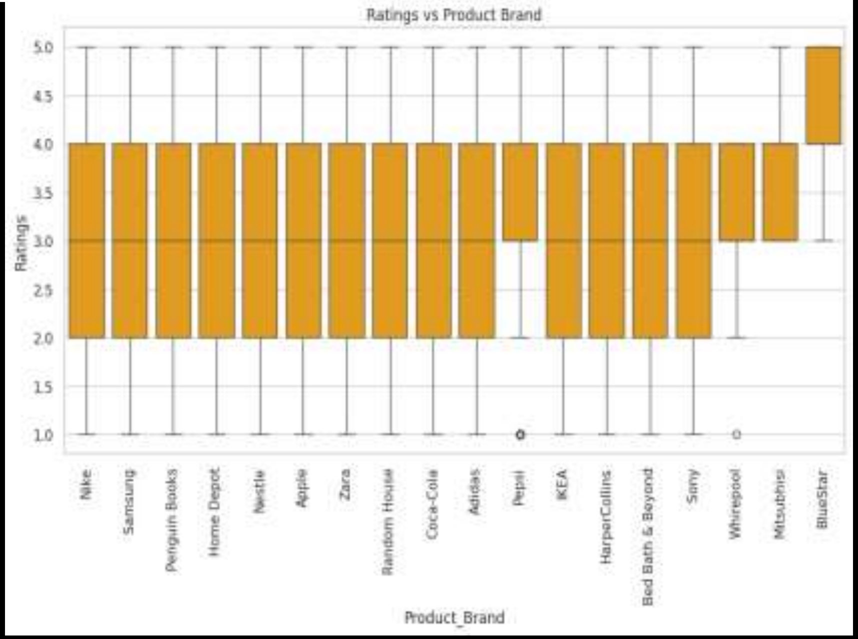


Age vs Payment Method (Box Plot):

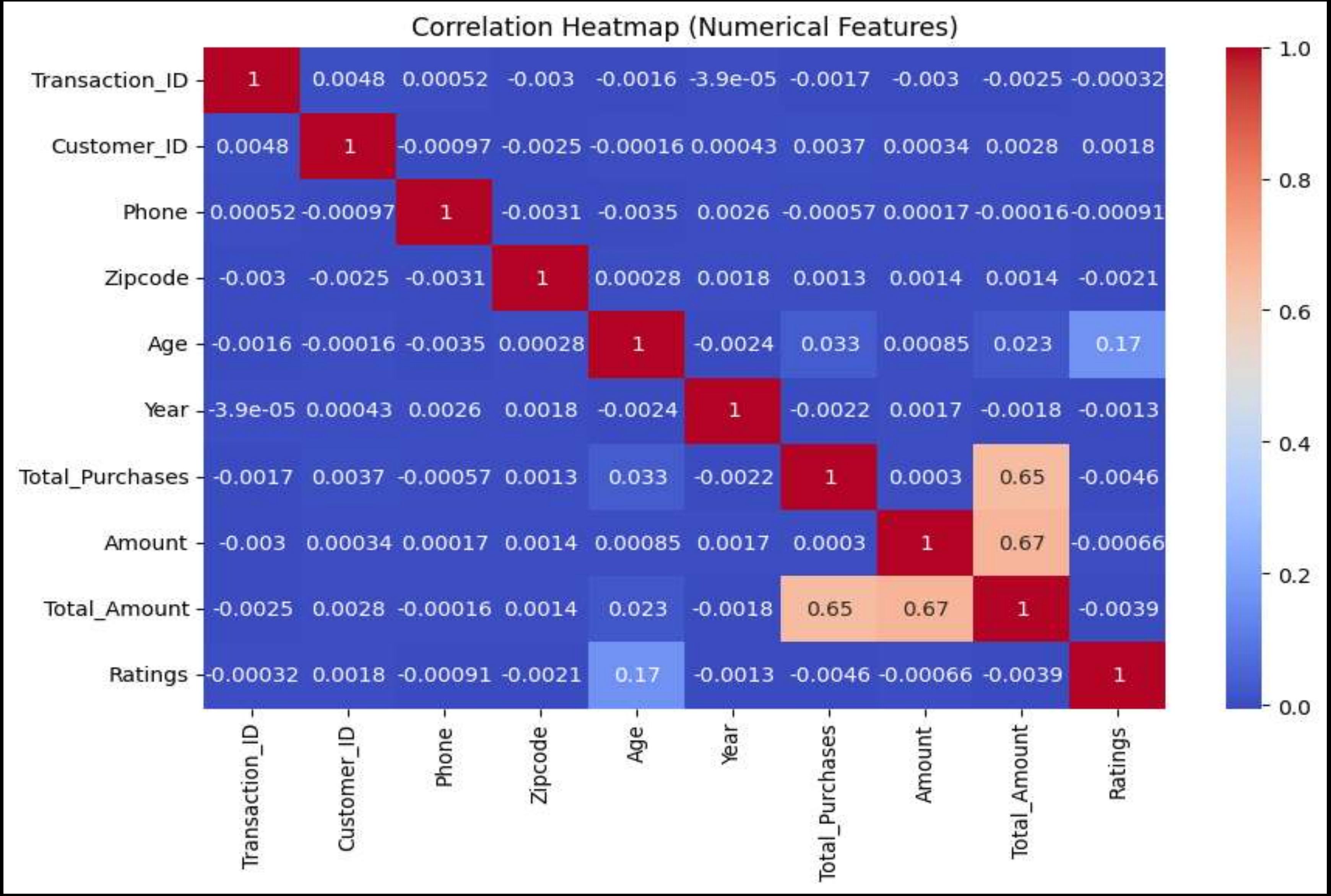
- The median age is similar for Debit Card, Credit Card, and Cash payments, roughly around the early 30s.
- Interestingly, the PayPal method shows a **lower median age**, suggesting younger people prefer using PayPal more than other methods.
- The age range for all methods spans from ~18 to ~70, indicating a wide usage across age groups

Ratings vs Product Brand box plot:

- Most product brands show a **wide range of ratings from 1 to 5**, indicating variability in customer satisfaction.
- The **BlueStar brand stands out** with consistently high ratings (most ratings are close to 5), suggesting better customer satisfaction compared to other brands.
- A few brands like **Pepsi and Whirlpool have some outlier low ratings (around 1)**, indicating that some customers were particularly dissatisfied.
- Overall, the median ratings for most brands seem to hover around 3 to 4, showing average customer satisfaction across the board.



3.Correlation Heatmap



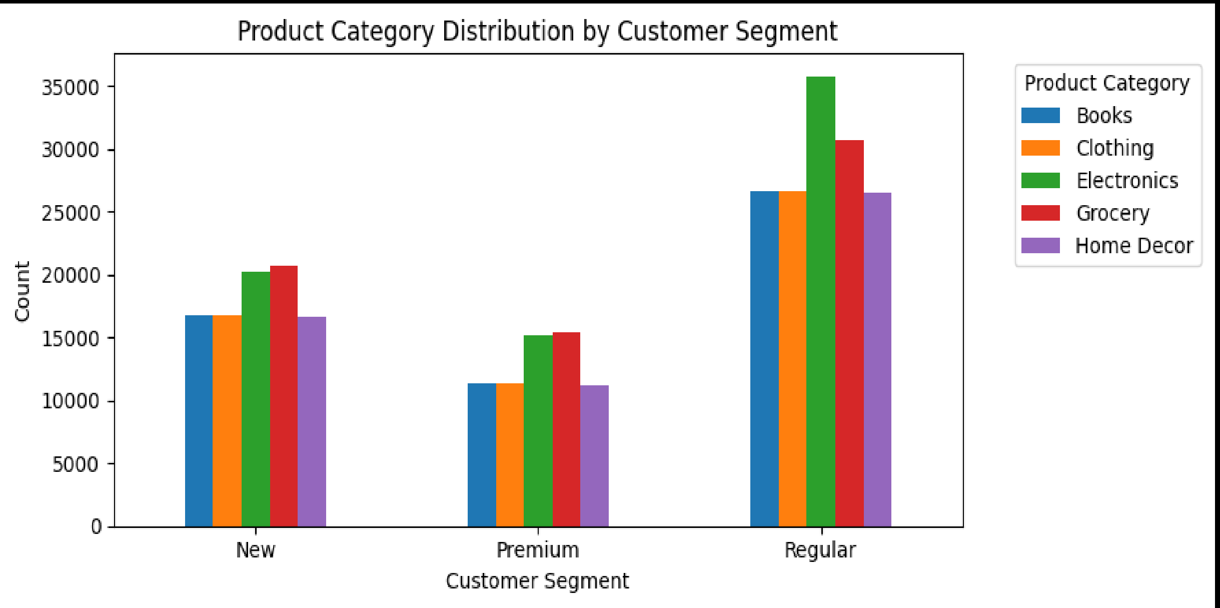
Insights from Correlation Heatmap:

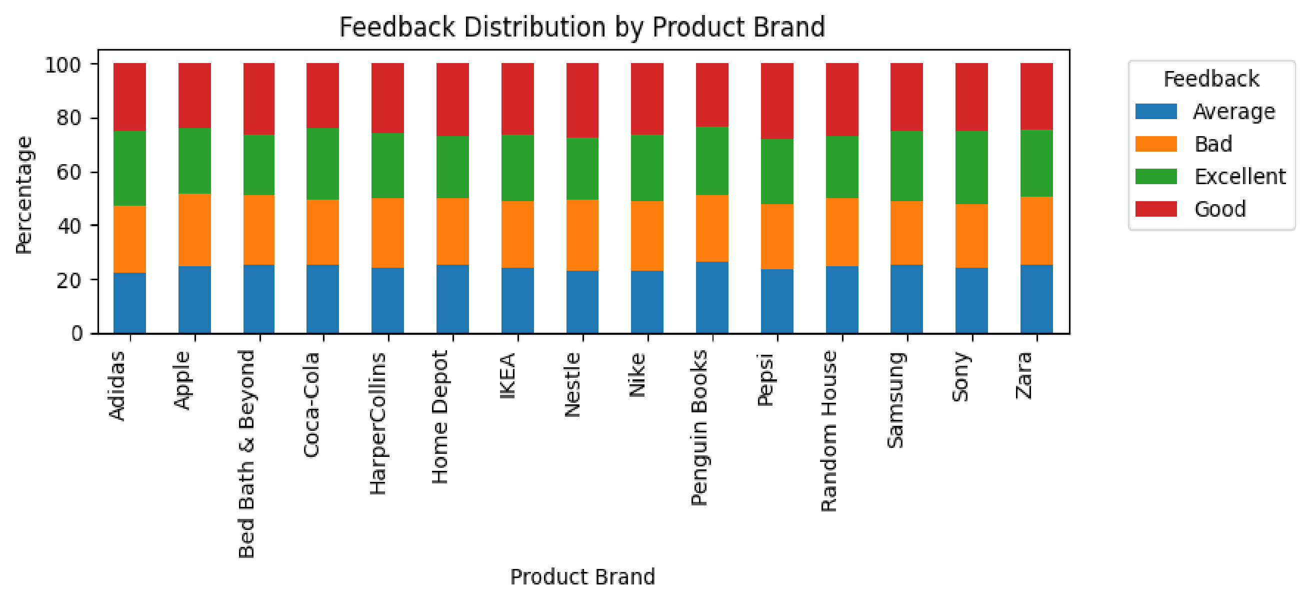
Spending-related features (Total_Purchases, Amount, Total_Amount) are strongly correlated, confirming that higher purchases directly drive revenue.

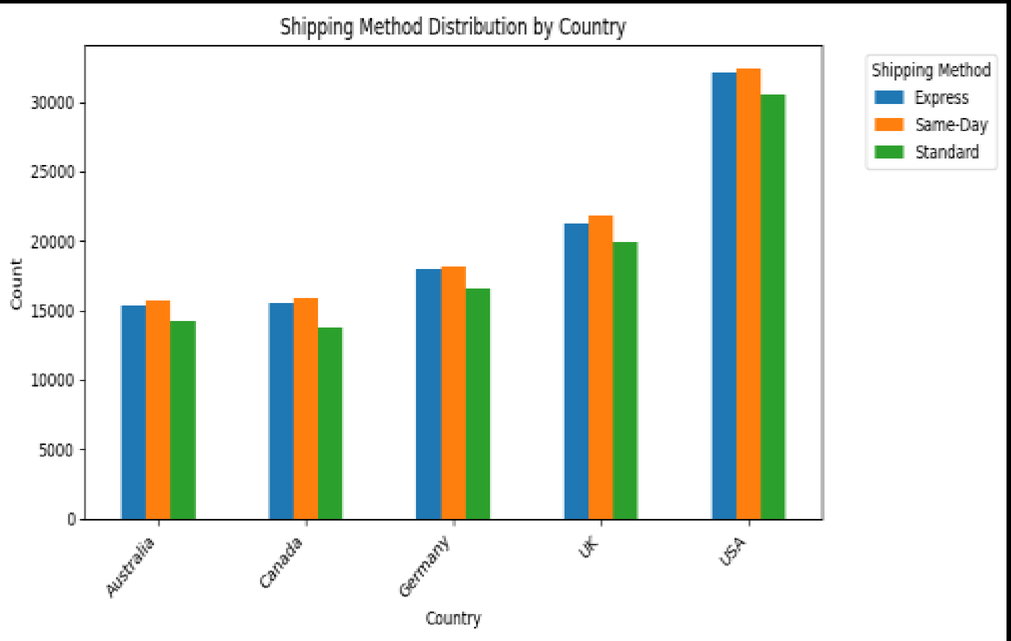
Customer Age shows only a weak positive relation with Ratings, while most demographic or ID fields have negligible correlations.

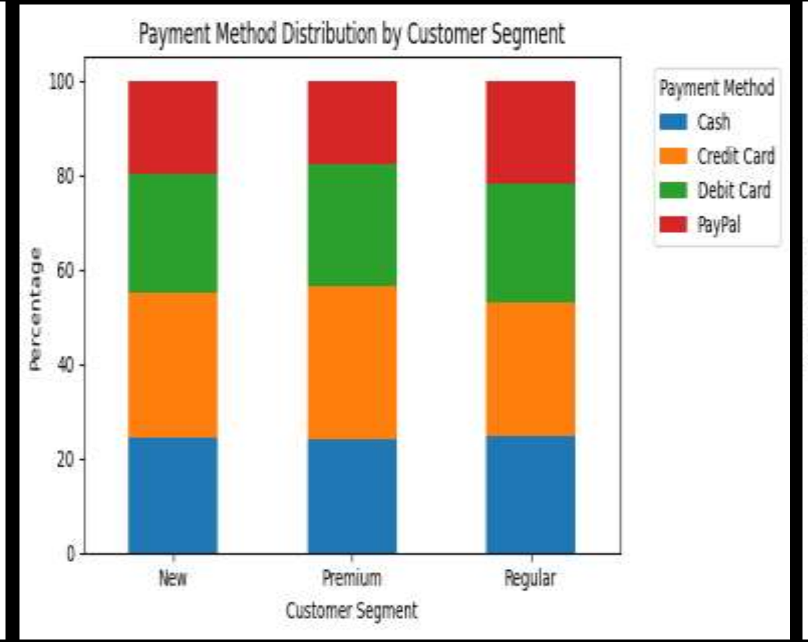
Overall, purchasing behavior drives the main numerical relationships, whereas satisfaction (ratings) and demographics play a smaller role.

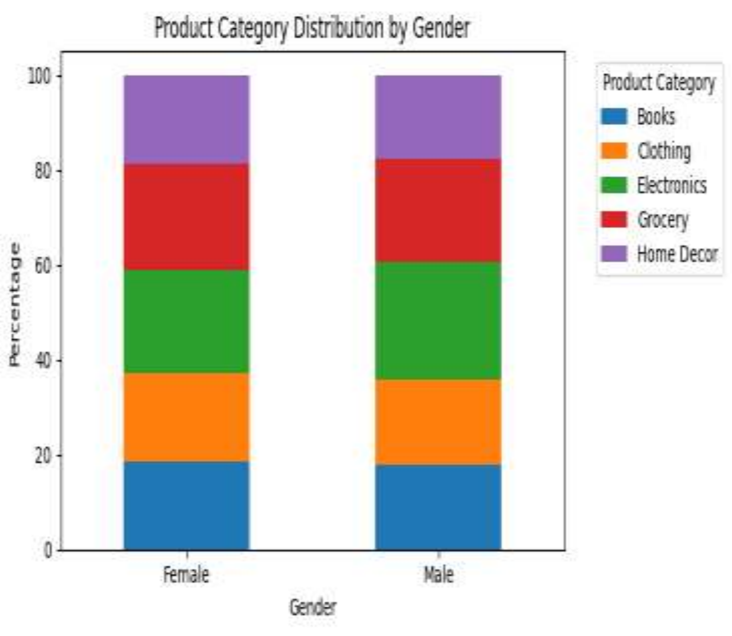
Bivariate Analysis: Categorical vs Categorical(Stacked,Clustered Barchart)











SUMMARY

PRODUCT CATEGORY x CUSTOMER SEGMENT

- This chart shows the count of purchases for each product category within each customer segment (New, Premium, and Regular). We can observe that:
- The Regular customer segment has the highest number of purchases across all product categories, which is expected as this is likely the largest segment.
- The distribution of purchases across product categories seems to be relatively consistent within each customer segment, although the overall volume differs.
- Electronics and Grocery categories appear to be the most popular across all segments in terms of the number of purchases

SHIPPING METHOD x COUNTRY

- This chart illustrates the count of different shipping methods used in each country. We can see that:
- The USA has the highest overall volume of shipments, followed by the UK, Germany, Canada, and Australia. This likely reflects the overall customer base size in these countries.
- Within each country, the distribution of shipping methods (Express, Same-Day, and Standard) appears to be relatively similar, with no single shipping

  method being overwhelmingly preferred in any specific country.
- "Same-Day" and "Express" shipping seem to be slightly more popular than "Standard" in most countries.

SUMMARY CONTD…..

PAYMENT METHOD x CUSTOMER SEGMENT

- This chart shows the count of purchases for each product category within each customer segment (New, Premium, and Regular).
- We can observe that : The Regular customer segment has the highest number of purchases across all product categories, which is expected as this is likely the largest segment
- .The distribution of purchases across product categories seems to be relatively consistent within each customer segment, although the overall volume differs . Electronics and Grocery categories appear to be the most popular across all segments in terms of the number of purchases

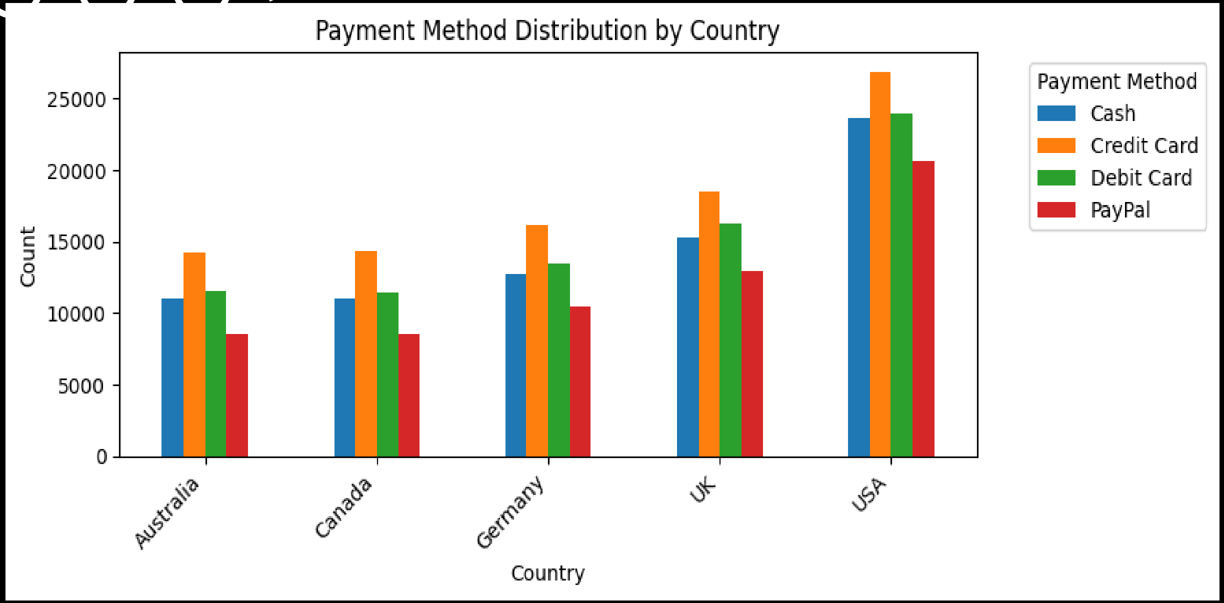
PRODUCT CATEGORY x GENDER

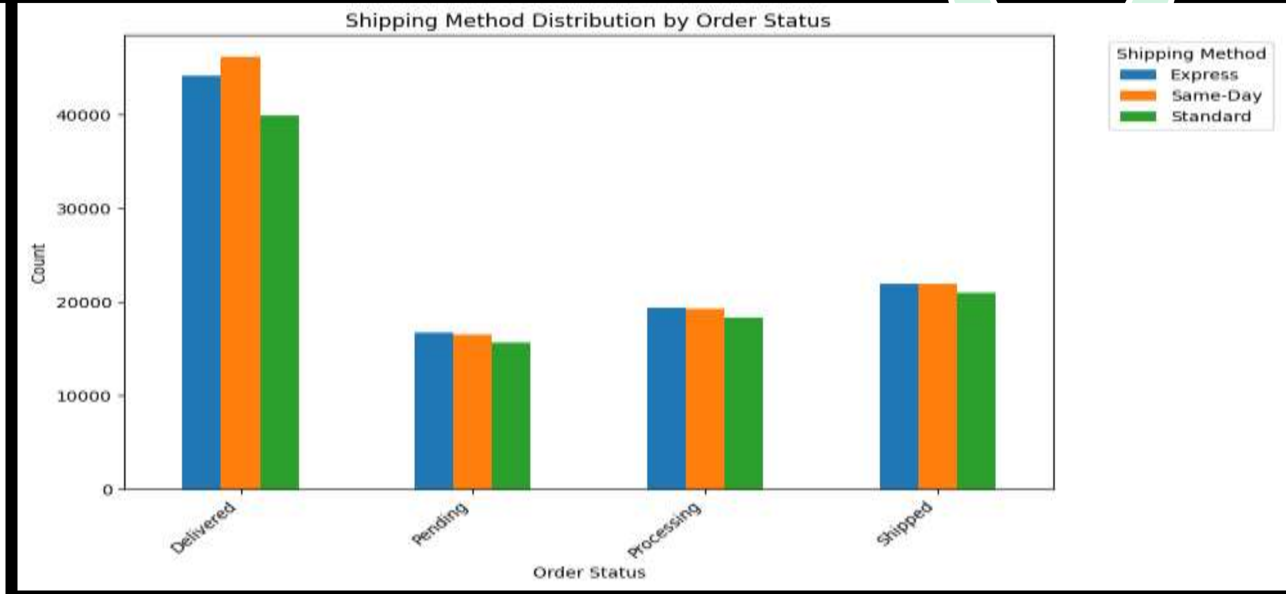
- This chart shows the distribution of product categories purchased by gender. We can see that the distribution of product categories appears to be quite similar between male and female customers.
- Both genders show a similar proportion of purchases across all product categories (Books, Clothing, Electronics, Grocery, and Home Decor). This suggests that, based on this dataset, there are no strong gender-based preferences for specific product categories.

FEEDBACK x PRODUCT BRAND

- This chart shows the distribution of feedback across different product brands. By examining the stacked bars for each brand, we can see the proportion of "Average," "Bad," "Excellent," and "Good" feedback they received. This allows us to:
- Compare the feedback profiles of different brands. Identify brands that consistently receive more positive feedback ("Excellent" and "Good"). Identify brands that have a higher percentage of negative feedback ("Average" and "Bad").
- For example, we can see that for some brands, the green and red portions (Excellent and Good) are larger, indicating more positive sentiment. For other brands, the blue and orange portions (Average and Bad) might be more significant, suggesting areas for improvement.

Multivariate Analysis

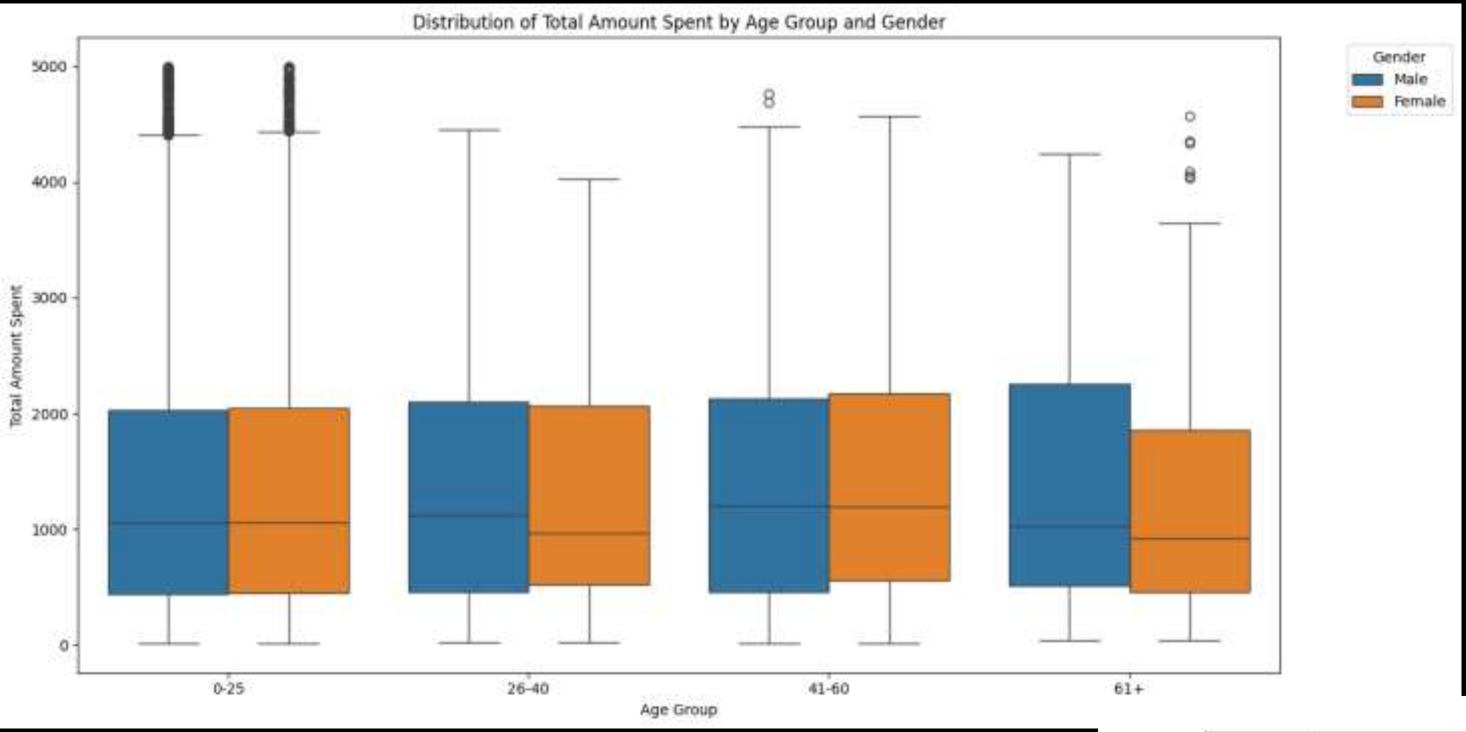




- This chart illustrates the count of different payment methods used in each country. We can see that:
- The USA has the highest overall number of transactions for each payment method, followed by the UK, Germany, Canada, and Australia.
- Within each country, Credit Card appears to be the most frequently used payment method, followed by Cash and Debit Card.
- PayPal seems to be the least used payment method across all countries compared to the other options.

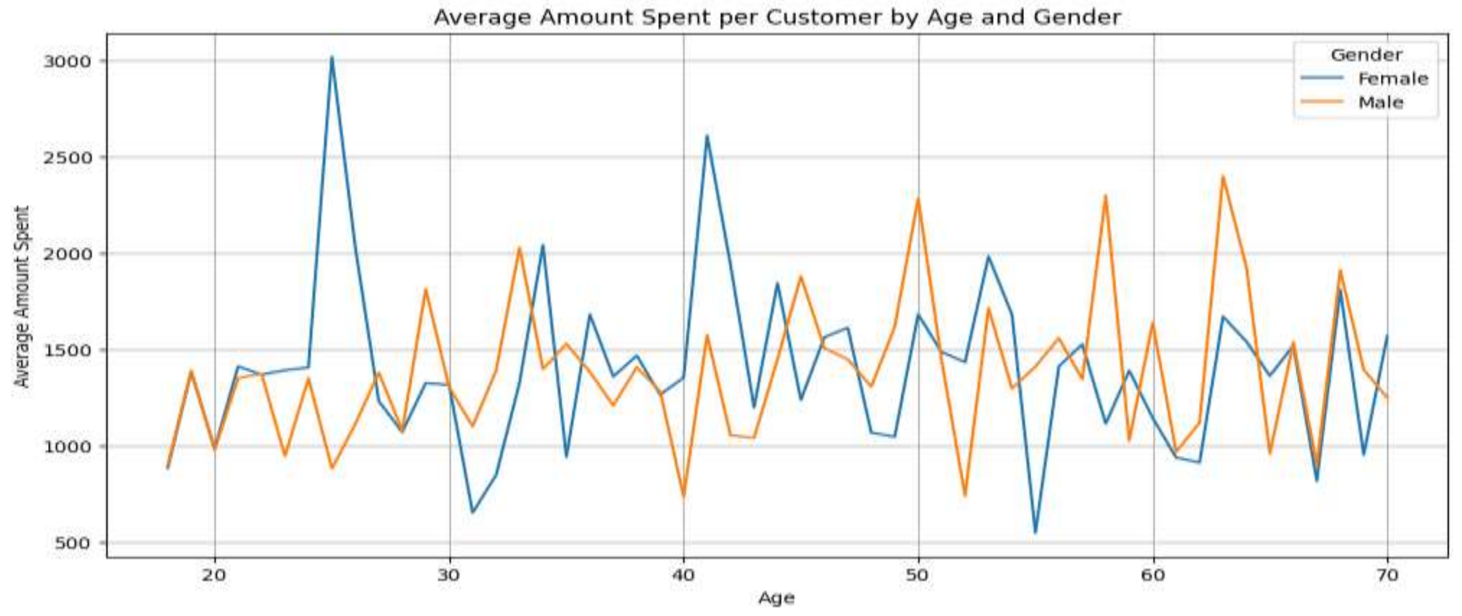
- This chart illustrates the distribution of shipping methods for different order statuses. We can see that:
- For Delivered orders, "Same-Day" shipping is slightly more frequent than "Express" and "Standard" shipping. This might indicate that customers who choose faster shipping methods are more likely to have their orders delivered successfully.
- For Pending, Processing, and Shipped orders, the distribution of shipping methods appears to be relatively similar, with no single shipping method dominating these statuses.

Multivariate Analysis

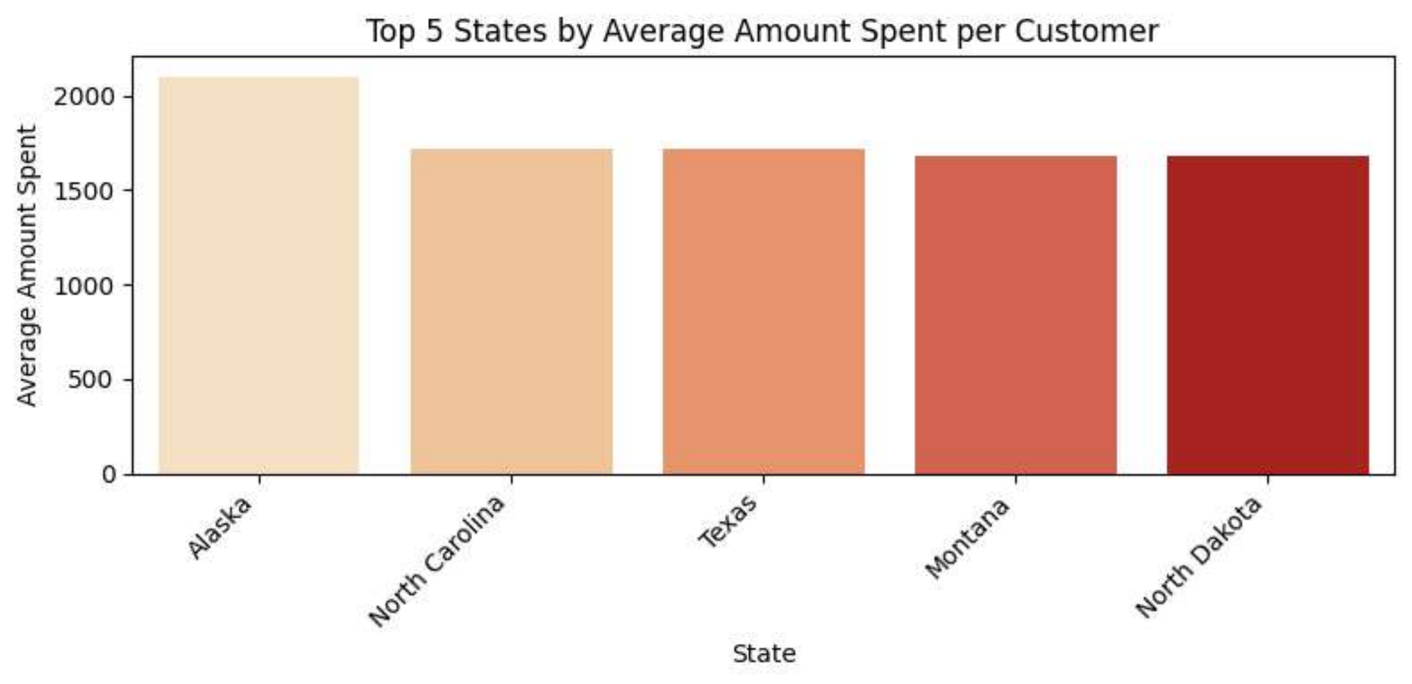


- Here are some observations from the plot:
- The median total amount spent (the line within the box) appears to be relatively similar across most age groups for both genders, except for the 61+ age group where the median for females seems lower than for males.
- The spread of spending (the size of the box) also seems somewhat consistent across age groups.
- The whiskers (lines extending from the box) show the range of typical spending within each group.
- There are several outliers (individual points) in all age groups, especially in the higher spending range, indicating some customers spend significantly more than the typical amount for their age and gender group.
- The 61+ age group for females shows a tighter distribution of spending with fewer high-value outliers compared to other groups.
- This visualization helps to understand the spending patterns within different age and gender segments and identify where the high-value transactions are occurring.

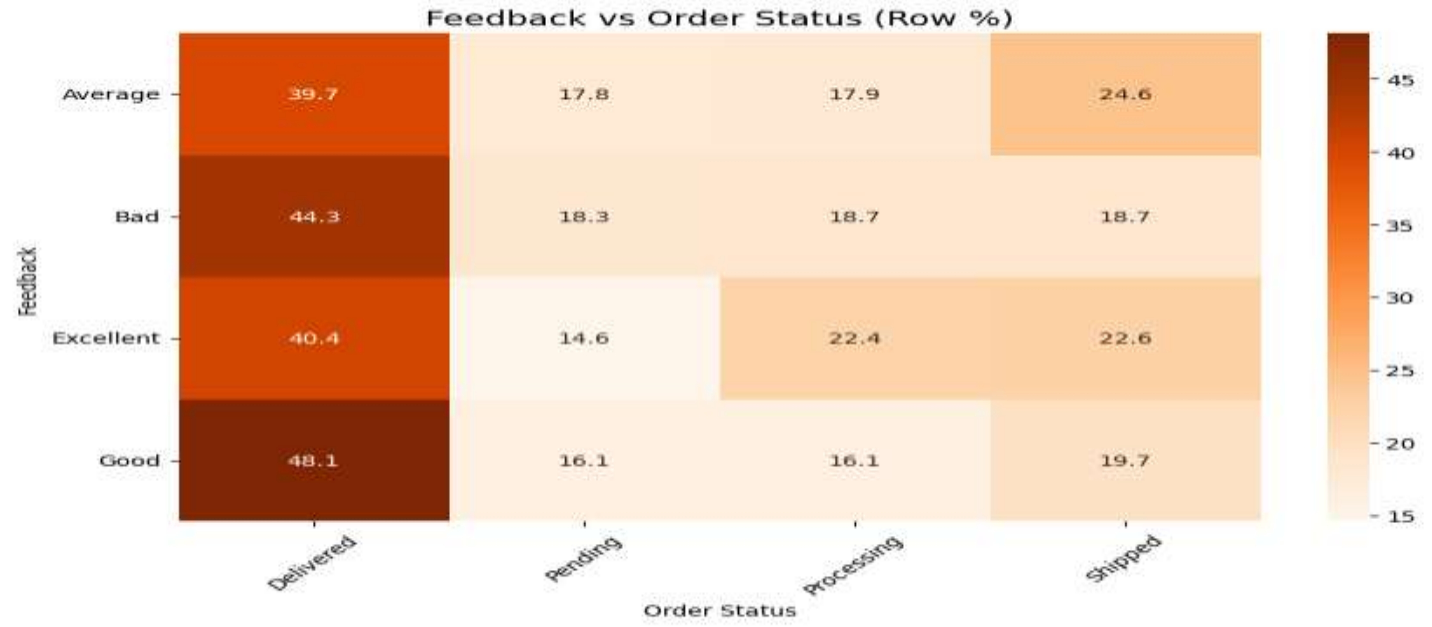
The average amount spent fluctuates across different age groups for both genders.
There doesn't appear to be a clear linear trend of increasing or decreasing spending with age for either gender.
In some age ranges, there are noticeable differences in average spending between males and females, while in others, the averages are quite similar.
There are some spikes and dips in the average spending lines, suggesting that certain age groups within each gender might have significantly higher or lower average spending.



Multivariate Analysis

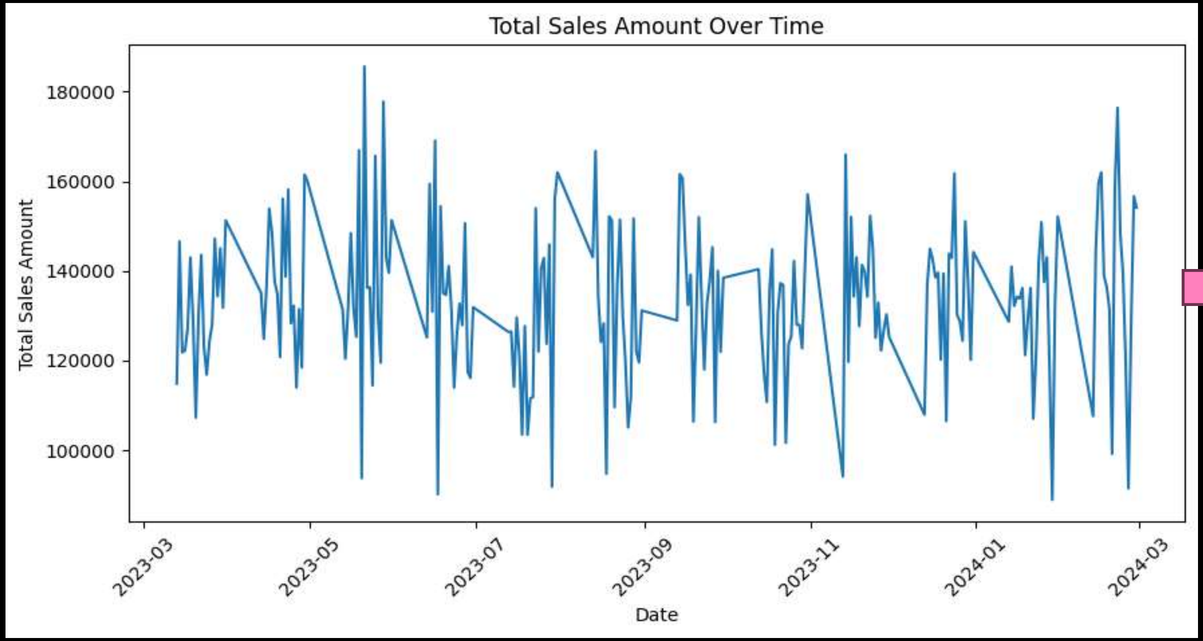


> Based on the bar plot, the top 5 states with the highest average amount spent per customer are: Alaska North Carolina Texas Montana North Dakota Alaska stands out with a significantly higher average amount spent compared to the other states in the top 5. The other four states have relatively similar average spending amounts.

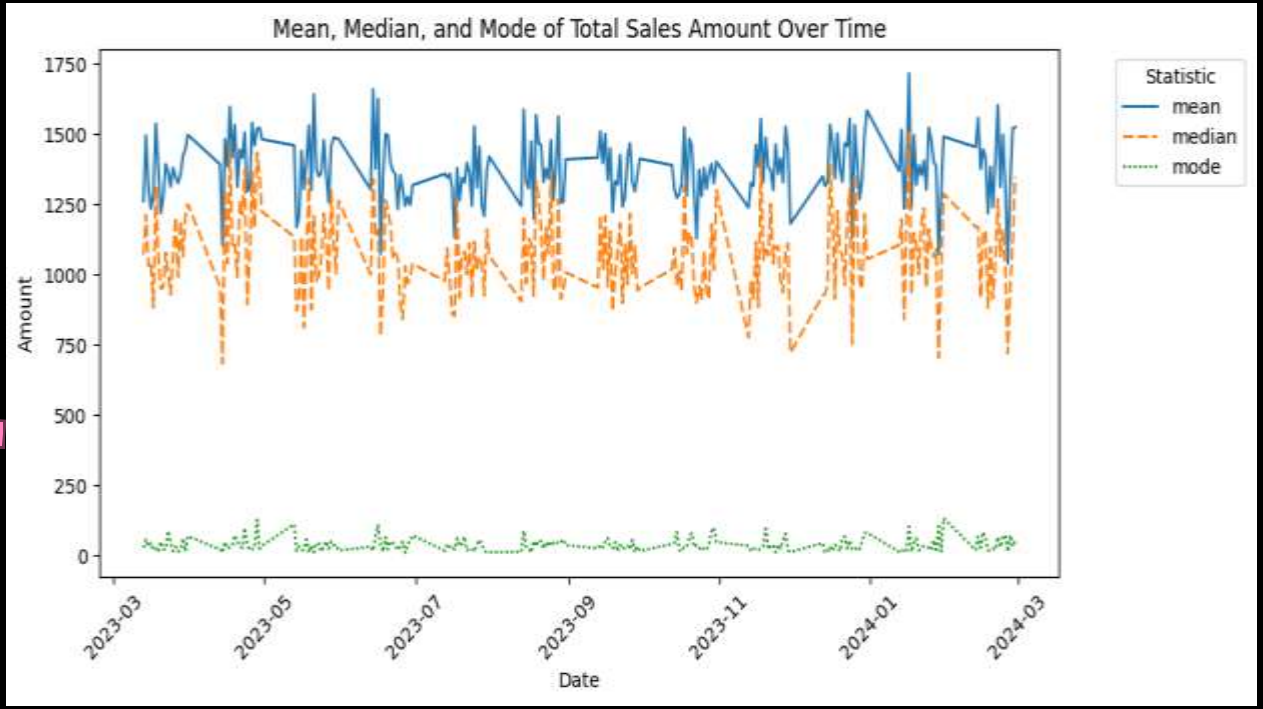


> Links between customer feedback and order completion/delivery status.
Example expectations:
Shipped/Delivered → More Excellent/Good feedback.
Processing/Delayed/Cancelled → More Average/Bad feedback.

Time Series Analysis

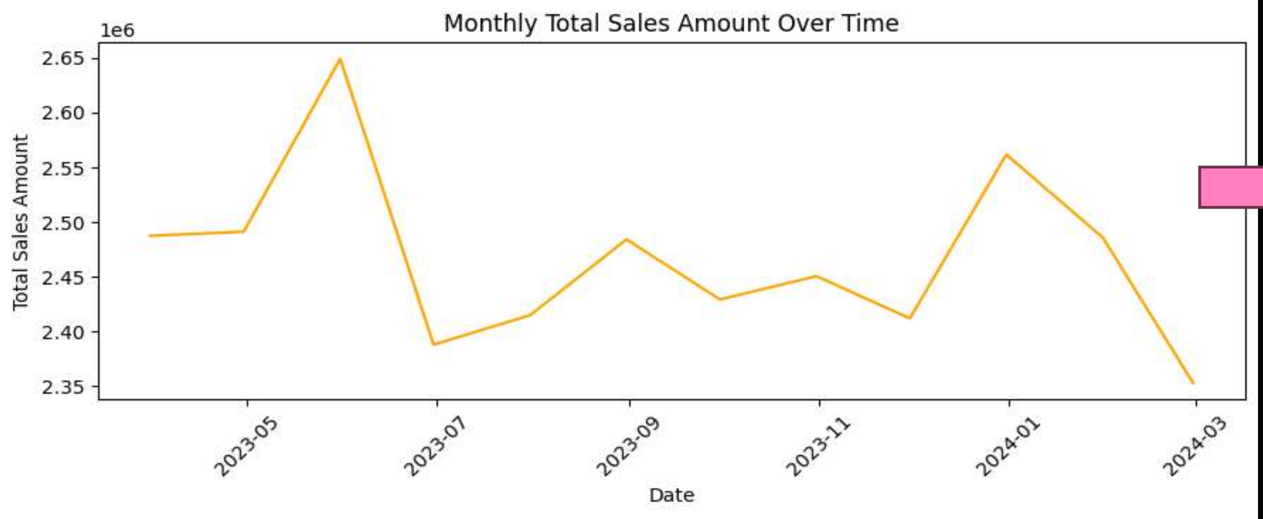


The total sales amount over time shows a generally consistent range of values, with no strong upward or downward trend within the plotted period.There are noticeable fluctuations and spikes in the sales data, suggesting potential daily or weekly patterns or events that influence sales volume.Identifying the specific dates of the peaks and valleys could help understand the underlying reasons for these variations, such as promotions, holidays, or other external factors.

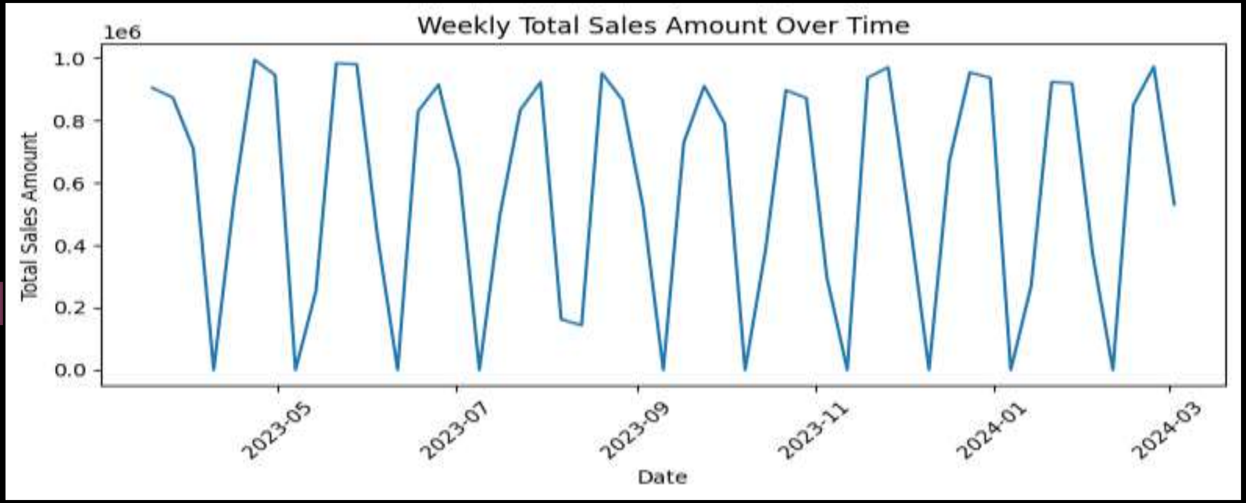


The mean and median lines generally follow similar patterns of fluctuations, indicating that the distribution of total sales amount is somewhat symmetrical most of the time.The mode line is consistently much lower than the mean and median. This suggests that the most frequent total sales amount is relatively low compared to the average and middle values. This could be due to a large number of smaller transactions.The gaps between the mean and median lines, and the large gap between the mode and both the mean/median, highlight that the distribution of total sales amount is likely skewed, with a tail towards higher values (as also suggested by the outliers in 'Total_Amount' we found earlier).

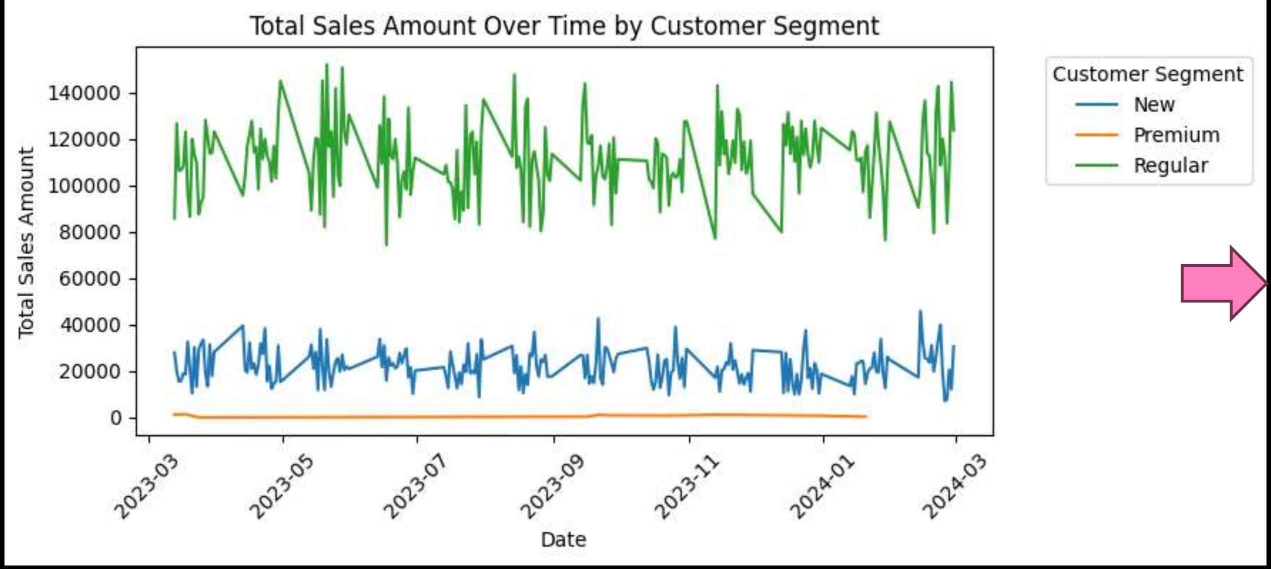
Time Series Analysis



The plot shows fluctuations in monthly total sales amount throughout the observed period.There is a noticeable peak in sales around May 2023, followed by a dip in June and July 2023.Sales seem to pick up again towards the end of 2023, with another peak around December 2023 or January 2024, which could be influenced by holiday seasons.There is a decrease in sales in February 2024 compared to the previous months.

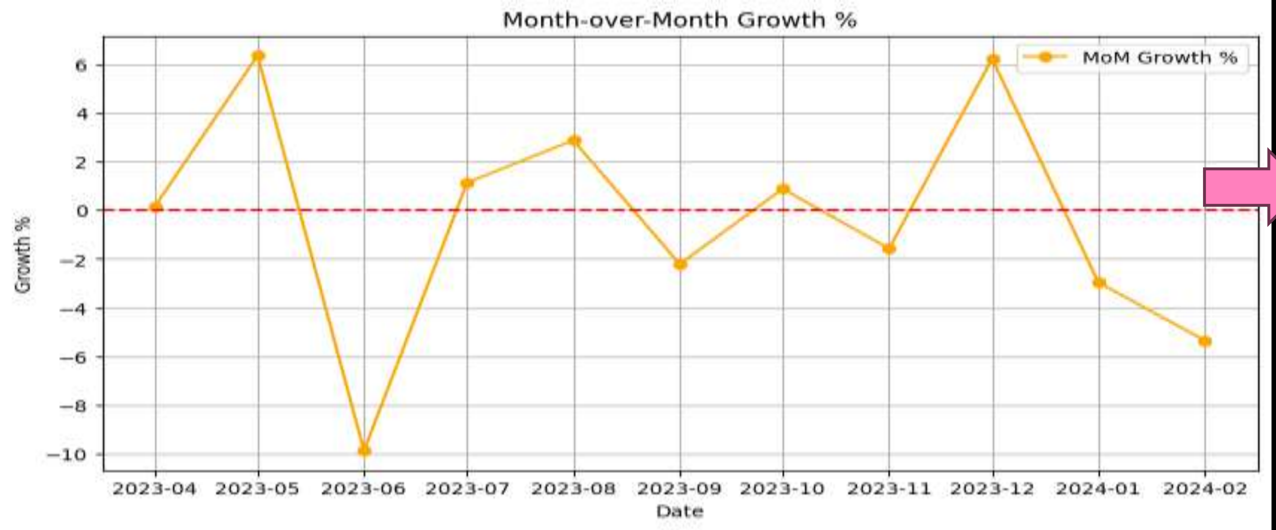


The weekly sales trend shows a consistent pattern of high and low sales within each week, indicating a strong weekly seasonality.

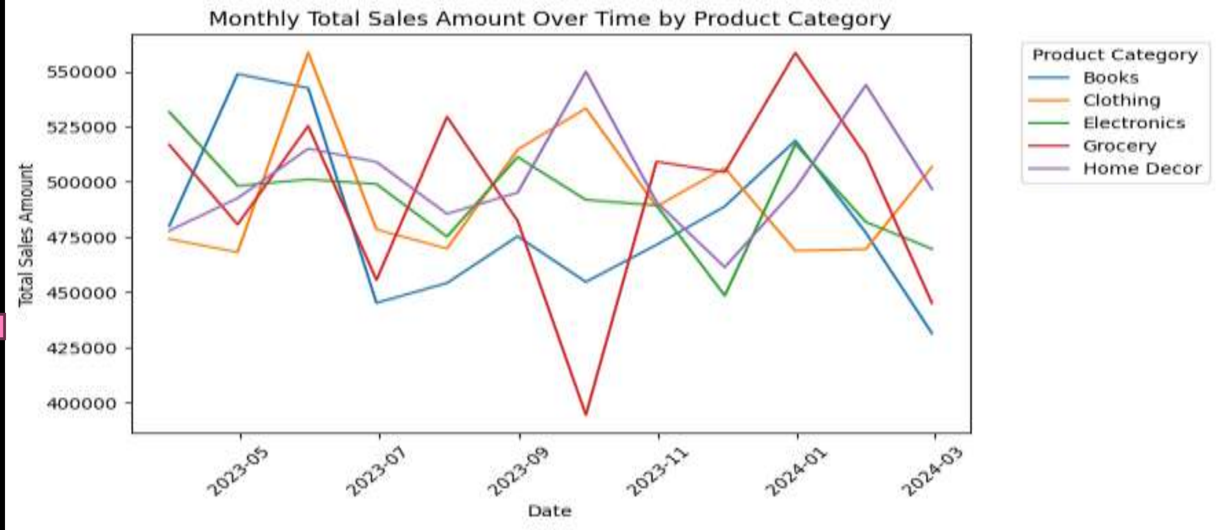


Total Sales Amount Over Time by Customer Segment: The "Regular" customer segment consistently generates the highest total sales amount over time, followed by the "New" customer segment. The "Premium" customer segment has significantly lower total sales compared to the other two segments. This suggests that the "Regular" segment is the most valuable in terms of overall revenue.

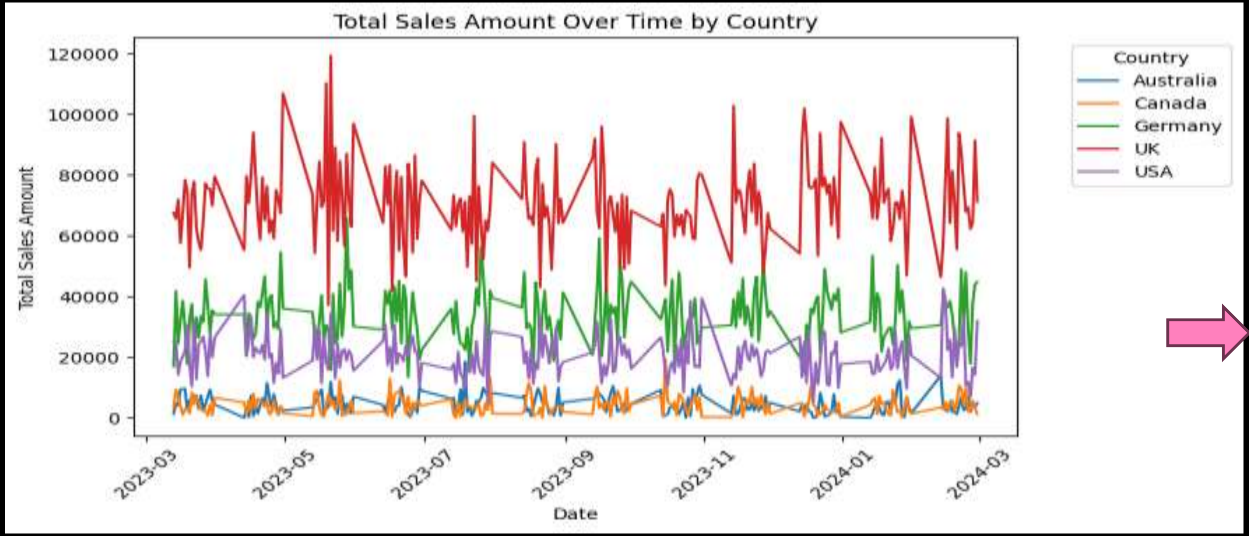
Time Series Analysis



Above 0 line → sales grew compared to the previous month.Below 0 line → sales dropped.Sharp spikes → promotions, discounts, or special campaigns likely boosted sales.Sharp dips → off-season demand or operational issues. This plot helps you spot short-term momentum and identify months that underperformed

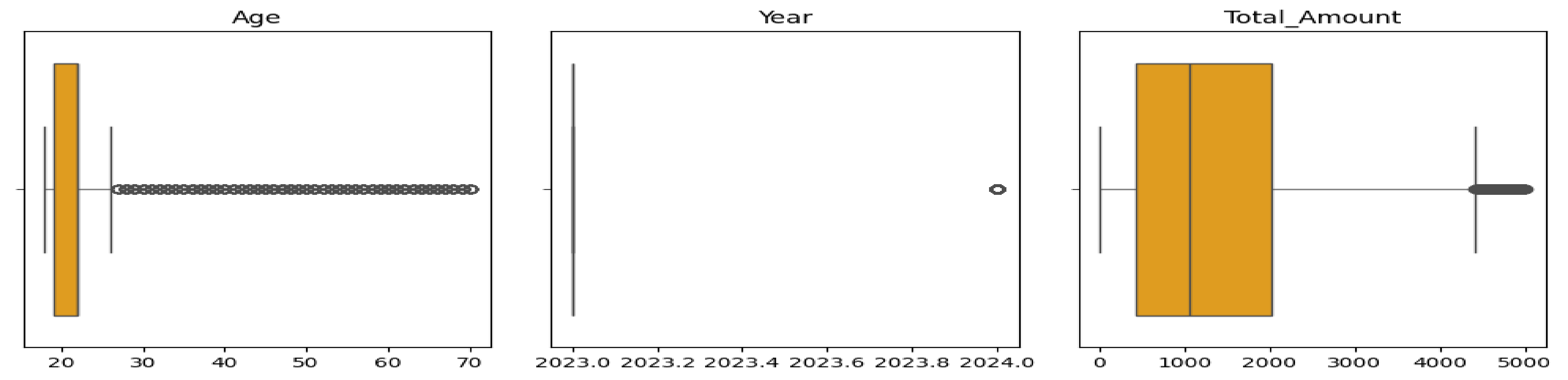


Monthly sales for all product categories fluctuate together, suggesting overall market trends or seasonality affect all categories similarly.



The USA consistently shows the highest total sales amount over time compared to other countries.The UK also contributes significantly to the total sales, generally ranking second after the USA.Germany, Canada, and Australia have lower total sales amounts compared to the USA and UK, and their sales trends appear to be relatively similar to each other.All countries exhibit similar patterns of fluctuations in sales over time, suggesting that overall trends or seasonality affect sales across all regions.

OUTLIERS ANALYSIS



Age:
The box plot for 'Age' clearly shows a number of data points below the lower whisker. These points represent the younger ages that were identified as outliers by the IQR method. The majority of the data is concentrated in a younger age range, as indicated by the box.

Year:
The box plot for 'Year' shows a tight distribution around the year 2023, with a few data points extending to 2024. These 2024 data points are identified as outliers, likely because they fall outside the main cluster of data in 2023.

Total_Amount:
The box plot for 'Total_Amount' shows several data points extending far above the upper whisker. These points represent the transactions with significantly higher amounts that were identified as outliers. The box itself is relatively compact compared to the spread of these outliers, highlighting their extreme values.

It allows US to compare the feedback profiles of different brands and identify brands that consistently receive more positive or negative feedback. For example, some brands might have a higher percentage of "Excellent" feedback, while others might have a notable percentage of "Bad" or "Average" reviews.

SUMMARY AND RECOMMENDATIONS

SUMMARY:

The analysis shows that regular customers are the primary source of sales across all product categories, especially electronics and groceries.

- Spending is generally consistent among younger customers with occasional high-value outliers, and most feedback is positive, though ratings trend low to mid-range, indicating potential dissatisfaction.
- The USA leads in transactions and sales, while Chicago is the top city and Alaska the highest-spending state.
- Debit and credit cards dominate payments, with PayPal and cash less prevalent, and shipping is most often done via same-day or express methods.
- Sales are seasonal, spiking in April, May, and December, with similar patterns across all regions and product categories.

RECOMMENDATION:

- Focus marketing and loyalty initiatives on retaining and up-selling regular customers, as they are the backbone of revenue.
- Address potential causes of lower customer ratings through targeted improvements in customer service, product quality, and feedback follow-up.
- Capitalize on peak sales months with special promotions—especially in April, May, and December—to maximize revenue.
- Expand successful shipping and payment options to more regions, ensuring same-day/express shipping and convenient digital payment methods remain widely available.
- Use advanced analytics to identify and engage high-value outliers (potential VIPs) for premium loyalty programs or personalized offers.
- Monitor states and cities with high average spend (e.g., Alaska, Chicago) for localized campaigns and inventory planning.
- Regularly review product and brand feedback to improve offerings and focus on brands/products that consistently receive excellent ratings.

THANK YOU