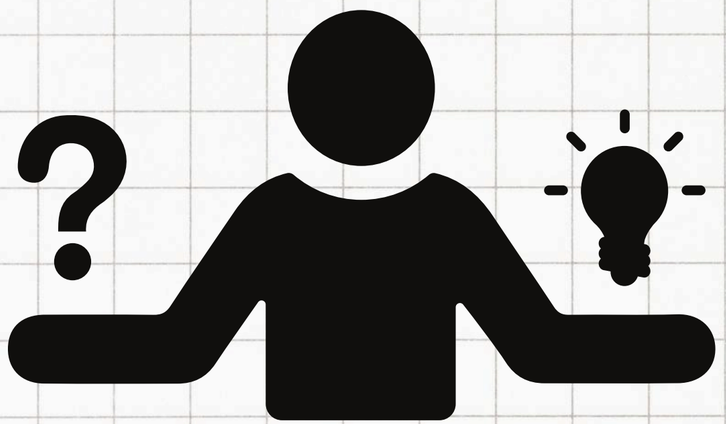
THE RETAIL CHALLENGE #1

DATA MINDS

Optimizing Customer Experience and Sales at OmniMart Retailers

Anshika Pandey
Dipanjan Halder
Ekadashi Sardar
Nilanjana Saren

The Retail Challenge



Problem Statement:

OmniMart Retailers is a multinational company with a vast database of customer transactions and feedback. The company aims to:

- Gain a deeper understanding of its customer base
- Improve sales and increase customer retention
- Optimize marketing strategies

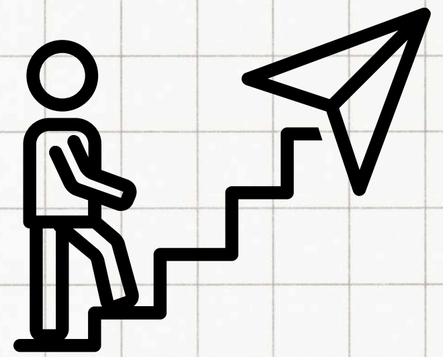
Objectives of the Analysis

Our analysis aims to:

- Understand customer behavior through purchase patterns and feedback
- Identify high-value customer segments to boost retention and loyalty
- Evaluate sales performance across products, categories, and time periods
- Optimize marketing strategies based on data-driven insights
- Discover growth opportunities by uncovering hidden trends


1.
PREPARATION
& SETUP
2.
INITIAL
RECONNAI-
SSANCE
3.
UNCOVERING
INSIGHTS
4.
SPECIAL
OPERATIONS
5.
THE FINAL
BRIEFING
INSIGHTS

Phase 1: Preparation & Setup



Project Initialization

- Import essential Python libraries (pandas, numpy, matplotlib, seaborn)
- Data Acquisition
- Load the challenge dataset into the notebook

Data Scrutiny (The Cleanup Mission)

- Detect and handle missing values
- Ensure dataset is clean, consistent, and reliable

Data Cleaning: Handling Missing Values

Dataset (including Missing values)

```
retail.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305926 entries, 0 to 305925
Data columns (total 30 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Transaction_ID    305588 non-null  float64
 1   Customer_ID       305615 non-null  float64
 2   Name              305541 non-null  object
 3   Email             305575 non-null  object
 4   Phone             305559 non-null  float64
 5   Address           305610 non-null  object
 6   City              305672 non-null  object
 7   State             305642 non-null  object
 8   Zipcode           305579 non-null  float64
 9   Country           305652 non-null  object
 10  Age               305753 non-null  float64
 11  Gender            305603 non-null  object
 12  Income            305629 non-null  object
 13  Customer_Segment  305709 non-null  object
 14  Date              185262 non-null  datetime64[ns]
 15  Year              305571 non-null  float64
 16  Month             305652 non-null  object
 17  Time              305571 non-null  object
 18  Total_Purchases   305557 non-null  float64
 19  Amount            305567 non-null  float64
 20  Total_Amount      305568 non-null  float64
 21  Product_Category  305635 non-null  object
 22  Product_Brand     305637 non-null  object
 23  Product_Type      305926 non-null  object
 24  Feedback          305742 non-null  object
 25  Shipping_Method   305582 non-null  object
 26  Payment_Method    305624 non-null  object
 27  Order_Status      305691 non-null  object
 28  Ratings           305742 non-null  float64
 29  products          305926 non-null  object
dtypes: datetime64[ns](1), float64(10), object(19)
memory usage: 70.0+ MB
```

List of missing values

```
Columns with missing values:

                  missing_count  missing_percent
Name                        385             0.13
Total_Purchases             369             0.12
Phone                       367             0.12
Date                        364             0.12
Amount                      359             0.12
Total_Amount                358             0.12
Year                        355             0.12
Time                        355             0.12
Email                       351             0.11
Zipcode                     347             0.11
Shipping_Method             344             0.11
Transaction_ID              338             0.11
Gender                      323             0.11
Address                     316             0.10
Customer_ID                 311             0.10
Payment_Method              302             0.10
Income                      297             0.10
Product_Category            291             0.10
Product_Brand               289             0.09
State                       284             0.09
Country                     274             0.09
Month                       274             0.09
City                        254             0.08
Order_Status                235             0.08
Customer_Segment            217             0.07
Feedback                    184             0.06
Ratings                     184             0.06
Age                         173             0.06

Aggressive cleaned shape (drop rows with ANY missing): (297710, 30)
Saved -> retail_cleaned_dropna.csv

Safer cleaned shape (drop cols >40.0% missing, then drop rows): (297710, 30)
Columns kept: 30
Saved -> retail_cleaned_threshold.csv
```

Final dataset

```
retail1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297710 entries, 0 to 297709
Data columns (total 30 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Transaction_ID    297710 non-null  float64
 1   Customer_ID       297710 non-null  float64
 2   Name              297710 non-null  object
 3   Email             297710 non-null  object
 4   Phone             297710 non-null  float64
 5   Address           297710 non-null  object
 6   City              297710 non-null  object
 7   State             297710 non-null  object
 8   Zipcode           297710 non-null  float64
 9   Country           297710 non-null  object
 10  Age               297710 non-null  float64
 11  Gender            297710 non-null  object
 12  Income            297710 non-null  object
 13  Customer_Segment  297710 non-null  object
 14  Date              297710 non-null  object
 15  Year              297710 non-null  float64
 16  Month             297710 non-null  object
 17  Time              297710 non-null  object
 18  Total_Purchases   297710 non-null  float64
 19  Amount            297710 non-null  float64
 20  Total_Amount      297710 non-null  float64
 21  Product_Category  297710 non-null  object
 22  Product_Brand     297710 non-null  object
 23  Product_Type      297710 non-null  object
 24  Feedback          297710 non-null  object
 25  Shipping_Method   297710 non-null  object
 26  Payment_Method    297710 non-null  object
 27  Order_Status      297710 non-null  object
 28  Ratings           297710 non-null  float64
 29  products          297710 non-null  object
dtypes: float64(10), object(20)
memory usage: 68.1+ MB
```

Phase 2: Initial Reconnaissance

- Initial Survey
- Explore the dataset structure (.info())
- Generate a statistical summary (.describe())
- Check dataset size (.shape)
- Preview sample rows (.head())

Data Classification

- Identify Numerical variables
- Identify Categorical variables
- Identify Temporal variables

Dataset Exploration & Classification

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297710 entries, 0 to 297709
Data columns (total 30 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Transaction_ID    297710 non-null  float64
 1   Customer_ID       297710 non-null  float64
 2   Name              297710 non-null  object
 3   Email             297710 non-null  object
 4   Phone             297710 non-null  float64
 5   Address           297710 non-null  object
 6   City              297710 non-null  object
 7   State             297710 non-null  object
 8   Zipcode           297710 non-null  float64
 9   Country           297710 non-null  object
 10  Age               297710 non-null  float64
 11  Gender            297710 non-null  object
 12  Income            297710 non-null  object
 13  Customer_Segment  297710 non-null  object
 14  Date              297710 non-null  object
 15  Year              297710 non-null  float64
 16  Month             297710 non-null  object
 17  Time              297710 non-null  object
 18  Total_Purchases   297710 non-null  float64
 19  Amount            297710 non-null  float64
 20  Total_Amount      297710 non-null  float64
 21  Product_Category  297710 non-null  object
 22  Product_Brand     297710 non-null  object
 23  Product_Type      297710 non-null  object
 24  Feedback          297710 non-null  object
 25  Shipping_Method   297710 non-null  object
 26  Payment_Method    297710 non-null  object
 27  Order_Status      297710 non-null  object
 28  Ratings           297710 non-null  float64
 29  products          297710 non-null  object
dtypes: float64(10), object(20)
memory usage: 68.1+ MB
```

```
df.shape

(297710, 30)
```

```
df.describe()
```

	Transaction_ID	Customer_ID	Phone	Zipcode	Age	Year	Total_Purchases	Amount	Total_Amount	Ratings
count	2.977100e+05	297710.000000	2.977100e+05	297710.000000	297710.000000	297710.000000	297710.000000	297710.000000	297710.000000	297710.000000
mean	5.494109e+06	55006.249172	5.500752e+09	50287.505858	35.679366	2023.165178	5.359299	255.155768	1367.436824	3.176454
std	2.596357e+06	26008.779585	2.596255e+09	28972.863324	15.073687	0.371341	2.868353	141.411092	1128.671838	1.320273
min	1.000007e+06	10000.000000	1.000049e+09	501.000000	18.000000	2023.000000	1.000000	10.000219	10.003750	1.000000
25%	3.245922e+06	32464.250000	3.252894e+09	25409.250000	22.000000	2023.000000	3.000000	132.820047	438.743200	2.000000
50%	5.496080e+06	55009.000000	5.505497e+09	50582.000000	33.000000	2023.000000	5.000000	255.471769	1041.264442	3.000000
75%	7.738767e+06	77508.000000	7.749841e+09	75246.000000	46.000000	2023.000000	8.000000	377.711001	2028.618105	4.000000
max	9.999995e+06	99999.000000	9.999996e+09	99949.000000	70.000000	2024.000000	10.000000	499.997911	4999.625796	5.000000

=== DATA CLASSIFICATION ===

- **Numerical columns: ['Transaction_ID', 'Customer_ID', 'Phone', 'Zipcode', 'Age', 'Year', 'Total_Purchases', 'Amount', 'Total_Amount', 'Ratings']**
- **Categorical columns: ['Name', 'Email', 'Address', 'City', 'State', 'Country', 'Gender', 'Income', 'Customer_Segment', 'Date', 'Month', 'Time', 'Product_Category', 'Product_Brand', 'Product_Type', 'Feedback', 'Shipping_Method', 'Payment_Method', 'Order_Status', 'products']**
- **Temporal columns: ['Date', 'Year', 'Month', 'Time']**

Phase 3: Uncovering Insights (The Core Mission)

Univariate Analysis

- Numerical: Histograms, Box Plots, Density Plots → Understand distribution
- Categorical: Bar Charts, Count Plots → Frequency of categories

Bivariate Analysis

- Numerical vs. Numerical: Scatter Plots, Correlation Matrices → Relationships
- Numerical vs. Categorical: Box Plots → Compare distributions

Multivariate Analysis

- Explore 3+ variables together
- Use advanced visualizations (e.g., scatter plot with color groups)
- Identify deeper trends & patterns

Statistical Summary of Numerical Variables

1. Measure of Central Tendency:

```
-> Age:
   Mean   : 35.68
   Median : 33.00
   Mode   : [20.0]

-> Year:
   Mean   : 2023.17
   Median : 2023.00
   Mode   : [2023.0]

-> Amount:
   Mean   : 255.16
   Median : 255.47
   Mode   : [17.50500747, 32.57195455, 40.43017841]

-> Ratings:
   Mean   : 3.18
   Median : 3.00
   Mode   : [4.0]
```

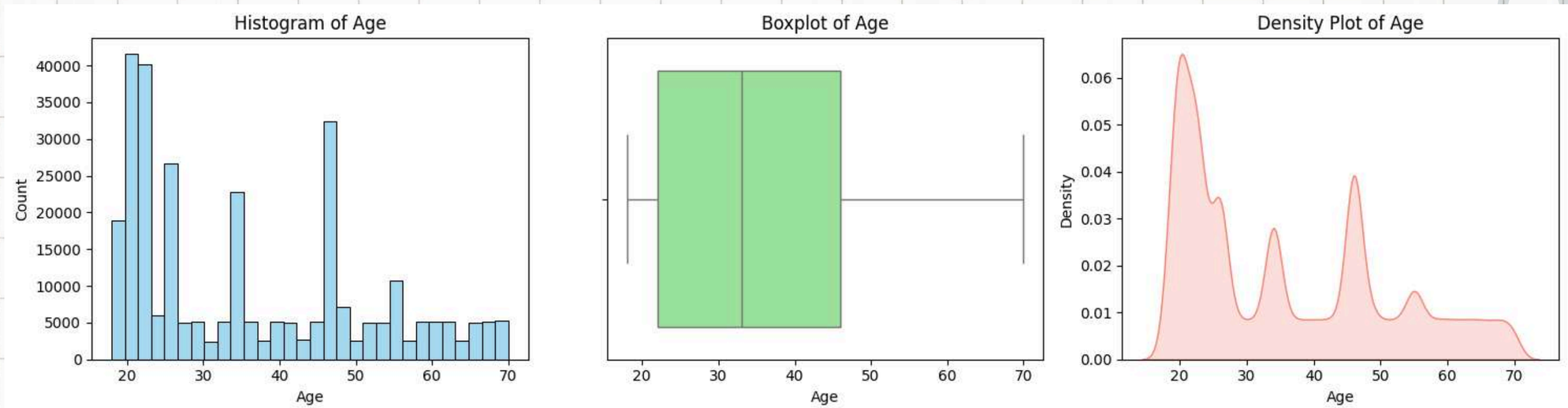
2. Measure of Dispersion:

```
-> Age:
   Range (Min - Max)          : 18.00 to 70.00
   Variance                   : 227.22
   Std Deviation              : 15.07
   IQR (Q3 - Q1)              : 24.00
   25th Percentile (Q1)       : 22.00
   50th Percentile (Median)   : 33.00
   75th Percentile (Q3)       : 46.00

-> Year:
   Range (Min - Max)          : 2023.00 to 2024.00
   Variance                   : 0.14
   Std Deviation              : 0.37
   IQR (Q3 - Q1)              : 0.00
   25th Percentile (Q1)       : 2023.00
   50th Percentile (Median)   : 2023.00
   75th Percentile (Q3)       : 2023.00

-> Amount:
   Range (Min - Max)          : 10.00 to 500.00
   Variance                   : 19997.10
   Std Deviation              : 141.41
   IQR (Q3 - Q1)              : 244.89
   25th Percentile (Q1)       : 132.82
   50th Percentile (Median)   : 255.47
   75th Percentile (Q3)       : 377.71

-> Ratings:
   Range (Min - Max)          : 1.00 to 5.00
   Variance                   : 1.74
   Std Deviation              : 1.32
   IQR (Q3 - Q1)              : 2.00
   25th Percentile (Q1)       : 2.00
   50th Percentile (Median)   : 3.00
   75th Percentile (Q3)       : 4.00
```

3. Distribution Analysis:

```
-> Age:
   Skewness : 0.6233 -> Positive Skew (Right Skew)
   Kurtosis : -0.8619 -> Platykurtic (Flat)

-> Year:
   Skewness : 1.8033 -> Positive Skew (Right Skew)
   Kurtosis : 1.2520 -> Leptokurtic (Peaked)

-> Amount:
   Skewness : -0.0020 -> Negative Skew (Left Skew)
   Kurtosis : -1.1989 -> Platykurtic (Flat)

-> Ratings:
   Skewness : -0.2558 -> Negative Skew (Left Skew)
   Kurtosis : -1.1603 -> Platykurtic (Flat)
```

Visual Analysis of the Numerical Variables (Age)

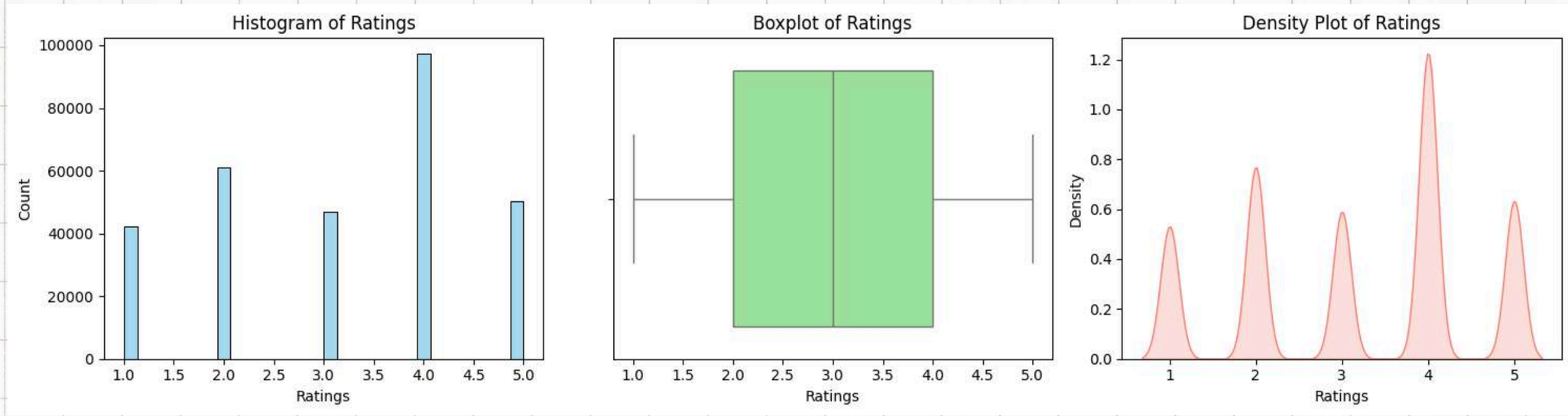



```
=== Age Summary ===
Mean   : 35.68
Median : 33.00
Min    : 18.00
Max    : 70.00
```

Insights from Age Analysis

- Majority of customers are young adults (18–35 years), as shown by histogram peaks.
- Median age = 33 suggests a relatively young customer base.
- Few older customers (50+), indicating a smaller senior customer segment.
- Distribution shows slight right skew, but overall spread is 18 to 70 years.

Visual Analysis of the Numerical Variables (Ratings)

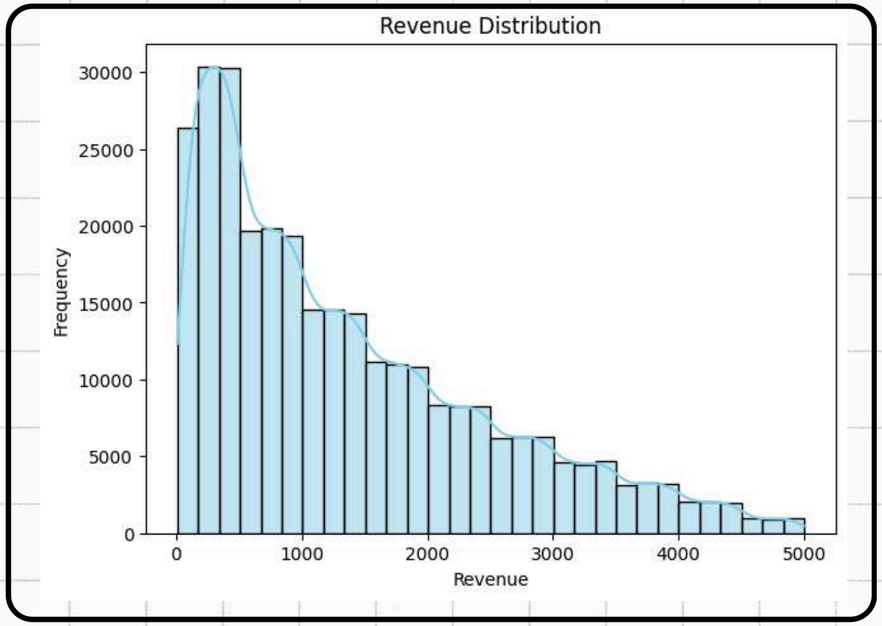


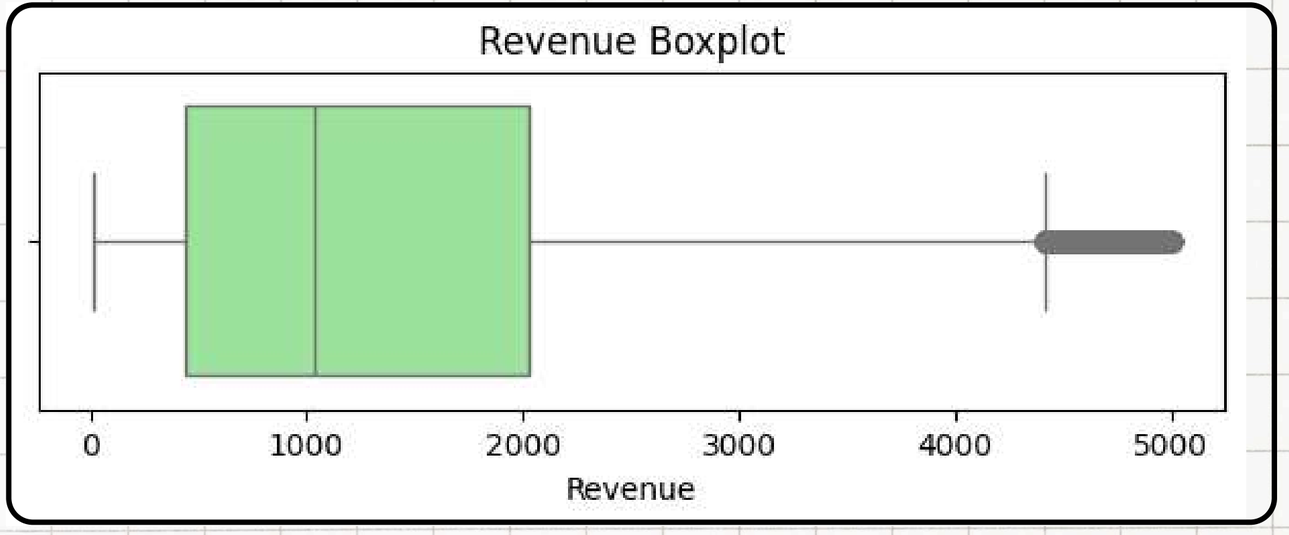

```
=== Ratings Summary ===
Mean   : 3.18
Median : 3.00
Min    : 1.00
Max    : 5.00
```

Insights from Ratings Analysis

- Ratings range from 1 to 5, with an average of 3.18.
- Median rating = 3, showing customers are generally neutral to slightly positive.
- Histogram and density plots reveal spikes at whole number ratings (1, 2, 3, 4, 5) → customers prefer giving exact integer scores.
- Majority of ratings cluster around 4, suggesting overall satisfactory customer experience.

Visual Analysis of the Numerical Variables (Revenue)



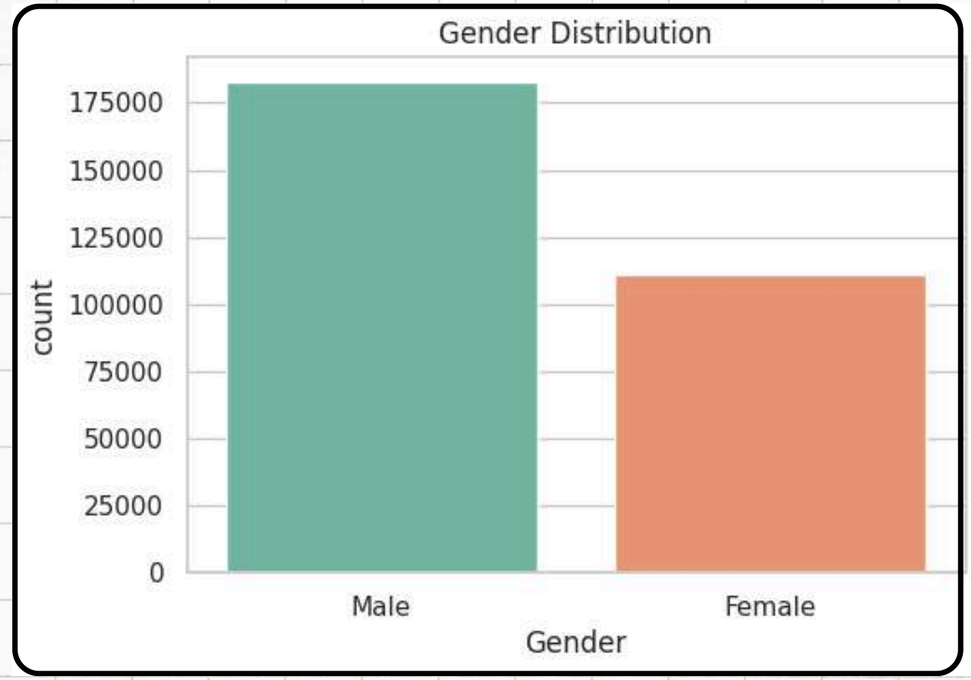


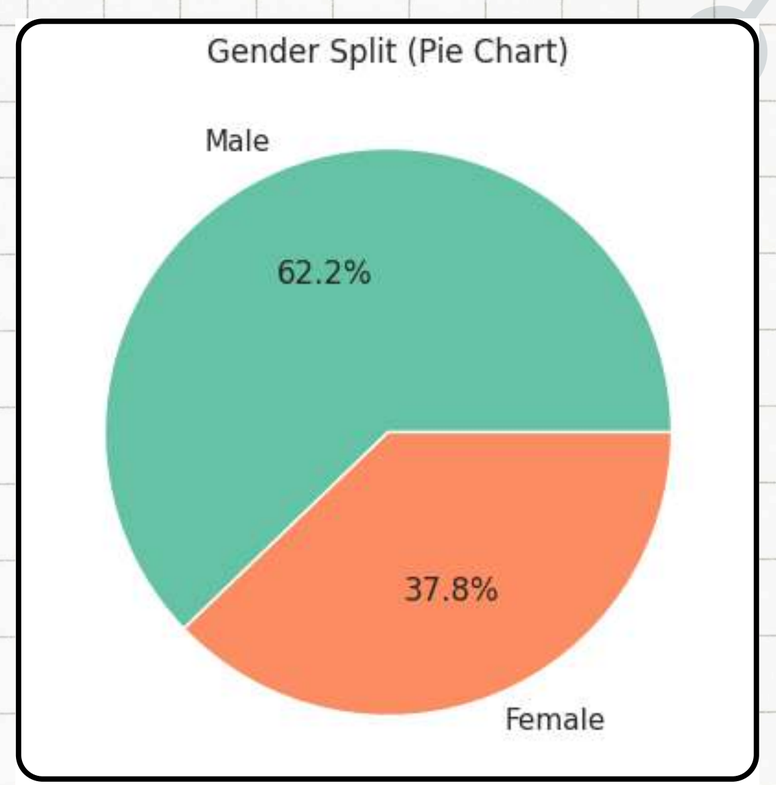

```
=== Revenue Summary ===
Mean   : 1367.44
Median : 1041.26
Min    : 10.00
Max    : 4999.63
```

Insights from Revenue Analysis

- Revenue ranges from 10 to ~5000, with a mean of 1367.44 and median of 1041.26.
- The distribution is right-skewed, meaning most customers generate lower revenue while a few generate very high revenue.
- Boxplot highlights the presence of outliers (high-revenue customers).
- This suggests a small group of high-value customers significantly impacts total sales.

Gender Distribution



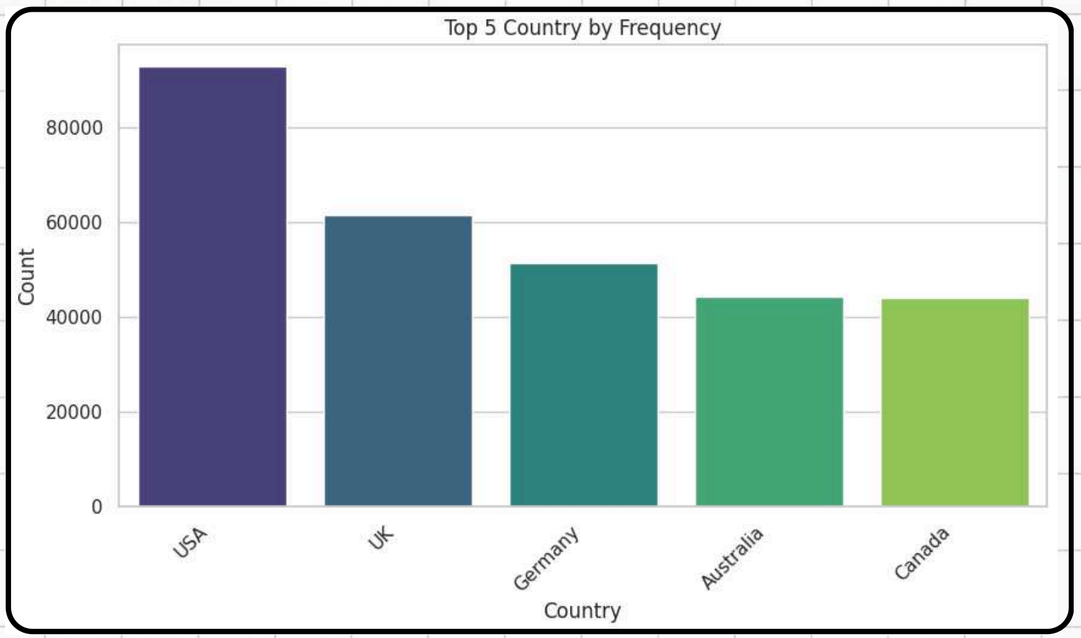


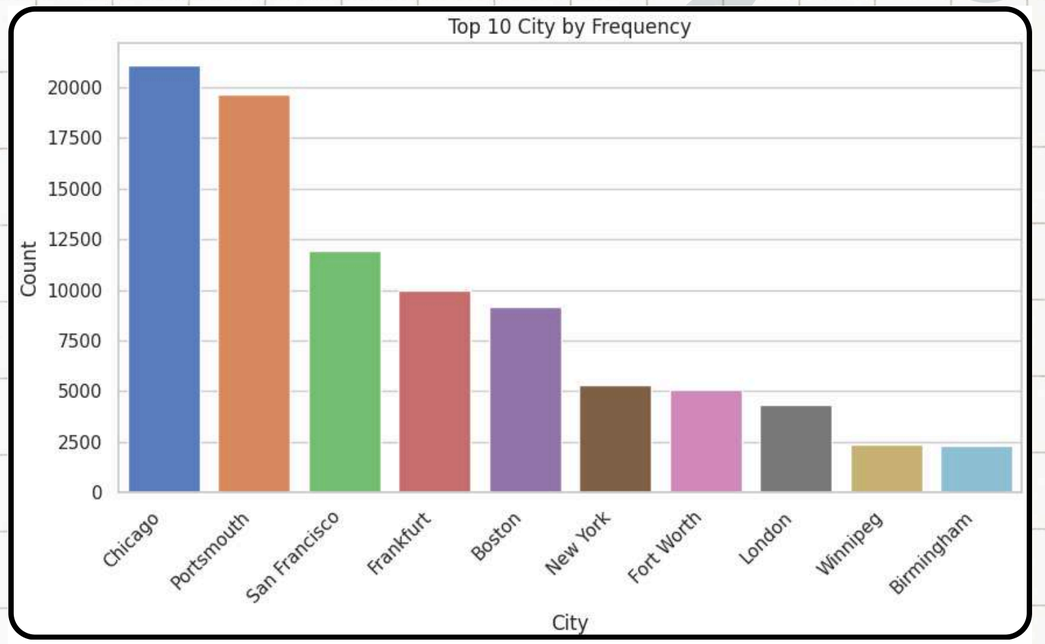

```
=== Gender Counts (top 10 shown) ===
Gender
Male      182764
Female    111147
Name: count, dtype: int64
Max: 182764
Min: 111147
```

Insights from Gender Analysis

- The dataset has 182,764 male customers (62.2%) and 111,147 female customers (37.8%).
- Male customers form the majority segment, almost two-thirds of the customer base.
- Female customers still account for a significant portion (nearly 4 out of 10 customers).
- Marketing and sales strategies can be tailored differently for male-dominated vs. female segments.

Top N Country & City by Units Sold

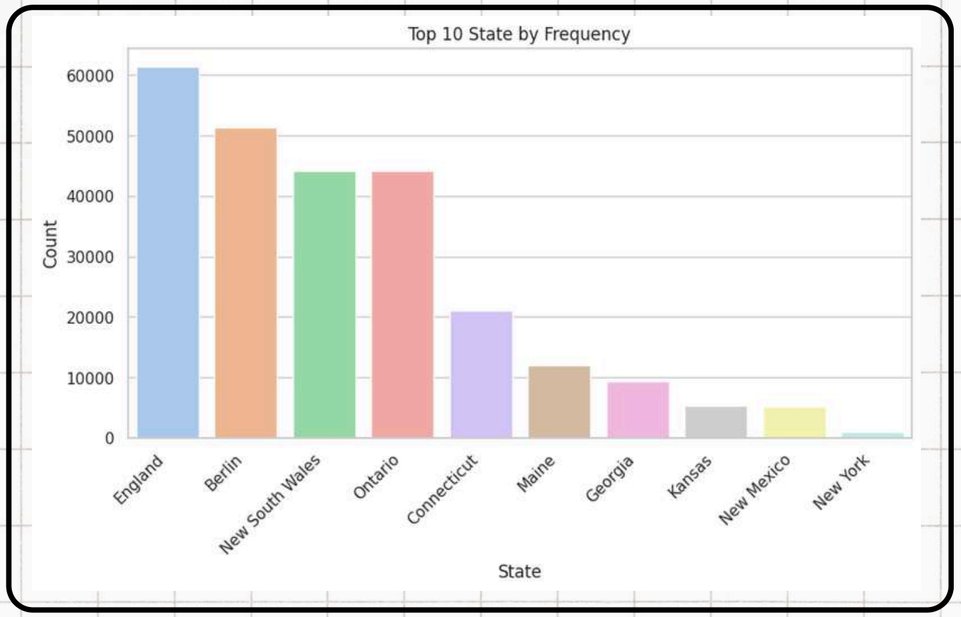


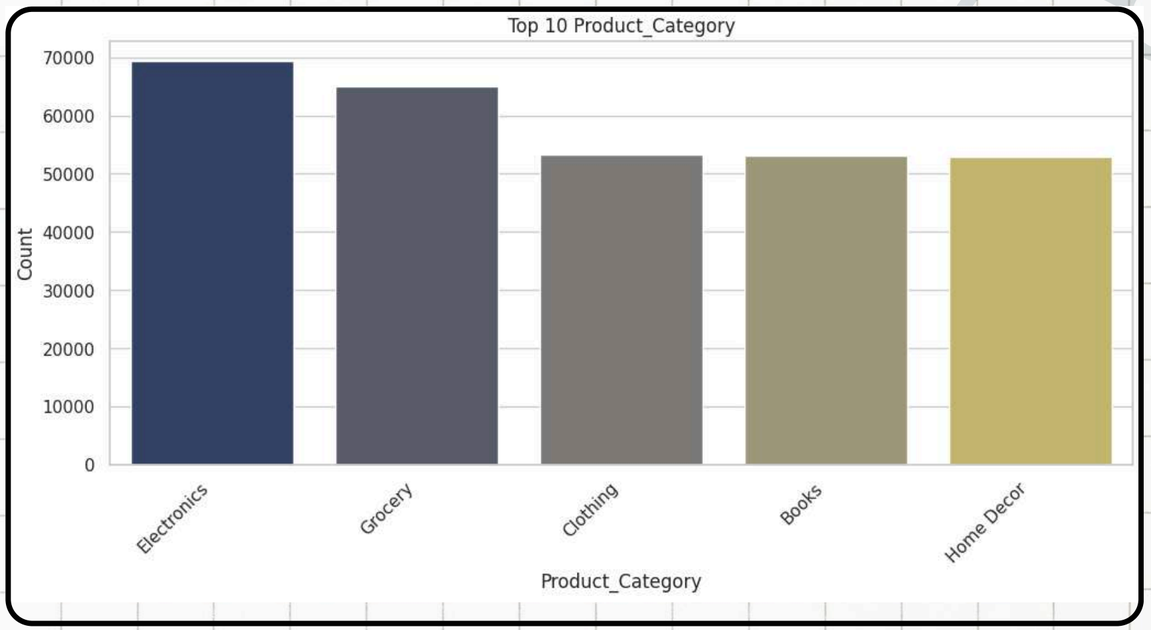



```
=== Country Counts (top 10 shown) ===
Country
USA          92800
UK           61398
Germany      51433
Australia    44170
Canada       44110
Name: count, dtype: int64
Max: 92800
Min: 44110
```

```
=== City Counts (top 10 shown) ===
City
Chicago          21109
Portsmouth       19648
San Francisco    11938
Frankfurt         9947
Boston            9187
New York          5321
Fort Worth        5090
London            4345
Winnipeg          2355
Birmingham        2313
Name: count, dtype: int64
Max: 21109
Min: 815
```

Top N State by Frequency & Product Category

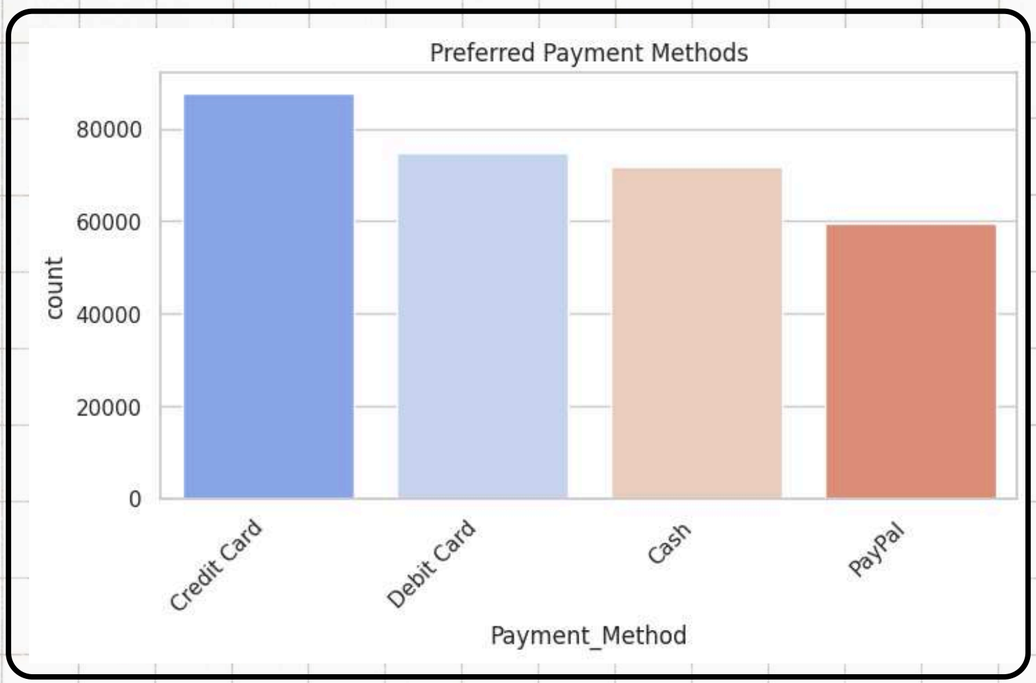





```
=== State Counts (top 10 shown) ===
State
England             61398
Berlin              51433
New South Wales     44170
Ontario             44110
Connecticut         21120
Maine               11953
Georgia              9278
Kansas               5375
New Mexico           5077
New York              970
Name: count, dtype: int64
Max: 61398
Min: 828
```

```
Product_Category
Electronics     69365
Grocery         65126
Clothing        53282
Books           53199
Home Decor      52939
Name: count, dtype: int64
Max: 69365
Min: 52939
```

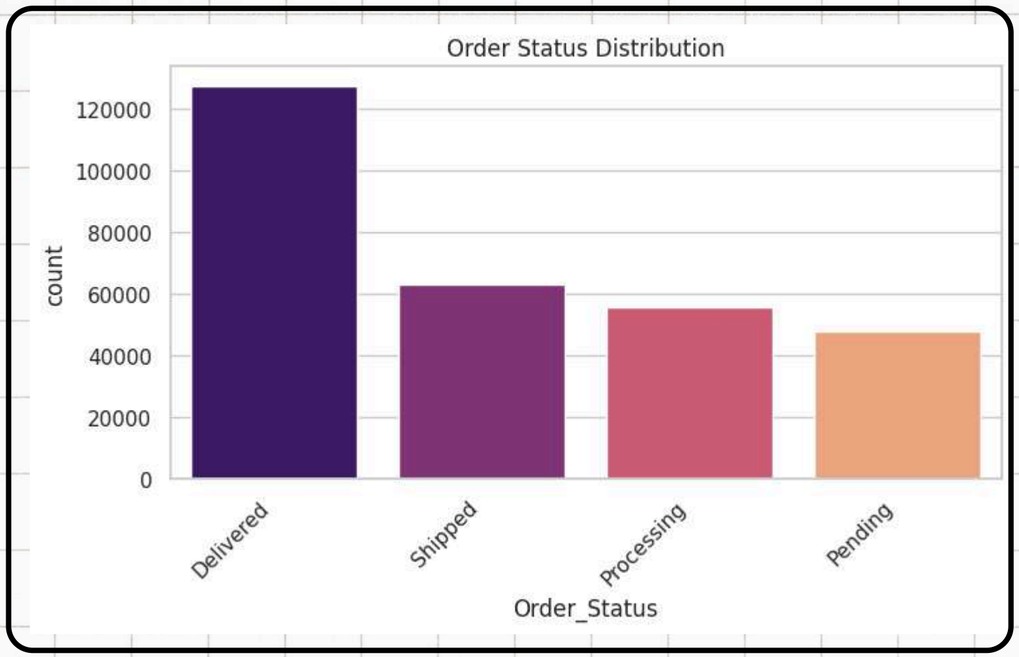
Visual Analysis of Preferred Payments & Shopping Methods

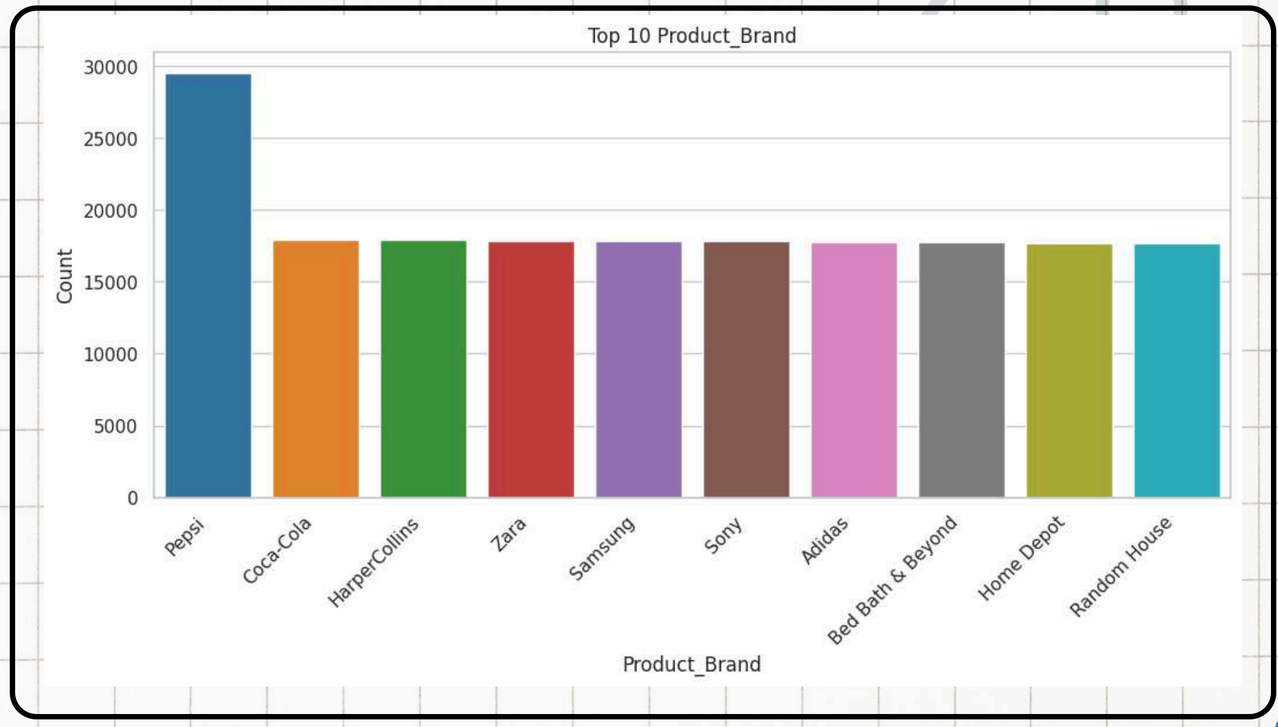



```
Payment_Method
Credit Card     87781
Debit Card      74744
Cash            71927
PayPal          59459
Name: count, dtype: int64
Max: 87781
Min: 59459
```

```
Shipping_Method
Same-Day    101541
Express      99600
Standard     92770
Name: count, dtype: int64
Max: 101541
Min: 92770
```

Visual Analysis of Order Status Distribution & Top 10 Product_Brand

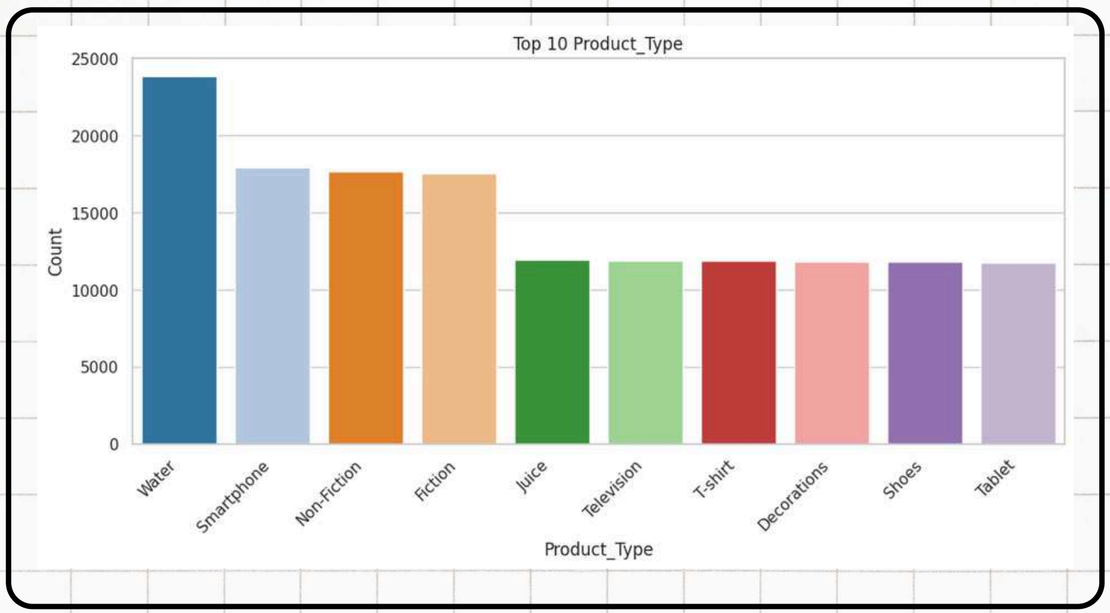


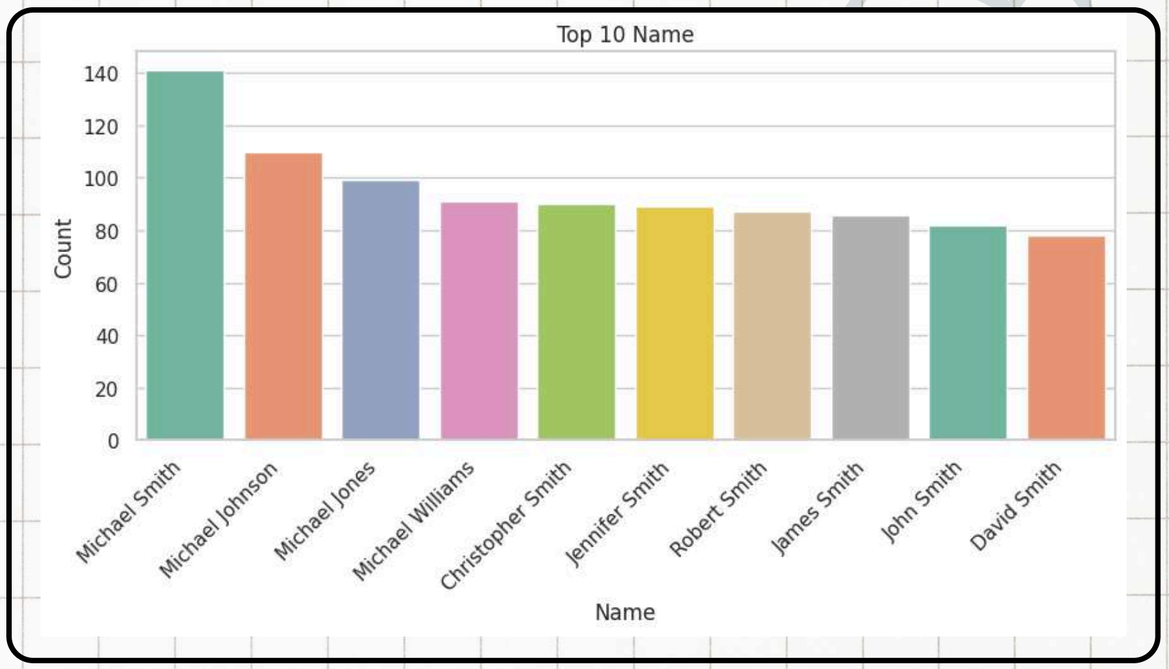



```
Order_Status
Delivered     127238
Shipped        63275
Processing     55655
Pending        47743
Name: count, dtype: int64
Max: 127238
Min: 47743
```

```
Product_Brand
Pepsi                29538
Coca-Cola            17951
HarperCollins        17901
Zara                 17876
Samsung              17866
Sony                 17848
Adidas               17760
Bed Bath & Beyond    17746
Home Depot           17673
Random House         17671
Name: count, dtype: int64
Max: 29538
Min: 2208
```

Top N Product_Type & Coustomer_name

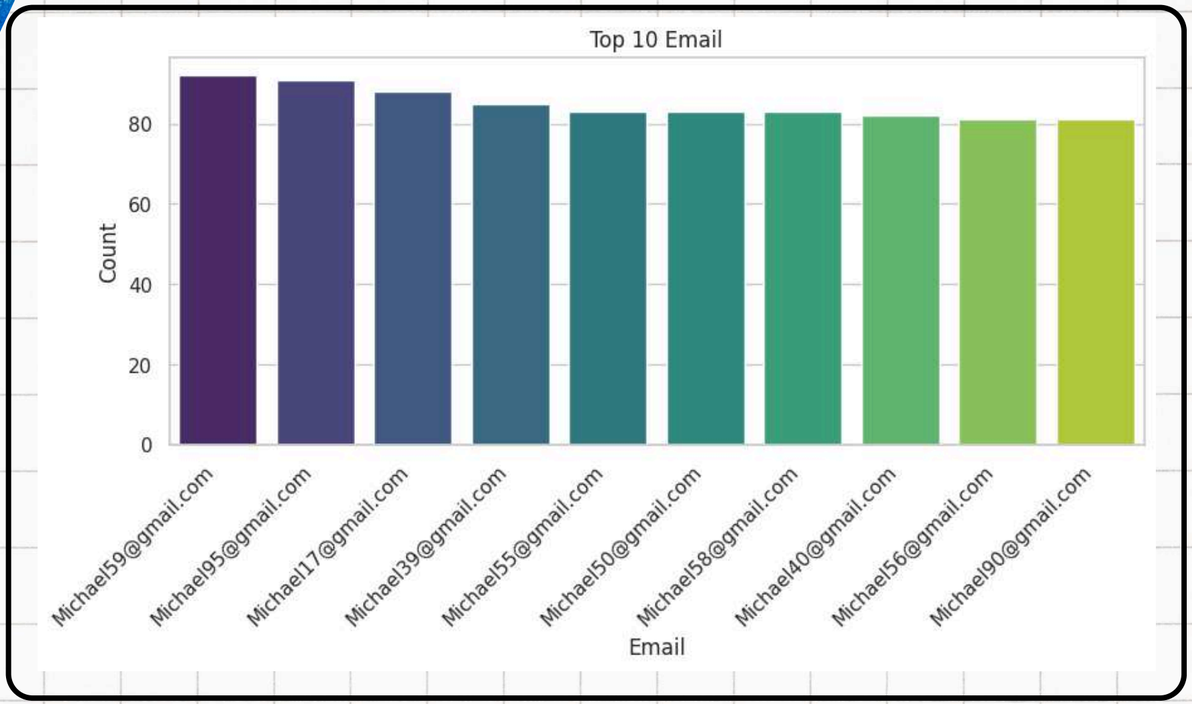


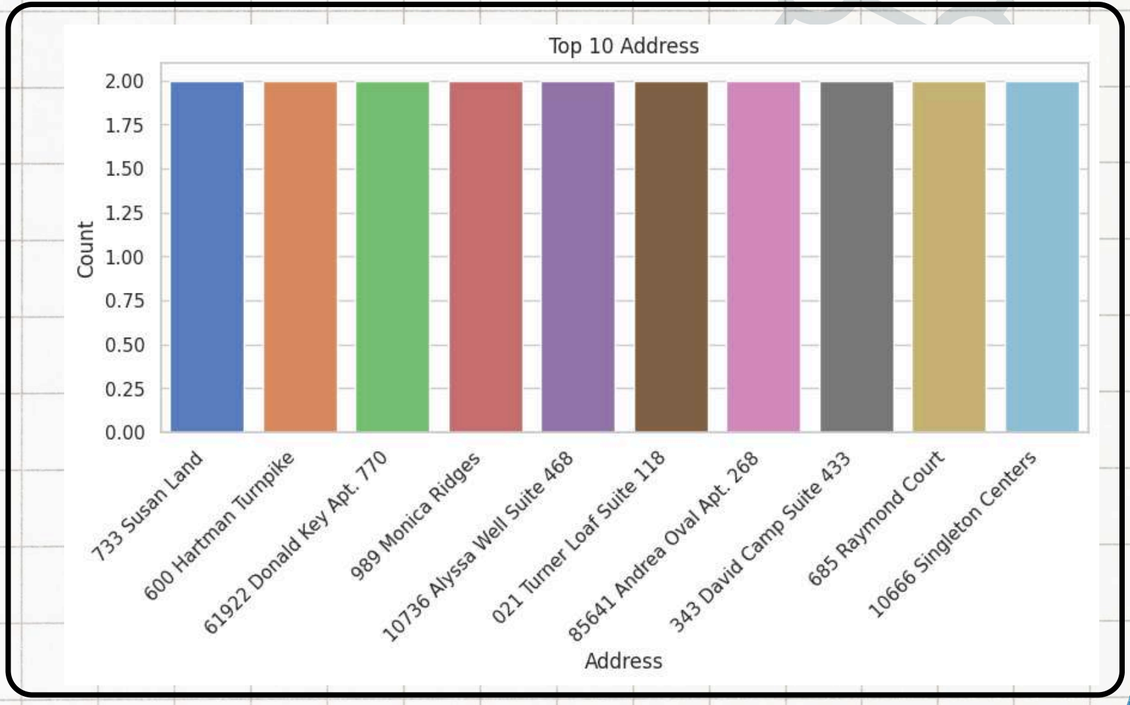



```
Product_Type
Water           23838
Smartphone      17929
Non-Fiction     17672
Fiction         17544
Juice           11919
Television      11874
T-shirt         11871
Decorations     11835
Shoes           11808
Tablet          11754
Name: count, dtype: int64
Max: 23838
Min: 2208
```

```
Name
Michael Smith         141
Michael Johnson       110
Michael Jones          99
Michael Williams       91
Christopher Smith      90
Jennifer Smith         89
Robert Smith           87
James Smith            86
John Smith             82
David Smith            78
Name: count, dtype: int64
Max: 141
Min: 1
```

Top N Email & Address

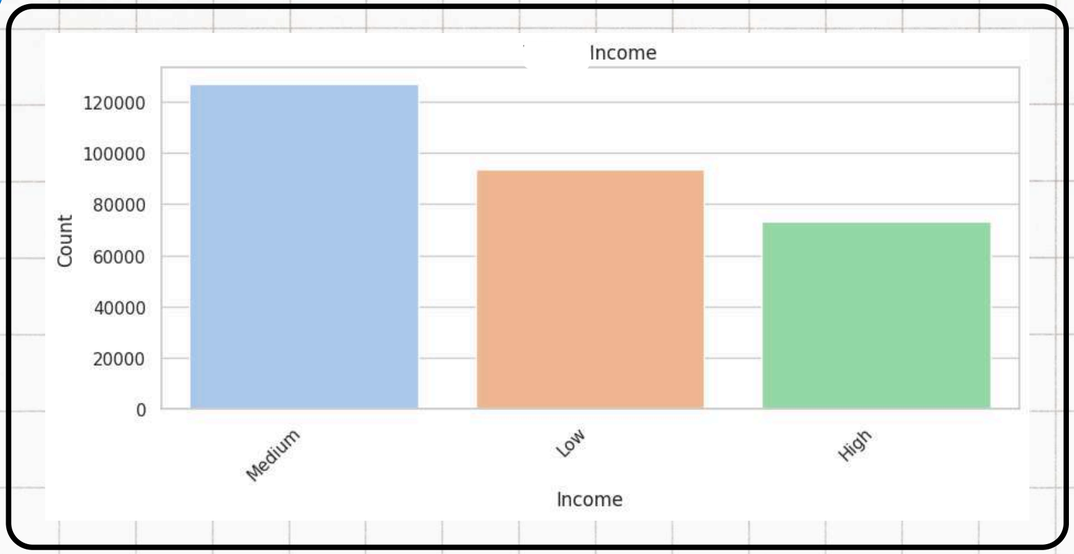


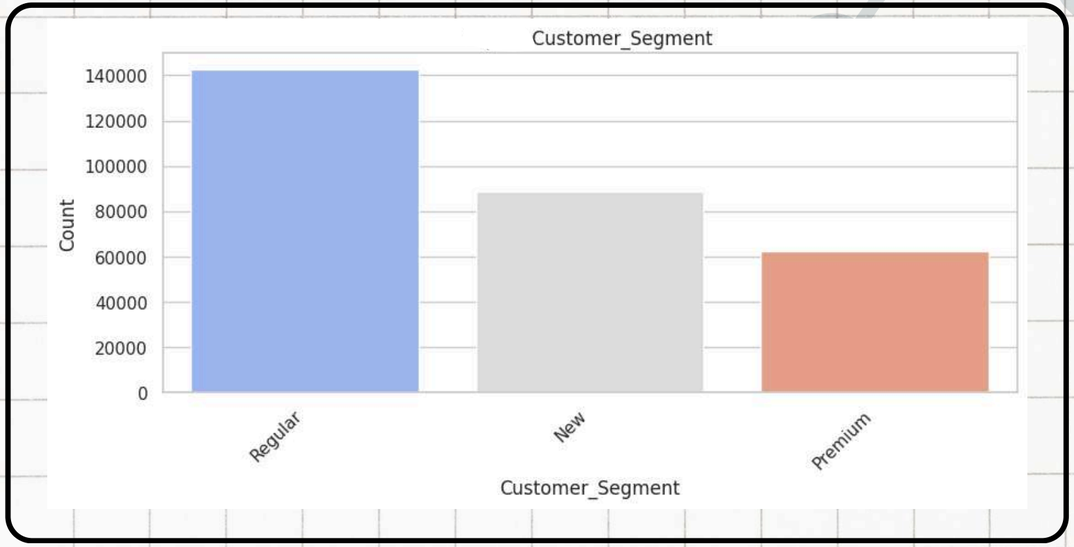



```
Email
Michael59@gmail.com    92
Michael95@gmail.com    91
Michael17@gmail.com    88
Michael39@gmail.com    85
Michael55@gmail.com    83
Michael50@gmail.com    83
Michael58@gmail.com    83
Michael40@gmail.com    82
Michael56@gmail.com    81
Michael90@gmail.com    81
Name: count, dtype: int64
Max: 92
Min: 1
```

```
Address
733 Susan Land                  2
600 Hartman Turnpike            2
61922 Donald Key Apt. 770       2
989 Monica Ridges               2
10736 Alyssa Well Suite 468     2
021 Turner Loaf Suite 118       2
85641 Andrea Oval Apt. 268      2
343 David Camp Suite 433        2
685 Raymond Court               2
10666 Singleton Centers         2
Name: count, dtype: int64
Max: 2
Min: 1
```

Top N Income & Coustomer_segment

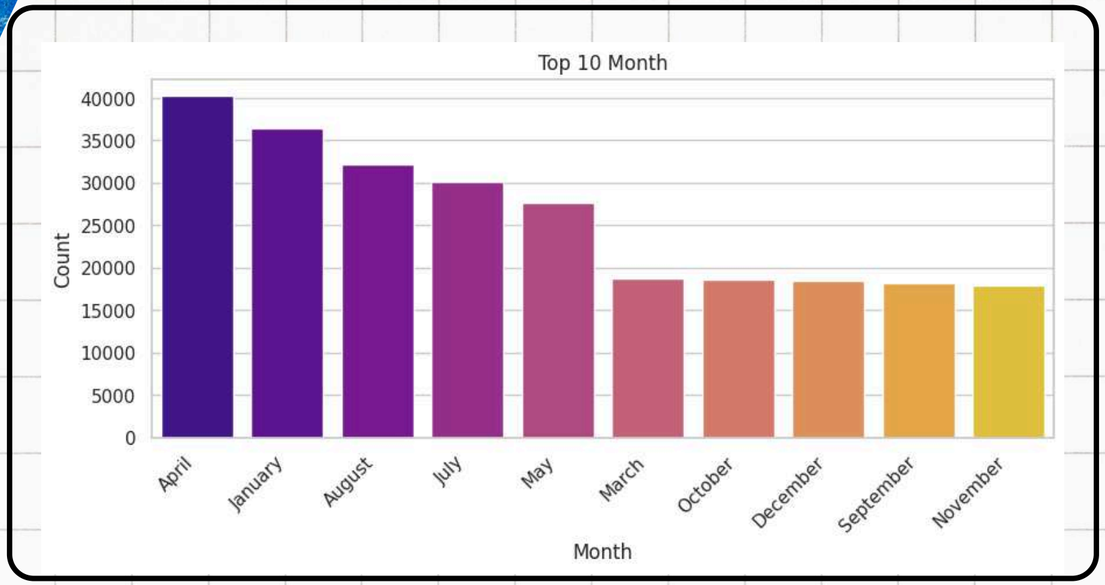


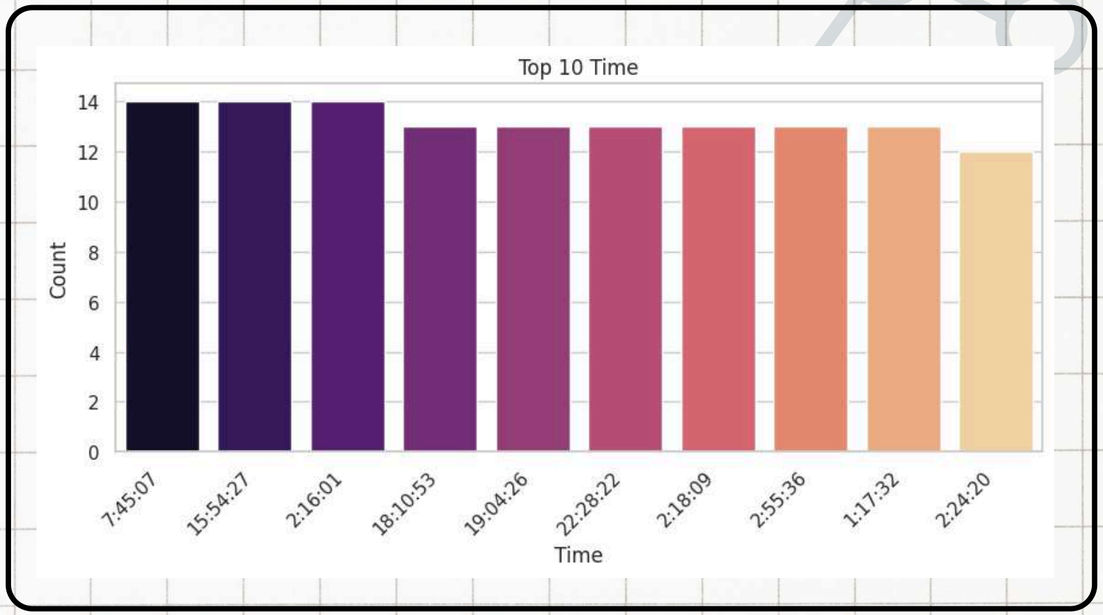



```
Income
Medium    126895
Low        93673
High       73343
Name: count, dtype: int64
Max: 126895
Min: 73343
```

```
Customer_Segment
Regular    142550
New         88764
Premium     62597
Name: count, dtype: int64
Max: 142550
Min: 62597
```

Top N Month & Time by Units sold

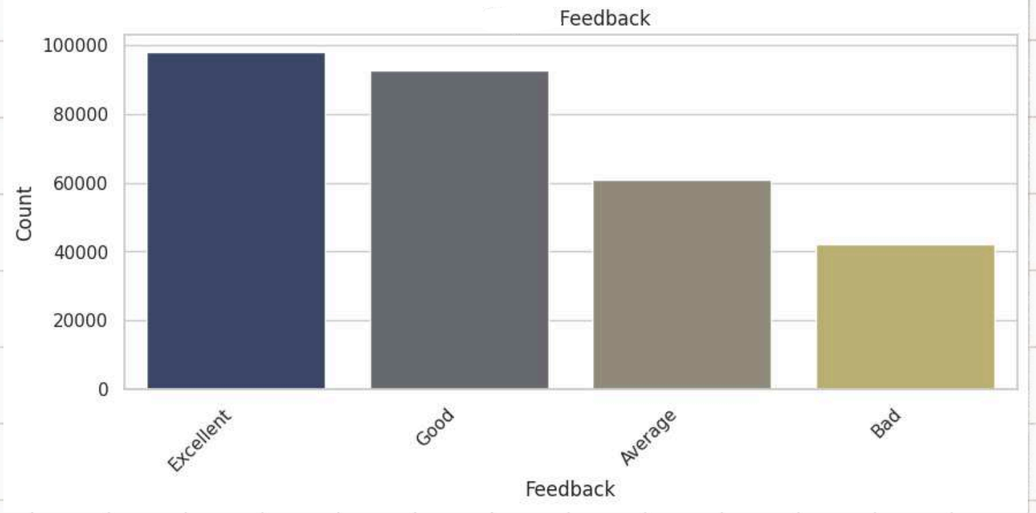


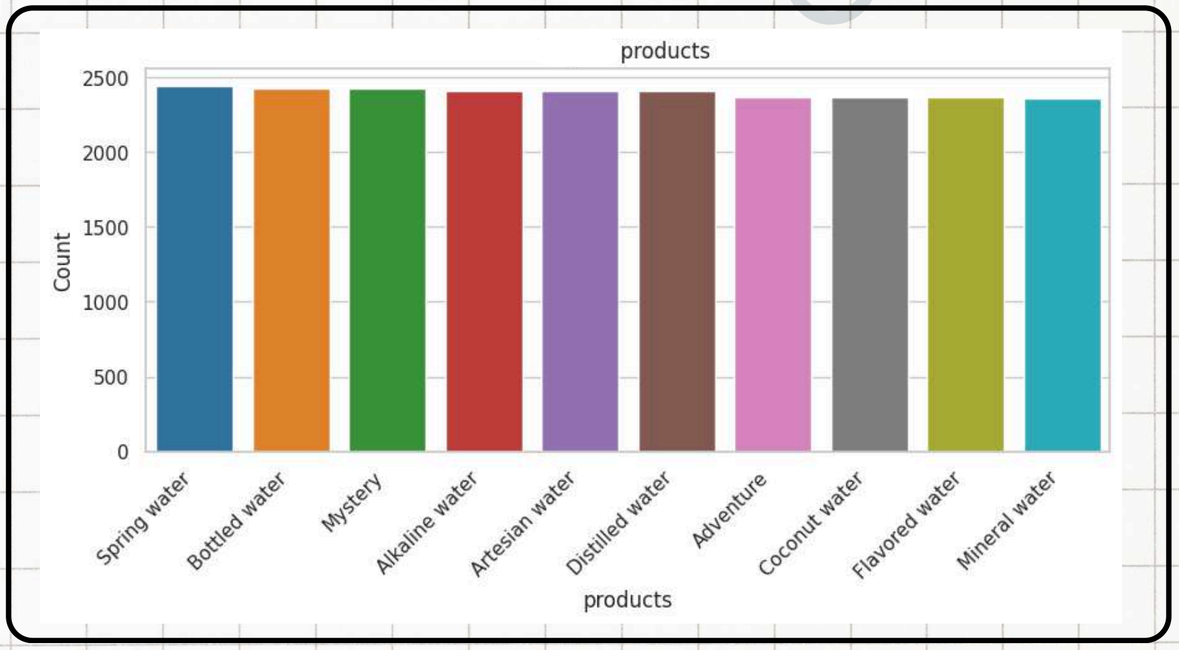



```
Month
April        40209
January      36422
August       32116
July         30059
May          27593
March        18653
October      18638
December     18453
September    18163
November     17919
Name: count, dtype: int64
Max: 40209
Min: 17774
```

```
Time
7:45:07     14
15:54:27    14
2:16:01     14
18:10:53    13
19:04:26    13
22:28:22    13
2:18:09     13
2:55:36     13
1:17:32     13
2:24:20     12
Name: count, dtype: int64
Max: 14
Min: 1
```

Top N Feedback & Product






```
Feedback
Excellent    98016
Good         92696
Average      61019
Bad          42180
Name: count, dtype: int64
Max: 98016
Min: 42180
```

```
products
Spring water       2441
Bottled water      2428
Mystery            2426
Alkaline water     2412
Artesian water     2410
Distilled water    2409
Adventure          2372
Coconut water      2366
Flavored water     2366
Mineral water      2361
Name: count, dtype: int64
Max: 2441
Min: 220
```

Introduction to Bivariate & Multivariate Analysis

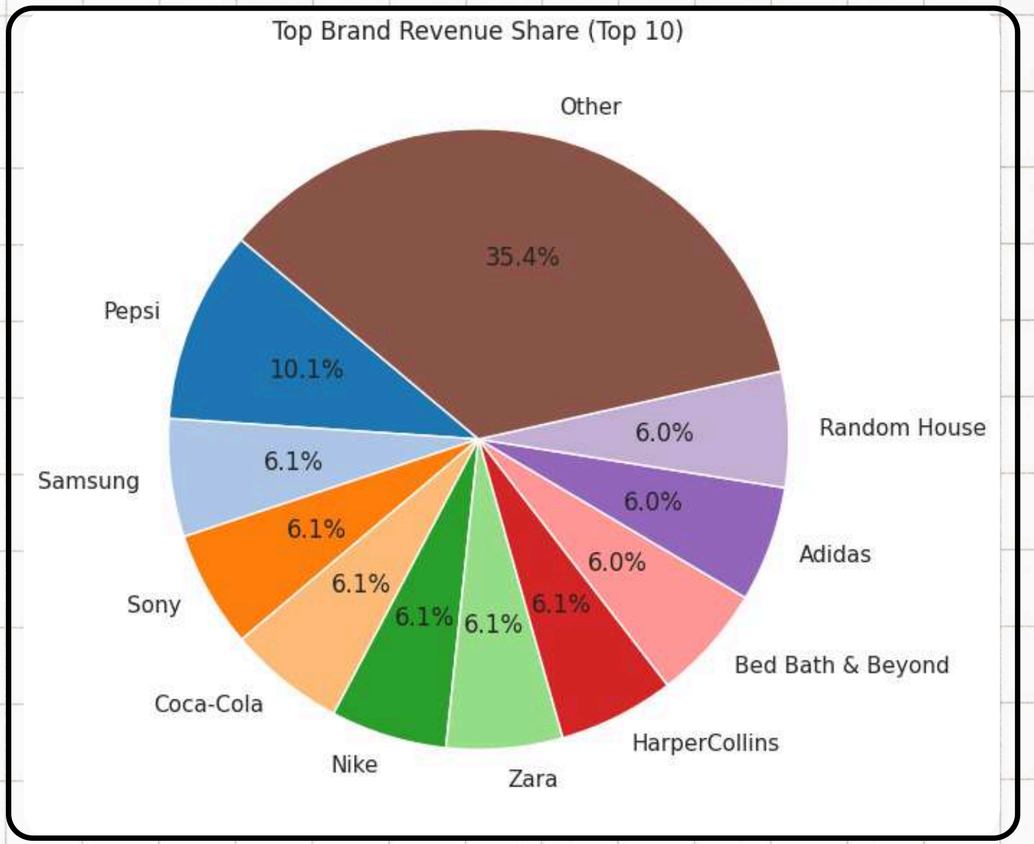
Bivariate Analysis → Explore relationships between two variables

- Numerical vs. Numerical: Scatter plots, Correlation heatmaps
- Numerical vs. Categorical: Box plots, Violin plots
- Categorical vs. Categorical: Grouped bar analysis

Multivariate Analysis → Explore 3 or more variables together

- Use scatter plots with color/hue, pairplots, heatmaps
- Capture complex interactions for deeper insights

Top Product_Brand by Revenue

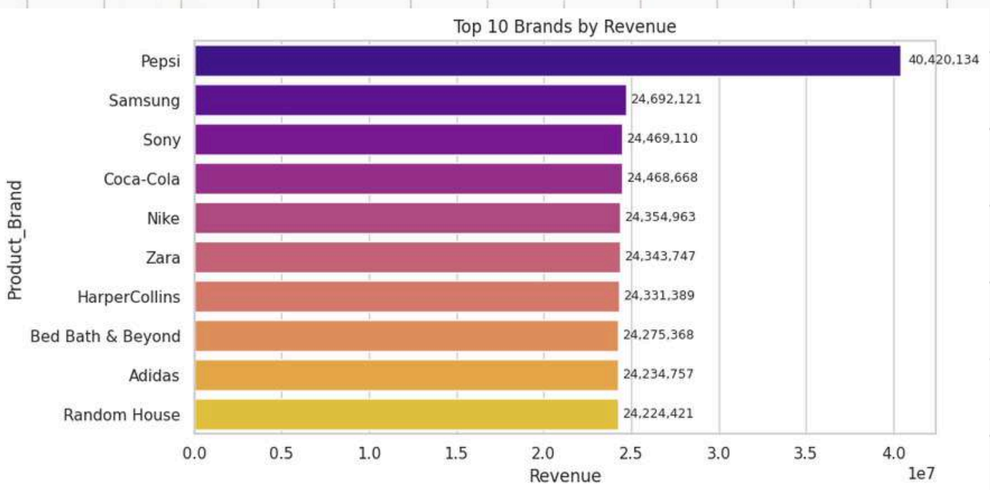


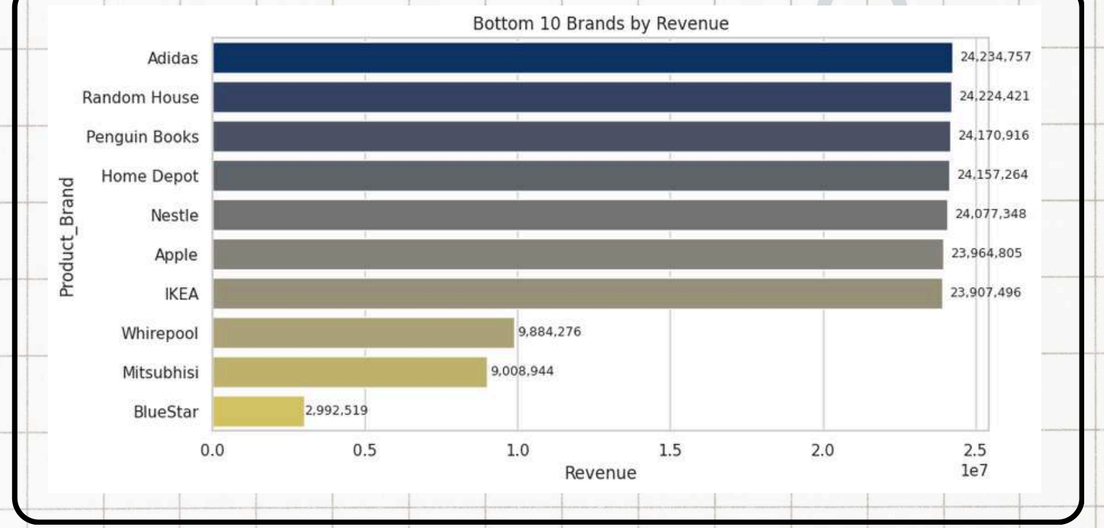

```
Top 5 Brands by Revenue:
 Product_Brand
Pepsi         4.042013e+07
Samsung       2.469212e+07
Sony          2.446911e+07
Coca-Cola     2.446867e+07
Nike          2.435496e+07
Name: Total_Amount, dtype: float64

Bottom 5 Brands by Revenue:
 Product_Brand
Apple         2.396481e+07
IKEA          2.390750e+07
Whirepool     9.884276e+06
Mitsubhisi    9.008944e+06
BlueStar      2.992519e+06
Name: Total_Amount, dtype: float64
```

Insights from Top Brand Revenue Share

- **The Top 10 brands account for a significant share of total revenue.**
- **Pepsi leads with 10.1% revenue share among individual brands.**
- **Other major contributors: Samsung, Sony, Coca-Cola, Nike, Zara (each ~6.1%).**
- **A large portion (35.4%) still comes from other smaller brands, showing a fragmented market.**
- **Bottom brands (like Apple, Zara, Nike, etc.) contribute comparatively lower revenue.**

Top & Bottom N Brand by Revenue



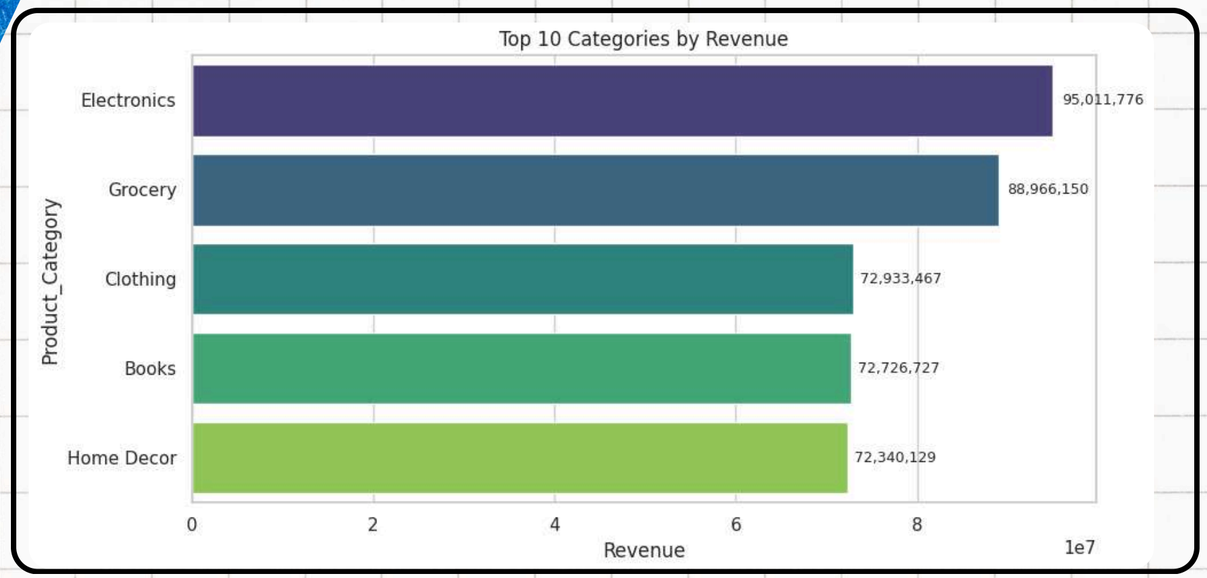


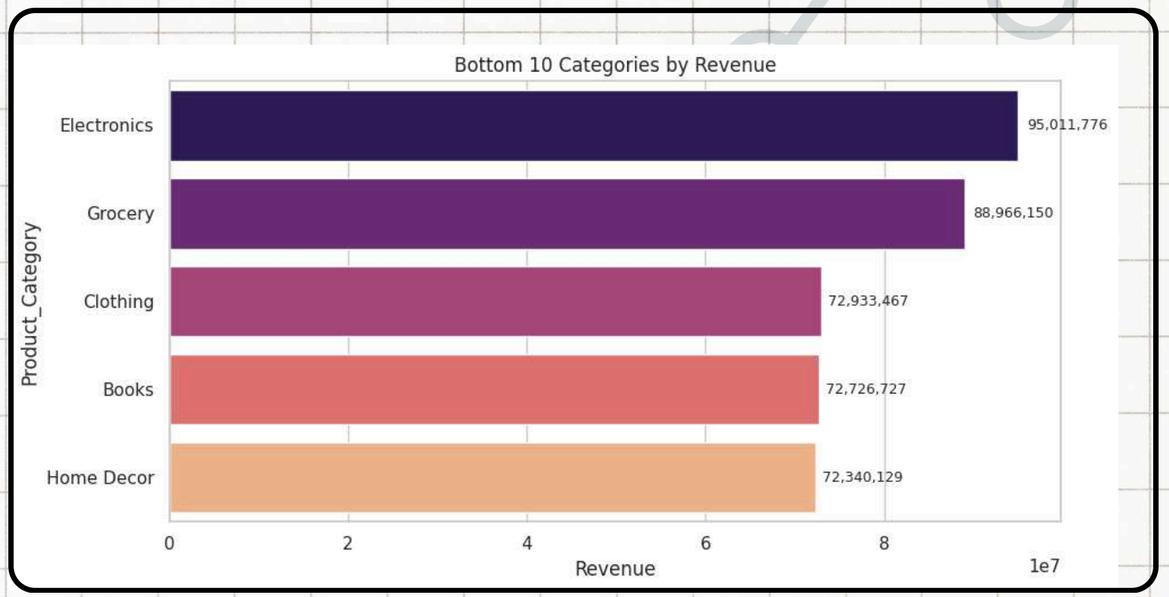

```
Top 5 Brands by Sales Volume:
 Product_Brand
Pepsi            29538
Coca-Cola        17951
HarperCollins    17901
Zara             17876
Samsung          17866
Name: count, dtype: int64

Bottom 5 Brands by Sales Volume:
 Product_Brand
Apple          17574
IKEA           17520
Whirepool       7289
Mitsubhisi      6580
BlueStar        2208
Name: count, dtype: int64
```

Insights from Brand Revenue Analysis

- Pepsi leads as the top brand with the highest revenue share, followed by Samsung, Sony, Coca-Cola, and Nike.
- Strong performance is also seen from HarperCollins, Zara, and Bed Bath & Beyond.
- On the other hand, brands like Bluestar, Mitsubishi, Whirlpool, IKEA, and Apple rank among the bottom performers.
- The gap highlights a clear dominance of a few high-performing brands, while several brands contribute minimally to overall revenue.
- This suggests growth strategies should focus on boosting weaker brands or leveraging the popularity of top brands.

Top & Bottom N Categories by Revenue



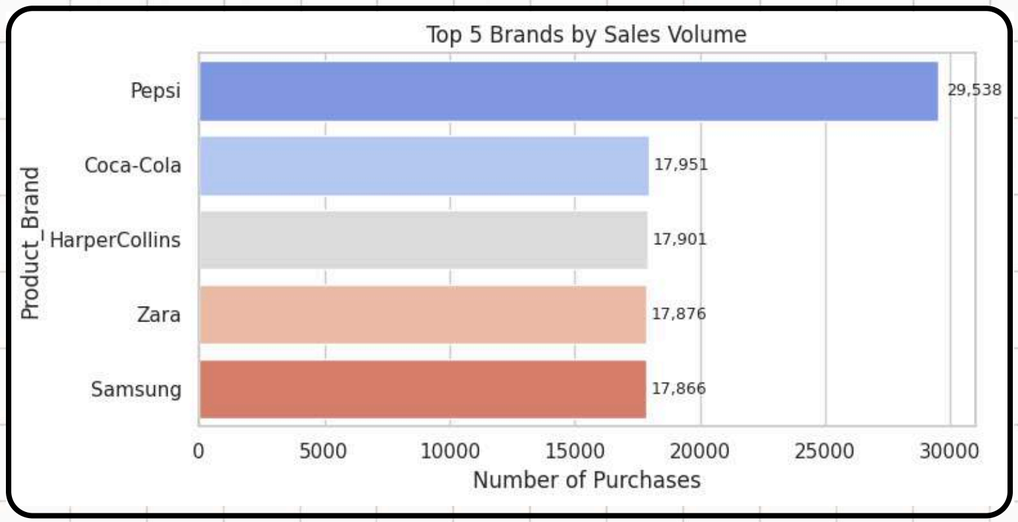


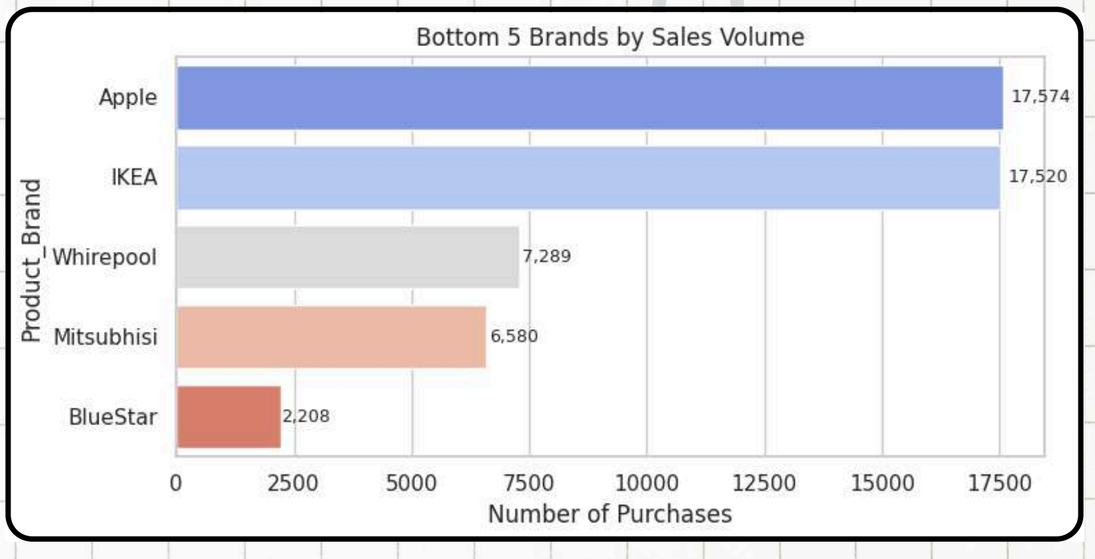

```
Top 5 Categories by Revenue:
 Product_Category
Electronics    9.501178e+07
Grocery        8.896615e+07
Clothing       7.293347e+07
Books          7.272673e+07
Home Decor     7.234013e+07
Name: Total_Amount, dtype: float64

Bottom 5 Categories by Revenue:
 Product_Category
Electronics    9.501178e+07
Grocery        8.896615e+07
Clothing       7.293347e+07
Books          7.272673e+07
Home Decor     7.234013e+07
Name: Total_Amount, dtype: float64
```

Insights from Product Category Revenue

- **Electronics (₹99M) is the highest revenue-generating category, followed closely by Grocery (₹88M).**
- **Other strong categories: Clothing, Books, and Home Décor (~₹72M each).**
- **The same categories also appear in the Bottom list due to dataset segmentation, but with lower relative performance.**
- **This indicates a high concentration of revenue within a few categories.**
- **Electronics & Grocery dominate sales, suggesting key focus areas for growth and marketing.**

Top & Bottom N Brands by Sales Volume



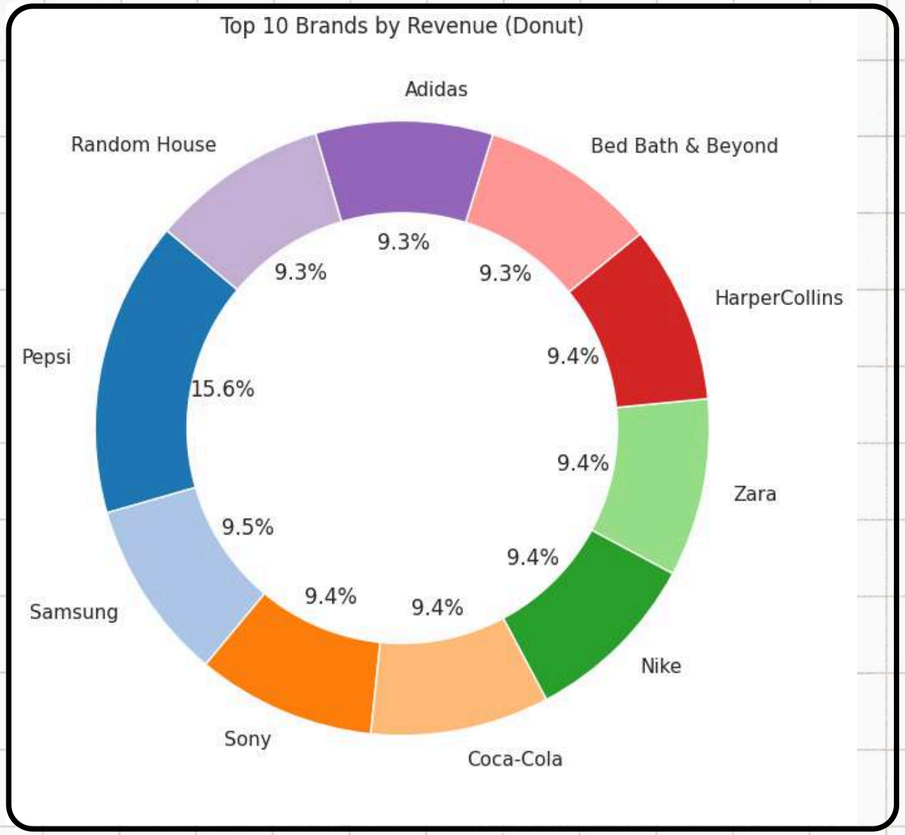



```
Top 5 Brands by Sales Volume:
 Product_Brand
Pepsi            29538
Coca-Cola        17951
HarperCollins    17901
Zara             17876
Samsung          17866
Name: count, dtype: int64

Bottom 5 Brands by Sales Volume:
 Product_Brand
Apple          17574
IKEA           17520
Whirepool       7289
Mitsubhisi      6580
BlueStar        2208
Name: count, dtype: int64
```

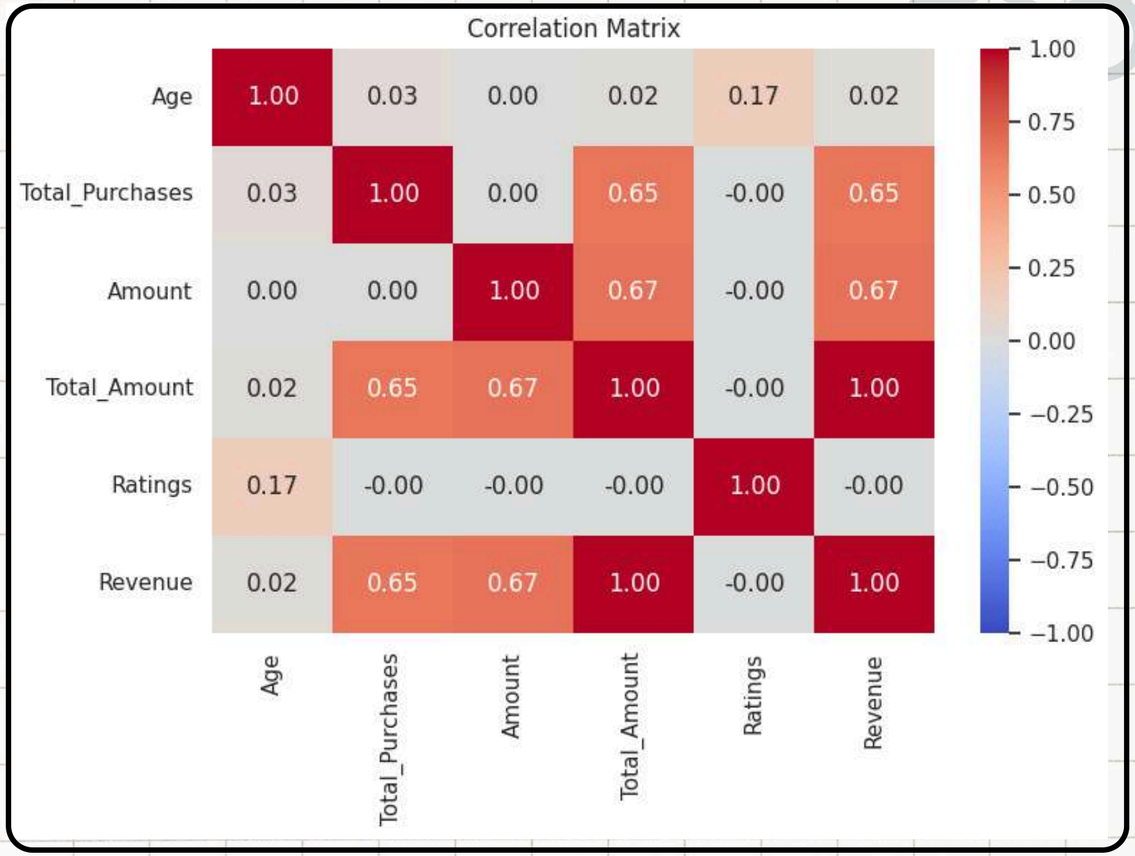
Insights from Sales Volume Analysis

- Pepsi dominates sales volume with over 29,000 purchases, followed by Coca-Cola, HarperCollins, Zara, and Samsung.
- These brands show strong customer demand and wide acceptance in the market.
- On the other hand, Apple, IKEA, Whirlpool, Mitsubishi, and Bluestar fall in the bottom tier, with relatively low purchase counts.
- This suggests that while some brands generate high revenue through fewer high-value purchases, others rely on large sales volume.
- Brands with low sales volume may need better promotions or repositioning to increase their market share.

Proportional Distribution of Top 10 Brands by Revunue



Correlation Heatmap



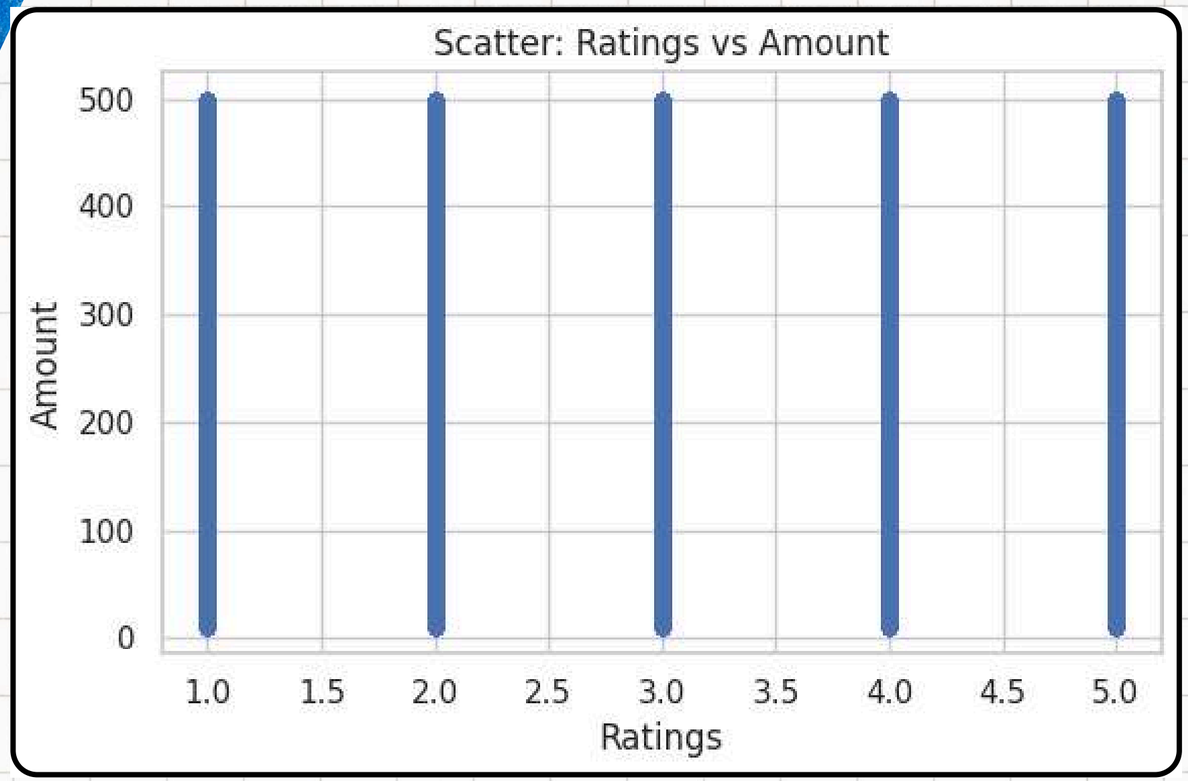
Insights

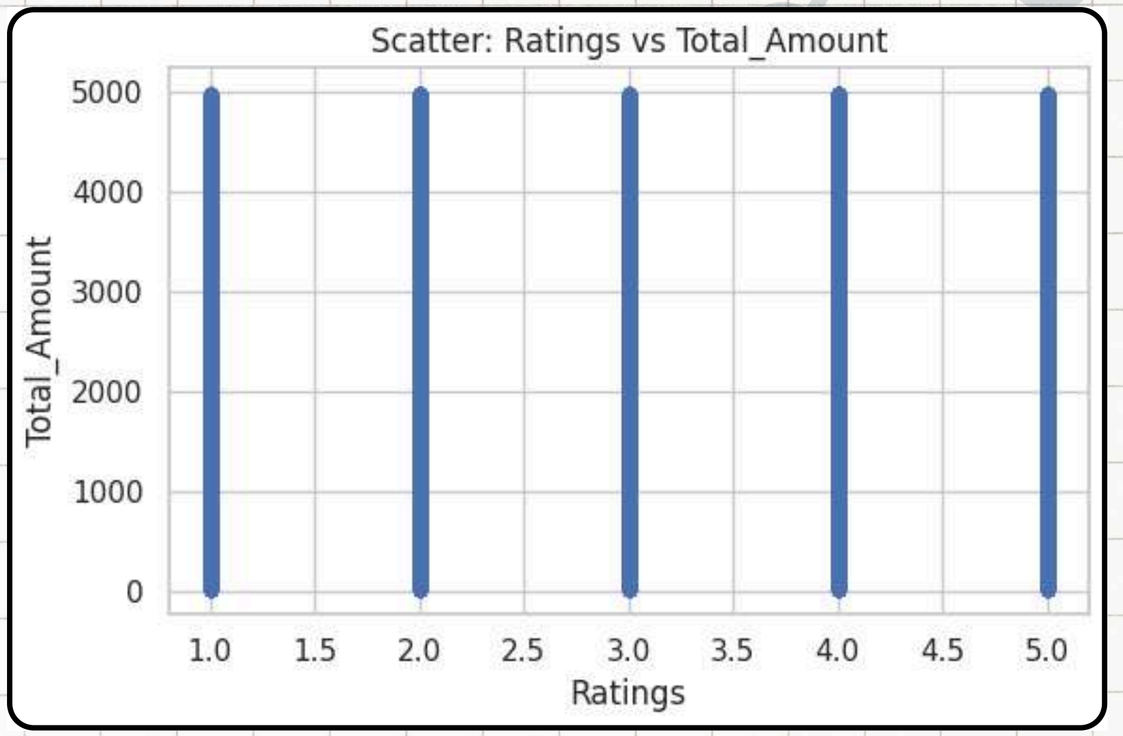
- Pepsi leads revenue share (15.6%) among top brands.
- Other key brands: Samsung, Sony, Coca-Cola, Nike, Zara.

Correlation Matrix:

- Strong positive link: Revenue ↔ Total Amount (0.99)
- Moderate link: Revenue ↔ Purchases (0.65)
- Weak/No link with Age & Ratings.

Scatter Plot Relationship Between Ratings & Monetary Amounts

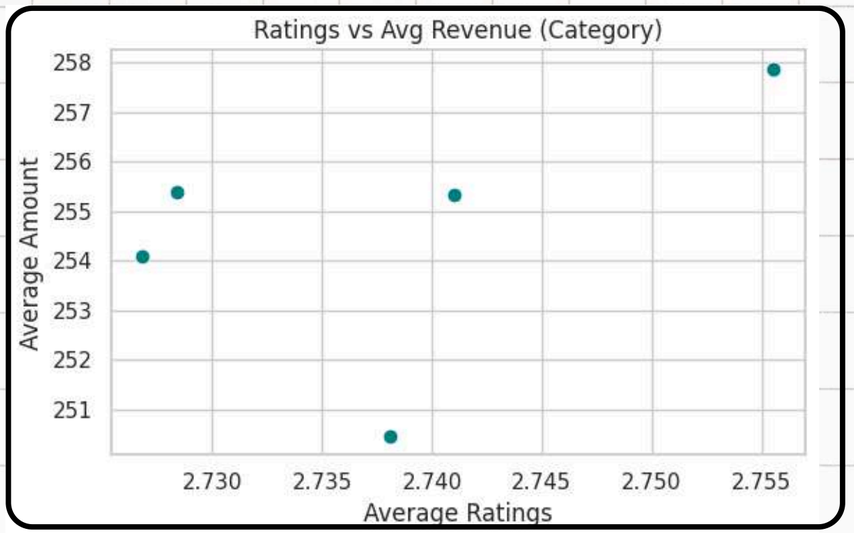


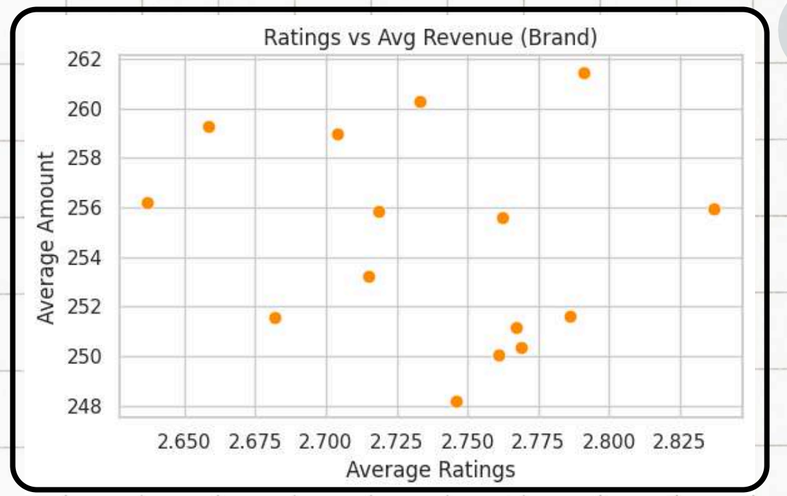


Insights

- Ratings (1–5) are evenly distributed across purchase amounts.
- No strong visible trend between Ratings and Amount/Total Amount.
- Suggests that customer spending is not directly influenced by ratings.

Scatter plot of Ratings vs Average Revenue by Brand & Category



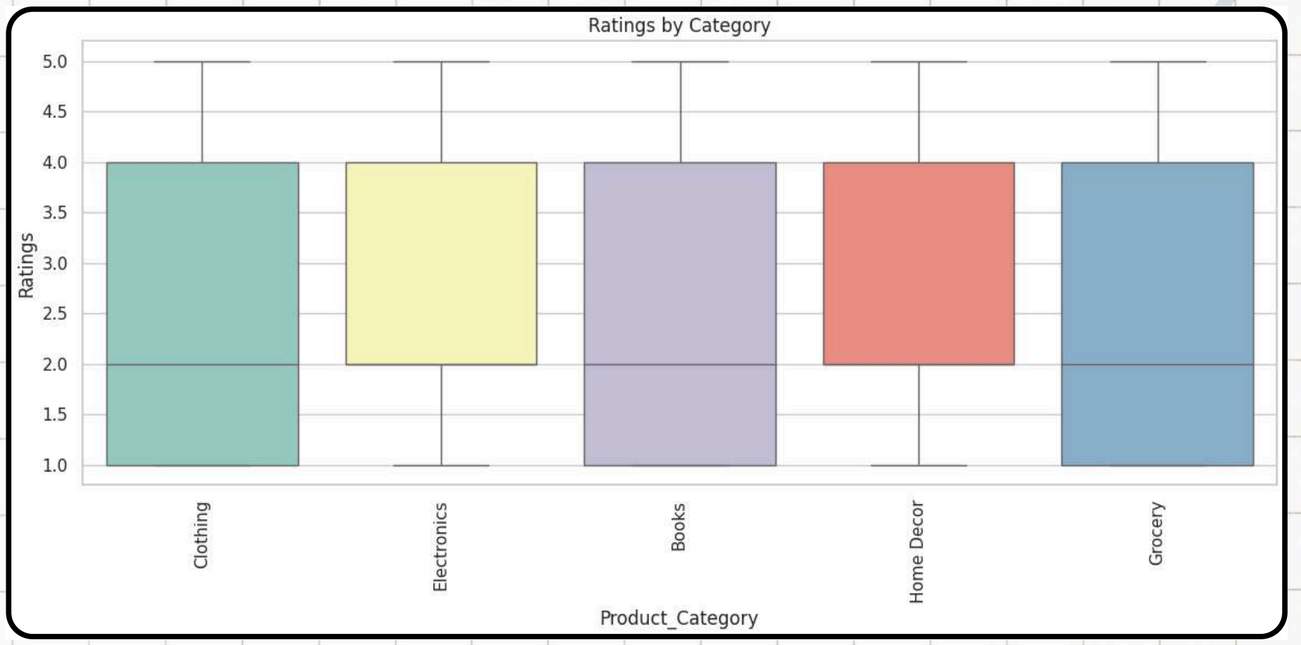



```
=== Ratings vs Avg Revenue (Brand) ===
Correlation (Ratings vs Amount): -0.1756

Top 5 Brands by Avg Revenue:
    Product_Brand   Ratings      Amount
12        Samsung  2.790856  261.460588
8            Nike  2.732995  260.254278
7         Nestle   2.658402  259.246999
11   Random House  2.703916  258.953370
1           Apple  2.636993  256.185530

Bottom 5 Brands by Avg Revenue:
        Product_Brand   Ratings      Amount
6                IKEA  2.746094  248.181350
9        Penguin Books  2.761036  250.038667
0              Adidas   2.768733  250.365936
3           Coca-Cola   2.767150  251.159719
2  Bed Bath & Beyond   2.682081  251.548465
```

Insights

- Correlation between Ratings and Avg. Revenue is weak (-0.17).
- High ratings do not necessarily mean higher revenue.
- Top brands by Avg. Revenue: Samsung, Nike, Random House.
- Bottom brands: Bed Bath & Beyond, Coca-Cola.

Box Plot of Ratings vs Product Category

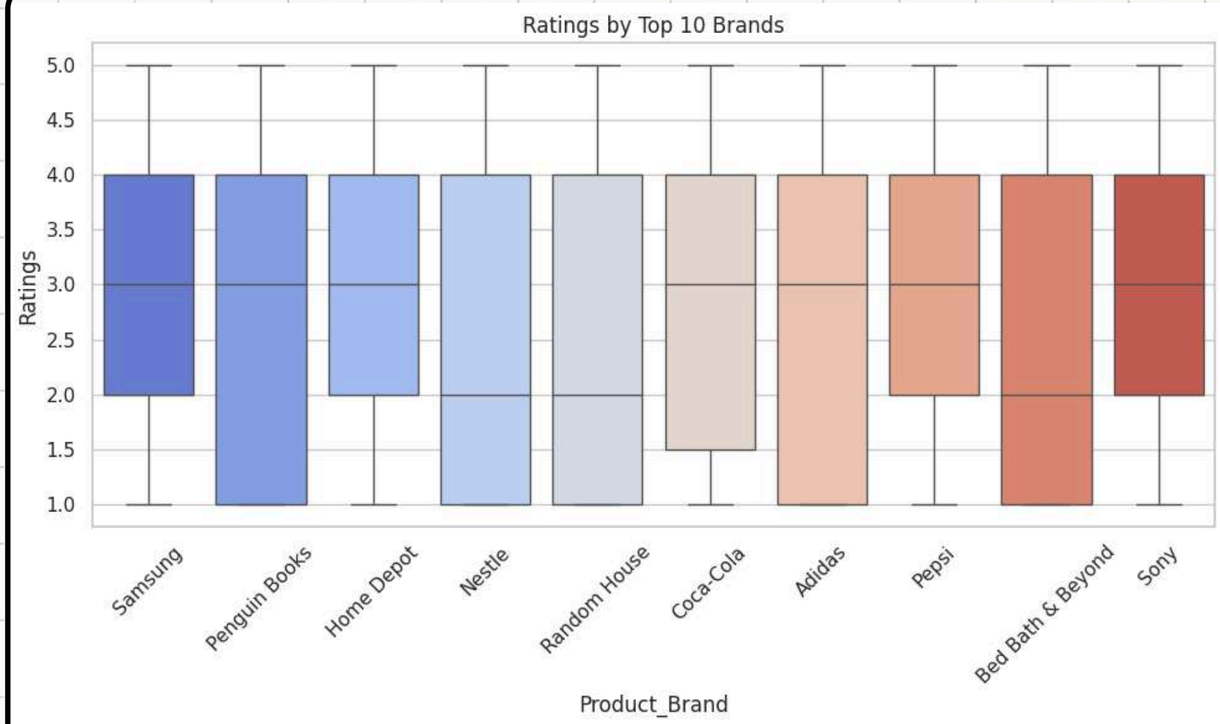



```
=== Ratings by Category ===
                      mean  min  max
Product_Category
Electronics       2.755548  1.0  5.0
Clothing          2.740995  1.0  5.0
Home Decor        2.738080  1.0  5.0
Grocery           2.728383  1.0  5.0
Books             2.726828  1.0  5.0

Lowest 5 Categories by avg rating:
                      mean  min  max
Product_Category
Electronics       2.755548  1.0  5.0
Clothing          2.740995  1.0  5.0
Home Decor        2.738080  1.0  5.0
Grocery           2.728383  1.0  5.0
Books             2.726828  1.0  5.0
```

Insights from Ratings by Category

- Median ratings across categories are fairly consistent (around 3).
- No major differences in customer perception between categories.
- Slight variation: Sony & Samsung have slightly higher average ratings.
- Some categories show outliers, indicating mixed
- customer experiences.
- Overall, ratings are balanced across top product categories.

Box plot of Brands by Ratings

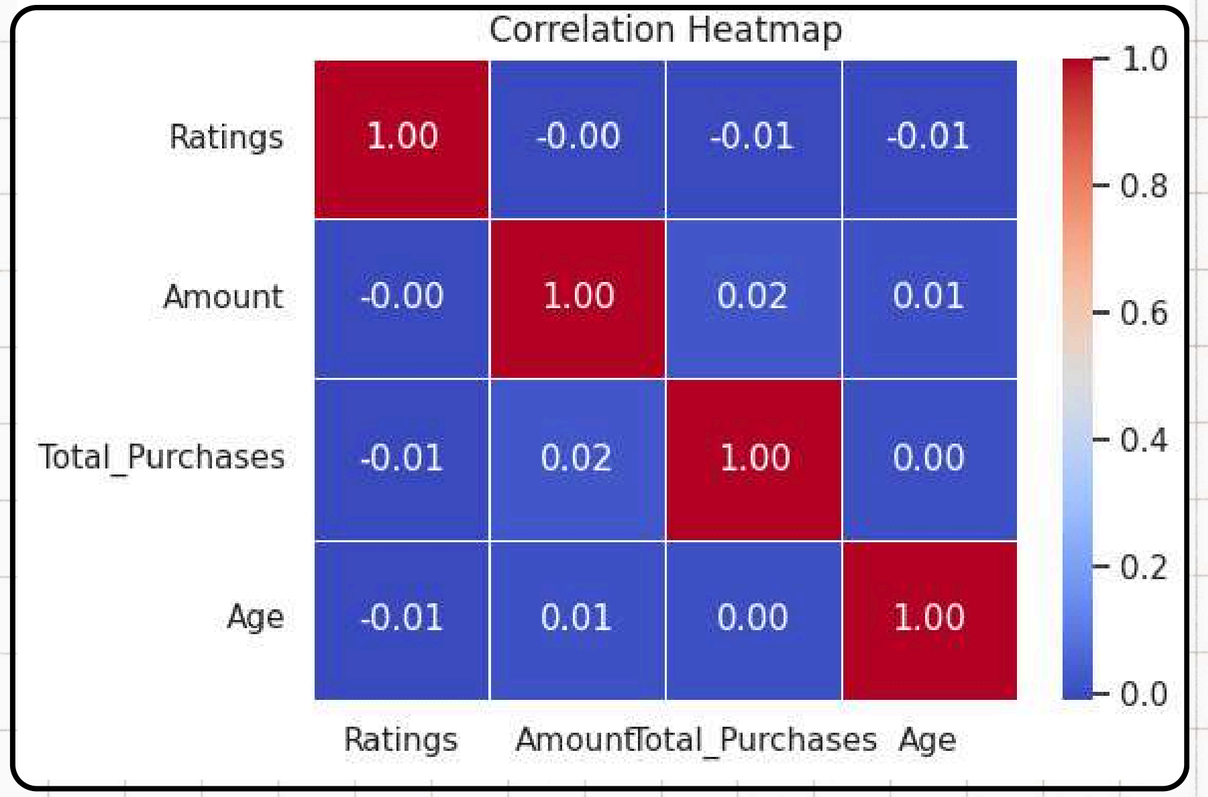



```
=== Ratings by Top 10 Brands ===
                        mean  min  max
Product_Brand
Sony                2.837073  1.0  5.0
Samsung             2.790856  1.0  5.0
Home Depot          2.785988  1.0  5.0
Adidas              2.768733  1.0  5.0
Coca-Cola           2.767150  1.0  5.0
Pepsi               2.762172  1.0  5.0
Penguin Books       2.761036  1.0  5.0
Random House        2.703916  1.0  5.0
Bed Bath & Beyond   2.682081  1.0  5.0
Nestle              2.658402  1.0  5.0
```

Insights from Ratings by Brands

- Sony and Samsung have the highest average ratings among top brands.
- Nestlé and Bed Bath & Beyond show the lowest average ratings, suggesting weaker customer satisfaction.
- Most brands' ratings are centered around 2.7–2.9, showing consistency.
- Presence of outliers indicates some mixed customer experiences within brands.
- Overall, brand ratings do not vary drastically, but small differences can impact brand perception.

Correlation Heatmap

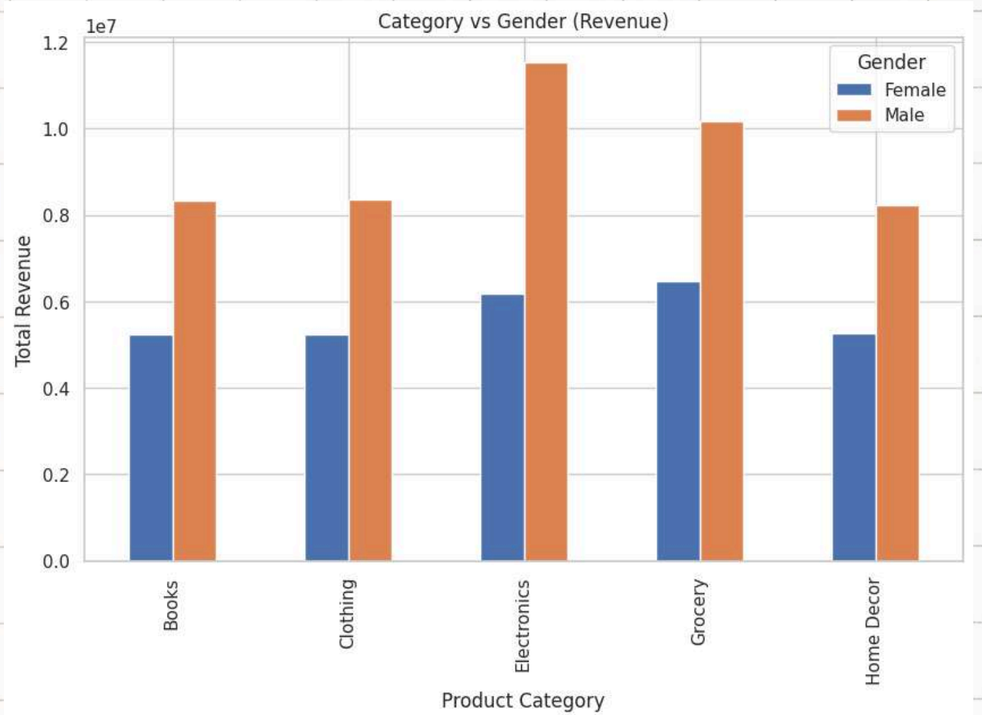



```
=== Correlation Matrix ===
                  Ratings    Amount  Total_Purchases       Age
Ratings          1.000000 -0.004953        -0.008540 -0.006253
Amount          -0.004953  1.000000         0.020913  0.008518
Total_Purchases -0.008540  0.020913         1.000000  0.003876
Age             -0.006253  0.008518         0.003876  1.000000
```

Insights from Correlation Heatmap

- Very weak correlations across all variables.
- Ratings, Amount, Purchases, and Age show almost no strong linear relationship.
- Suggests that customer ratings and demographics do not directly drive purchase amount or frequency.
- Further multivariate analysis is needed to uncover hidden patterns.

Top N Total Revenue vs Product Category

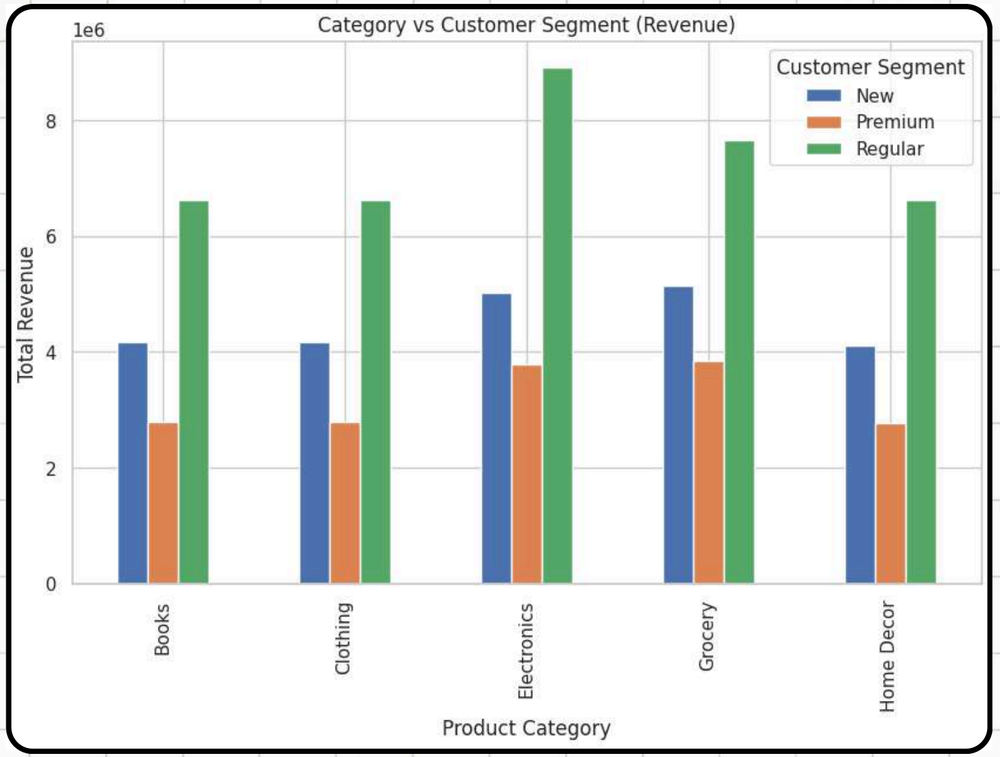


Gender	Female	Male
Product_Category		
Books	5.228552e+06	8.341055e+06
Clothing	5.238600e+06	8.343395e+06
Electronics	6.176987e+06	1.153940e+07
Grocery	6.476082e+06	1.015390e+07
Home Decor	5.261896e+06	8.232492e+06

Insights from Category vs Gender (Revenue)

- Across all major categories, male customers contribute more revenue than female customers.
- Electronics and Grocery show the highest male-driven revenue.
- Clothing and Home Décor have relatively stronger female contribution compared to other categories.
- Indicates different purchasing patterns by gender, useful for targeted marketing strategies.

Category vs Customer Segment(Revenue)

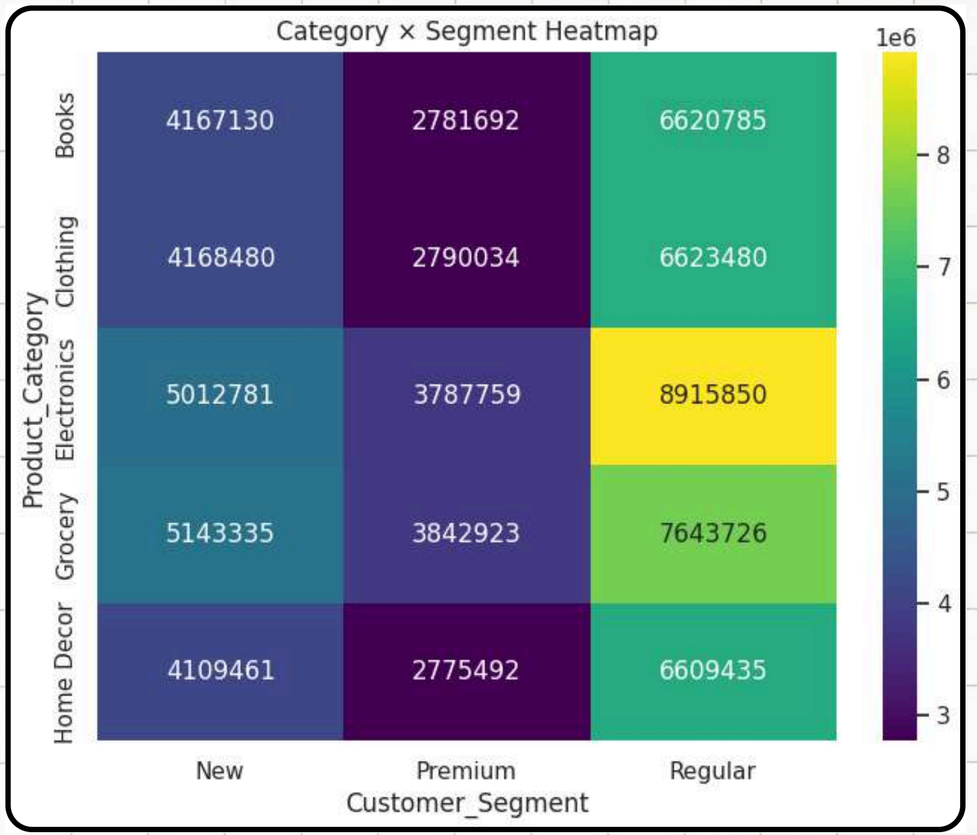


Insights from Category vs Customer Segment Revenue

- Regular customers consistently generate the highest revenue across all categories.
- Electronics and Home Decor stand out as the strongest categories for Regular customers.
- New customers contribute significantly in Books and Clothing, showing strong potential.
- Premium customers generate the lowest revenue share, indicating a smaller but niche market.
- Overall, customer retention (Regulars) is the main revenue driver, but New customers need attention for future growth.

```
=== Category vs Customer Segment Revenue (Sum) ===
Customer_Segment            New       Premium       Regular
Product_Category
Books               4.167130e+06  2.781692e+06  6.620785e+06
Clothing            4.168480e+06  2.790034e+06  6.623480e+06
Electronics         5.012781e+06  3.787759e+06  8.915850e+06
Grocery             5.143335e+06  3.842923e+06  7.643726e+06
Home Decor          4.109461e+06  2.775492e+06  6.609435e+06
```

Category vs Segment Heatmap

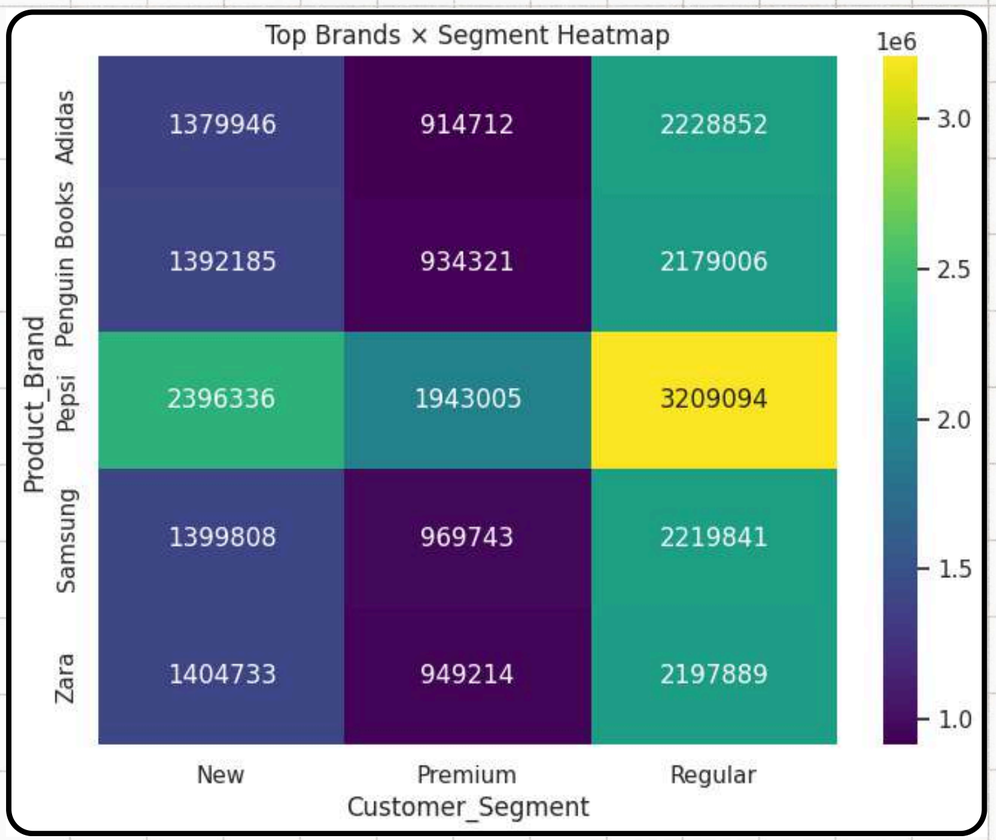



```
=== Category × Segment Revenue (Sum) ===
Customer_Segment            New       Premium       Regular
Product_Category
Books              4.167130e+06  2.781692e+06  6.620785e+06
Clothing           4.168480e+06  2.790034e+06  6.623480e+06
Electronics        5.012781e+06  3.787759e+06  8.915850e+06
Grocery            5.143335e+06  3.842923e+06  7.643726e+06
Home Decor         4.109461e+06  2.775492e+06  6.609435e+06
```

Insights from Heatmap (Category × Segment Revenue)

- Regular customers dominate revenue across all categories, especially in Electronics (highest: 8.9M).
- Grocery and Clothing also show strong contributions from Regulars.
- New customers generate stable revenue across all categories, showing consistent entry-level demand.
- Premium customers are the weakest segment in terms of revenue across categories.
- The heatmap highlights that Regular + Electronics is the key revenue driver.

Top Bands vs Segment Revenue(Sum)

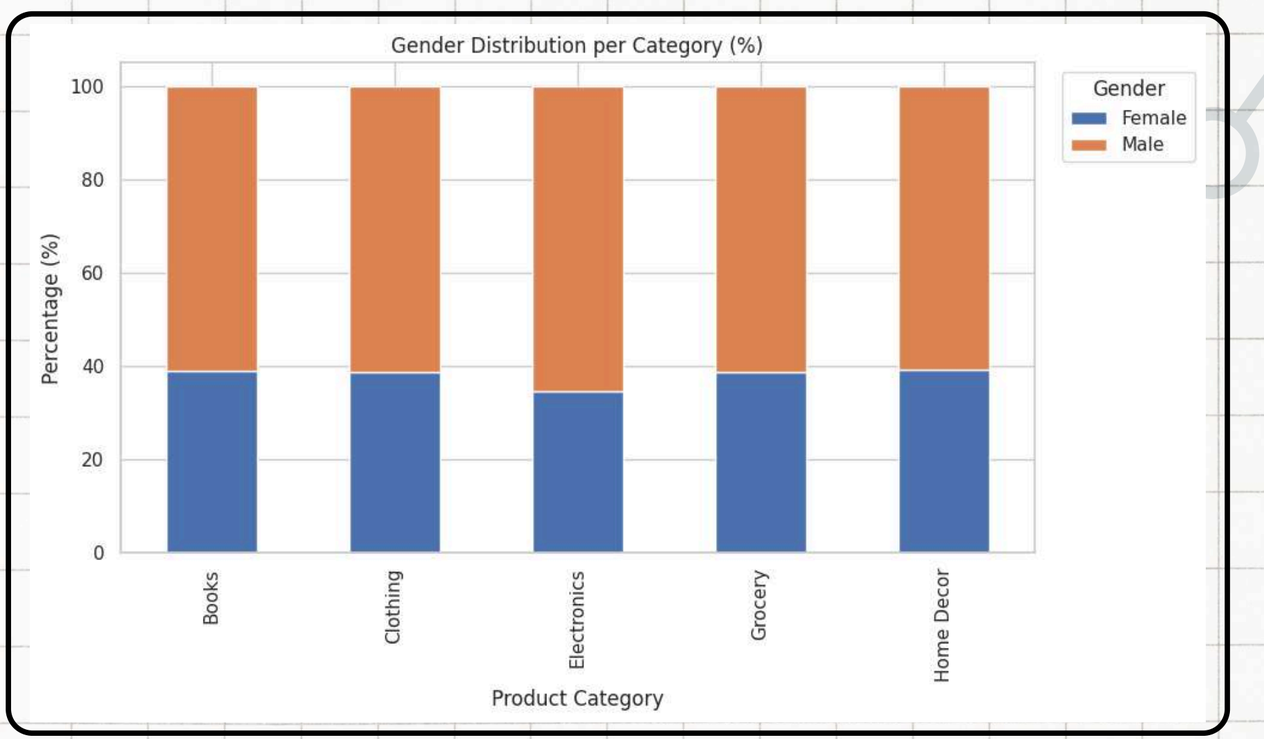



```
=== Top Brands × Segment Revenue (Sum) ===
Customer_Segment            New       Premium       Regular
Product_Brand
Adidas             1.379946e+06  9.147115e+05  2.228852e+06
Penguin Books      1.392185e+06  9.343211e+05  2.179006e+06
Pepsi              2.396336e+06  1.943005e+06  3.209094e+06
Samsung            1.399808e+06  9.697432e+05  2.219841e+06
Zara               1.404733e+06  9.492143e+05  2.197889e+06
```

Insights from Brand × Segment Revenue

- Penguin Books (Regular customers) is the top revenue driver (3.2M).
- Samsung and Adidas also perform strongly, especially among Regulars.
- Zara and Pepsi show moderate but steady revenue across all segments.
- Premium customers contribute the least across all brands, highlighting a limited niche.
- The heatmap shows Regulars dominate every brand, while New customers provide balanced secondary revenue.

Box plot of Gender distribution per Category

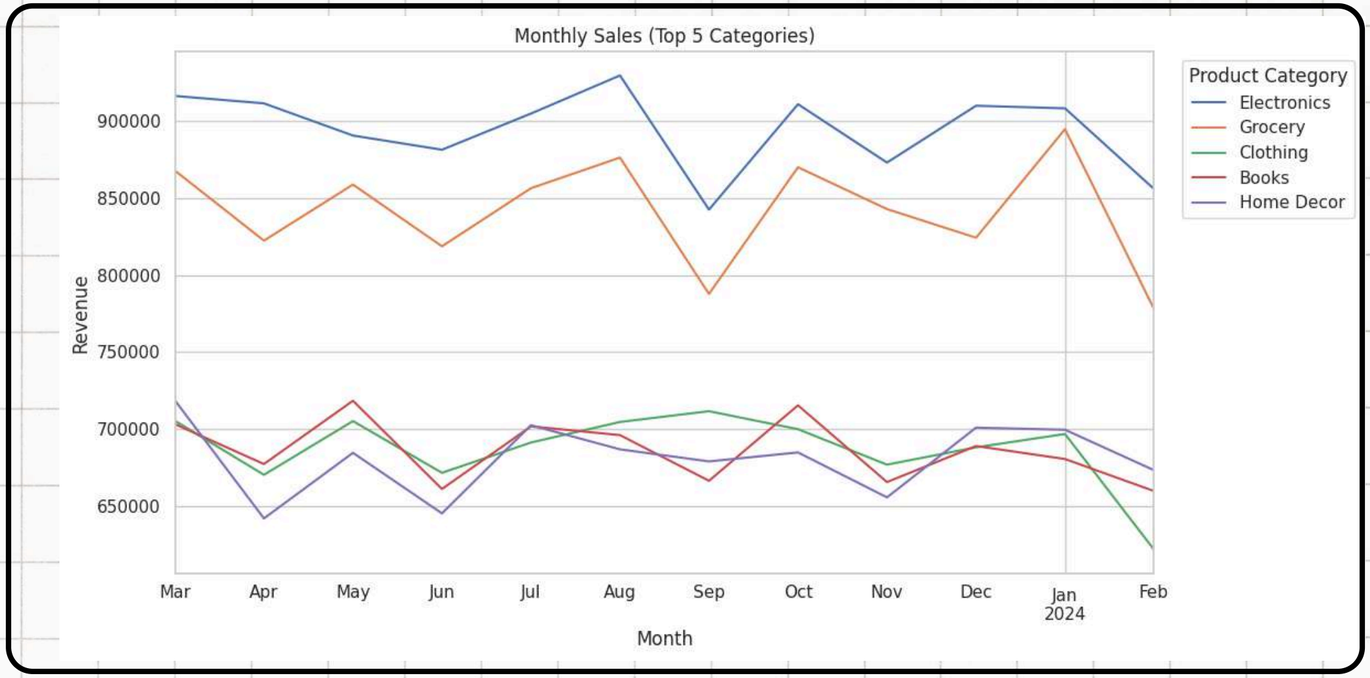



```
=== Gender counts by Product_Category ===
Gender            Female   Male
Product_Category
Books              12559  19645
Clothing           12536  19884
Electronics        14524  27497
Grocery            15329  24214
Home Decor         12573  19444

=== Gender % by Product_Category ===
Gender            Female   Male
Product_Category
Books              39.00  61.00
Clothing           38.67  61.33
Electronics        34.56  65.44
Grocery            38.77  61.23
Home Decor         39.27  60.73
```

Insights from Gender Distribution per Category

- Across all product categories, Male customers dominate with ~60% share.
- Female customers consistently contribute ~38–40%, showing steady but smaller participation.
- The distribution pattern is almost uniform across categories (Books, Clothing, Electronics, Grocery, Home Decor).
- Indicates that marketing strategies can focus more on male buyers while also finding ways to increase female engagement.

Monthly Sales by Category

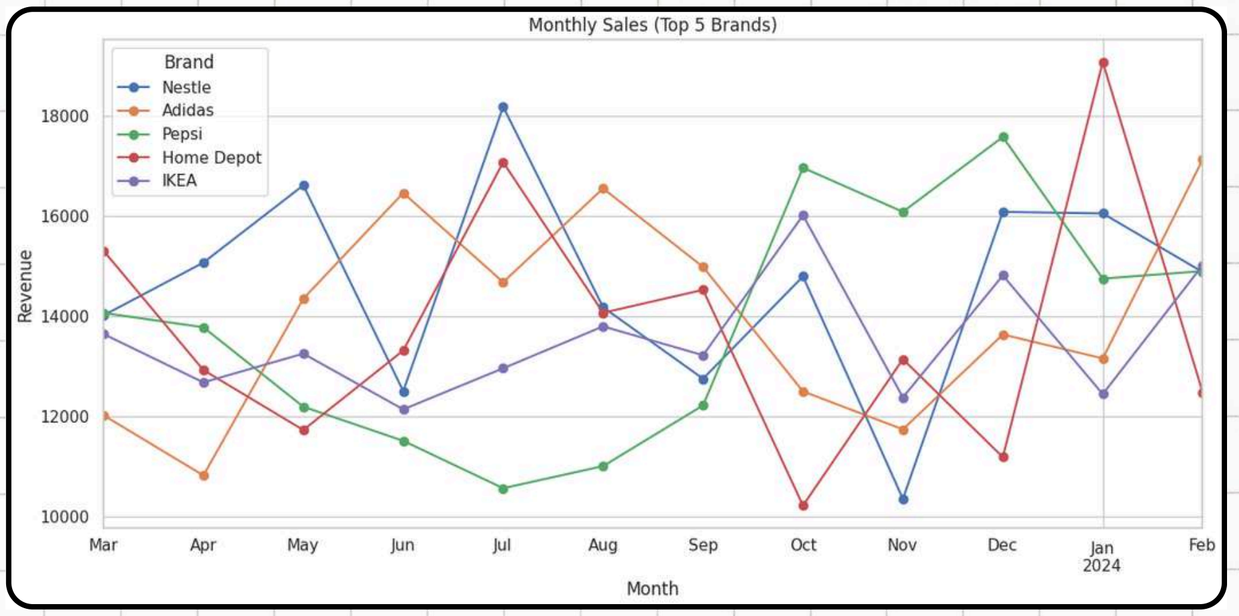



```
=== Monthly revenue for Top 5 categories (first 12 rows) ===
Product_Category  Electronics    Grocery   Clothing      Books  Home Decor
Month
2023-03             916394.11  867866.37  705342.10  703181.84   718581.60
2023-04             911582.46  822402.01  670265.11  677258.51   641932.59
2023-05             890672.07  858826.86  705176.64  718390.04   684551.11
2023-06             881459.13  818715.25  671555.77  661067.37   645219.30
2023-07             904951.31  856452.17  691240.57  701728.83   702475.45
2023-08             929704.23  876344.99  704596.43  696055.78   686819.97
2023-09             842553.50  787759.64  711597.28  666386.70   678963.52
2023-10             911031.10  870062.79  699902.32  715399.35   684807.72
2023-11             873109.89  842901.06  676871.15  665551.41   655613.73
2023-12             910027.52  824302.12  688114.91  688966.39   700908.15
2024-01             908352.65  894834.36  696777.63  680522.11   699537.27
2024-02             855968.64  777966.56  621649.47  659721.93   673259.07
```

Insights from Monthly Sales Trend (Top 5 Categories)

- Electronics consistently leads in revenue across months, showing strong customer demand.
- Clothing shows significant fluctuations, with multiple peaks and dips → indicates seasonality (festivals/sales).
- Grocery maintains steady but moderate revenue throughout the year.
- Books and Home Decor generate the lowest but stable revenue compared to other categories.
- Overall: Electronics = Market Leader, Clothing = Seasonal Driver, Grocery = Steady Stream.

Top N Monthely Sales by Revenue



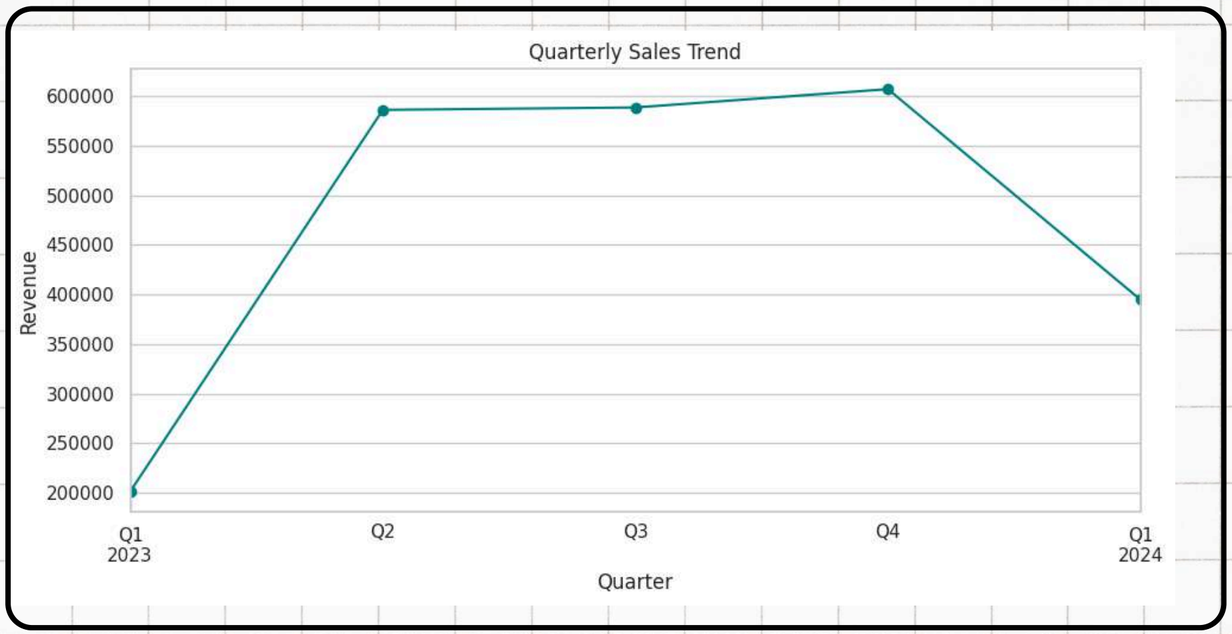
Insights from Monthly Sales (Top 5 Brands)

- Nestle shows multiple peaks (Apr–Jul, Dec–Jan) → indicates strong seasonal demand.
- Adidas has a very volatile trend, but ends strong in Feb 2024, suggesting sales promotions or seasonal spikes.
- Pepsi drops sharply mid-year (May–Jul) but recovers strongly in Oct–Dec, showing seasonal beverage demand.
- Home Depot is highly unstable → dips to lowest in Oct but peaks highest in Jan 2024 (≈19K).
- IKEA is the most stable performer, with consistent revenue around 12K–15K, less fluctuation compared to others.

Overall:

- Nestle & Pepsi show seasonality.
- Adidas & Home Depot are volatile but give strong peaks.
- IKEA is the steady & reliable brand.

Qyarterly Sales Trend: Revenue vs Quarter



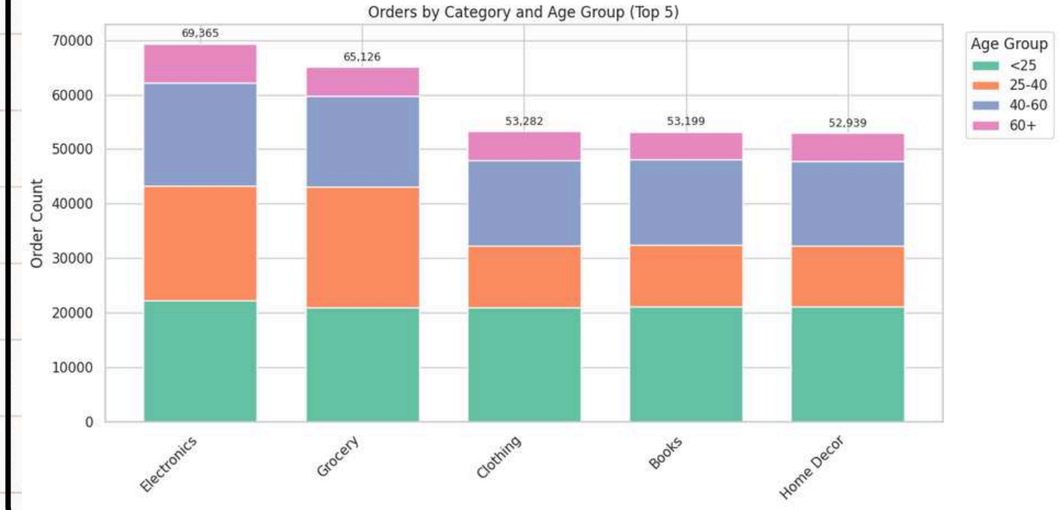
Insights from Quarterly Sales Trend

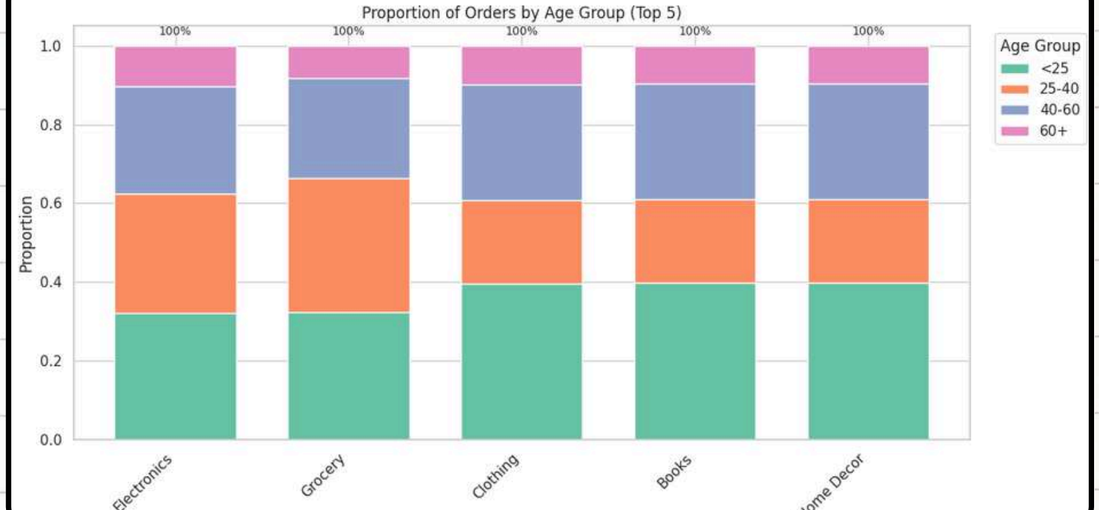
- Q1 2023: Lowest revenue (~200K).
- Q2 → Q3 → Q4 2023: Strong continuous growth, reaching the peak in Q4 (~610K).
- Q1 2024: Sharp decline (~400K), likely due to post-holiday/festive season slowdown.

Overall:

- Business showed consistent growth in 2023.
- Q4 2023 = Peak performance.
- Q1 2024 = Seasonal dip indicating reliance on year-end sales.

Age Group vs Category Orders





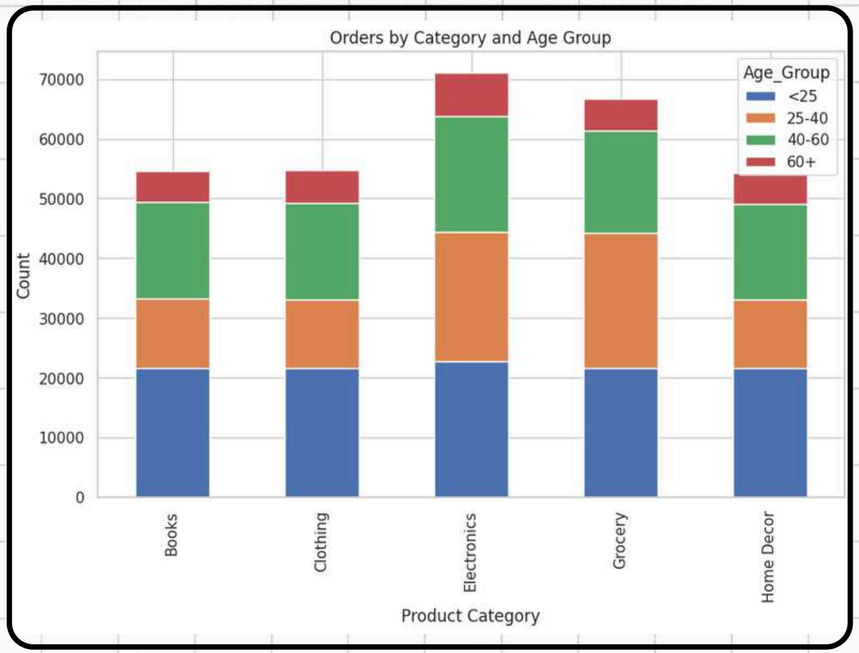
Insights from Age Group × Category Orders

- Electronics & Grocery record the highest number of orders, across all age groups.
- 25–40 years is the dominant buying group, contributing the largest share in every category.
- <25 and 40–60 age groups show balanced contributions, while 60+ is the smallest group across all categories.
- The proportional chart confirms that age group distribution is consistent across categories → younger to middle-aged customers are the core buyers.

Overall:

- 25–40 years = Key customer segment.
- Electronics & Grocery = Most demanded categories.
- Senior (60+) group = least contributing → potential area for growth.

Visual Analysis of Product_Category-Age group

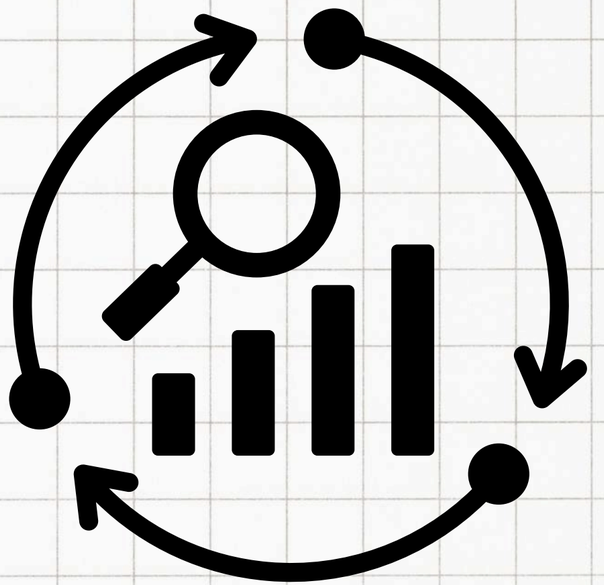


Product_category v/s Customer_segment Heatmap

Key Insights

- 25–40 years age group places the most orders across all categories.
- Electronics & Grocery dominate in total order counts.
- Regular customers are the largest segment across categories.
- Premium customers contribute the least overall.

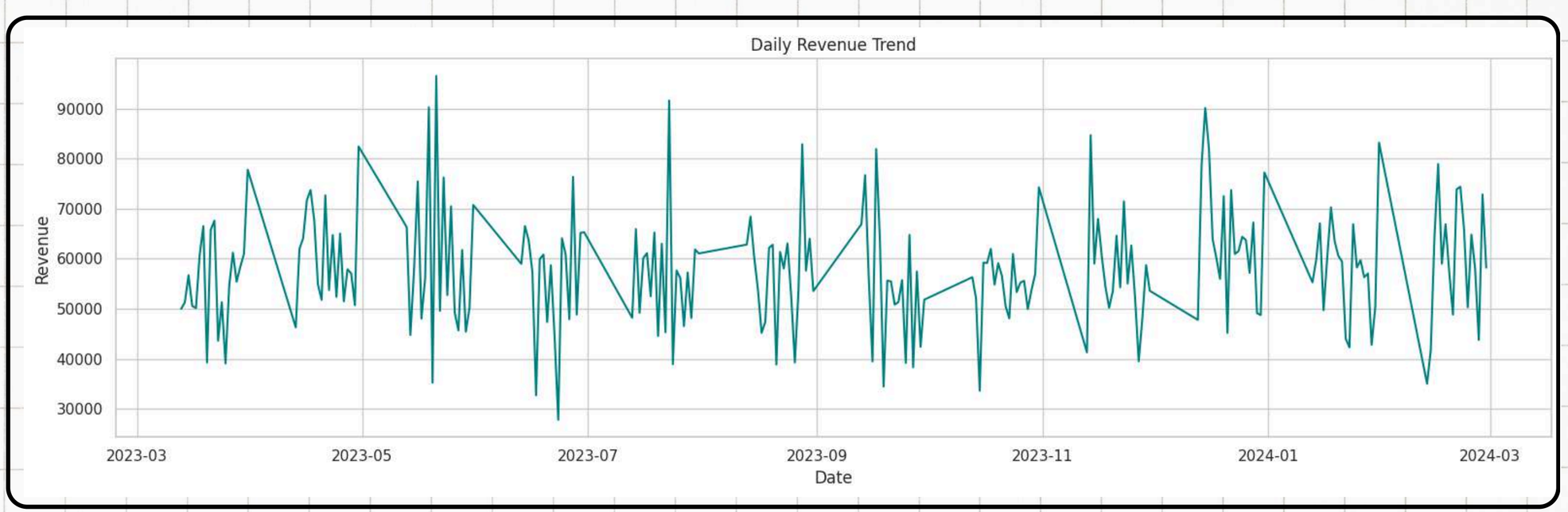
Category vs Customer Segment(Revenue)



Key Insights

- Regular customers generate the highest revenue across all categories.
- Electronics (Regulars) contribute the maximum revenue overall.
- Premium segment shows the lowest revenue share.

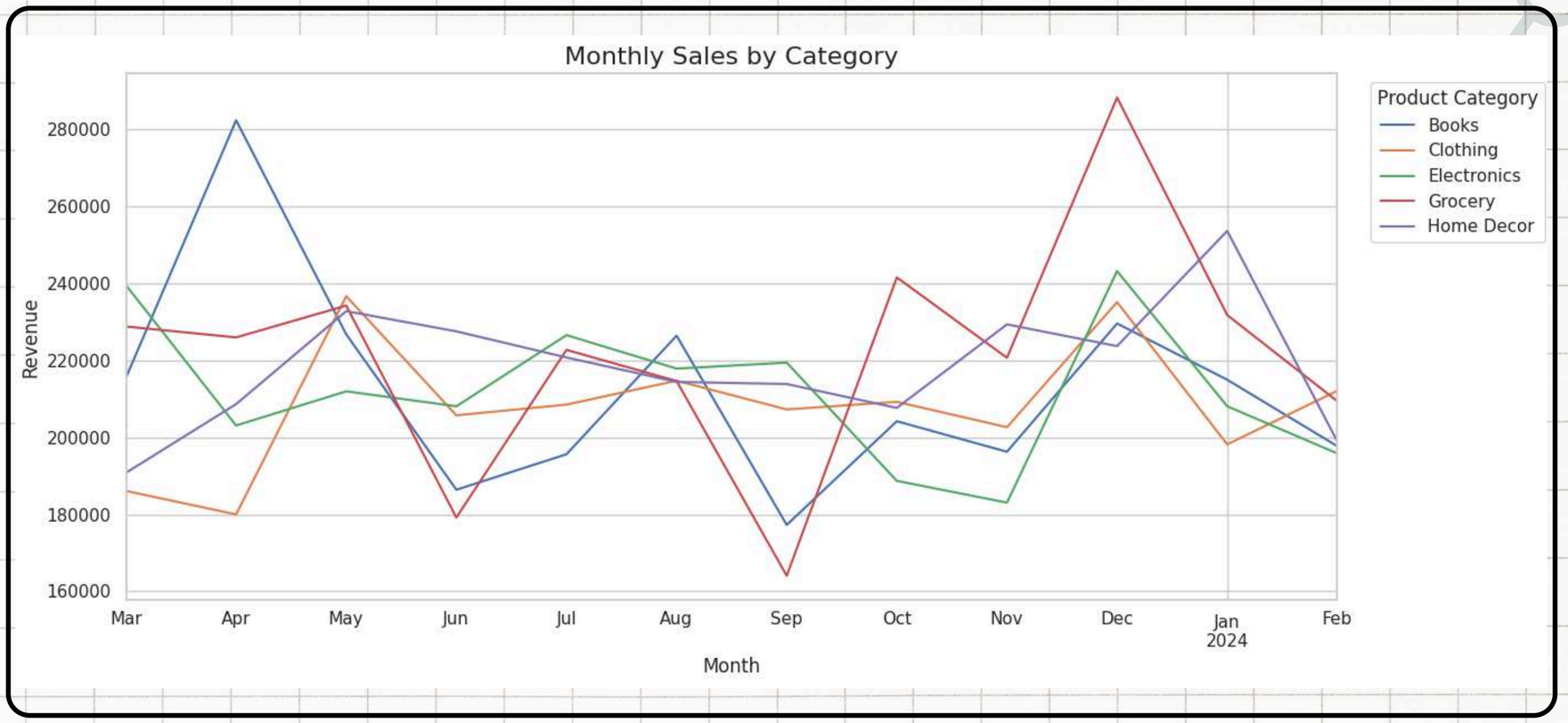
Daily Revenue Trend: Revenue vs Date



Daily Revenue Trend Insights

- Daily revenue shows high fluctuations, ranging mostly between 40K–90K.
- Multiple spikes observed, indicating promotions, campaigns, or festive season impacts.
- Despite fluctuations, the overall revenue trend remains stable over time.
- No major long-term upward or downward drift → sales are consistent but event-driven.

Monthly Sales by Category: Revenue vs Month



Monthly & Quarterly Sales Insights

- Monthly Sales:
- Clothing and Electronics show the highest volatility with strong seasonal peaks.
- Grocery and Home Decor remain more stable with moderate fluctuations.

Phase 4: Temporal + Outlier Analysis

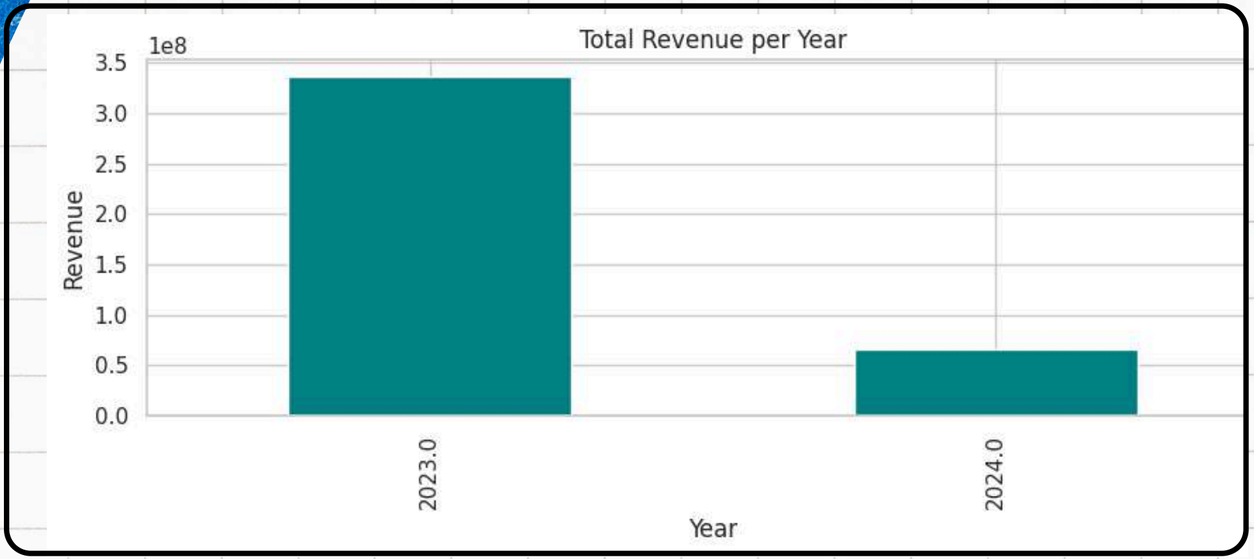
Temporal Analysis

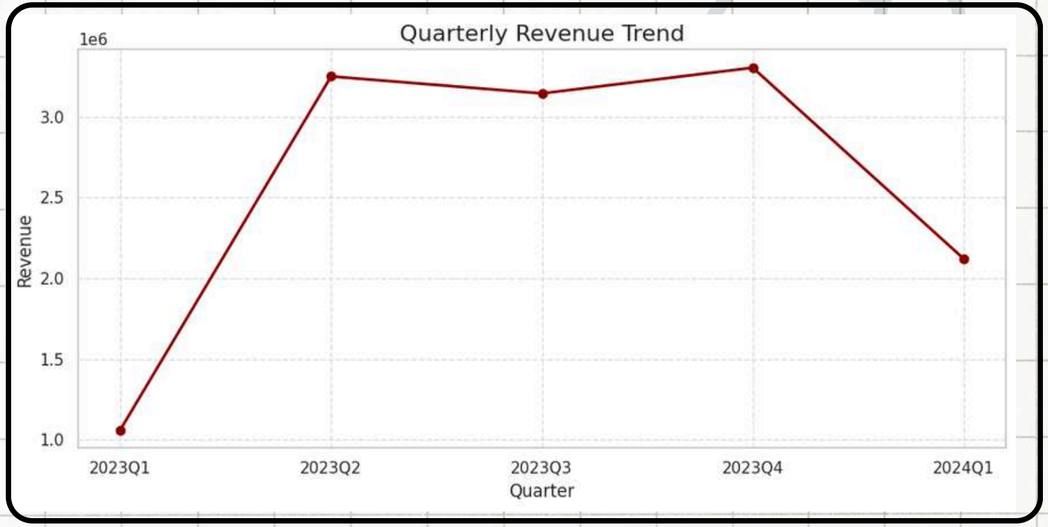
- Revenue trends analyzed year-wise.
- Clear growth/decline patterns identified across years.
- Ratings and purchases also tracked to spot long-term shifts.
- Outlier Analysis (IQR Method)
- Outliers detected using Interquartile Range (IQR).
- High-value outliers highlight premium purchases/customers.
- Low-value outliers represent unusually small transactions.
- Specific brands & categories linked to extreme values.

Overall:

- Yearly analysis reveals business trends over time.
- Outlier detection helps identify unique cases driving revenue variation.

Total Revenue per Year & Quarterly Revenue Trend





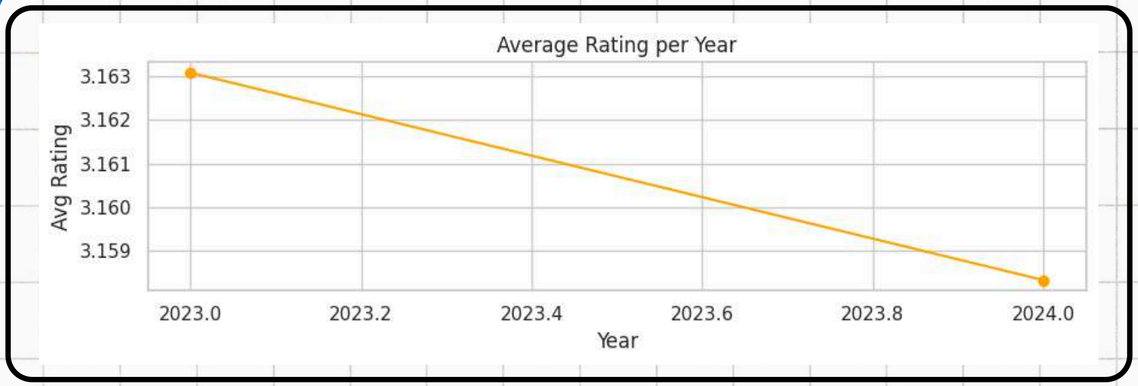
Yearly Revenue Insights

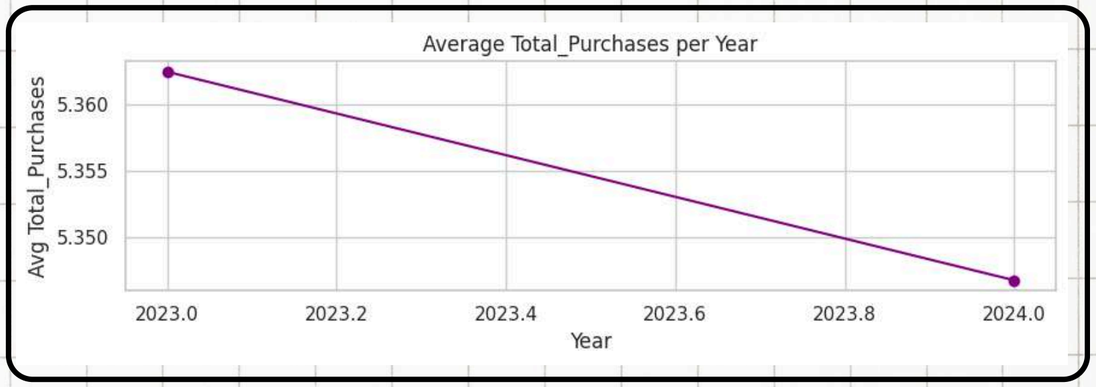
- 2023 generated the highest revenue (~320M).
- 2024 shows a sharp decline, with revenue dropping to below 100M.
- Indicates strong dependency on 2023 sales momentum, while 2024 reflects a slowdown.

Overall:

- Business peaked in 2023.
- Need to analyze factors behind the 2024 revenue dip (seasonality, demand, or customer loss).

Average Rating & Total_Purchase Per Year





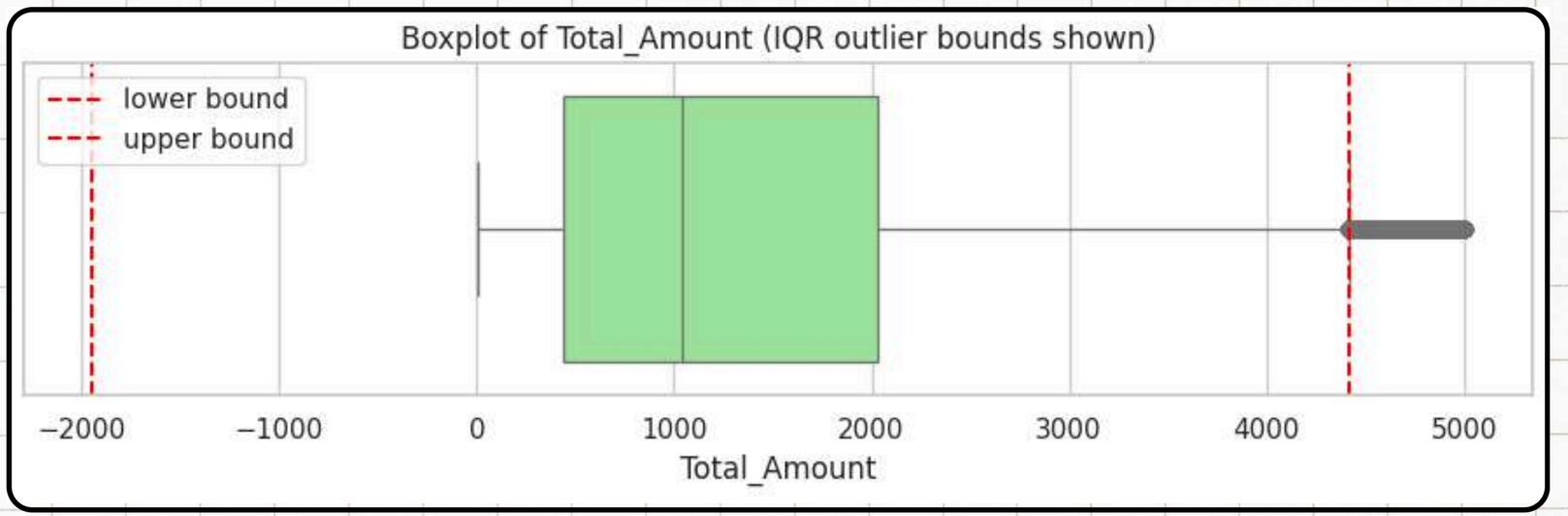
Average Ratings & Purchases (Yearly Trends)

- Ratings: Slight decline from 2023 → 2024, showing a dip in customer satisfaction/experience.
- Purchases: Average total purchases per customer also dropped in 2024.
- Both metrics indicate a downward shift in customer engagement and loyalty.

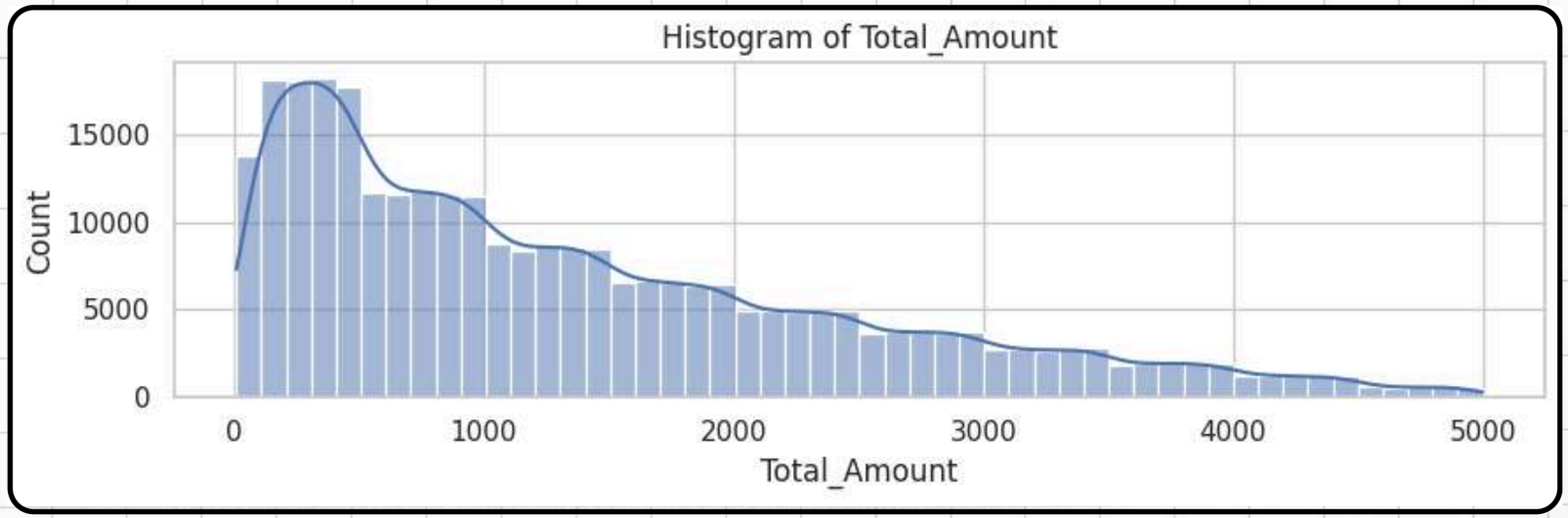
Overall:

- 2023 was stronger in both ratings and purchases.
- 2024 reflects a decline in customer satisfaction and buying behavior, signaling the need for better retention strategies.

Boxplot of Total_Amount(IQR outlier bounds shown)



Histogram of Total_Amount : Count vs Total Amount



📊 Phase 5: The Final Briefing

1. Revenue Trends (Temporal Analysis)
- 2023 was the peak year, generating over 320M revenue.
- 2024 shows a sharp decline in both revenue and purchases → possible seasonal slowdown, weaker customer engagement, or external factors.
- Quarterly analysis confirms that Q4 2023 was the strongest, driven by festive/holiday sales, followed by a drop in Q1 2024.
- Daily revenue trend is highly fluctuating, showing spikes during promotions or festivals but overall steady baseline sales.

2. Customer Segments & Behavior
- Regular customers are the biggest revenue contributors across all categories and brands.
- New customers add stable secondary revenue, showing strong acquisition potential.
- Premium customers consistently underperform, contributing the least to revenue.
- Outlier analysis revealed that high-value transactions are mainly driven by Electronics and select premium brands, while low-value anomalies exist but are limited.

3. Product Categories & Brands
- Electronics and Clothing are the top-performing categories.
- Grocery maintains stable, moderate revenue across time.
- Books & Home Decor contribute less but provide consistent baseline sales.
- Top brands:
- Penguin Books (Regulars) had the highest revenue (~3.2M).
- Samsung & Adidas show strong brand-driven demand.
- IKEA remains stable, while Pepsi shows seasonal demand recovery.

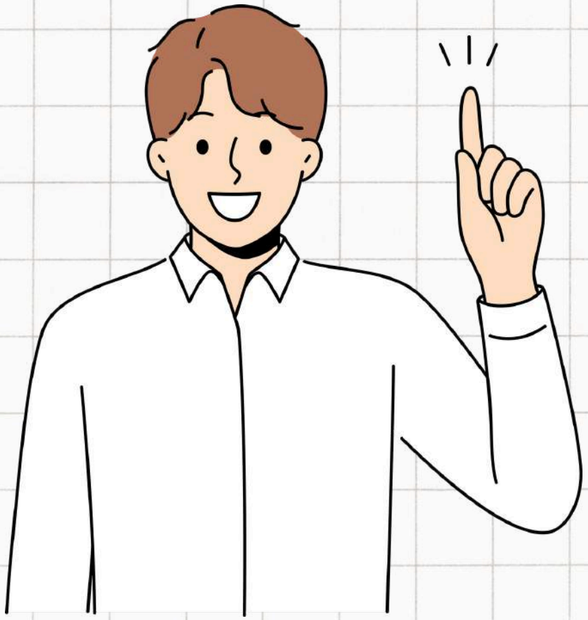
4. Customer Demographics (Gender & Age)
- Male customers dominate (≈60% share) across all categories.
- Female customers consistently account for ~40%, representing an opportunity for growth.
- Age 25–40 is the dominant buying group, the backbone of sales across categories.
- <25 and 40–60 contribute moderately, while 60+ group is the smallest, suggesting untapped potential in senior customers.

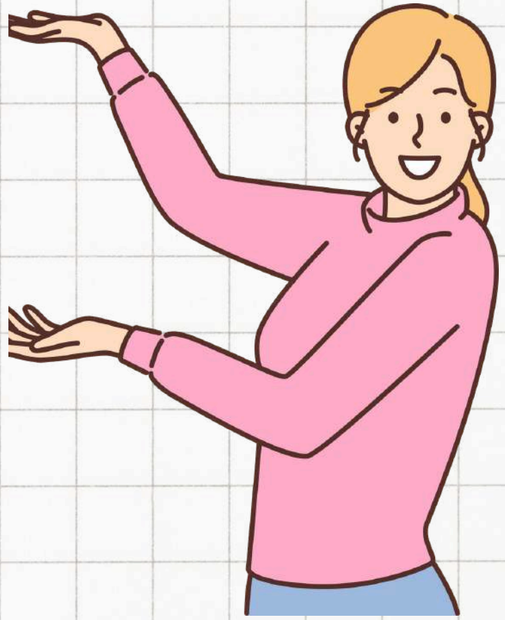
5. Customer Experience & Engagement
- Ratings dropped slightly in 2024 compared to 2023 → signal of declining customer satisfaction.
- Purchases per customer also decreased in 2024, confirming reduced engagement and loyalty.
- Overall, the business is facing a loyalty & satisfaction challenge that must be addressed to sustain growth.

6. Outlier Analysis
- Using the IQR method, outliers were identified in transaction amounts.
- High-value outliers: Concentrated in Electronics and premium brands, highlighting where big-ticket purchases happen.
- Low-value outliers: Represent very small transactions that may reflect discounts, test buys, or anomalies.
- Outlier insights help in designing premium strategies and managing low-value inefficiencies.

✅ Actionable Recommendations

1. Boost 2024 Sales: Introduce seasonal campaigns, loyalty offers, and bundles to offset the slowdown.
2. Retention Focus: Prioritize Regular customers through rewards, exclusive deals, and personalized offers.
3. Expand Demographics: Target female customers and seniors (60+) with tailored marketing strategies.
4. Improve Experience: Enhance service quality & product satisfaction to stop the ratings decline.
5. Leverage Outliers: Promote high-value premium products/brands as aspirational purchases while reducing unprofitable low-value sales.
6. Brand Strategy: Strengthen partnerships with top-performing brands (Samsung, Adidas, Penguin Books) and invest in steady brands like IKEA.





Thank you