E-commerce Customer Data

For

Behaviour Analysis

Introduction

This report presents an exploratory data analysis (EDA) of the E-commerce Customer Behaviour and Purchase Dataset, a synthetic dataset generated using the Faker Python library. The dataset captures various facets of customer behaviour and purchase history within a digital marketplace, offering insights for data analysis and predictive modeling in e-commerce. Key applications of this dataset include customer churn prediction, market basket analysis, recommendation systems, and trend analysis.

The analysis begins with data cleaning to ensure dataset integrity by addressing missing values and verifying data types for accurate analysis. Feature engineering is then employed to enrich the dataset, allowing deeper exploration of customer behaviours and purchase patterns.

Using statistical techniques and visualization tools, this report aims to identify trends in purchase behaviour, spending patterns, and customer retention. Additionally, it examines relationships between variables such as customer demographics, product categories, and payment methods. Through this exploratory analysis, we aim to uncover actionable insights that can inform strategies for customer retention and revenue growth in e-commerce.

Data Overview

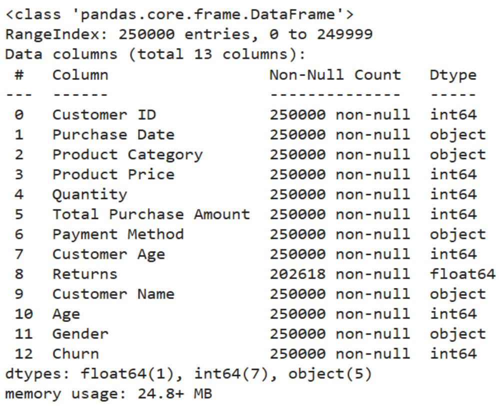
The E-commerce Customer Behaviour and Purchase Dataset provides a synthetic yet comprehensive view of customer interactions and transactions within a digital marketplace. Below is a detailed description of the dataset's structure and columns:

1. Customer ID: A unique identifier for each customer, enabling individual-level analysis and tracking.
2. Customer Name: The name of the customer, generated using the Faker library, providing context and personalization.
3. Customer Age: The age of the customer, which helps in demographic segmentation and understanding age-based trends in purchases.
4. Gender: The gender of the customer, allowing for gender-based analysis of buying patterns.
5. Purchase Date: The date of each purchase made by the customer, crucial for analyzing temporal trends in purchasing behaviour.

6. Product Category: The category or type of the purchased product, enabling classification and analysis of product preferences.
7. Product Price: The price of the purchased product, a key metric for revenue and pricing strategy evaluation.
8. Quantity: The quantity of the product purchased, reflecting the scale of individual transactions and bulk purchase patterns.
9. Total Purchase Amount: The total amount spent by the customer in each transaction, a critical variable for revenue analysis and customer value assessment.
10. Payment Method: The method of payment used by the customer (e.g., credit card, PayPal), offering insights into payment preferences and trends.
11. Returns: A binary column indicating whether the customer returned any products from the order (0 for no return, 1 for return), useful for understanding return rates and their impact on profitability.
12. Churn: A binary column representing whether the customer has churned (0 for retained, 1 for churned), essential for customer retention analysis and predictive modeling.

Dataset Characteristics

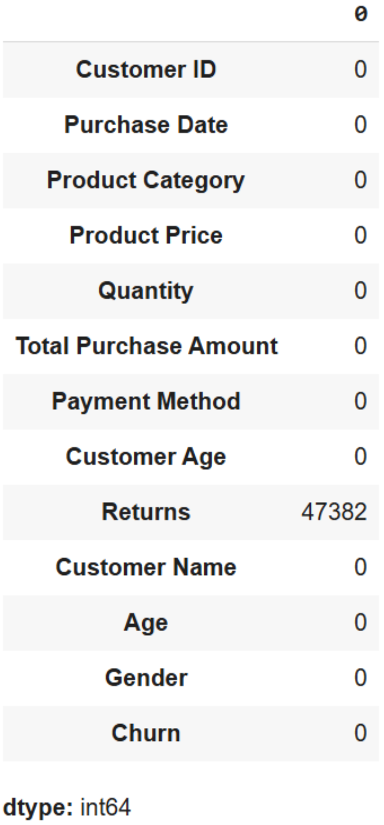
- **Total Entries:** 250,000
- **Total Columns:** 13
- **Non-null Counts:** Most columns contain complete data; however, certain columns, such as Returns, have missing values, indicating potential areas for further data cleaning or imputation to ensure analytical accuracy.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                 Non-Null Count   Dtype  
---  ------                 --------------   -----  
 0   Customer ID            250000 non-null  int64  
 1   Purchase Date          250000 non-null  object 
 2   Product Category       250000 non-null  object 
 3   Product Price          250000 non-null  int64  
 4   Quantity               250000 non-null  int64  
 5   Total Purchase Amount  250000 non-null  int64  
 6   Payment Method         250000 non-null  object 
 7   Customer Age           250000 non-null  int64  
 8   Returns                202618 non-null  float64
 9   Customer Name          250000 non-null  object 
 10  Age                    250000 non-null  int64  
 11  Gender                 250000 non-null  object 
 12  Churn                  250000 non-null  int64  
dtypes: float64(1), int64(7), object(5)
memory usage: 24.8+ MB
```

Data Cleaning Process

1. **Handling Missing Values:** The Returns column contained null values, which were addressed by either removing rows using `dropna()` or imputing them as "No Returns" when applicable. This ensured the dataset remained consistent and free from ambiguous records.



	0
Customer ID	0
Purchase Date	0
Product Category	0
Product Price	0
Quantity	0
Total Purchase Amount	0
Payment Method	0
Customer Age	0
Returns	47382
Customer Name	0
Age	0
Gender	0
Churn	0

dtype: int64

2. **Checking for Duplicates:** A thorough check confirmed that no duplicate entries were present in the dataset. This finding guarantees that the data is not artificially inflated, ensuring accurate analysis.
3. **Data Type Conversion**: The Purchase Date column, initially stored as an object, was converted to the datetime format using `pd.to_datetime(df['Purchase Date'])`. This change enables temporal analyses, such as identifying peak purchase periods, seasonal trends, and monthly or weekly sales distributions.

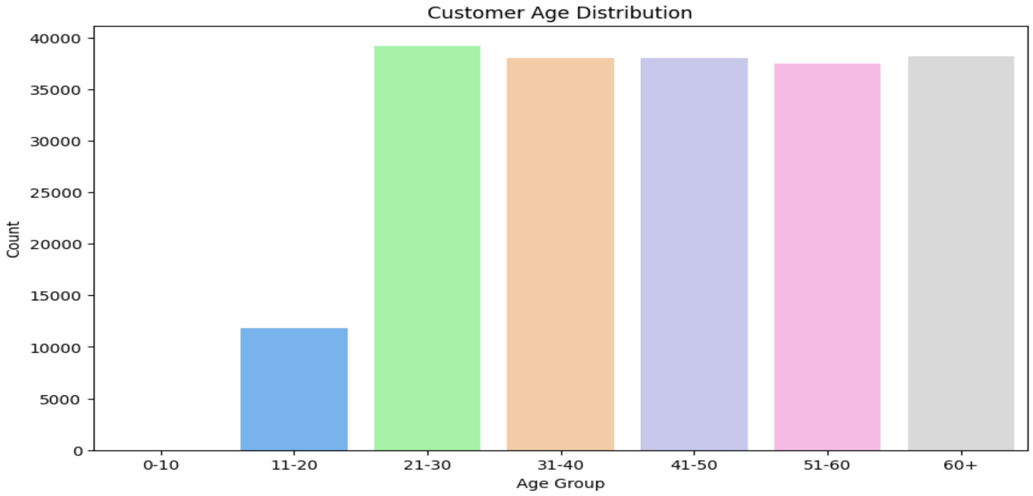
Feature Engineering

- Extracted Temporal Features: Derived Year, Month, and Day from the Purchase Date column to capture detailed temporal information for identifying seasonal trends in purchasing behaviour.
- Analyzed Weekly Patterns: Extracted the Day of Week from the Purchase Date, assigning numerical values (0 for Monday to 6 for Sunday) to identify purchasing patterns across the week.
- Classified Weekends: Created an Is Weekend feature to distinguish weekend purchases from weekdays, allowing insights into differences in buying behaviour.

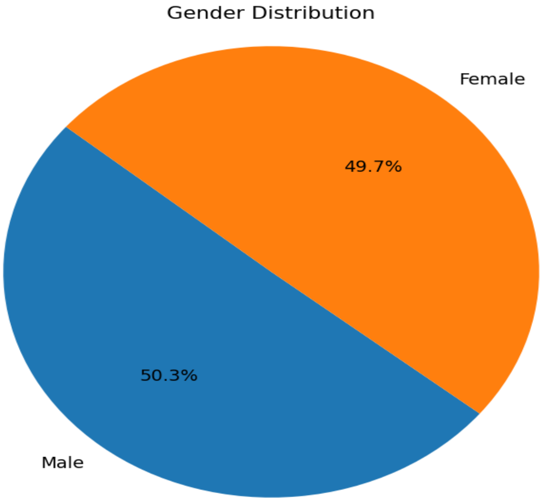
Data Analysis

1. Demographic and Revenue Analysis

- **Age Group Analysis:**
  - The 0-10 age group has no customers, while the 11-20 age group has fewer customers.
  - The 21-30 age group has the highest count of customers, significantly outpacing the others.
  - The above 30 age groups show similar levels of engagement.
  - Key Insight: The largest segment is customers aged 21–30, with the rest showing comparable counts among older age groups.



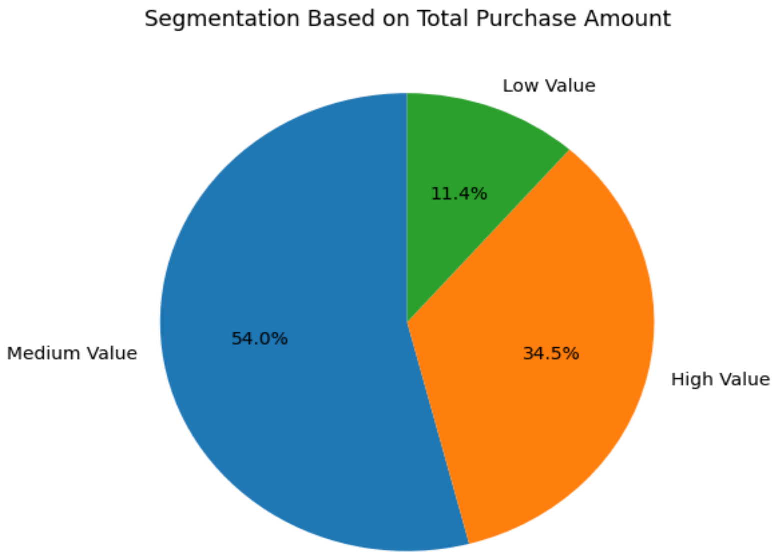
- **Gender Distribution Analysis:**
    - The distribution of customers is almost balanced, with 50.3% identifying as Male and 49.7% as Female.
    - This slight male majority suggests that marketing strategies are effective for both genders, but could also be refined to ensure that female customers are equally targeted.
    - **Key Insight:** There is a nearly even split between male and female customers, with males slightly ahead at 50.3%.



- **Total Revenue:**
    - The total revenue generated amounts to $552,299,723.00. This substantial figure reflects the company's successful sales strategy and customer engagement.

2. Segmentation Based on Total Purchase Amount

- **Customer Segment Distribution:**
    - The customer segmentation reveals that there are 109,494 medium-value customers, accounting for 54.0% of the total. High-value customers number 70,004, making up 34.5%, while low-value customers total 23,120, representing 11.4% of the customer base.



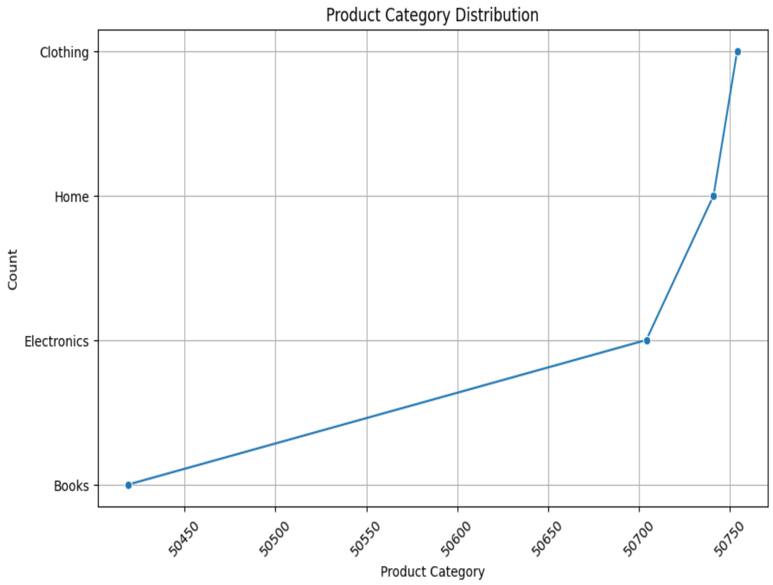
- **Key Insights:** The Medium Value category leads with 54.0% (109,494 customers), offering a significant opportunity for upselling and cross-selling to elevate them to High Value. The High Value segment, representing 34.5% (70,004 customers), is a key revenue contributor. Meanwhile, the Low Value segment (11.4%) highlights the need for enhanced customer retention and engagement strategies to convert these customers into higher value segments.

3. Product Category Analysis

- **Product Category Distribution:**

Product Category	**Count**
Clothing	50754
Home	50741
Electronics	50704
Books	50419

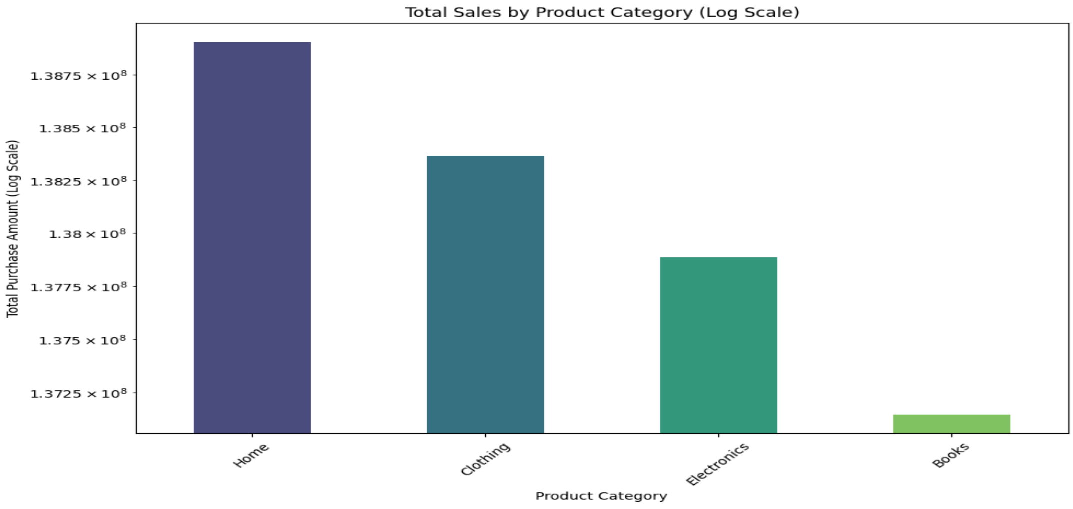
- The line plot visualizes the distribution of products across different categories, showing that the counts are quite similar across categories, with Clothing and Home leading slightly.



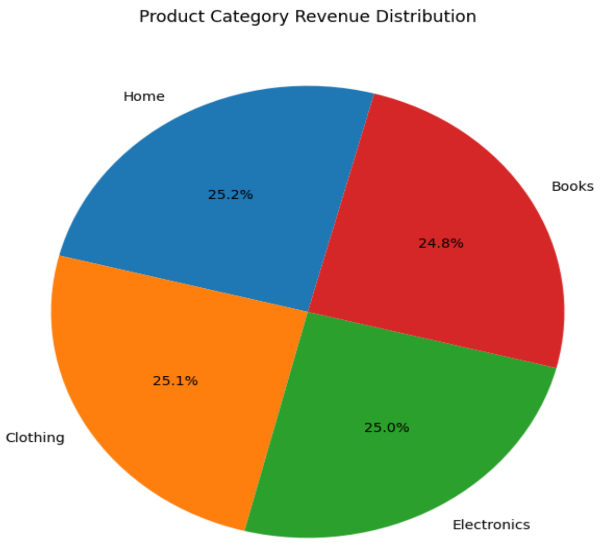
- **Total Sales by Product Category:**

Product Category	**Total Purchase Amount**
Home	138906211
Clothing	138363540
Electronics	137884886
Books	137145086

  - The bar plot displays total sales with a log scale, indicating that Home products generate the highest revenue, followed closely by Clothing and Electronics. The log scale helps in visualizing the small differences between categories more effectively.



- **Product Category Revenue Distribution:**
  - The pie chart reveals the revenue contribution of each product category, showing a balanced distribution among them, with Home products leading slightly.

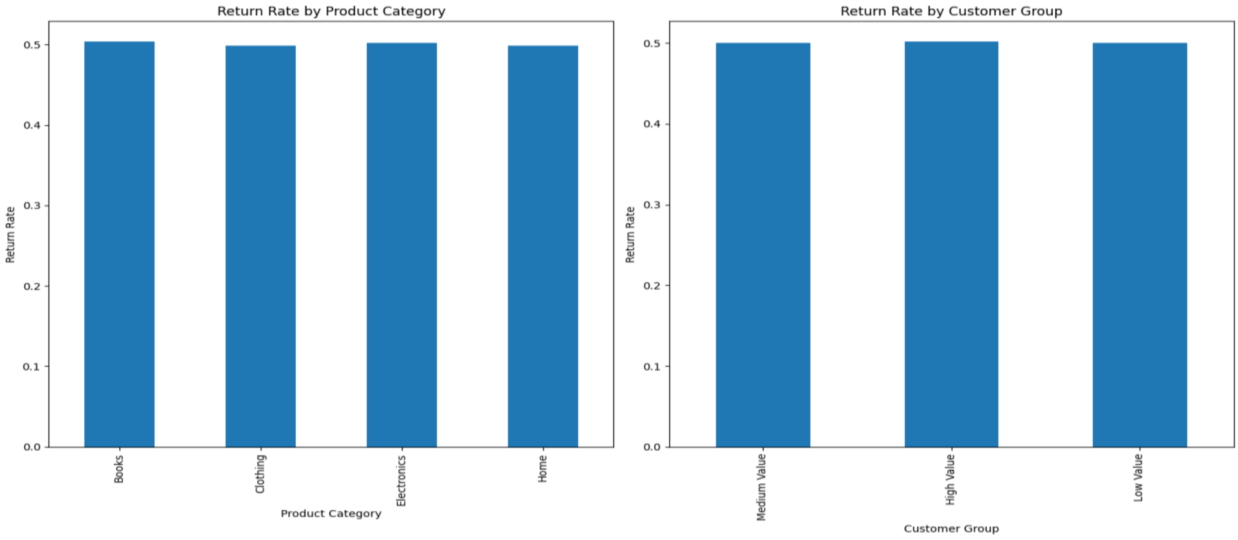


- **Key Insights:** Home products are the highest revenue contributors, indicating strong demand and marketing potential. Competitive sales in Clothing and Electronics suggest promotional strategies could enhance returns, while targeted marketing for Books may improve their performance amid balanced category contributions.

4. Returns Analysis

- **Return Rate:**
  - The overall return rate is 50.08%, indicating that more than half of the purchases are returned. This high percentage is significant as it impacts both revenue and customer satisfaction.
- **Total Lost Revenue Due to Returns:**
  - The total lost revenue from returns amounts to $276,753,766.00. This figure represents a substantial financial loss that could be mitigated by addressing the factors contributing to returns.

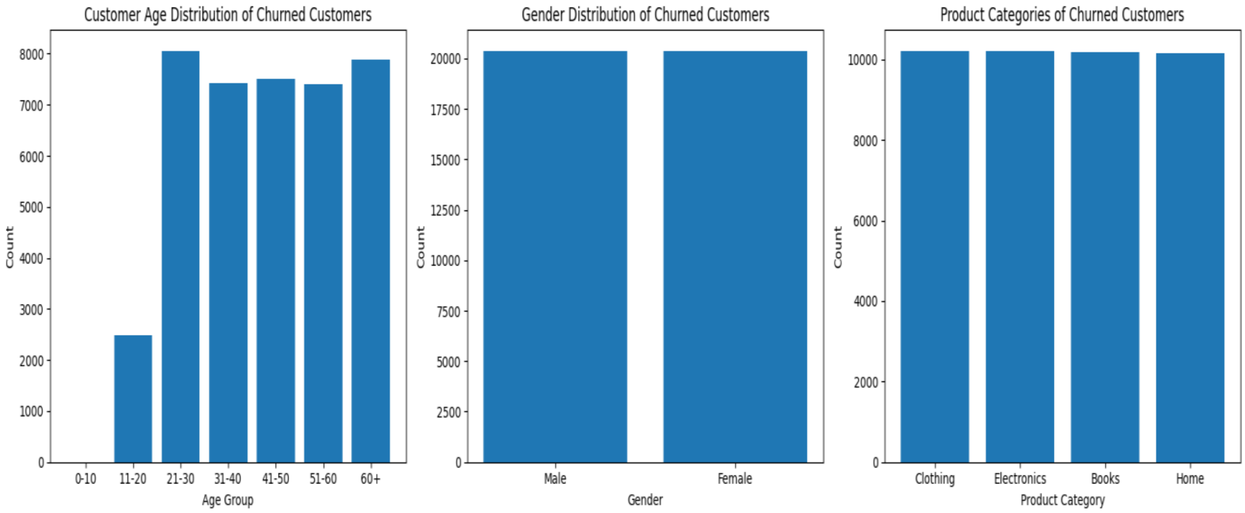
- **Return Rates by Product Category:**
        - Books: 50.39% (highest return rate)
        - Electronics: 50.19%
        - Clothing: 49.85%
        - Home: 49.90%
    - Return rates are fairly consistent across all categories, indicating that improvements to reduce returns should apply broadly rather than targeting specific product categories.
- **Return Rates by Customer Segments:**
        - High-Value Customers: 50.19% (highest return rate)
        - Medium-Value Customers: 50.03%
        - Low-Value Customers: 50.01%
    - High-value customers, despite their importance to the business, are returning a significant portion of their purchases. This points to potential dissatisfaction or unmet expectations within this segment



- **Key Insights:** The return rate exceeds 50%, indicating a significant area for improvement to enhance revenue and customer loyalty. This consistent issue across categories, coupled with higher return rates among high-value customers, suggests unmet expectations and highlights the need for strategies to reduce returns and improve customer experiences.

5. Churn Analysis

- **Churn Rate:**
  - The churn rate is calculated at 20.11%, indicating that about one-fifth of our customers have stopped making purchases. This metric is crucial for evaluating our customer retention efforts and identifying areas for improvement.
- **Understanding Churned Customers:**
  - We filter the dataset to analyze customers with a churn status of 1. This allows us to focus on the characteristics of churned customers, including their total purchase amounts, age, gender, and product category preferences.
- **Customer Age Distribution:**
  - Analyzing the age distribution reveals how churn varies across different age groups. Identifying age groups with higher churn rates can help tailor retention strategies to these demographics.
- **Gender Distribution:**
  - Examining churned customers by gender provides insights into whether certain genders are more likely to churn. This analysis can inform marketing and engagement strategies aimed at retaining these customer segments.
- **Product Category Analysis:**
  - Understanding which product categories are commonly associated with churned customers helps identify specific offerings that may not meet customer expectations. This can guide product improvements and targeted marketing efforts.



6. Customer Lifetime Value (CLV) Analysis

- **CLV Calculation:**
  - The **Customer Lifetime Value (CLV)** is a crucial metric that estimates the total revenue a customer is expected to generate throughout their relationship with the business.
  - The calculated CLV is **$13,555.36**, indicating the average expected value per customer over their lifetime with the platform.
- **Segmenting Customers for CLV Analysis:**
  - Customers are segmented based on behaviour and characteristics to better understand their CLV. The following CLV values are calculated for each segment:

Customer Segment	**CLV**
High Value	$21,679.53
Medium Value	$10,690.43
Low Value	$2,524.58

  - Each segment shows significant potential for long-term revenue generation, with all segments having CLV values above **$1,000**.
- **Key Insights:** Segmenting customers into 'High-Value,' 'Mid-Value,' and 'Low-Value' groups revealed that 'High-Value' customers have the highest Customer Lifetime Value (CLV), making them a top priority for retention and engagement efforts.

7. Payment Method Analysis

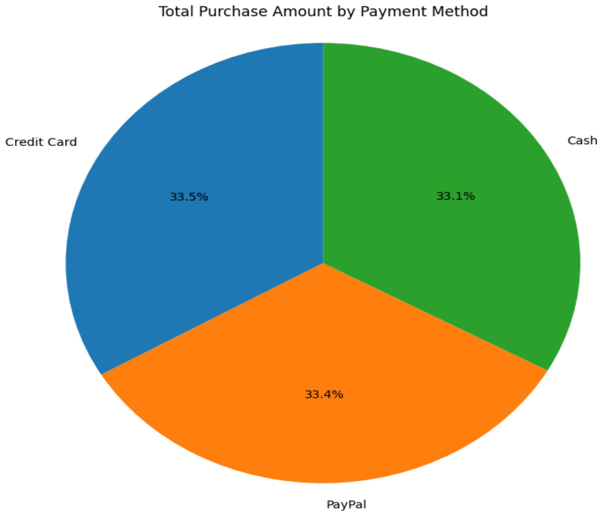
- Payment Method Distribution:
  - The analysis reveals the following counts of transactions by payment method:

Payment Method	Count
PayPal	67,811
Credit Card	67,517
Cash	67,290

- Total Purchase Amount by Payment Method:
  - The total purchase amounts aggregated by payment method are:

Payment Method	Total Purchase Amount
Credit Card	$184,799,917
PayPal	$184,585,493
Cash	$182,914,313

  - This data indicates that while the counts of transactions for PayPal and Credit Card are similar, Credit Cards slightly edge out in total sales volume.

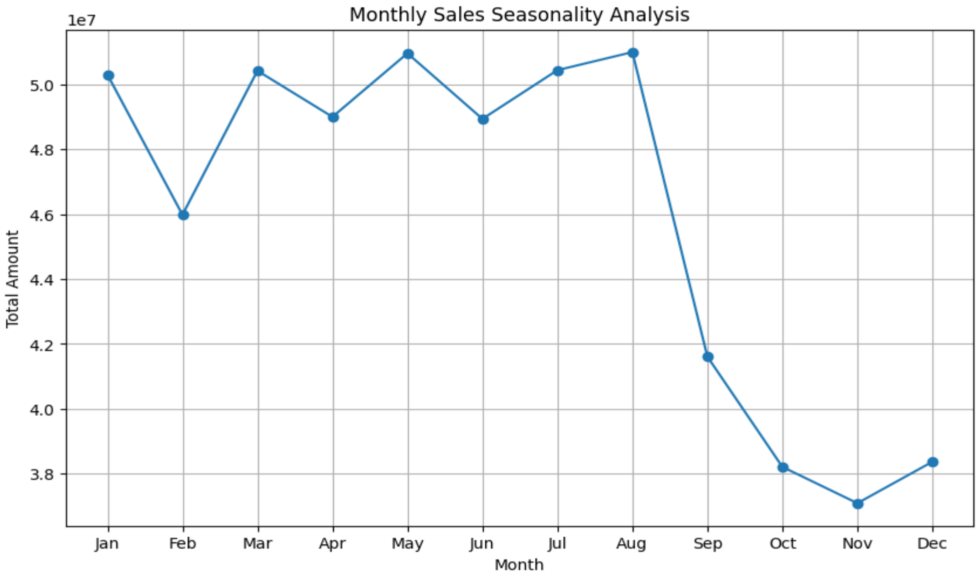


- **Key Insights:** Customers show a strong preference for PayPal and Credit Cards for transactions, which could inform future marketing and promotional strategies. The slightly higher revenue from Credit Cards suggests potential opportunities for further incentivizing credit card use (e.g., through loyalty programs or discounts).

8. Monthly Sales Seasonality Analysis

- Monthly Sales Data:
  - Monthly sales totals were calculated by grouping the dataset by month, which provides insight into seasonal purchasing patterns.

- A line plot visualizes the total purchase amount over the months, showing sales trends and potential seasonality. The X-axis represents the months, while the Y-axis reflects the total sales amount.



Conclusion

The exploratory data analysis of the E-commerce Customer Behaviour and Purchase Dataset revealed several actionable insights. Young adults (ages 21–30) form the largest customer segment, indicating strong engagement, while low engagement from younger age groups (0–20) presents growth opportunities through targeted campaigns. Revenue analysis showed that Home, Clothing, and Electronics are the leading product categories, but the high return rate of 50.08% across all categories highlights a critical area for improvement, especially among high-value customers who exhibit significant churn tendencies.

To address these challenges, strategies such as enhancing product descriptions, improving quality control, and providing tailored incentives for high-value customers could help reduce returns and improve retention. Additionally, further segmentation and analysis of churned customers by demographics and purchasing behaviours may yield more targeted retention strategies.

These findings provide a roadmap for prioritizing customer engagement, reducing losses from returns, and optimizing the product portfolio, positioning the business for sustained growth and profitability.