Customer Segmentation & Product Recommendation System

A Data Science Project on Customer Insights and Recommendations

manasa yedla
FLIPKART.PVT.LTD

1.Introduction:

The goal of this project is to analyze customer purchase history, segment customers into meaningful groups, and provide personalized product recommendations. This system helps marketing teams target specific customer segments and improves sales by suggesting products that align with customer preferences.

Objective:

Analyze customer purchase history to segment buyers into meaningful groups and create a basic recommendation engine that suggests relevant products to each segment.

2. Dataset Description:

- For this project, we have used the dataset **expanded_sample_transactions.csv**, which contains transaction-level information for Flipkart customers.
- The dataset provides details about individual purchases made by customers, including the product bought, the category it belongs to, the amount spent, and the date of purchase.

Key Columns in the Dataset:

Column	Description
CustomerID	Unique identifier for each customer
ProductID	Unique identifier for each product
Category	Category to which the purchased product belongs
PurchaseAmount	Amount spent by the customer in that transaction
PurchaseDate	Date when the transaction occurred

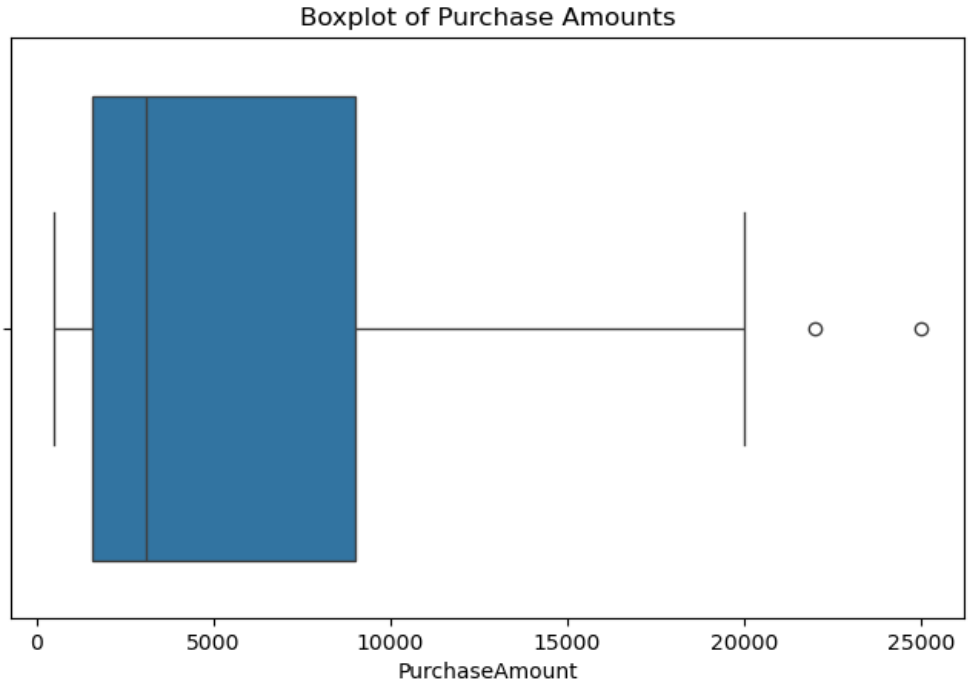
3. Data Cleaning

Before performing any analysis, the dataset was examined for **quality issues** to ensure accuracy and consistency.

Steps Taken:

1. **Checked for missing values:**
    - Used df.isnull().sum() to identify any missing entries in the dataset.
    - Missing values can affect calculations of spending, frequency, and clustering, so they must be handled appropriately.

2. **Checked for duplicate records:**
    - Used df.duplicated().sum() to detect repeated rows.
    - Removed duplicates to avoid double-counting transactions.
3. **Checked for outliers:**
    - A boxplot of PurchaseAmount (Figure 1) was used to detect extremely high or low values that could skew the analysis.
    - No extreme outliers were removed at this stage, but they were noted for further exploration.
4. **Saved cleaned dataset:**
    - After cleaning, the processed dataset was saved as cleaned_sales_data.csv for use in feature engineering and analysis.



Summary:
The dataset is now **clean, consistent, and ready** for feature engineering and exploratory data analysis.

Here is the cleaned Data Set :

"C:\Users\manasa\Downloads\cleaned_sales_data.csv"

4. Feature Engineering

Feature engineering was done to convert the raw transactional data into meaningful "customer-level features". This helped in understanding the overall behavior of each customer and prepared the data for customer segmentation. The main goal was to combine all purchase records of each customer into one summary record.

Steps Performed:

1. **Customer Aggregation:**

The dataset was grouped by **CustomerID** to calculate important details such as:

- Total amount spent by each customer
- Average amount spent per purchase
- Number of purchases made

In addition, other features were also included like: 1.Preferred Product Category

2.First Purchase Date and 3.Last Purchase Data.

```
customer_features = df.groupby('CustomerID').agg(
    TotalSpend=('PurchaseAmount', 'sum'),
    AvgSpend=('PurchaseAmount', 'mean'),
    PurchaseFrequency=('PurchaseAmount', 'count'),
    PreferredCategory=('Category', lambda x: x.mode()[0]),
    FirstPurchaseDate=('PurchaseDate', 'min'),
    LastPurchaseDate=('PurchaseDate', 'max')
).reset_index()
```

2. **New Features Created:**

Feature Name	Description
TotalSpend	Total amount spent by each customer
AvgSpend	Average spending per purchase
PurchaseFrequency	Number of transactions made by the customer
PreferredCategory	Most frequently purchased product category
FirstPurchaseDate	Date of the first purchase
LastPurchaseDate	Date of the last purchase

3. **Merging Enhanced Features:**

After calculating these metrics, the numerical features were standardized and merged with other attributes to prepare for clustering and segmentation.

```
customer_numeric_features = df.groupby('CustomerID').agg({
    'PurchaseAmount': ['sum', 'mean', 'count']
}).reset_index()
```

4. **Final Dataset for Segmentation:**

The final dataset contained **one record per customer** with all the summarized details, which made it ready for customer segmentation and recommendation analysis.

5. Exploratory Data Analysis (EDA)

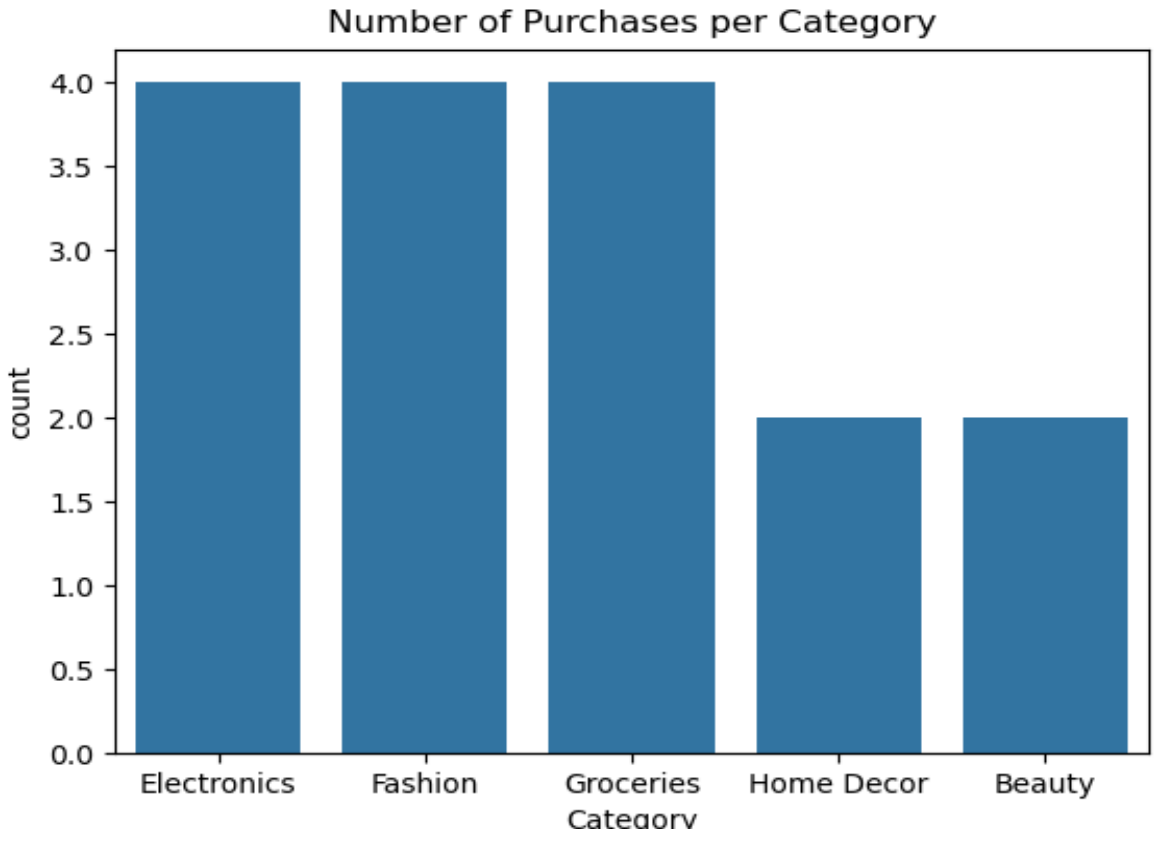
Exploratory Data Analysis was done to understand the dataset better and to identify key patterns, trends, and relationships within the data. It was performed at **two levels**:

1. Transaction-Level EDA
2. Customer-Level EDA

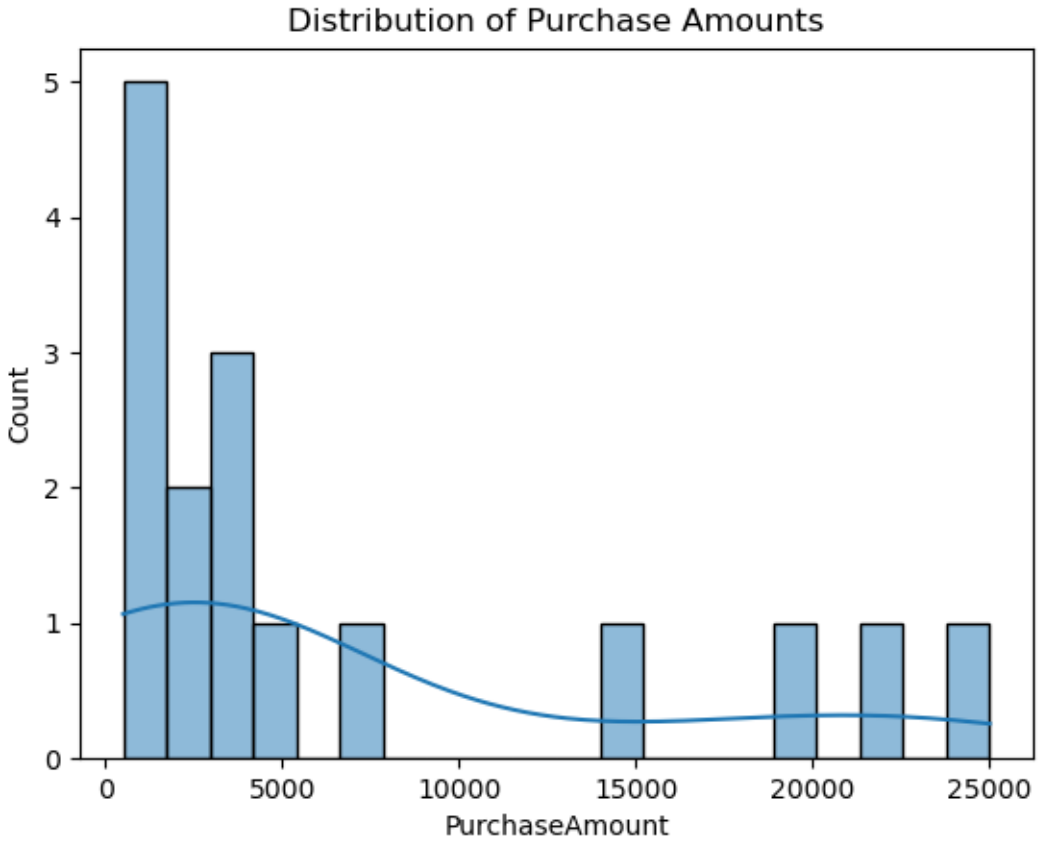
1. Transaction-Level EDA

At this level, the analysis focused on understanding overall sales patterns and product trends.
The following visualizations and observations were made:

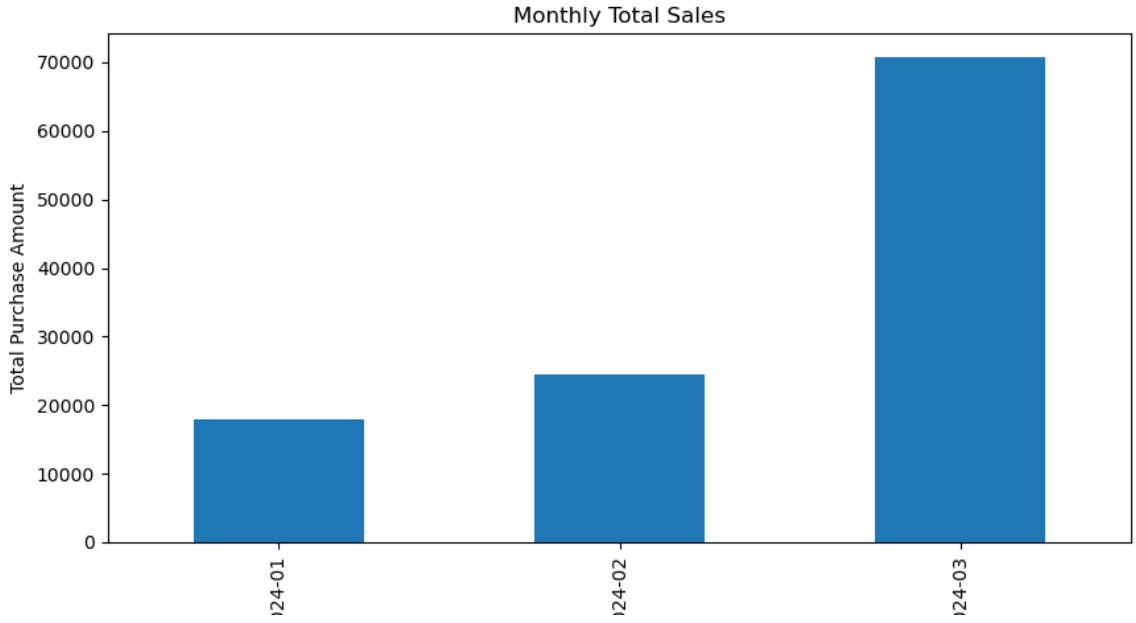
- **Number of Purchases per Category:**
  A count plot was created to see which product categories were purchased the most.



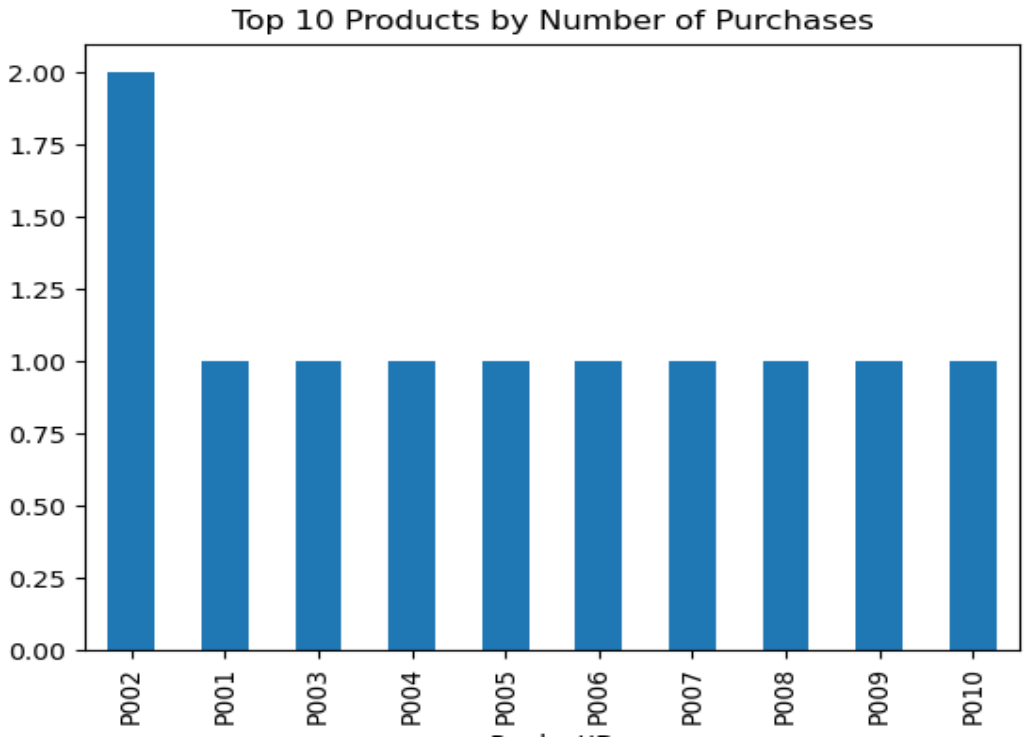
- **Distribution of Purchase Amounts:**
  A histogram and KDE plot were used to show how the purchase amounts varied. This helped to detect spending ranges and identify outliers.



- **Monthly Sales Trend:**
  Monthly total purchase amounts were visualized to understand sales trends over time.



- **Top 10 Products:**
  A bar chart was plotted to display the top 10 products based on the number of purchases.

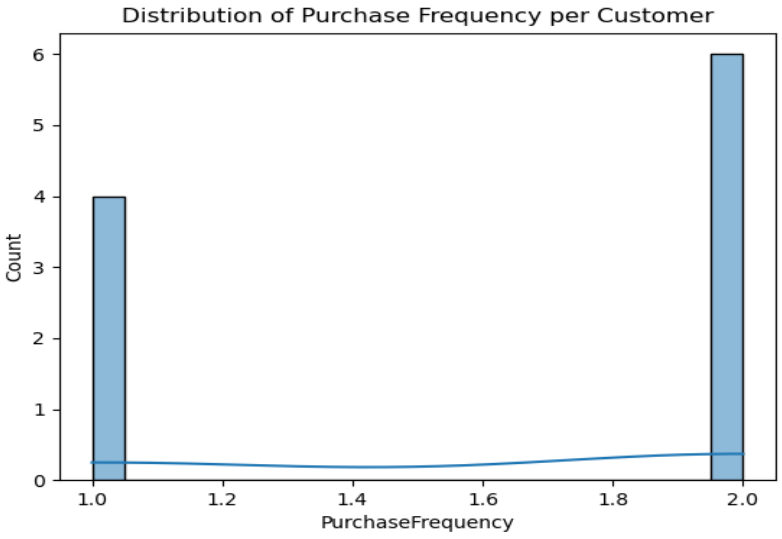


These analyses helped to identify popular categories, spending behavior, and seasonal trends in transactions.

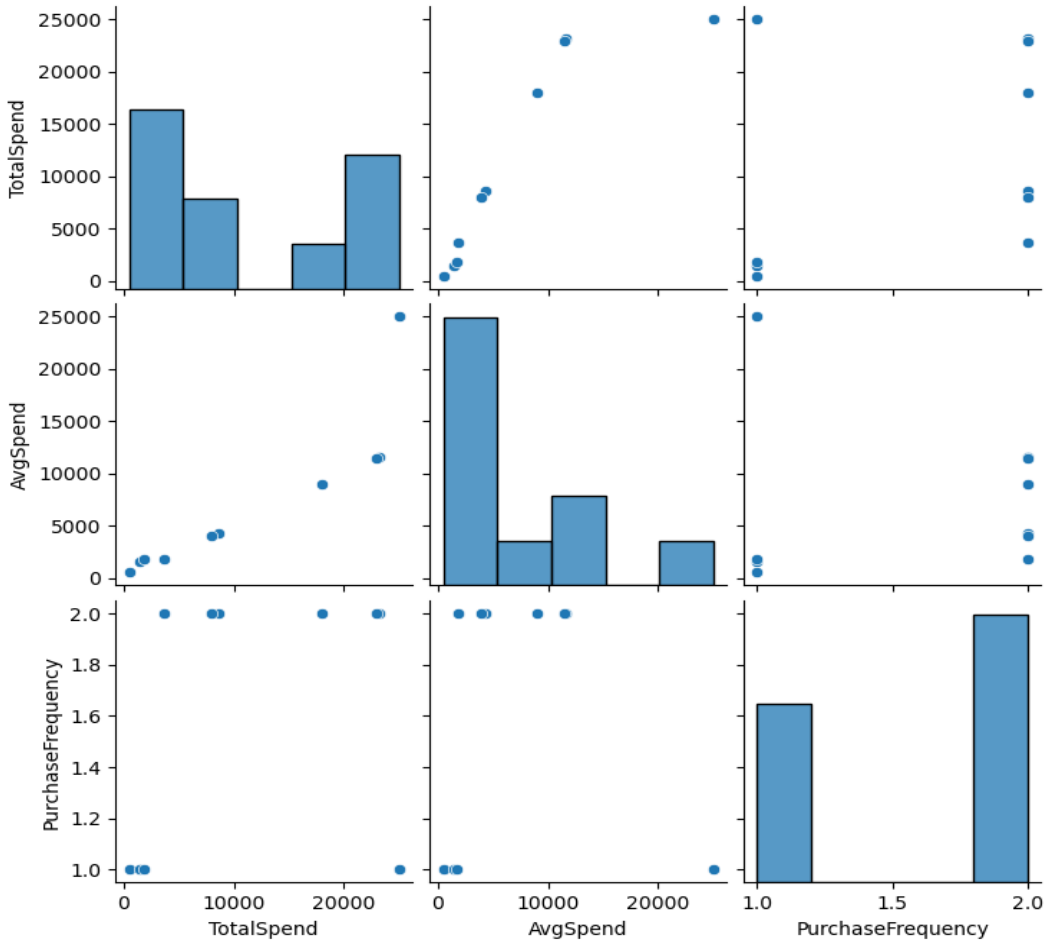
2. Customer-Level EDA

After summarizing transactions into customer-level data (in the feature engineering step), EDA was performed again to study customer behavior.
This included:

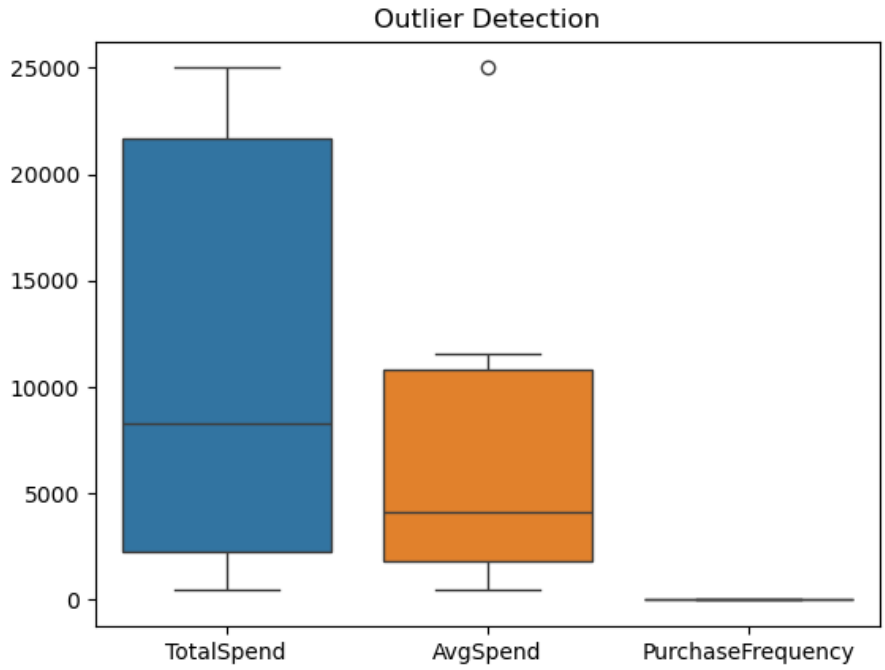
- **Distribution of Total Spend and Purchase Frequency:**
  Histograms were plotted to visualize how much customers spend and how frequently they make purchase.



- **Pairplot and Correlation Heatmap:**
  Relationships between TotalSpend, AvgSpend, and PurchaseFrequency were examined using pairplots and a heatmap.



- **Outlier Detection:**
  Boxplots were used to check for customers with unusually high or low spending.



Outcome:

The EDA phase provided a clear understanding of the data, helped in detecting trends and patterns, and confirmed that the dataset was ready for clustering.

6.Customer Segmentation

Objective

Segment customers based on their purchasing behavior to identify distinct groups for targeted marketing and product recommendations.

1. Why Customer Segmentation?

- Not all customers behave the same way.
- Segmenting customers helps us:
    - Identify **high-value customers**.
    - Design **personalized marketing campaigns**.
    - Recommend products that each group is likely to buy.

2. Features Used for Clustering

For each customer, we considered:

Feature	Description
TotalSpend	Total amount spent by the customer
AvgSpend	Average spend per purchase
PurchaseFrequency	Number of purchases made

These features capture the customer's spending behavior and engagement.

3. Clustering Method

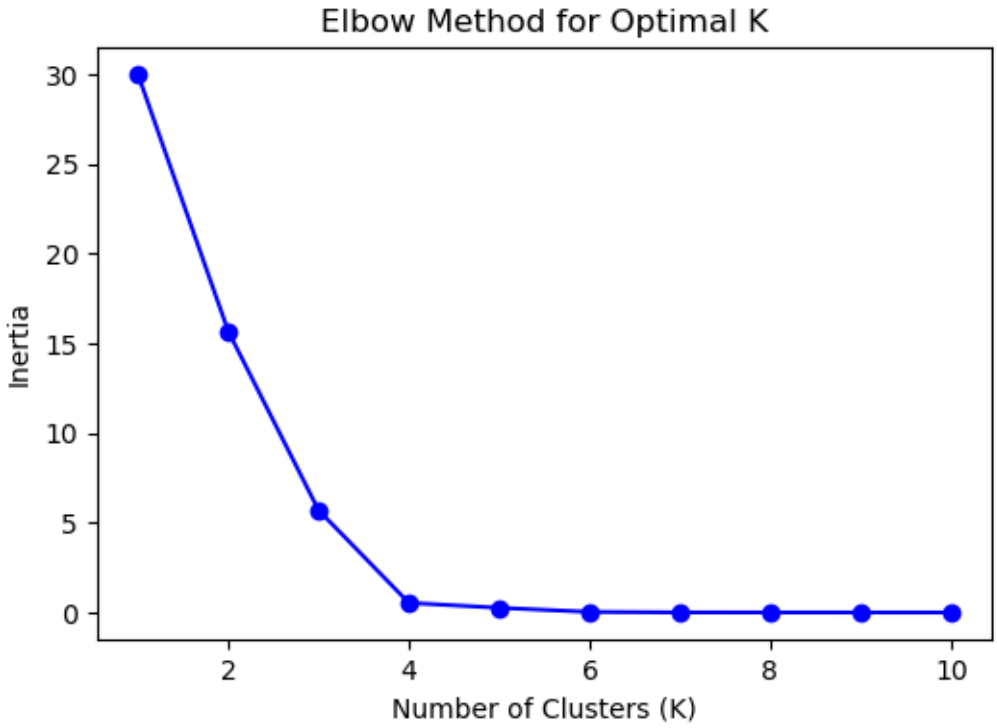
- We used **K-Means clustering**, which groups customers based on similarities in their spending behavior.
- Before clustering, we **scaled the data** so that all features have equal importance.
- We used the **Elbow Method** to choose the number of clusters. For our dataset, we chose **2 clusters** for simplicity.

After running K-Means, we got 2 customer segments:

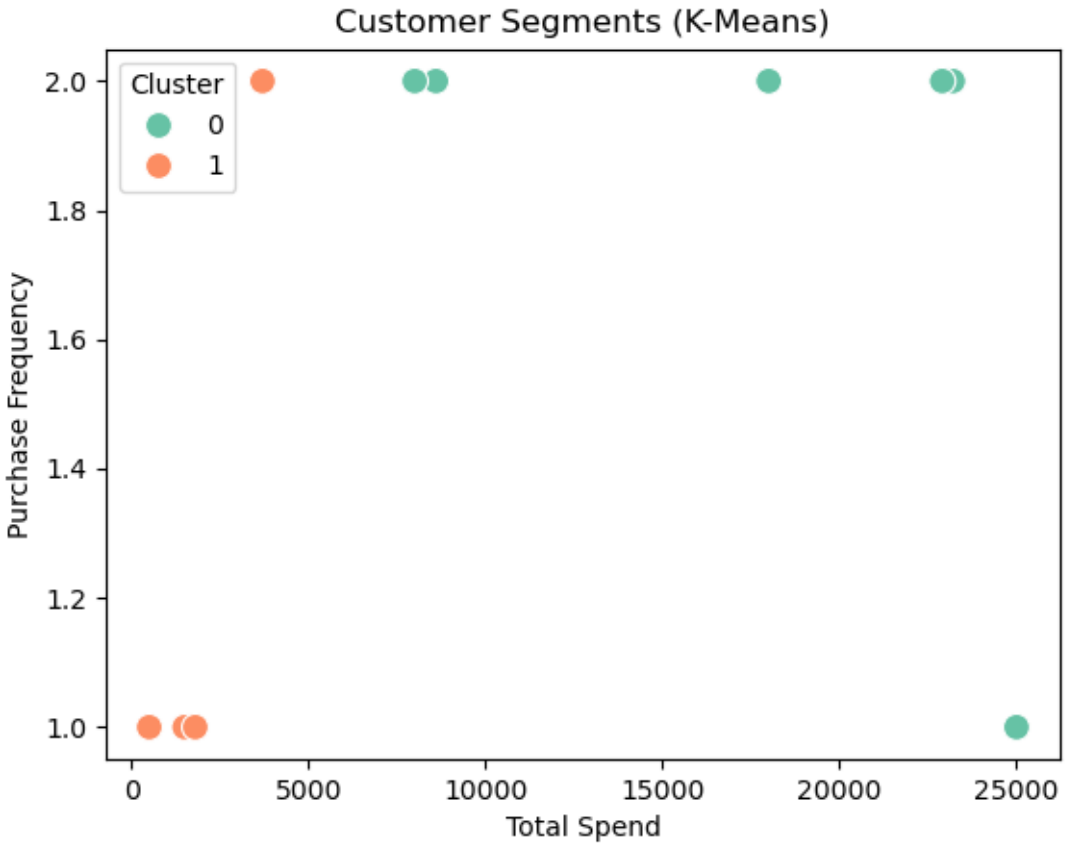
- Cluster 0 – High-Value Customers
- Cluster 1 – Frequent, Low-Value Customers

4. Visualization

- **Scatterplot:** Shows Total Spend vs Purchase Frequency with clusters colored differently.



Boxplots: Show Total Spend, Avg Spend, and Purchase Frequency across clusters.



5. Key Insights

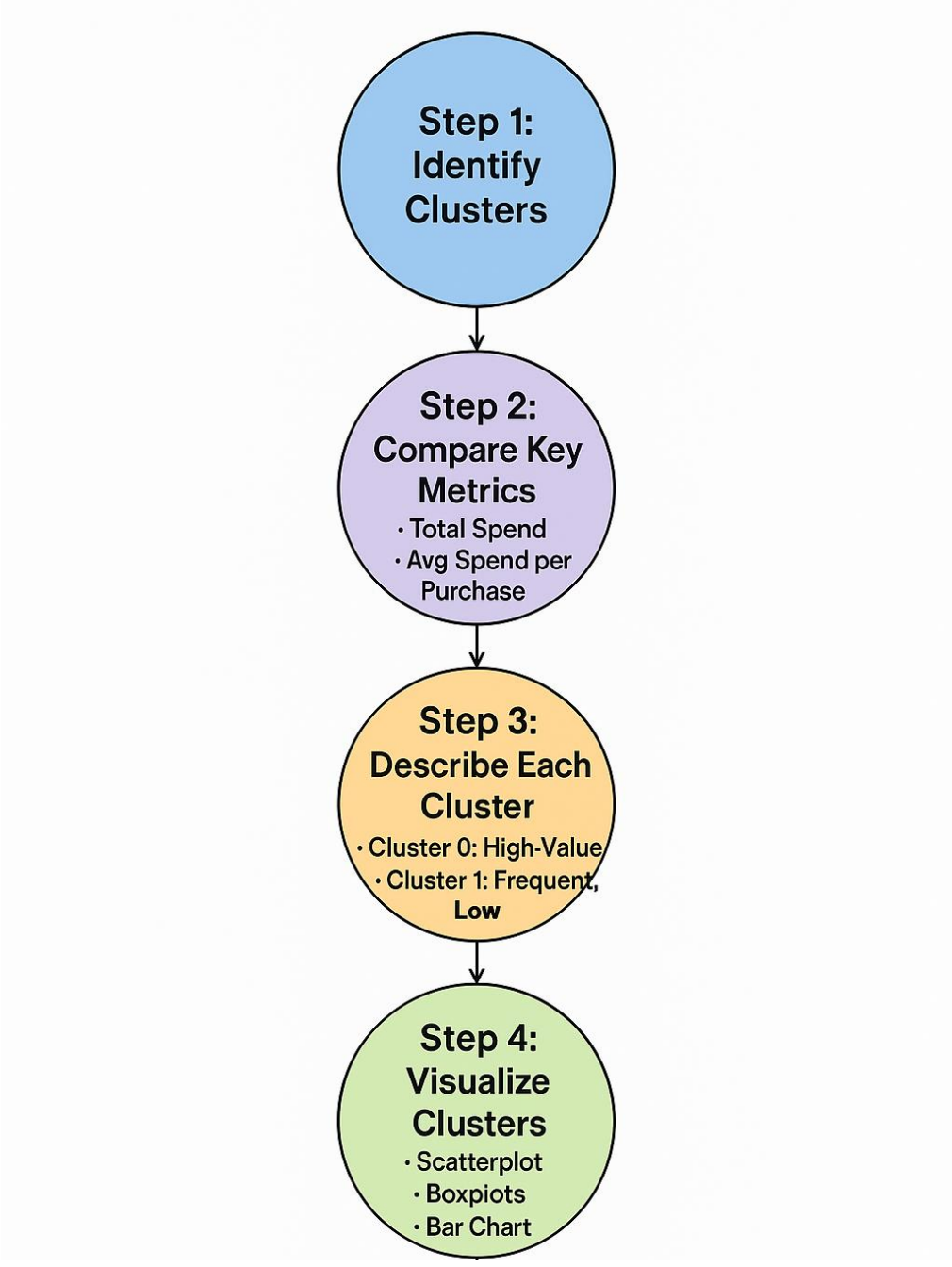
- High-value customers contribute significantly to revenue.
- Frequent buyers are opportunities for cross-selling.
- Segmentation allows **personalized marketing and recommendations**, which can increase sales and customer loyalty.

Here is the Customer segments saved as **'customer_segments.csv'**

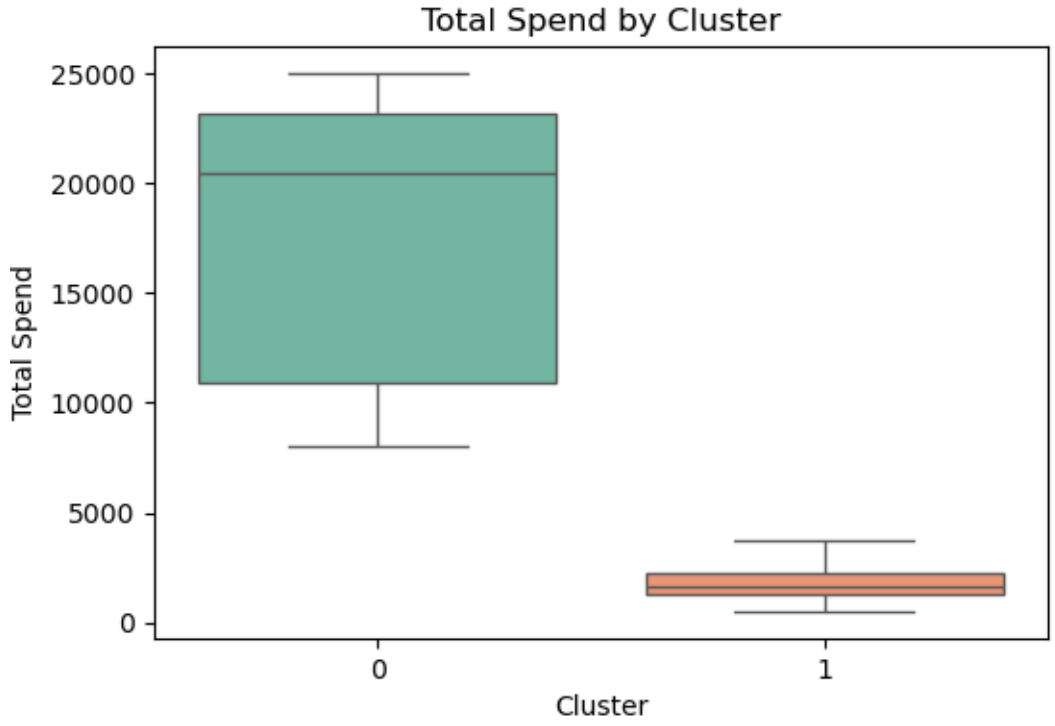
"C:\Users\manasa\Downloads\customer_segments.csv"

7. Cluster Profile Analysis – Customer Behavior & Spending

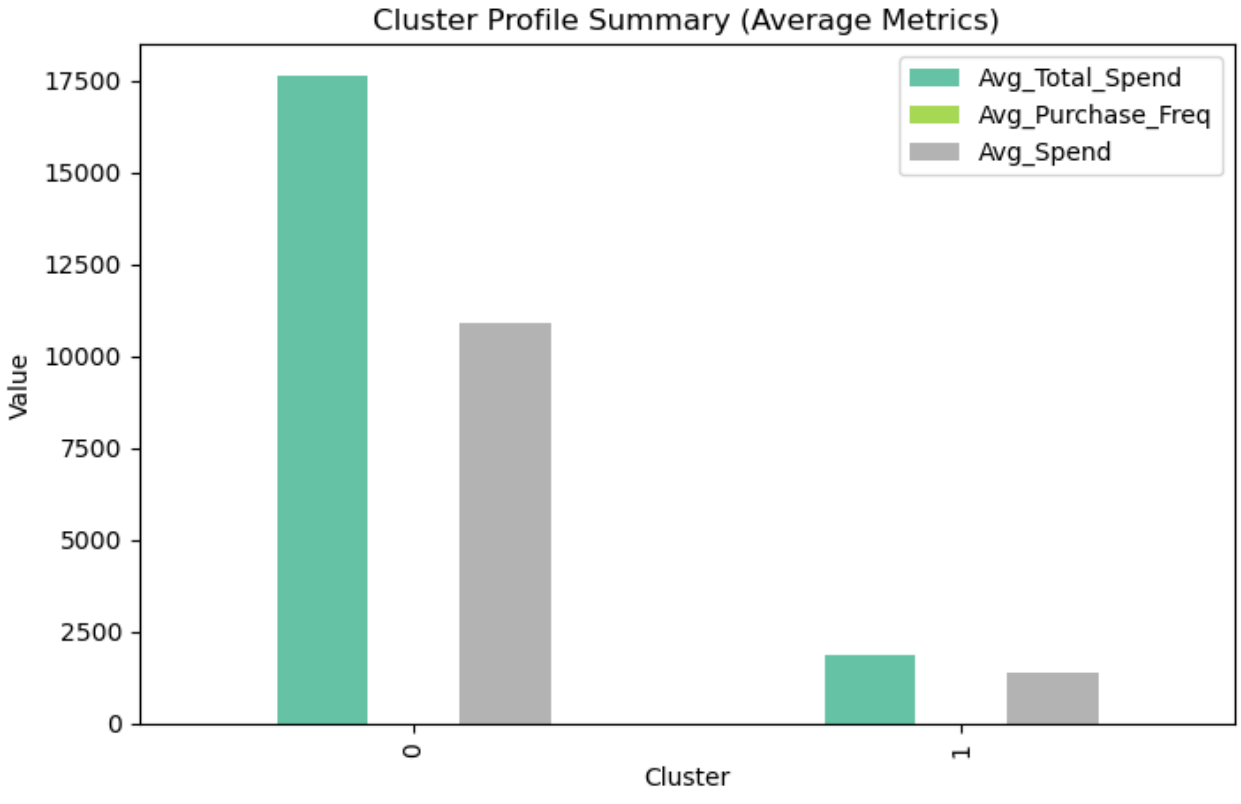
Here is the simple diagram which explain the steps in the process in a simple way:



Scatterplot: Shows Total Spend vs Total spend by clusters colored differently.



Bar Chart: Compares the average metrics for each cluster, giving a clear summary.



Key Insights:

- Cluster 0 are **high-value customers** – target them with premium offers and loyalty programs.
- Cluster 1 are **frequent, low-value customers** – target them with discounts, bundles, and promotions.
- This analysis helps us **understand customer behavior and plan marketing strategies** effectively.

Here is the Cluster profile summary saved as 'cluster_profile_summary.csv'

"C:\Users\manasa\Downloads\cluster_profile_summary.csv"

8. Recommendation System

1.Objective:

Build a personalized product recommendation system that suggests relevant products to each customer based on the purchasing behavior of similar customers (identified through customer segmentation).

2. Approach Used:

- A **Cluster-Based Collaborative Filtering** approach was used.
- Instead of recommending products purely based on individual history, this system recommends **popular products from the customer's cluster** (group of similar customers).

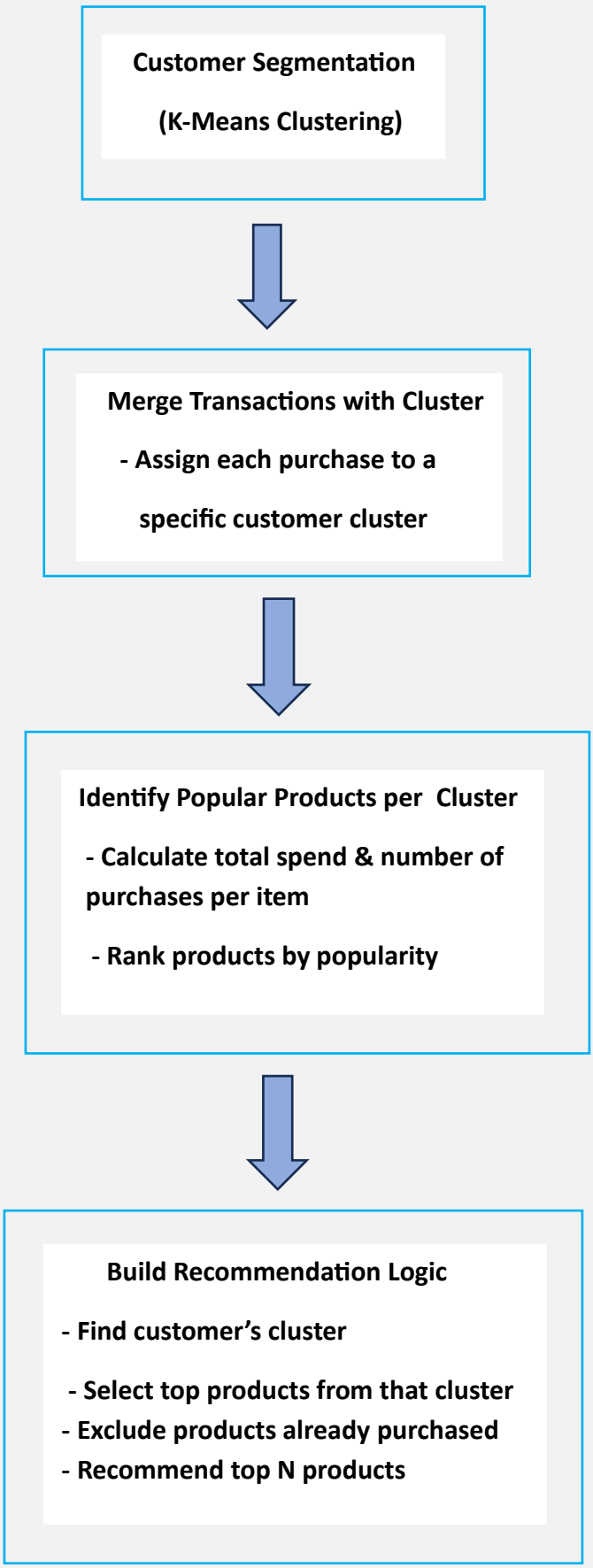
3. Data Used:

Used the following data after merging transaction details with cluster labels:

Feature	Description
CustomerID	Unique identifier for each customer
Cluster	Cluster label assigned during segmentation
ProductID	Product purchased
PurchaseAmount	Amount spent on the product

4. Methodology (Diagrammatic Representation):

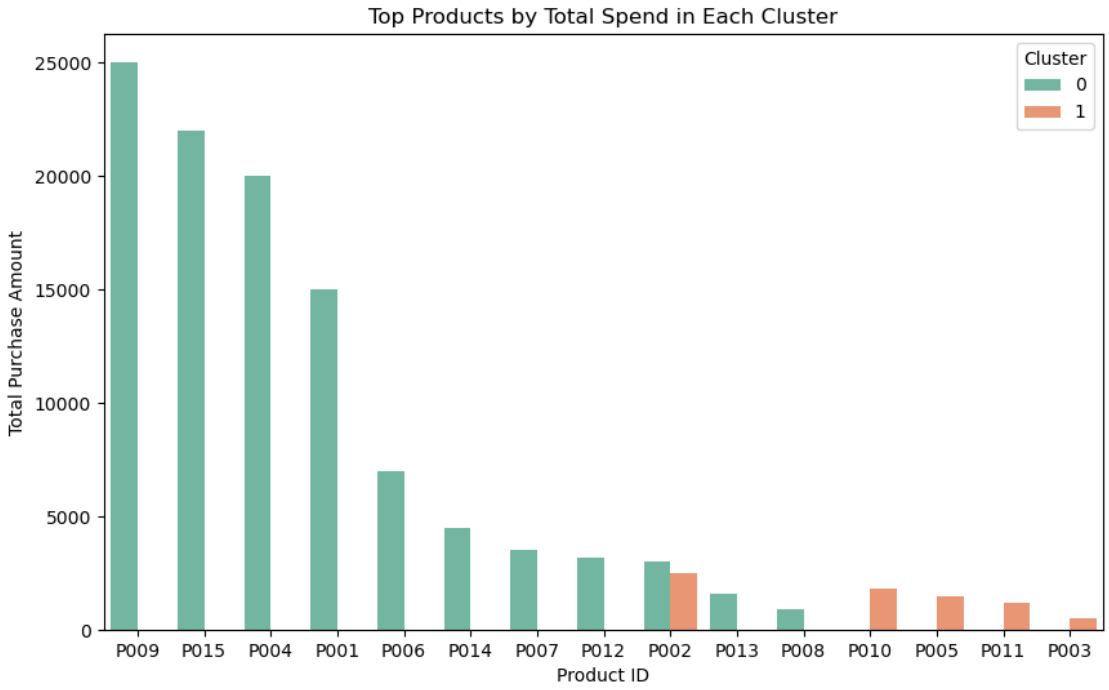
🧩 Cluster-Based Recommendation System Workflow

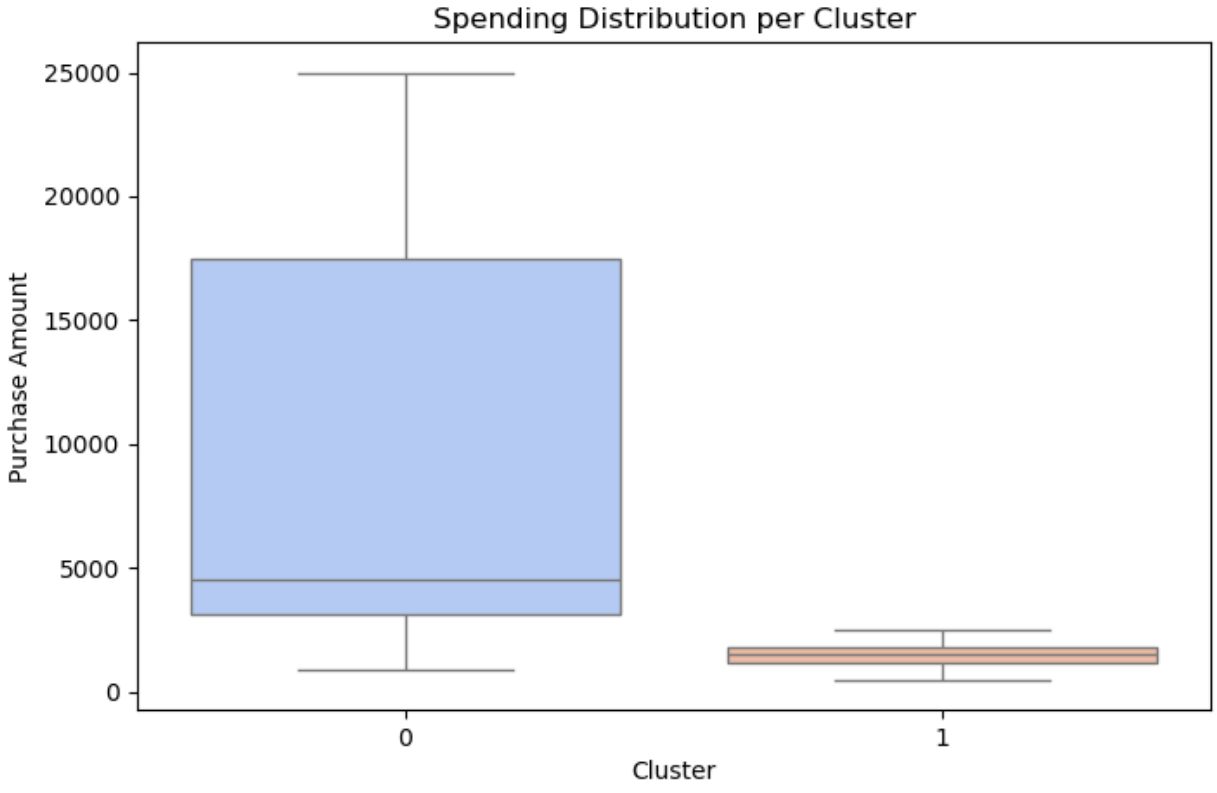


5. Visualization

- **Bar Chart:** Top Products by Total Spend per Cluster
- **Boxplot:** Spending Distribution per Cluster

These visualizations highlight which products and clusters generate the most sales and how spending patterns vary across groups.





6. Insights

- Customers in the same cluster tend to buy similar types of products.
- Recommending top products from a customer's cluster increases the chance of purchase.
- Helps businesses identify **cross-selling** and **up-selling** opportunities.

Conclusion

This project successfully created a **Customer Segmentation and Product Recommendation System** using customer purchase data.
By grouping customers with *K-Means clustering*, we identified different spending and buying patterns.
The *Recommendation System* then suggested popular products from each customer's group, making the suggestions more personalized.
Overall, the system helps businesses *understand customers better*, improve marketing, and increase sales through targeted recommendations.