Dataset Overview:

Total Customers: 500

Total Products: 50

Total Transactions: 5,000

Key Attributes:

- Customer ID
- Product ID
- Product Category
- Purchase Amount
- Purchase Date

Product Analysis

Top 10 Products by Total Sales

The most profitable products (by total dollar value) are concentrated among a small subset of unique products. These top performers drive a large proportion of overall revenue, indicating **product concentration**.

Top 10 Products by Units Sold

These differ slightly from the top revenue generators—some high-volume products sell many units but at lower prices. This suggests there's a **distinction between high-margin and high-volume products**.

Total Sales by Category

Some standout categories (e.g., *Electronics*) dominate total sales, implying strong customer demand or higher product pricing.

Units Sold by Category

Grocery and *Clothing* may have lower individual price points but rank high in unit volume, reflecting **repeat purchases or essential goods**.

Customer Analysis

Total and Average Customer Spending

- **Average total spend per customer** and **average purchase value** were calculated.
- A histogram of average purchase amounts shows a **right-skewed distribution**, meaning most customers spend modest amounts, but a few spend significantly more.

Purchase Frequency vs. Spending

A scatterplot of purchase frequency against total spend reveals a **positive correlation**—frequent buyers tend to spend more.

Insight: Encourage repeat purchases through loyalty programs or personalized recommendations.

Top Spenders

- Contribute disproportionately to total revenue.

- Favor specific product categories (frequency by category was analyzed for this group).

These insights suggest that we have to create customer personas based on spending and frequency to tailor marketing efforts.

Customer Segments

Dataset Overview

- Total Customers: 500
- Features Used:
  - `Total_Spent` *(numeric)*
  - `Purchase_Count` *(numeric)*
  - `Product_Category` *(categorical)*
- Clustering Method: K-Prototypes (ideal for mixed data types)
- Optimal Clusters: 3 (determined via Elbow Method)

Cluster 0 – High Spenders

Avg Total Spent: $1,809

Avg Purchases: 13

Dominant Category: Grocery

Behavior:

Top-tier customers with higher-value transactions.

Cluster 1 – Frequent Shoppers

Avg Total Spent: $1085

Avg Purchases: 10

Dominant Category: Grocery

Behavior:

Consistent shoppers with regular activity across common-use categories.

Cluster 2 – Occasional Shoppers

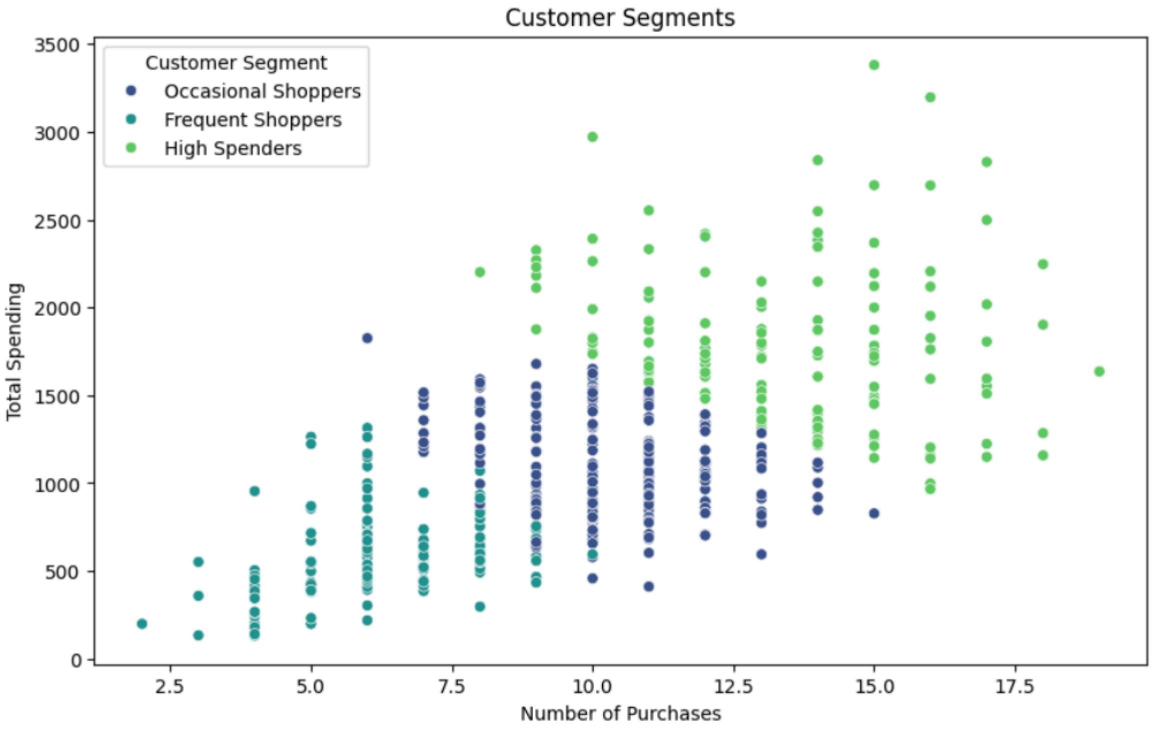
Avg Total Spent: $601

Avg Purchases: 6

Dominant Category: Clothing

Behavior:

These customers make occasional purchases, typically of lower-cost, everyday items. They're likely browsing or need-based shoppers.



Recommendations (collaborative filtering)

Methodology and Testing

Methodology: Cosine Similarity

- To recommend similar products for Customer X, we look at Customer X's cluster and identify 3 most similar customers in that cluster. Products from similar customers are recommended.

For example, let's take C020.

C020 Cluster: Occasional Buyers
Popular category for this cluster: Grocery
Recommended Products: ['P016', 'P027', 'P021', 'P015', 'P001']

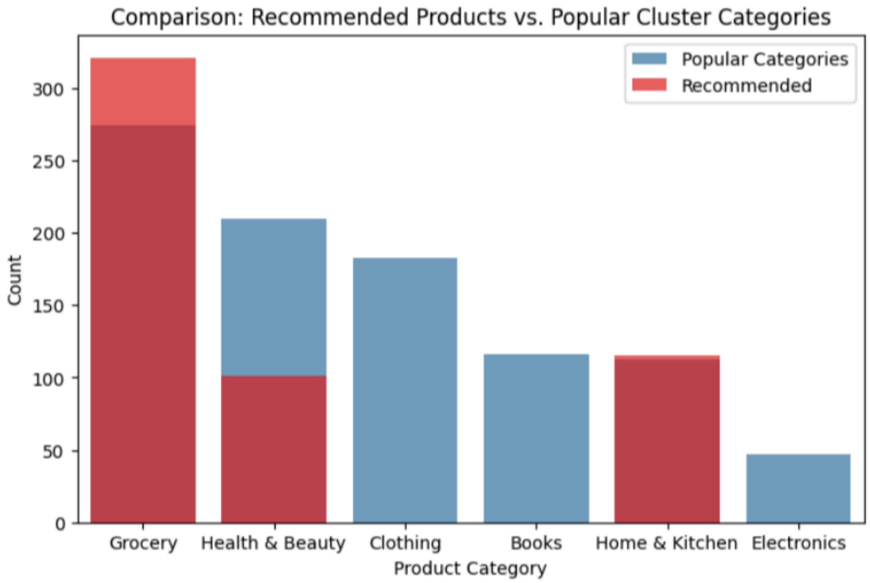
Similar customers: C014, C225, C434

- All three similar customers are from the same cluster

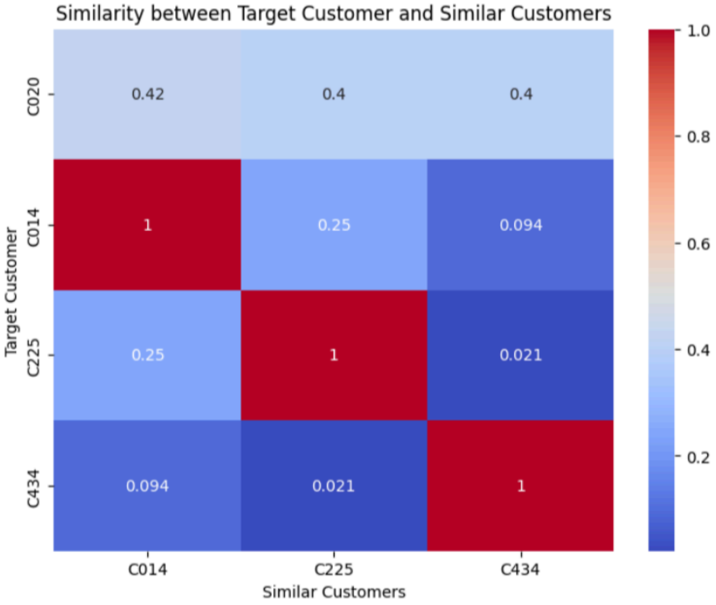
These are the products purchased by these customers:

	Customer ID	Product ID	Product Category
182	C014	P004	Home & Kitchen
361	C014	P016	Grocery
615	C014	P035	Clothing
702	C434	P039	Clothing
1022	C014	P027	Grocery
1155	C014	P015	Grocery
1847	C225	P038	Health & Beauty
2142	C014	P021	Health & Beauty
2492	C225	P004	Home & Kitchen
2683	C225	P024	Clothing
2918	C434	P011	Clothing
2992	C434	P040	Clothing
3040	C014	P001	Home & Kitchen
3554	C434	P016	Grocery
3867	C225	P047	Books
4011	C014	P003	Health & Beauty
4568	C434	P048	Health & Beauty
4609	C225	P035	Clothing

These are the categories for the recommendations for C020:
'P016' - grocery ($29.39)
 'P027' - grocery ($47.56)
'P021' - health & beauty ($41.17)
'P015’ - grocery ($26.46)
'P001' - home & kitchen ($121.58)



Average Cost for recommended products: $53
Average Purchase for C020: $85
Top category for C020: Clothing



Summary

The recommendations for this customer matches the cluster similarity more than it matches the customer’s previous purchases. There are a few reasons that can be attributed to this:

1. The purchase data for this customer might be limited, so no strong similarity could be found.
2. The purchase data for this customer might not suggest a strong preference. If this is the case, the algorithm considers “global” preferences, which is the preference of the cluster
3. The data, by nature, does not reflect real-life biases or distributions, which makes it harder to find a stronger pattern

Future Modifications

Cluster refinement: Consider multi-dimensional clustering that accounts for both purchase frequency and category preferences.

Purchase Recency: Consider adding timestamps and giving more weight to recent purchases, as they may better reflect current preferences.