# Customer Segmentation & Recommendation Report

## Introduction

The goal of this analysis is to segment customers based on purchasing behavior (Recency, Frequency, Monetary - RFM analysis) and to develop an effective product recommendation system using a hybrid collaborative and content-based filtering approach. The insights derived aim to improve targeted marketing and customer engagement.

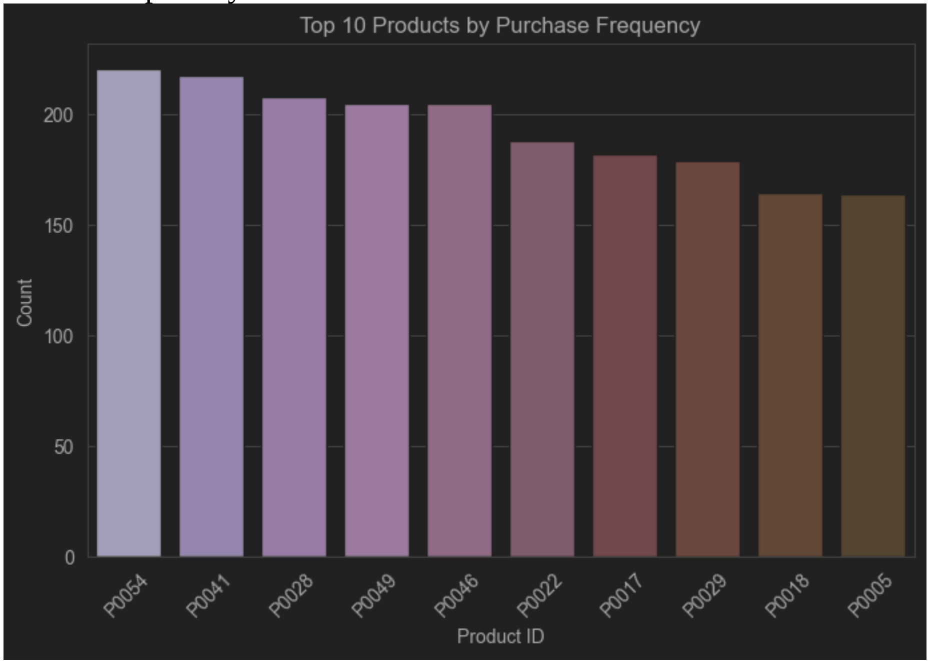
## Exploratory Data Analysis (EDA)

### Key Insights

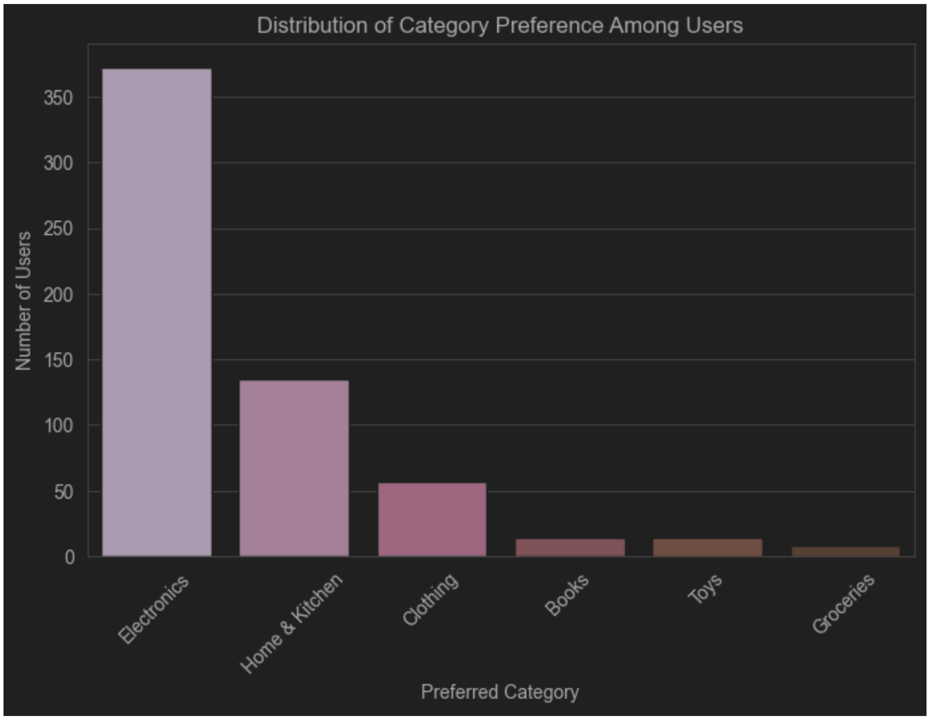
- **Popular Products:** Products `P0054` and `P0041` are the most popular, with approximately 250 purchases each and 163 unique users purchased `P0054`
- **Category Preferences:** Electronics is the most preferred category overall, while Groceries is least preferred.
- **Purchase Amount Distribution:** Heavily right-skewed, indicating most purchases are lower in value, with a minority making high-value purchases.
- **Purchase Frequency:** Customer purchases are evenly distributed, with most customers having moderate frequency.
- Top Selling Category: Electronics is the top selling category by purchase amount but Books are the top selling category by sold product count which is 1466 transactions.
- Average Spending Per Customer: $4349.53

## Visualizations Included:

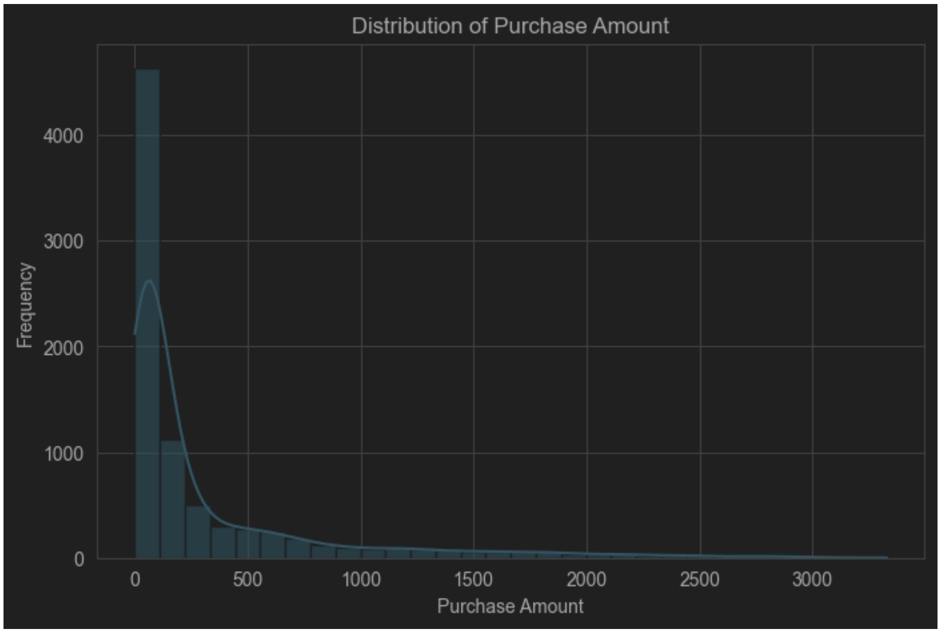
- Product Popularity Distribution



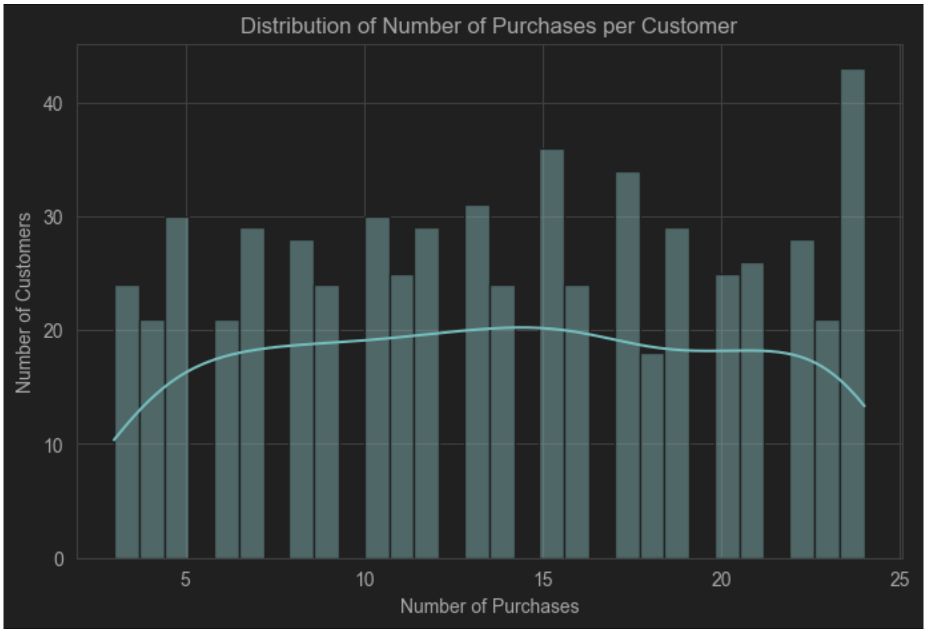
- Product Category Frequency



- Purchase Amount Distribution



- Number of Purchases per Customer



# Customer Segmentation Analysis

## RFM Feature Definitions

- **Recency:** Days since last purchase.
- **Frequency:** Total purchases made.
- **Monetary:** Total spending amount.

## Clustering Methodology

- **Model:** K-Means clustering on scaled RFM features & prodct preference
- **Cluster Selection:** 6 clusters chosen for enhanced recommendation granularity based on PCA visualization and the Elbow Method.

## Cluster Summary & Labels

| Cluster | Recency (days) | Frequency | Monetary ($) | Category Preference |
|---|---|---|---|---|
| 0 | 9.98 | 13.49 | 2834.67 | Books |
| 1 | 77.56 | 5.17 | 1519.75 | Books |
| 2 | 13.66 | 6.29 | 1606.55 | Clothing |
| 3 | 11.21 | 19.58 | 12724.86 | Electronics |
| 4 | 37.47 | 11.00 | 3607.41 | Clothing |
| 5 | 8.25 | 20.61 | 5015.20 | Clothing |

## Cluster Descriptions

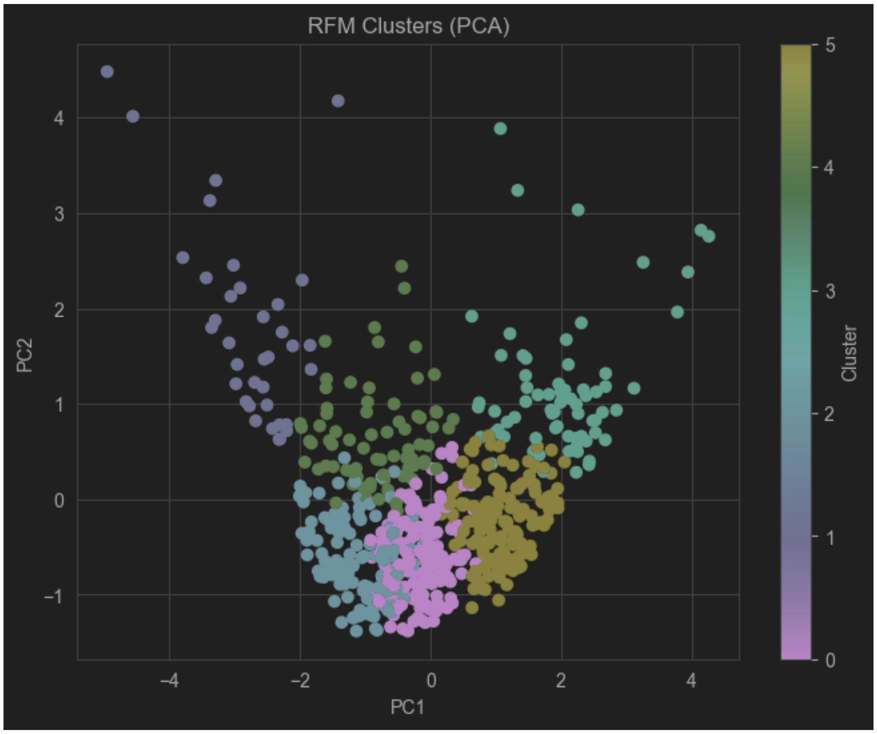
- **Moderate Buyers (0):** Regular, moderate spenders.
- **Potentially Churned Low-Value Customers (1):** Infrequent and low spend, high churn risk.
- **Occasional Moderate Buyers (2):** Moderate spend but infrequent.
- **High-Value Customers (3):** Frequent, recent, high spenders.
- **At-Risk Customers (4):** Declining purchasing trends.
- **Frequent Buyers (5):** High frequency, moderate spenders.

## Key Cluster Insights

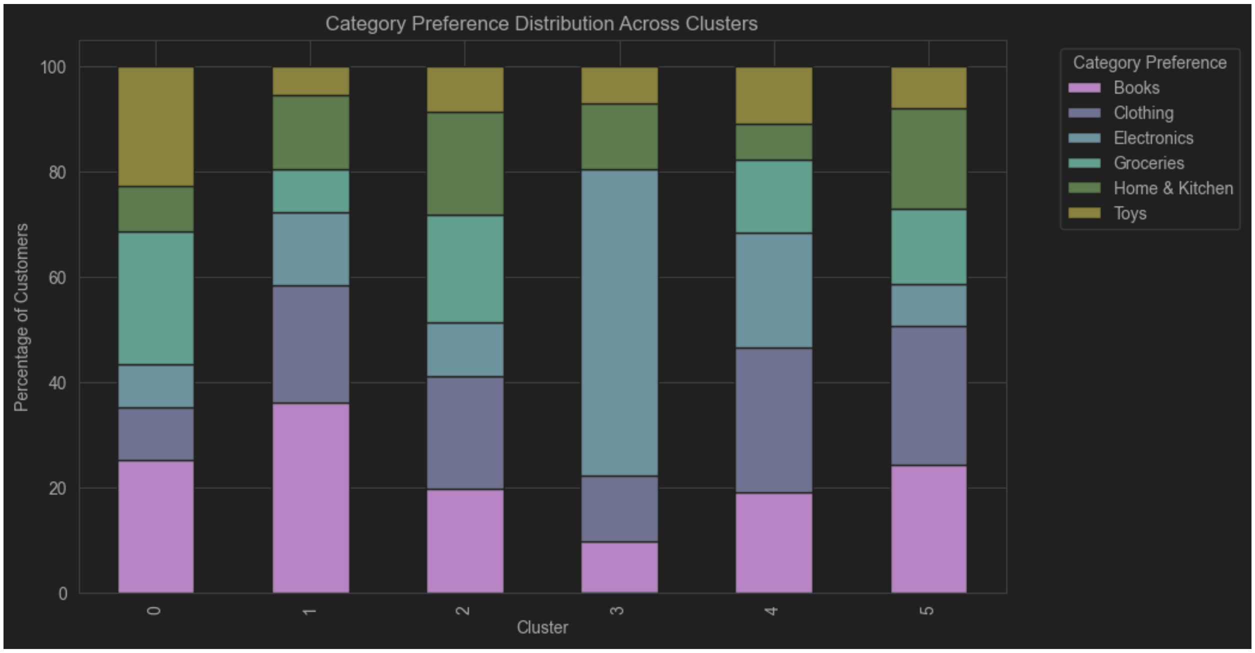
- High-value customers strongly prefer Electronics (142% more than second-best).
- Potentially churned customers prefer Books; At-risk customers prefer Clothing—both could benefit from diversified recommendations.
- Toys are notably preferred by Moderate Buyers (~22%).
- PCA analysis showed clear segmentation, particularly distinguishing high-value, at-risk, and moderate buyers.

## Visualizations Included:

- PCA Visualization of Clusters



- Category Preference Distribution Across Clusters



# Recommendation Logic

## Overview

The recommendation system utilizes a hybrid approach combining user segmentation (RFM-based clustering) with content-based filtering to generate personalized product recommendations. This approach aims to leverage both behavioral patterns (user purchase history) and product characteristics (category similarity).

## Detailed Recommendation Steps

### 1. User Segmentation

Each customer is first assigned to a specific segment based on their purchasing behavior, defined by:

- **Recency** (how recently the customer made a purchase)
- **Frequency** (how often the customer purchases)
- **Monetary** (the total amount spent)

For example, Customer `C0011` was categorized into **Cluster 0 (Moderate Buyers)** due to their moderate frequency, spending, and recency.

### 2. Cluster-Based Product Popularity

Products frequently purchased within the user's assigned cluster are identified. The most popular product (highest total revenue) in the user's cluster is selected as the **"seed product"** for generating recommendations. For instance, in Cluster 0, product `P0046` was identified as the top seed product.

### 3. Content-Based Filtering with TF-IDF

To ensure recommendations are not purely popularity-driven, TF-IDF vectorization is used on product categories to calculate product embeddings. Cosine similarity is then computed between these embeddings to identify products that are most similar in terms of category attributes.

### 4. Generating Initial Recommendations

Using the seed product, the system retrieves the top N products with the highest similarity scores within the cluster (excluding the seed product itself). An initial recommendation list might look like:

```
[‘P0046’, ‘P0041’, ‘P0022’, ‘P0049’, ‘P0028’]
```

### 5. Ensuring Diversity through Randomization

To avoid overly narrow recommendations and maintain user engagement, the model introduces diversity through two randomization techniques:

- **10% Chance:** One recommendation is replaced with a random product completely outside the user’s assigned cluster.
- **10% Chance:** One recommendation is replaced with a product from the cluster most similar to the user’s assigned cluster (determined by Euclidean distance between cluster centers).

For example, introducing diversity might adjust the recommendations to:

```
🔀 Introducing a product from closest cluster 4: P0019
Recommended Products for User C0011: [‘P0046’, ‘P0041’, ‘P0022’, ‘P0019’,
‘P0049’]
```

## Benefits of This Hybrid Approach

- **Personalization:** Tailored to specific customer segments based on purchasing behavior.
- **Relevance:** Combines behavioral insights with product similarity to ensure meaningful recommendations.
- **Diversity:** Periodic inclusion of products from outside clusters prevents recommendation monotony, maintaining customer interest.

This structured yet flexible recommendation process effectively balances customer behavior and product attributes, ensuring high relevance and sustained user engagement.

---

## Evaluation Methodology

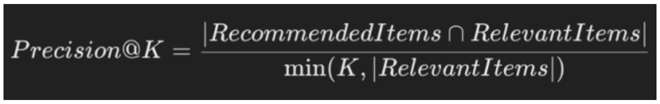
To assess the quality of the recommendation system, we employed a **precision and recall at K (Precision@K and Recall@K)** approach, using a carefully constructed train-test split tailored specifically for recommendation tasks.

### Train-Test Splitting Approach

- **User-Level Splitting:**
  Each user’s purchase history is individually partitioned into training and test subsets. This ensures realistic evaluation, as recommendations aim to predict future or unseen user purchases.
- **Fraction-Based Splitting:**
  For every user, a fixed proportion (e.g., 20%) of transactions is randomly selected to form the test set, while the remaining data constitutes the training set.
  *(More details on the splitting method can be found in the provided evaluation code.)*

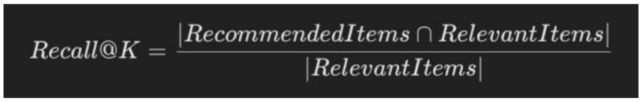
**Metrics Used**

- **Precision@K:**
  Measures the fraction of recommended items that were actually relevant (i.e., items the user truly purchased in the test set).



$$Precision@K = \frac{|RecommendedItems \cap RelevantItems|}{\min(K, |RelevantItems|)}$$

- **Recall@K:**
  Measures the fraction of relevant items (actual items in the test set) successfully recommended by the model.



$$Recall@K = \frac{|RecommendedItems \cap RelevantItems|}{|RelevantItems|}$$

**Evaluation Procedure**

1. For each user:
   - Generate top-K recommendations using the trained recommendation system.
   - Compare the recommendations against actual items present in the user’s test set.
2. Calculate Precision@K and Recall@K individually for each user.
3. Aggregate the metrics across all users to determine the average performance.

This methodology provides clear insight into both the **accuracy and coverage** of recommendations.

*Further implementation details and the complete evaluation procedure are provided in the accompanying evaluation code.*