Task_02: Customer Segmentation Analysis
Team: Cyferlink

Customer Segmentation Analysis Report

Executive Summary

This report presents a comprehensive analysis of customer segmentation for an e-commerce platform. The analysis was conducted to identify distinct customer segments based on behavioral patterns. Using clustering techniques, we successfully identified three primary customer segments:

Bargain Hunters, High Spenders, and Window Shoppers.

The segmentation process involved exploratory data analysis, preprocessing, model selection, and evaluation. K-Means clustering was selected as the primary model due to its interpretability and performance metrics. The final model revealed meaningful segments that align with the expected customer personas.

This segmentation provides actionable insights that can enhance targeted marketing strategies, optimize product recommendations, and improve customer retention rates.

Introduction

Customer segmentation is a critical strategy for e-commerce businesses seeking to understand their customer base better. By categorizing customers into distinct groups based on their behavior, businesses can develop targeted marketing campaigns, personalize user experiences, and allocate resources more efficiently.

The dataset was expected to contain three distinct customer segments: Bargain Hunters, High Spenders, and Window Shoppers. Each segment exhibits unique behavioral patterns that can inform business strategies.

Exploratory Data Analysis

Dataset Overview

The analysis began with loading and examining the dataset structure. Initial inspection revealed the dataset's dimensions, basic statistics, missing values, and potential duplicates.

Dataset Shape: (999, 6)
Number of duplicate rows: 0
Missing Values:

total_purchases	20
avg_cart_value	20
total_time_spent	0
product_click	20
discount_counts	0
customer_id	0

Feature Distributions

Understanding the distribution of each feature is crucial for identifying patterns and anomalies in customer behavior.

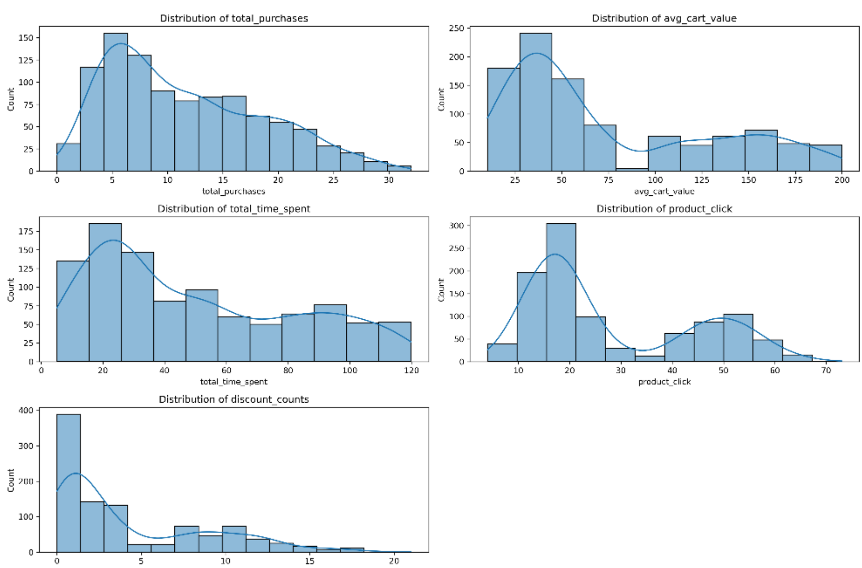

Figure 1: Histograms showing the distribution of all numeric feature

The histograms reveal several important characteristics:

- **total_purchases**: Right-skewed distribution, indicating most customers make fewer purchases while a small segment makes significantly more
- **avg_cart_value**: Relatively normal distribution with some high-value outliers
- **total_time_spent**: Bimodal distribution, suggesting two distinct patterns of engagement
- **product_click**: Right-skewed, with most customers viewing a moderate number of products
- **discount_counts**: Right-skewed, with many customers using few or no discounts

Outlier Analysis

Boxplots were used to identify potential outliers in each feature.

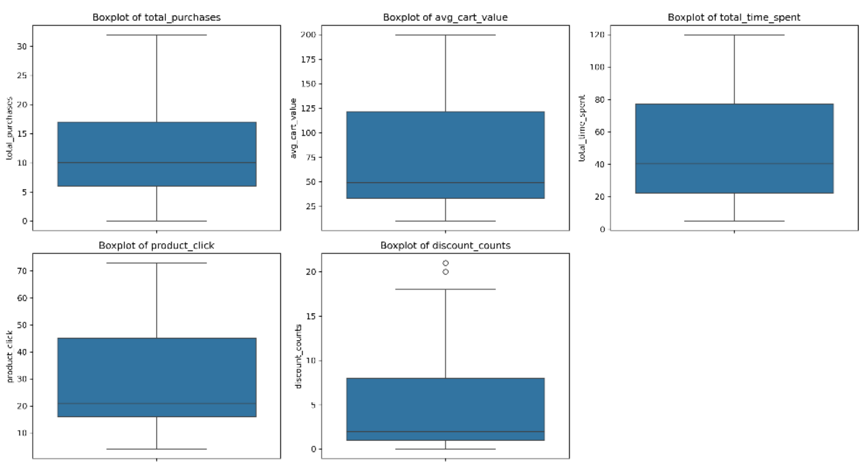

Figure 2: Boxplots showing the distribution and outliers for each feature

The boxplots revealed:

- Significant outliers in **avg_cart_value**, suggesting some customers make exceptionally high-value purchases
- Moderate outliers in **total_time_spent** and **product_click**
- Few outliers in **discount_counts**

Feature Correlations

Understanding relationships between features helps identify patterns that might influence clustering.

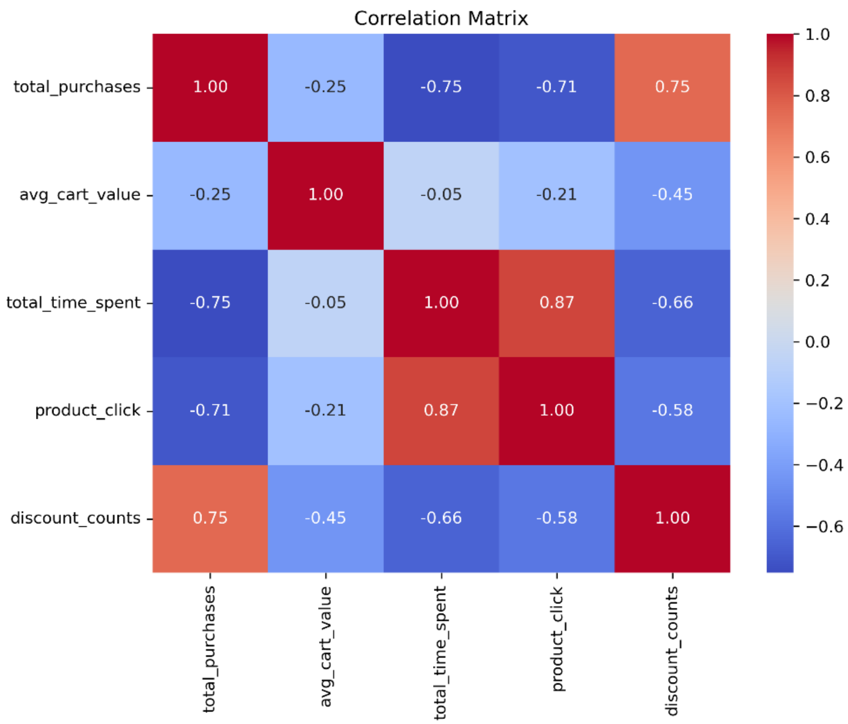

Figure 3: Correlation matrix showing relationships between features

Key correlations observed:

- Positive correlation between **product_click** and **total_time_spent** (0.87), suggesting customers who browse more also spend more time on the platform
- Moderate positive correlation between **total_purchases** and **discount_counts** (0.75), indicating customers who purchase more also tend to use more discounts
- Negative correlation between **avg_cart_value** and **discount_counts** (-0.45), suggesting customers who make high-value purchases use fewer discounts

Statistical Measures

Skewness and kurtosis measurements provided additional insights into the data distribution:

Skewness of Features:

total_purchases	0.639
avg_cart_value	0.783
total_time_spent	0.565
product_click	0.696
discount_counts	1.070

Kurtosis of Features:

total_purchases	-0.542
avg_cart_value	-0.807
total_time_spent	-0.944
product_click	-1.009
discount_counts	0.097

The skewness values confirm the right-skewed nature of most features, while the negative kurtosis for **total_time_spent** supports its observed bimodal distribution.

Data Preprocessing

Handling Missing Values

Missing values were addressed using KNN imputation with n_neighbors=2. This method preserves the relationships between features by filling missing values based on similar records.

Shape after cleaning: (999, 6)

The imputation process maintained the dataset's size while ensuring complete records for subsequent analysis.

Feature Standardization

Standardization was applied to ensure all features contributed equally to the clustering process. This step is crucial for distance-based algorithms like K-Means, which are sensitive to the scale of input features.

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

The standardization process transformed each feature to have zero mean and unit variance, allowing features like **avg_cart_value** (with larger values) to have equal influence as features like **discount_counts** (with smaller values).

Model Selection

Clustering Algorithm Selection

While multiple clustering algorithms were evaluated, K-Means was selected as the primary approach due to its:

- Simplicity and interpretability
- Scalability for larger datasets
- Efficiency in identifying spherical clusters
- Compatibility with the expected number of segments

Alternative algorithms considered included:

- DBSCAN (density-based spatial clustering)
- Hierarchical Clustering (agglomerative approach)

Parameter Selection

For K-Means clustering, the number of clusters (k=3) was determined based on:

1. Business context (three expected segments)
2. Elbow method analysis
3. Silhouette score optimization

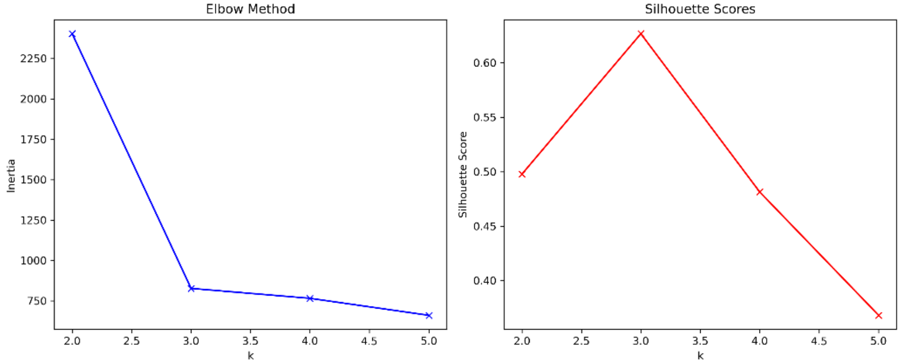

Figure 4: Elbow method and silhouette analysis for optimal cluster selection

The elbow method showed diminishing returns after k=3, while silhouette scores peaked at k=3, confirming the optimal number of clusters aligns with business expectations.

Model Evaluation

Quantitative Metrics

Several evaluation metrics were used to assess clustering quality:

K-Means Inertia	827.129
K-Means Silhouette Score	0.627
Davies-Bouldin Score	0.549
Calinski-Harabasz Score	2509.403

The metrics indicated:

- **Silhouette Score** of 0.627 suggests reasonably well-separated clusters
- **Davies-Bouldin Score** of 0.549 (lower is better) indicates good cluster separation
- **Calinski-Harabasz Score** of 2509.403 (higher is better) suggests well-defined clusters

Comparative Analysis

The primary K-Means model was compared with alternative approaches:

DBSCAN Number of Clusters: 4
DBSCAN Number of Noise Points: 93
DBSCAN Silhouette Score: 0.368

Hierarchical Clustering Silhouette Score: 0.627

K-Means outperformed both alternatives based on silhouette scores, validating its selection as the primary model.

Silhouette Analysis

A detailed silhouette plot provided visual confirmation of cluster quality.

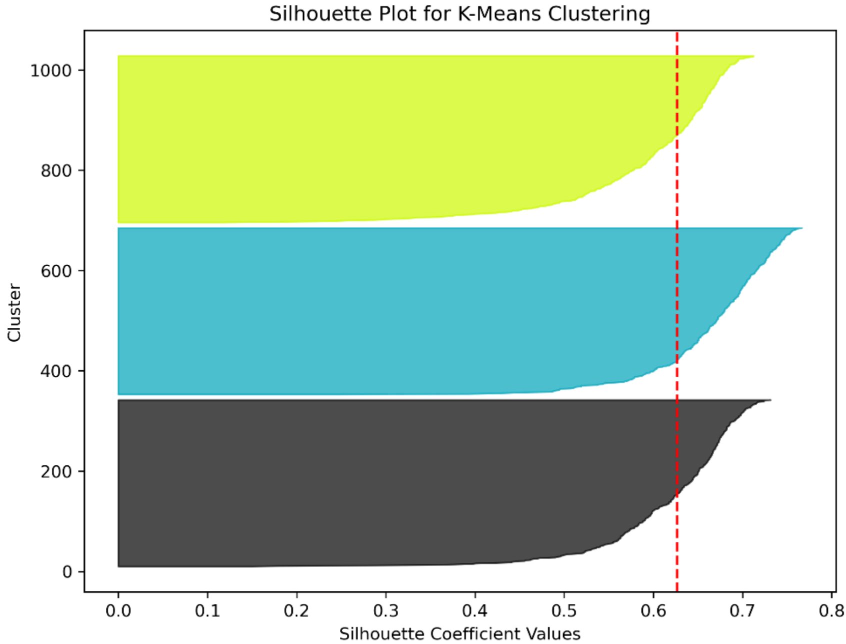

Figure 5: Silhouette plot showing the quality of individual cluster assignments

Most samples showed positive silhouette values well above the average (red dashed line), indicating good cluster assignment.

Hierarchical Structure Analysis

A dendrogram was generated to visualize the hierarchical relationship between data points.

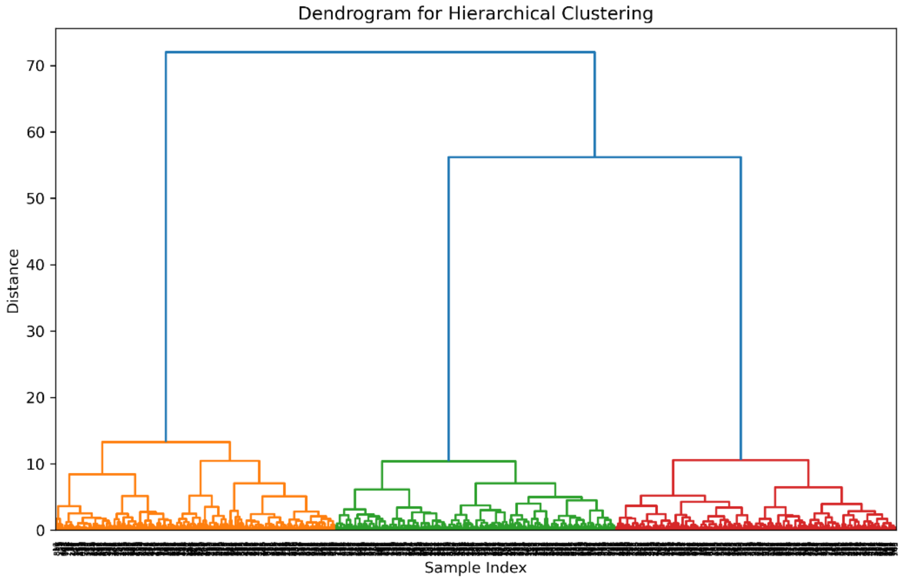

Figure 6: Dendrogram showing hierarchical relationships in the data

The dendrogram supported a natural separation into three main clusters, consistent with the K-Means results.

Identifying Clusters

Cluster Profiles

After application of K-Means clustering, three distinct segments emerged:

Cluster Profiles (Mean Values):

Cluster	total_purchases	avg_cart_value	total_time_spent	product_click	discount_counts
0	10.18	147.06	40.39	19.90	1.95
1	4.86	49.05	90.14	49.71	1.02
2	19.66	30.43	17.51	14.92	9.97

Cluster Visualization

Multiple visualization techniques were used to examine cluster separation.

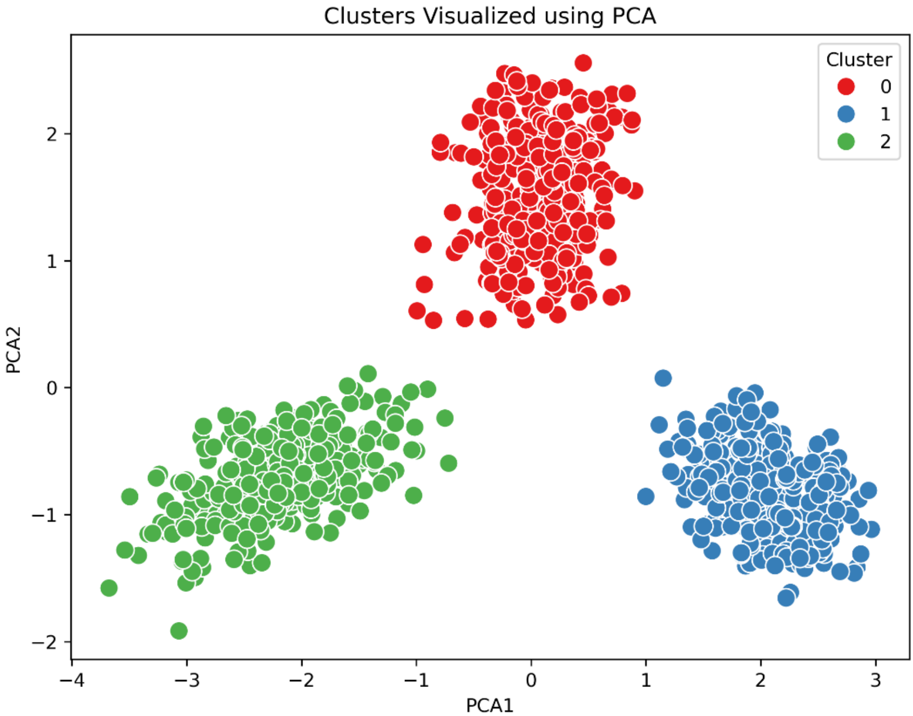

Figure 7: PCA visualization of the three identified clusters

The PCA explained variance ratio was [0.637, 0.261], indicating that 89.8% of the variance was captured by the first two principal components.

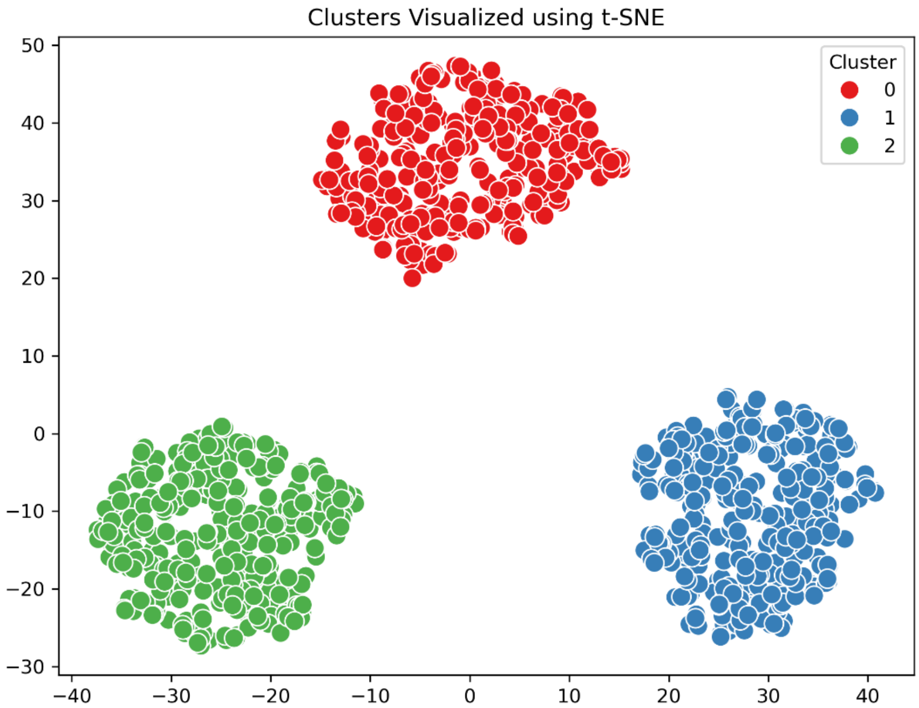

Figure 8: t-SNE visualization showing local structure of clusters

The t-SNE visualization revealed clear separation between clusters, with minimal overlap, confirming the effectiveness of the segmentation.

Feature Distributions by Cluster

Boxplots for each feature by cluster provided detailed insights into the distinguishing characteristics of each segment.

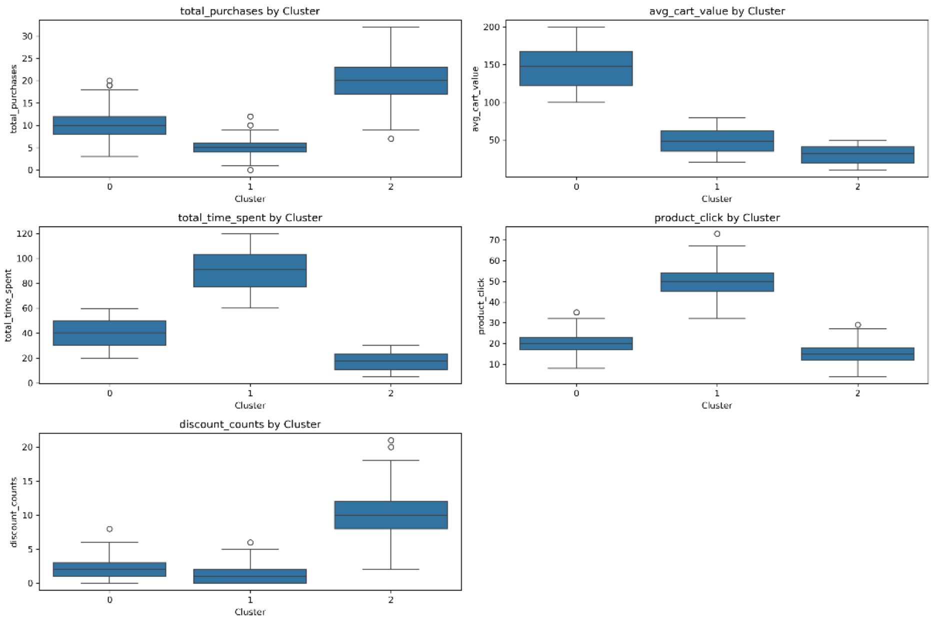

Figure 9: Feature distributions by cluster

The boxplots revealed:

- **Cluster 0(High Spenders)** showed moderate purchases, high cart value, moderate time spent, and low discount usage
- **Cluster 1(Window Shoppers)** showed low purchases, low cart value, high time spent, and high product clicks
- **Cluster 2(Bargain Hunters)** showed high purchases, moderate cart value, low time spent, and high discount usage

Statistical Validation

ANOVA tests confirmed significant differences between clusters for all features:

ANOVA for total_purchases: F-statistic= 1639.41, p-value=0.0000
ANOVA for avg_cart_value: F-statistic= 3244.94, p-value=0.0000
ANOVA for total_time_spent: F-statistic= 3035.21, p-value=0.0000
ANOVA for product_click: F-statistic= 4141.79, p-value=0.0000
ANOVA for discount_counts: F-statistic= 1833.10, p-value=0.0000

All p-values were extremely small (<0.0001), confirming that the differences observed between clusters are statistically significant.

Segment Mapping

Based on the cluster profiles, we mapped the clusters to the expected customer segments:

Cluster 0 → High Spenders (32.8% of customers)
Cluster 1 → Window Shoppers (37.5% of customers)
Cluster 2 → Bargain Hunters (29.7% of customers)

Segment Characteristics

1. **High Spenders (Cluster 0)**
   - Moderate purchases (10.18)
   - High cart value (147.06)
   - Moderate time spent (40.39)
   - Moderate product clicks (19.90)
   - Low discount usage (1.95)
2. **Window Shoppers (Cluster 1)**
   - Low purchases (4.86)
   - Low cart value (49.05)
   - High time spent (90.14)
   - High product clicks (49.71)
   - Moderate discount usage (1.02)
3. **Bargain Hunters (Cluster 2)**
   - High purchases (19.66)
   - Moderate cart value (30.43)
   - Low time spent (17.51)
   - Low product clicks (14.92)
   - High discount usage (9.97)

Business Recommendations

Based on the identified customer segments, we recommend the following targeted strategies:

1. High Spenders (Cluster 0)

- Develop premium loyalty programs with exclusive benefits
- Create personalized recommendations for high-value products
- Implement VIP customer service channels
- Design early access to new products or collections
- Focus on quality and exclusivity in marketing communications

2. Window Shoppers (Cluster 1)

- Implement retargeting campaigns to convert browsing to purchases
- Offer limited-time incentives to create urgency
- Improve user experience to streamline the purchase process
- Create engaging content to maintain their interest
- Develop abandoned cart recovery strategies with compelling offers

3. Bargain Hunters (Cluster 2)

- Design bundle deals and volume discounts
- Create a structured loyalty program that rewards frequent purchases
- Implement flash sales and limited-time offers
- Develop a targeted email campaign for promotions and discounts
- Offer exclusive early access to clearance events

Conclusion

This customer segmentation analysis successfully identified three distinct customer segments that align with the expected profiles. The K-Means clustering model demonstrated strong performance metrics and clear cluster separation.

The identified segments provide valuable insights into customer behavior that can inform targeted marketing strategies, product development, and customer experience enhancements. By tailoring approaches to each segment, the e-commerce platform can optimize customer engagement, increase conversion rates, and improve overall business performance.

Future work could include:

- Developing a real-time segmentation system to classify new customers
- Performing longitudinal analysis to track segment shifts over time
- Integrating additional behavioral or demographic data to refine segments further
- Conducting A/B testing of segment-specific marketing strategies