Task 02 - Customer Segmentation Report

Madhushankha De Silva

data_wizards

University of Moratuwa, Colombo, Sri Lanka

madhushankhadesgmail.com

March 9, 2025

Abstract

Customer segmentation is a crucial aspect of e-commerce businesses, enabling companies to tailor marketing strategies based on consumer behavior. This report presents an analysis of customer segmentation using clustering techniques. The study involves data preprocessing, exploratory data analysis, model selection, and evaluation to identify distinct customer groups. We leverage K-means and hierarchical clustering to identify behavioral patterns in customers.

1 Introduction

Customer segmentation involves dividing customers into distinct groups based on their behavior. In this project, we analyze an e-commerce dataset containing features such as total purchases, average cart value, and time spent on the platform. The objective is to identify three predefined customer segments: Bargain Hunters, High Spenders, and Window Shoppers. We employ both K-means and hierarchical clustering for this purpose.

2 Dataset Description

The dataset consists of customer behavioral data collected from an e-commerce platform. It contains the following six features:

- **customer_id**: Unique identifier for each customer.
- **total_purchases**: Total number of purchases made by the customer.
- **avg_cart_value**: Average value of items in the customer's cart.
- **total_time_spent**: Total time spent on the platform (in minutes).
- **product_click**: Number of products viewed by the customer.
- **discount_count**: Number of times the customer used a discount code.

2.1 Data Sample

Table 1: Sample Data from Customer Segmentation Dataset

total_purchases	avg_cart_value	total_time_spent	product_click	discount_count	customer_id
7.0	129.34	52.17	18.0	0.0	CM00000
22.0	24.18	9.19	15.0	7.0	CM00001
2.0	32.18	90.69	50.0	2.0	CM00002
25.0	26.85	11.22	16.0	10.0	CM00003
7.0	125.45	34.19	30.0	3.0	CM00004

The dataset contains three hidden clusters representing distinct customer segments: Bargain Hunters, High Spenders, and Window Shoppers.

3 Methodology

3.1 Data Cleaning

Before proceeding with the analysis, we addressed the issue of missing values in the dataset. For any missing values in numerical columns, we used the median of the respective column to perform imputation. The median was chosen as it is less sensitive

to outliers compared to the mean. This step ensures that the dataset remains complete, preventing the loss of valuable information due to missing data.

The following steps were undertaken for data cleaning:

- **Identifying Null Values:** We identified the columns with missing values and assessed the extent of missing data.
- **Imputation with Median:** For each column with missing values, the median of that column was computed and used to fill the missing entries.
- **Verification:** After imputation, we verified that there were no remaining null values in the dataset.

3.2 Data Preprocessing

We handled missing values by imputing them with their columns' median, identified outliers using pair plots, and standardized numerical features using the StandardScaler to ensure uniform feature scaling.

- **Customer ID Formatting:** The *customer_id* feature was converted into a numeric value to facilitate better processing and eliminate any categorical inconsistencies.
- **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset, retaining the most significant features that contribute to variance in customer behavior. This helped in improving clustering efficiency and visualization.

3.3 Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution of features. We visualized feature distributions using histograms and boxplots, and computed summary statistics. Correlation analysis revealed that total purchases and discount count have a strong relationship.

3.3.1 Identifying Outliers with Boxplots

Boxplots were used to detect outliers by observing points that extend beyond the whiskers. Outliers were identified in features such as *discount_counts*, which required further investigation and treatment using statistical methods like the z-score.

3.3.2 Understanding Distributions with Histograms

Histograms provided insight into the distribution of each feature. Features like *total_purchases* exhibited a right-skewed distribution, indicating the presence of a few customers making significantly more purchases than the average. The *avg_cart_value* showed a bimodal distribution, suggesting two different purchasing behaviors among customers.

3.3.3 Insights from Pair Plots

Pair plots revealed relationships between numerical features. For example, customers with high *total_purchases* often had a lower *avg_cart_value*, reinforcing the segmentation hypothesis that Bargain Hunters make frequent but low-value purchases.

3.3.4 Correlation Matrix Analysis

A correlation heatmap was generated to examine feature dependencies. Strong positive correlations were observed between *total_purchases* and *discount_count*, suggesting that frequent shoppers tend to use more discount codes. Conversely, *avg_cart_value* had a weak correlation with most other features, indicating it may independently define customer segmentation.

3.4 Model Selection

Model selection plays a crucial role in identifying the most suitable approach for clustering customers based on their behaviors. In this project, we experimented with two clustering techniques: K-means and hierarchical clustering. Each method offers distinct advantages, and the choice of technique was driven by the nature of the data and the specific requirements of customer segmentation.

3.4.1 K-means Clustering

K-means clustering is one of the most widely used clustering algorithms due to its efficiency in partitioning data into a predefined number of clusters. The key advantage of K-means lies in its simplicity and scalability, making it suitable for large datasets. However, the performance of K-means is highly dependent on the initial placement of centroids and the assumption that clusters are spherical and equally sized.

Ensuring the Optimal Number of Clusters To ensure the optimal number of clusters, we applied the following:

- **Elbow Method:** This method involves plotting the explained variation as a function of the number of clusters and identifying the point where the improvement in variance starts to diminish (the "elbow"). This provides a heuristic for selecting the optimal number of clusters.

it suggested that three clusters would be optimal for the customer segmentation task, aligning with the expected groups: Bargain Hunters, High Spenders, and Window Shoppers.

K-means Model Evaluation After fitting the K-means model, we used the silhouette score to evaluate the clustering performance. The silhouette score was found to be 0.597034032119, indicating reasonably well-separated clusters. We achieved satisfactory results, suggesting that K-means was an effective technique for this dataset.

3.4.2 Hierarchical Clustering

Hierarchical clustering, unlike K-means, does not require the number of clusters to be specified in advance. This makes it a more flexible approach, especially when the number of clusters is unknown. We used agglomerative hierarchical clustering, which starts by treating each point as its own cluster and iteratively merges the closest clusters.

Linkage Criteria For hierarchical clustering, we experimented with different linkage methods to determine the best approach:

- **Ward's Method:** This method minimizes the variance within clusters, ensuring

that merged clusters have the smallest possible increase in within-cluster variance. It is particularly useful when trying to create clusters with similar sizes.

- **Single Linkage and Complete Linkage:** These methods focus on the distance between clusters based on either the closest or the farthest points, respectively. However, these methods tend to produce chains of clusters that can sometimes lead to poor results when the data contains noise.

Hierarchical Clustering Model Selection To evaluate the results of hierarchical clustering, we used the same evaluation metrics as K-means (silhouette score). Since the silhouette score was slightly higher(0.5994942415582004) than that of K-means (0.597034032119), hierarchical clustering provided useful insights into the potential for alternative segmentation structures with different numbers of clusters. Visualizing the dendrogram also helped to understand the merging process and decide the optimal cut-off for the number of clusters.Therefore Agglomerative Clustering model was selected for evaluation.

3.5 Model Evaluation

The effectiveness of the clustering model was assessed through visualizations to examine the distribution and behavior of customer segments. The cluster distribution was fairly balanced, with no significant imbalances across the groups.

A scatter plot of 'total_time_spent' vs 'total_purchases' highlighted three distinct customer profiles:

- **Cluster 0 (Window Shoppers):** had high browsing time but low purchase frequency.
- **Cluster 1 (Bargain Hunters):** had low browsing time but high purchasing frequency.
- **Cluster 2 (High Spenders):** exhibited moderate values for both features.

Analysis of 'total_purchases' confirmed that Cluster 1 had the highest frequency of purchases, while Cluster 2 represented high-value transactions. Discount usage was highest

in Cluster 1, confirming the bargain-seeking behavior of this group. Product click analysis showed that Cluster 0 viewed more products, supporting the profile of customers who engage with the platform without converting to purchases. Lastly, Cluster 2 had the highest average cart value, indicating high spending behavior.

These findings confirm the clustering model's ability to effectively segment customers based on distinct behavioral patterns.

4 Results

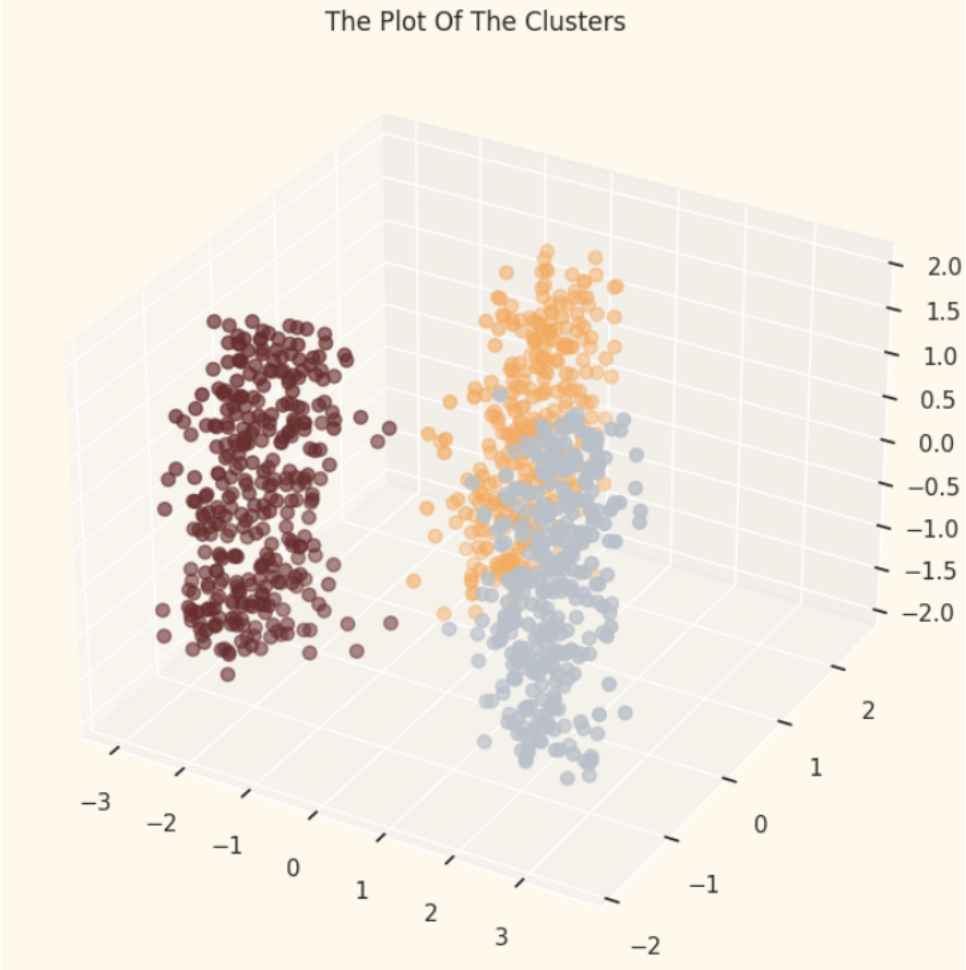

Figure 1: Customer Segmentation Clusters

The clustering model successfully identified three distinct customer groups. The Agglomerative Clustering approach resulted in a silhouette score of 0.597034032119383, confirming a well-separated cluster structure.

5 Discussion

The segmentation results align with the expected customer groups:

- Bargain Hunters: Frequent buyers with a preference for discounts.
- High Spenders: Customers with fewer but high-value purchases.
- Window Shoppers: Users who browse extensively but purchase rarely.

These insights enable businesses to tailor targeted marketing campaigns. High spenders can be encouraged with premium loyalty programs, while bargain hunters respond well to limited-time discounts.

6 Challenges and Future Improvements

During the analysis, we encountered several challenges:

- **Feature Scaling:** Ensuring that different features contributed equally required careful normalization.
- **Cluster Overlap:** Some customers exhibited behaviors spanning multiple segments, making classification less distinct.

To improve model performance, we suggest:

- Exploring additional clustering algorithms such as DBSCAN for better handling of noisy data.
- Incorporating advanced feature engineering techniques to derive new insights from existing data.
- Implementing deep learning-based clustering approaches to capture complex behavioral patterns.

7 Conclusion

This project demonstrated the effectiveness of clustering techniques in customer segmentation. The insights derived can help businesses develop targeted marketing strategies to improve customer engagement and retention.