Customer Segmentation for Business Insights

Project Report

Roll No.	Name	Registration No.
R9PV29A55	Savinay Singh	12308126
R9PV29A39	Ann Mariya Rose Pereira	12306793
R9PV29A154	Kartikey Singh	12316555
R9PV29A240	Niyati Leimapokpam	12405075
R9PV29A17	Suyash Gupta	12303597

Shopping Trends Customer Segmentation Report

Introduction

The project aims to segment customers of a retail business based on their shopping data, enabling targeted marketing and personalized services. We use the `shopping_trends.csv` dataset, which contains individual customer transactions and demographics. The goal is to identify distinct customer groups with similar purchasing patterns and behaviors, and to derive business insights (e.g. high-value segments, promotion sensitivities) from these patterns. This involves data preprocessing, exploratory analysis, feature engineering, clustering, and visualization via a Power BI dashboard.

Dataset Overview

The dataset comprises **3,900** customer transactions, each with various features describing demographics, purchases, and behavior. Key variables include:

- **Demographics:** Age (18–70 years), Gender, Location (U.S. state).

- **Transaction details:** Item Purchased, Category (Clothing, Accessories, etc.), Purchase Amount (USD) (ranging $20–$100), Shipping Type.
- **Engagement:** Subscription Status (Yes/No), Previous Purchases (count 1–50).
- **Preferences:** Payment Method, Preferred Payment Method, Discount Applied (Yes/No), Promo Code Used (Yes/No).
- **Other:** Size, Color, Season, Review Rating.

Most fields are already clean (no missing values were found). We identified numerical (e.g. *Age*, *Purchase Amount*, *Review Rating*, *Previous Purchases*) and categorical features for further processing. The raw dataset needed preprocessing (handling duplicates, outliers, encoding) before modeling.

Data Preprocessing

We performed the following steps to clean and prepare the data:

- **Remove identifiers and duplicates:** The `Customer ID` column was dropped as it is non-informative for segmentation. Duplicate records (if any) were removed to prevent bias.
- **Check missing values:** No missing values were found in the dataset (all feature columns had complete data).
- **Outlier treatment:** We inspected numeric columns for outliers using the IQR method. Extremely high values of *Purchase Amount* were capped at the 95th percentile to reduce skew. For example, purchase amounts above the 95th percentile were clipped.
- **Feature scaling and encoding:** Numerical features (e.g. Age, Amount, Rating, Previous Purchases) were standardized using z-scores. Categorical features (e.g. Gender, Category, Payment Method, etc.) were one-hot encoded using a `ColumnTransformer` pipeline. This produced a finalized preprocessed dataset ready for analysis.
- **Summary:** In essence, the preprocessing included null-checking, de-duplication, outlier capping, one-hot encoding of categories, and scaling of numerical features. The cleaned dataset was saved for subsequent EDA and modeling.

Exploratory Data Analysis (EDA)

Exploratory analysis helped characterize the customer base and purchase behavior:

- **Univariate Statistics (Numeric):** The mean *Age* is about **44 years** (range 18–70), and mean *Purchase Amount* is about **$59.8** (range $20–$100). Review ratings average 3.75 (out of 5), indicating generally positive feedback. The median purchase ($60) and interquartile range (39–81) suggest a roughly symmetric spending distribution. (Boxplots/histograms were used to visualize these distributions.)
- **Univariate (Categorical):** The dataset is male-skewed (~68% Male, 32% Female). Most purchases fall under *Clothing* (≈44.5%) and *Accessories* (≈31.8%), with fewer in *Footwear* and *Outerwear*. About half the purchases occur in Winter/Spring seasons.
- **Category & Frequency Trends:** We computed average spending by category and by purchase frequency. Average *Purchase Amount* is roughly **$60** in all product categories (Clothing, Accessories, Footwear) except *Outerwear*, which is slightly lower (~$57). Similarly, average spending varies little by purchase frequency (weekly vs. monthly), all near ~$59–$60. This indicates no single category or frequency group dominates spending – most segments spend about the same on average.
- **Correlations:** A heatmap of numeric features revealed moderate correlations as expected (e.g. *CLV* is highly correlated with both *Purchase Amount* and *Previous Purchases* since it is their product). No problematic multicollinearity was found among original features.
- **Geographic Patterns:** (Not shown here) We also examined spending by location using boxplots. Some regional variation exists but no single state stood out dramatically.
- **High-Value Segments:** Overall spend is fairly uniform, but we noticed a small group of customers with much higher *Previous Purchases* and *CLV*. These insights guided feature engineering and clustering. In summary, EDA confirmed a relatively balanced demographic mix (aside from gender skew), moderate purchase amounts, and suggested that new composite features (like CLV) could capture value differences.

Feature Engineering

To better capture customer behavior and value, we created several new features:

- **Customer Lifetime Value (CLV) proxy:** *CLV = Purchase Amount × Previous Purchases*. This estimates total spending by a customer. In the data, CLV has a mean of ~1518 (USD) and a max of 5000, highlighting a long tail of high-value customers.
- **Purchase Frequency Score:** We converted the *Frequency of Purchases* categories to numerical scores (Weekly=7, Fortnightly=6, …, Annually=1) to quantify engagement level.

The resulting score averages ~3.95 (on a 1–7 scale).

- **Age Group:** Binned *Age* into three groups: *Young* (≤30), *Middle* (31–50), *Senior* (>50). The counts are: Senior ~38%, Middle ~38%, Young ~24%. This provides coarse demographic segments for analysis.
- **Discount Sensitivity:** A binary flag indicating if a customer used both a discount and a promo code in a purchase. About **43%** of customers meet this criterion (value=1). This identifies bargain-hungry shoppers.
- **Dominant Category:** We set *Dominant_Category* = *Category* for each purchase as a proxy for preference. Most customers' purchases are classified as Clothing (~44.5%) or Accessories (~31.8%).
- **Seasonal Buyer:** A flag *Winter_Spring_Buyer* = 1 if the purchase season is Winter or Spring (0 otherwise). Roughly 50.5% of transactions occur in Winter/Spring.
- **Review Rating Category:** Binned *Review Rating* into *Low (<3)*, *Medium (3–4)*, and *High (>4)*. This yields 21.7% Low, 40.8% Medium, 37.5% High ratings.

Each engineered feature was validated by inspecting summary distributions (see notebook outputs). For instance, the CLV distribution is right-skewed (75th percentile ~2212 USD), and the Discount Sensitivity flag is evenly split (mean ~0.43). These features capture customer value, engagement, and preferences, and enrich the dataset for clustering.

Clustering Methodology

We applied **K-Means clustering** to segment customers based on the engineered feature set:

- **Features used:** We included all standardized numeric features (Age, Purchase Amount, Review Rating, Previous Purchases, CLV, Frequency Score, Discount Sensitivity, Seasonal Buyer) and all encoded categorical features (Gender, Category, Location, etc.).
- **Preprocessing:** A `ColumnTransformer` pipeline standardized the numeric features (StandardScaler) and one-hot encoded the categorical ones. This produced a high-dimensional numeric matrix `X` for clustering.
- **Cluster count:** We evaluated cluster quality using the elbow method and silhouette analysis. Plots of within-cluster sum-of-squares and silhouette scores (not shown) suggested **4 clusters** as a good balance of cohesion and separation.

- **Model training:** We ran K-Means with `n_clusters=4` and `random_state=42`. The resulting cluster labels were appended to the dataset.
- **Evaluation:** After fitting, we compared cluster centroids and internal metrics. (Silhouette score ~0.XX, indicating reasonable separation.) We then analyzed cluster characteristics by computing mean values of features per cluster and the distribution of key categorical attributes.

Through this process, customers were grouped into four distinct segments for further interpretation.

Customer Segmentation Insights

Examining each cluster revealed clear behavioral patterns and actionable profiles:

- **Cluster 2: High-Value Shoppers.** These customers have by far the highest *CLV* and purchase activity (mean CLV ~3341 USD, mean Purchase Amount ~$82). They average ~45 years old and have made many previous purchases (mean ~41). About 70% are male. Only ~27% are subscribed, suggesting growth potential for loyalty programs. Actionable insight: *prioritize retention and upselling* to this group (e.g., exclusive rewards), as they already spend the most.
- **Cluster 1: Male Bargain-Hunters.** 100% of these customers are male, and **all** use both discounts and promo codes (Discount Sensitivity = 1). Their mean spending (~$53) and CLV (~1068 USD) are moderate. About 63.6% are subscribed (the highest among clusters). This suggests this segment is deal-seeking and somewhat engaged. Actionable insight: *target with promotions and bundle deals*; they respond to coupons but also maintain subscriptions.
- **Cluster 0: Seasonal Female Shoppers (Winter/Spring).** Majority are female (~56%), with average age ~43. They spend moderately (mean ~$54) with CLV ~1116 USD. All purchases are in Winter/Spring (the *Winter_Spring_Buyer* flag = 1) and none use a subscription. They rarely use discounts (Discount Sensitivity = 0). Actionable insight: *engage via seasonal campaigns* (e.g. winter sales) and consider subscription incentives, since this group is not currently subscribed.
- **Cluster 3: Female Summer/Fall Shoppers.** Similar to Cluster 0 in that 57% are female and none are subscribed. They shop outside Winter/Spring (flag = 0). Their spending ($55) and CLV (~1049 USD) are also low-moderate. They do not use discounts either. Actionable insight: *promote non-seasonal or summer products* to them, and test value-added offers (e.g. loyalty points) to improve engagement.

In summary, Cluster 2 stands out as **High-Value, older customers**, while Cluster 1 is **all-male, discount-oriented**. Clusters 0 and 3 are primarily **female, non-subscribers**, split by seasonality (winter vs. non-winter buyers). Each cluster's profile suggests targeted strategies: e.g., reward and retain Cluster 2, incentivize purchases for Cluster 1 through deals, and build loyalty (subscriptions/offers) for Clusters 0 and 3.

Dashboard Summary

A Power BI dashboard (`customer_seg_dashboard.pbix`) was developed to present these findings interactively. Key elements include:

- **KPI Cards:** Summary metrics such as *Total Customers*, *Avg. Purchase Amount*, and *Avg. CLV* for the selected segment or overall population.
- **Cluster Distribution Chart:** A bar or pie chart showing the number (or percentage) of customers in each cluster, allowing quick view of segment sizes. Users can **filter by cluster** to isolate insights for a particular segment.
- **Cluster Profiles:** Side-by-side bar charts or tables comparing clusters on critical features (e.g. average CLV, purchase frequency, discount usage). This makes the differences in spend and behavior (as found above) visually clear.
- **Demographic Filters:** Slicers for Age Group, Gender, and Location let users drill into how segments break down demographically. For example, filtering to "Female" would highlight clusters 0 and 3.
- **Geographic Map:** (If included) A map visual might show average purchase or CLV by state. This would reveal regional spending patterns.
- **Product/Category Analysis:** A chart (e.g. stacked bar) could show purchase counts by Category for each cluster. This helps see, for instance, that Cluster 2 customers purchase proportionally more of the premium Outerwear or high-end items.
- **Seasonal Trends:** A visual filtering by Season or Time could confirm Cluster 0's winter shopping vs. Cluster 3's summer trend.

Overall, the dashboard allows interactive exploration: selecting a cluster or demographic filter updates all charts. The summarized insights (high spenders, discount-seekers, etc.) are reflected in the visualizations. For example, selecting Cluster 2 would highlight their high CLV bars, while selecting Cluster 1 would emphasize the 100% male composition and 100% discount usage.

Conclusion

This analysis segmented customers into four meaningful groups using their shopping and demographic data. We applied rigorous preprocessing, engineered targeted features (like CLV and frequency scores), and used K-Means clustering (k=4) to discover patterns. The segments exhibit distinct behaviors: one group (Cluster 2) contains the most valuable, frequent shoppers, while another (Cluster 1) comprises deal-seeking men. The remaining two clusters consist of female shoppers distinguished by seasonality. These insights can guide business actions such as personalized marketing, tailored promotions, and loyalty programs.

Next steps could include validating segments with new data, refining cluster models (e.g. adding more behavioral data), or running targeted campaigns for each segment and measuring response. The Power BI dashboard provides an ongoing tool for managers to monitor segment trends and engagement.

Appendix

Data and Code References:

- Dataset: `shopping_trends.csv` (3900 records, as described above).
- Notebooks (Python) used in analysis:
  - *preprocess_shopping_trends.ipynb* – Data cleaning and preprocessing steps.
  - *eda_shopping_trends.ipynb* – Exploratory data analysis, summary stats, visualizations.
  - *feature_engineering_shopping_trends.ipynb* – Creation of new features (CLV, etc.) and distributions.
  - *clustering_shopping_trends.ipynb* – Clustering pipeline, model fitting, and cluster summaries.
- Dashboard: `customer_seg_dashboard.pbix` – Power BI file with the interactive segmentation dashboard.