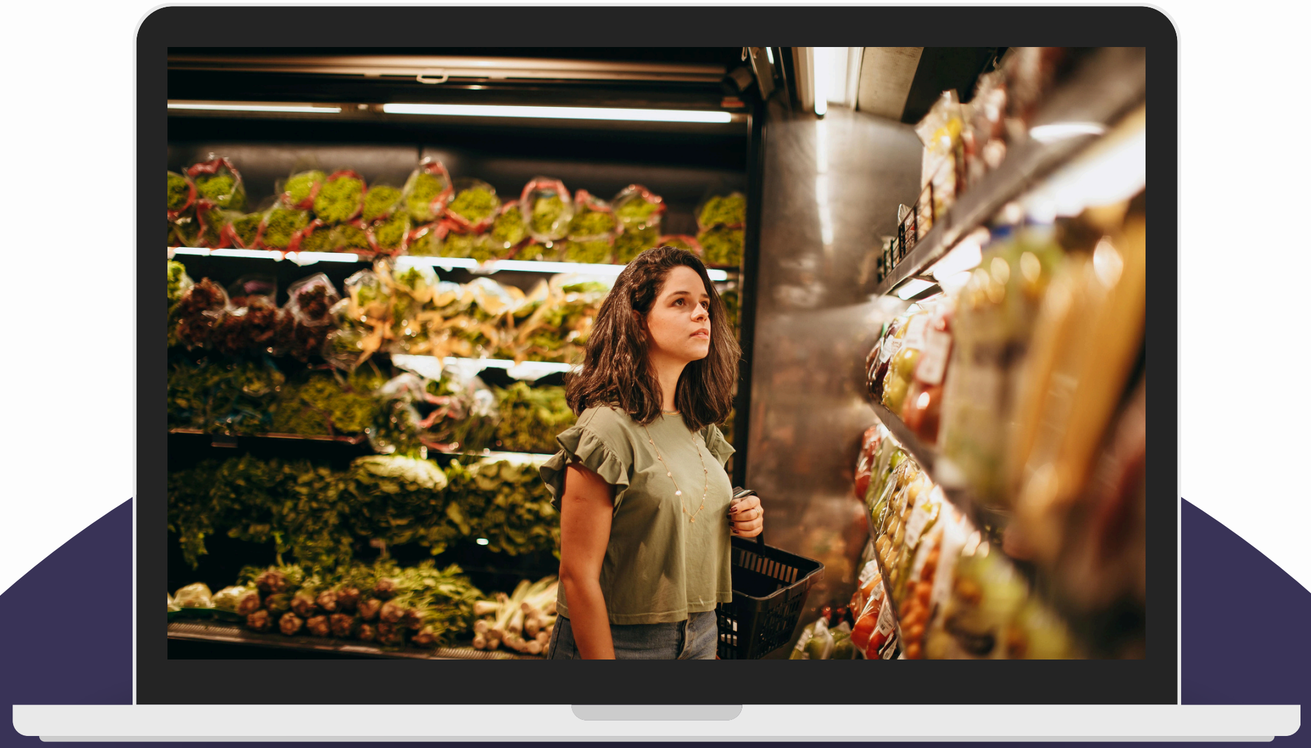
ASIAN INSTITUTE OF MANAGEMENT

Capstone Project | Christian Paul Firme

Customer Segmentation Using Unsupervised Learning



Problems

To drive more effective marketing and promotional decisions, companies need a deeper understanding of their customer base. While customers differ in demographics, purchasing behavior, loyalty, and promotion usage, they are currently treated as a largely homogeneous group—limiting the ability to personalize strategies and optimize spend.

Lack of Customer Differentiation

Customers with different behaviors, preferences, and value levels are treated similarly. This prevents the company from tailoring offers, messaging, and experiences to distinct customer needs.

Inefficient Marketing & Promotions

Without clear customer segments, campaigns tend to be generic and broadly targeted, increasing the risk of wasted promotional spend and lower campaign effectiveness.

No Data-Driven Segmentation Framework

There is no systematic, data-driven approach to identify and understand distinct customer groups, making it difficult to design actionable strategies and measure impact across segments.

Objectives

This project applies unsupervised learning techniques to uncover hidden structure within customer data. The focus is on identifying meaningful customer groupings without relying on predefined labels, enabling exploratory analysis and actionable insights for segmentation-driven decision making.

Identify Natural Customer Segments

Group customers into distinct clusters such that customers within the same cluster exhibit similar demographic and behavioral characteristics, while customers across clusters are meaningfully different.

Evaluate Multiple Clustering Algorithms

Apply and compare K-Means, Hierarchical Clustering, and DBSCAN to assess cluster quality, stability, scalability, and suitability for different data structures and density patterns.

Clarify Scope of the Machine Learning Task

Ensure the problem is correctly framed as clustering—not classification, regression, recommendation, or anomaly detection—while acknowledging that outliers and noise may emerge as useful by-products of the analysis.

Business Impact

The customer segmentation results are evaluated based on their ability to drive tangible business value. Rather than focusing solely on technical clustering quality, success is defined by how well the segments support better decision-making across marketing, promotions, and customer strategy.

Ultimately, the impact of this project lies in enabling more targeted, cost-efficient, and customer-centric actions across organizations.

01

Improved Promotion Targeting. Enable differentiated promotion strategies by customer segment (e.g., high-frequency vs low-frequency buyers), reducing reliance on one-size-fits-all campaigns.

02

Cost Savings from Reduced Blanket Promotions. Minimize inefficient mass promotions by focusing spend on segments most likely to respond, improving return on marketing and promotional investments.

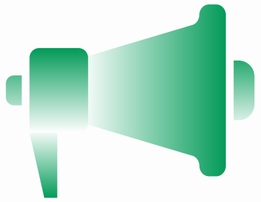
03

Increased Marketing Uplift through Personalization. Drive higher conversion and engagement by aligning offers, messaging, and incentives with the specific behaviors and preferences of each customer segment.

04

Better Visibility into Customer Value & Risk. Provide clearer insights into high-value, loyal, and at-risk customer segments—supporting loyalty programs, retention strategies, and long-term customer growth initiatives.

Success Metrics



Promotion Effectiveness & Targeting

Measure improvements in promotion targeting by comparing response rates, conversion, or uplift across customer segments (e.g., high-frequency vs low-frequency buyers), relative to historical blanket promotions.

Marketing Efficiency & ROI

Assess potential cost savings and incremental returns from reduced blanket promotions and more personalized offers, reflected in higher marketing ROI and improved customer satisfaction metrics.

Project-Specfic Metrics (Data Science)

Clusters are evaluated using internal validation metrics (Silhouette Score, Elbow Method, and Davies–Bouldin Index) to ensure strong cohesion and separation.

As a result of this project, companies shall be able to assess effectiveness of their campaigns that are designed for different customer groups.

Data & Features — Feature Overview

The dataset captures multiple dimensions of customer information, enabling segmentation based on demographics, economic capacity, behavior, and relationship indicators. Features are selected to reflect real-world drivers of customer purchasing patterns and engagement.

Identifier

- Customer ID (unique identifier; excluded from modeling)

Demographic Features

- Age
- Gender
- Education
- Region

Economic Features

- Income

Behavioral Features

- Purchase frequency
- Purchase amount
- Promotion usage

Relationship / Attitudinal Features

- Loyalty status
- Satisfaction score

Product Preference

- Primary product category

Modeling Note: *Only customer attributes relevant to similarity measurement are used for clustering; identifiers are excluded to prevent artificial separation.*

Data Quality Considerations

Clustering algorithms are sensitive to scale, noise, and distributional issues. The following data quality considerations are addressed prior to modeling:

Missing Values	Possible in income, education, and satisfaction score
Outliers	Income and purchase amount may contain extreme values
Categorical Imbalance	Some regions or product categories may be underrepresented
Scaling Sensitivity	Numeric variables vary significantly in magnitude (e.g., age vs income)

Why This Matters: Distance-based clustering methods (e.g., K-Means, Hierarchical) are highly sensitive to scale and noise, making preprocessing and normalization critical for meaningful segmentation.

EDA & Feature Engineering (Overview)

Prepare a clean, well-encoded, and well-scaled feature matrix suitable for distance-based clustering, while ensuring interpretability and robustness to noise and outliers.

Data Cleaning

- Removed duplicate customer records
- Imputed missing values (median for numeric, mode for categorical)
- Capped extreme values using IQR-based winsorization

Feature Engineering

- Ordinal encoding for ordered categories (e.g., loyalty status, purchase frequency)
- One-hot encoding for nominal categories (e.g., gender, region, product category)
- Robust scaling applied to numeric features

EDA & Clustering Readiness

- Distribution and outlier checks (histograms, boxplots)
- Correlation analysis among numeric features
- Hopkins statistic to assess clustering tendency

Dimensionality Reduction & Interpretability

- PCA for compact representation and visualization
- t-SNE / UMAP for 2D cluster visualization
- Surrogate-model + SHAP used to interpret cluster drivers

Model Implementation & Evaluation (Overview)

Implement and compare multiple unsupervised clustering algorithms to identify meaningful customer segments, balancing statistical performance with scalability, stability, and business interpretability.

Models Evaluated

- K-Means: Efficient, scalable baseline for customer segmentation
- Hierarchical Clustering: Interpretable structure, limited by memory constraints
- DBSCAN: Density-based clustering for irregular shapes and noise detection

---

Evaluation Criteria

- Primary metric: Silhouette Score
- Supporting metrics: Davies–Bouldin, Calinski–Harabasz
- Qualitative checks: Interpretability, cluster coverage, scalability

Metrics focus on cluster cohesion and separation, not predictive accuracy.

Model Comparison & Selection

The best-performing configuration from each clustering approach was compared using internal validation metrics and business usability considerations.

Model	Configuration	Clusters	Silhouette	Davies-Bouldin	Calinski-Harabasz	Noise Points
DBSCAN	eps = 0.3, min_samples = 10	5	0.69	0.37	624.33	99,937
Hierarchical	Ward linkage, k = 2	2	0.25	1.51	767.23	0
K-Means	k = 2	2	0.21	1.7	20,554.09	0

Model Comparison & Selection

K-Means

Efficient, scalable baseline for customer segmentation

- Evaluated for k = 2 to 10
- Selected using Elbow Method and Silhouette Score
- Best result at k = 2, with highest silhouette score
- Strong scalability and full customer coverage

Hierarchical Clustering

Interpretable structure, limited by memory constraints

- Tested using PCA-reduced data and subsampling
- Best configuration: Ward linkage, k = 2
- Not selected due to memory constraints and weaker separation

DBSCAN

Density-based clustering for irregular shapes and noise detection

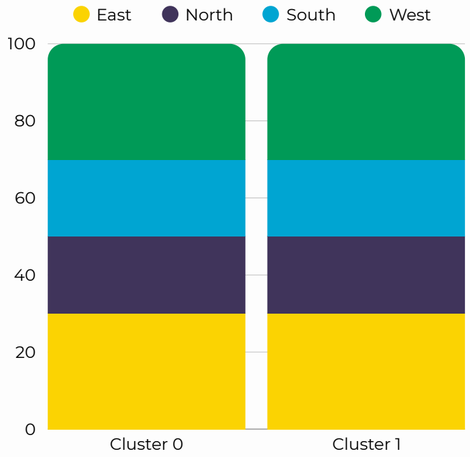
- Tuned using k-distance plot and parameter sweeps
- Identified dense clusters but labeled a large share of customers as noise
- Not suitable for full customer segmentation use case

Final model selected: K-Means. Chosen for its balance of: cluster quality, scalability, interpretability, business usability

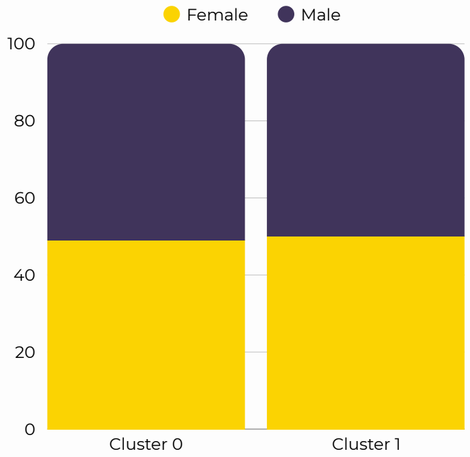
Key Results - Demographic Features

Cluster assignment is highly consistent across region, gender, age groups, and education, indicating that segmentation is **not driven by demographic attributes**.

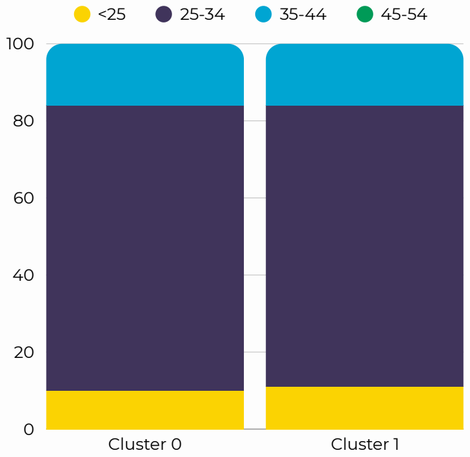
Region



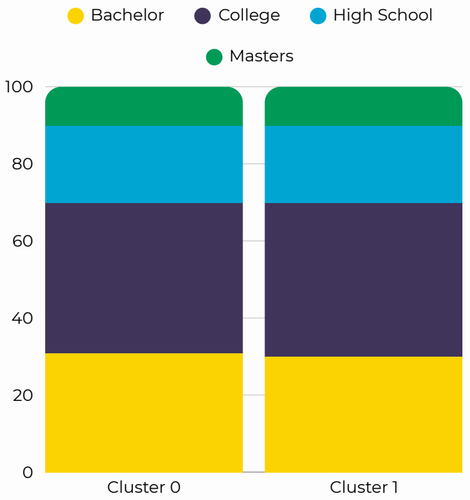
Gender



Age Group

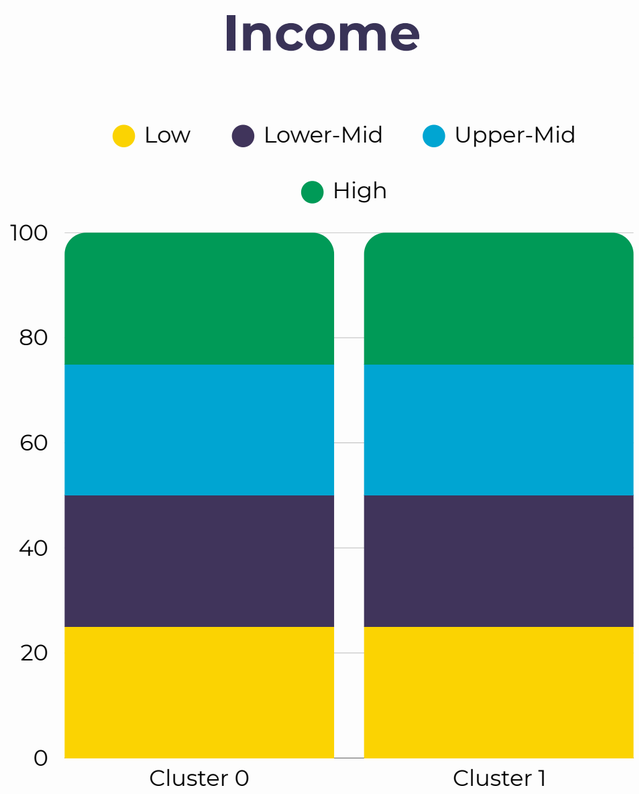


Education



Key Results - Economic Feature

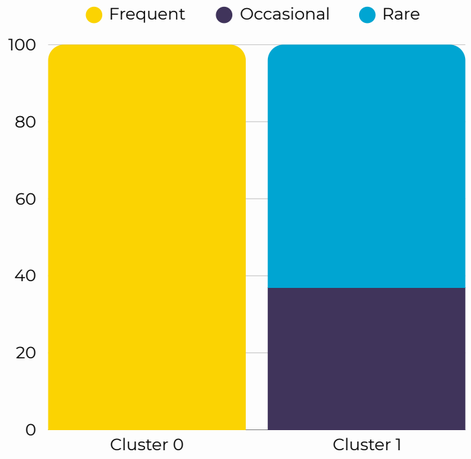
Cluster assignment is highly consistent across income bands, indicating that segmentation is **not driven by economic capacity**.



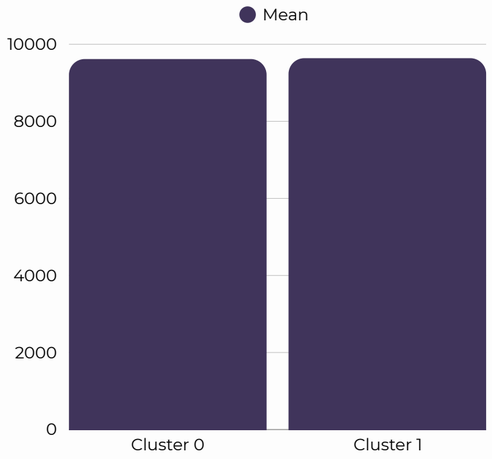
Key Results - Behavioral Features

Clusters are clearly differentiated by purchasing behavior, indicating that segmentation is **driven primarily by behavioral patterns** rather than demographic or economic attributes.

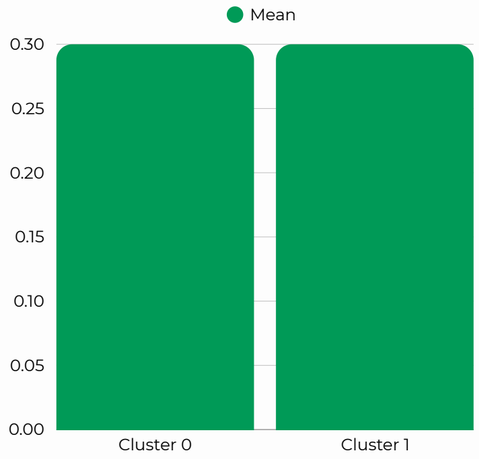
Purchase Frequency (%)



Ave. Purchase Amount



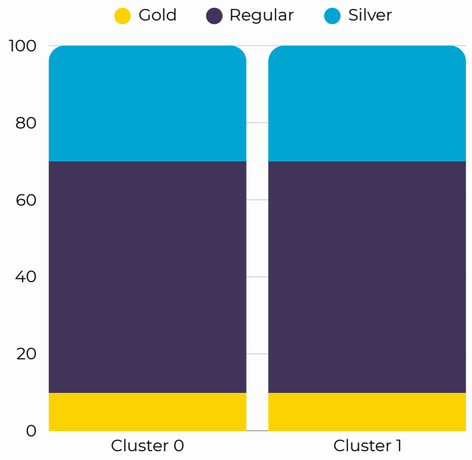
Ave. Promo Usage



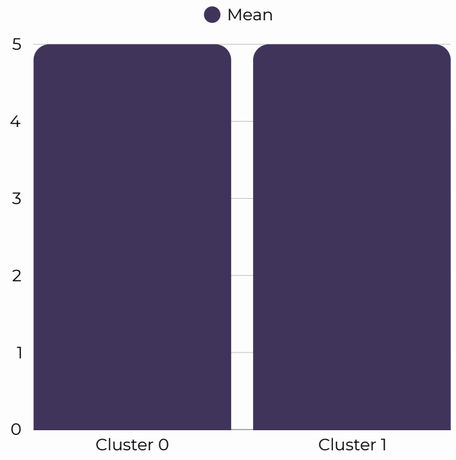
Key Results - Relationship Features

Relationship indicators show consistency across clusters, reinforcing the behavioral segmentation.

Loyalty Status (%)



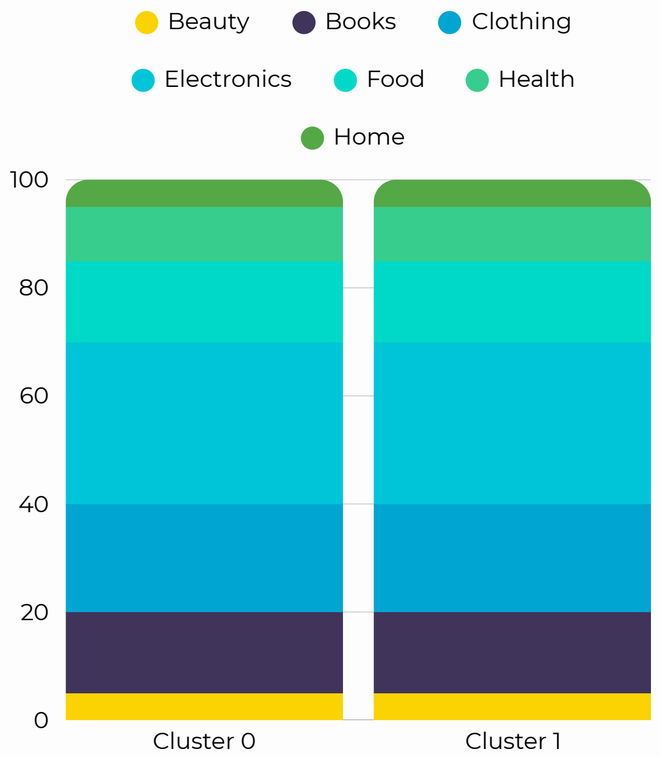
Ave. Satisfaction Score



Key Results - Product Preference Feature

Product category preferences are highly consistent across clusters, indicating that segmentation is **not driven by product choice.**

Product Preference



Segmentation Outcome

Customer segmentation resulted in two distinct clusters that are **clearly differentiated by purchasing behavior**, while remaining consistent across demographic, economic, relationship, and product preference attributes.

KEY FINDINGS

Behavioral Features (Primary Driver)

- Clusters differ meaningfully in purchase frequency, spending patterns, and promotion usage, indicating behavior-driven segmentation.

Demographic & Economic Features (Not Drivers)

- Cluster assignment is highly consistent across age, gender, region, education, and income bands, showing no demographic or economic bias.

Relationship & Product Preference Features (Not Drivers)

- Loyalty status, satisfaction scores, and product category preferences are evenly distributed across clusters and do not drive segmentation.

Business Implication: The resulting customer segments provide a behavior-based foundation for targeted marketing, promotion optimization, and personalization—without relying on sensitive or static customer attributes.