Capstone Project

Post-Graduate Diploma in Artificial Intelligence & Machine Learning

CUSTOMER SEGMENTATION

A Data-Driven Approach to Customer Profiling and Strategic Targeting

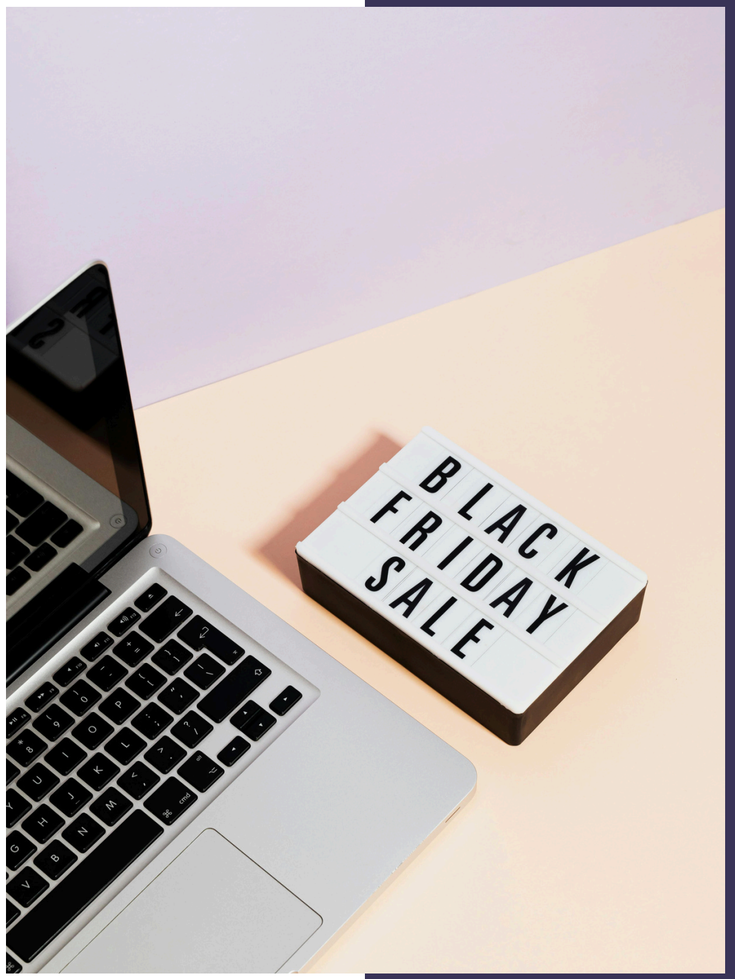
FRANZ FANGONILO

BUSINESS **PROBLEMS**

The business seeks to better understand its customer base to enable more targeted marketing, improve customer experience, and optimize promotional and retention spending.

However, customers are currently treated as a largely homogeneous group, despite exhibiting meaningful differences in:

- Purchasing behavior (frequency, spend, product mix),
- Sensitivity to discounts and promotions, and
- Loyalty and repeat purchasing patterns.



BUSINESS PROBLEMS

Undifferentiated Customers

Customers are treated as a single homogeneous group, despite clear differences in purchasing behavior, preferences, and engagement patterns. This limits the business's ability to tailor strategies to distinct customer needs.

Inefficient Marketing Spend

Promotions and campaigns are broadly deployed without understanding which customers are most responsive. This leads to diluted impact and suboptimal allocation of marketing budgets.

Low Campaign Effectiveness

Generic campaigns fail to resonate with diverse customer segments, resulting in lower conversion rates, weaker engagement, and reduced return on marketing investments.

Missed Personalization & CLV Growth

The lack of customer-level insights prevents personalized experiences, hindering loyalty building, repeat purchases, and long-term customer lifetime value growth.

Without data-driven segmentation, marketing investments are spread thin instead of targeted where they create the most value.

ASIAN INSTITUTE OF MANAGEMENT

OBJECTIVES

Identify Distinct Customer Segments

Identify distinct and actionable customer segments that exhibit similar purchasing behaviors, preferences, and engagement patterns.

Evaluate and Select Clustering Approaches

Apply and compare multiple clustering algorithms to assess segmentation quality, stability, and interpretability across varying data distributions and density patterns.

Ensure Correct Framing of the ML Task

Clearly define the problem as a clustering task, not classification or prediction, ensuring appropriate evaluation metrics and interpretation.

Translate Segments into Actionable Insights

Interpret and profile the resulting customer segments to enable targeted marketing, personalization strategies, and improved allocation of marketing and retention resources.

ENABLE DATA-DRIVEN DECISION MAKING FOR

BUSINESS IMPACT

More Effective Marketing & Promotions
Customer segmentation enables targeted campaign design, allowing the business to tailor offers, messaging, and promotions to distinct customer groups—leading to higher relevance, engagement, and conversion rates

Improved Customer Experience & Loyalty
Segment-driven insights support personalized customer experiences, increasing satisfaction, repeat purchases, and long-term customer lifetime value (CLV).

Optimized Marketing Spend
By identifying high-value and promotion-responsive segments, marketing and retention budgets can be allocated more efficiently, reducing wasted spend on low-impact customers and improving overall return on investment (ROI).

Better Strategic Decision-Making
The segmentation framework provides a data-driven foundation for downstream decisions such as pricing strategies, product bundling, loyalty programs, and customer prioritization.

SUCCESS METRICS

Business KPIs

Segment Distinctiveness

Clear differences across segments in:

- Average sales / spend
- Purchase frequency
- Discount dependence
- Customer lifetime indicators

Segment Actionability

- Each segment maps to a clear business strategy
- (e.g., high-value loyalists, price-sensitive customers, low-engagement segments)

Expected Business Impact

- Improved marketing efficiency
- Higher promotional ROI
- Better customer targeting and personalization

Data Science Metrics

Cluster Quality & Separation

- Silhouette Score (cluster separability)
- Within-Cluster Sum of Squares (compactness, for K-Means)

Model Robustness

- Cluster stability across data samples
- Coverage (avoiding overly small or noisy clusters)

Model Selection Support

- Elbow method and comparative evaluation across algorithms

DATASET & ANALYTICAL SCOPE

Dataset Source

- Superstore Sales Dataset (public retail transactions dataset from Kaggle)
- Transaction-level data with customer, product, pricing, discount, and profit attributes
- Widely used for customer analytics and segmentation benchmarking

Unit of Analysis

- Raw data: Transaction-level
- Modeling data: Customer-level aggregation (each row represents a unique customer)

Why This Dataset Works

- Captures repeat purchasing behavior
- Contains value, frequency, discount, and product mix signals
- Suitable for identifying behavioral and value-based segments

FEATURE ENGINEERING & DATA PREPARATION

Customer-Level Feature Design

Transactional data was transformed into behavioral customer features, including:

- Value: total & average spend, profit contribution
- Frequency: number of orders
- Diversity: product category breadth
- Promotion Sensitivity: sales-weighted discount usage

Stable attributes (e.g., region, segment) were aggregated using modal values for interpretability.

Data Quality & Treatment

- Minimal missing data; robust imputation applied
- High-value customers retained, not removed
- Outliers capped using IQR-based winsorization
- Numerical features standardized to ensure fair distance computation

This balances robustness with business realism.

EDA INSIGHTS & CLUSTER READINESS

Key Exploratory Findings

- Strong right-skew in customer value → few customers drive most revenue
- High heterogeneity in purchase frequency and basket diversity
- Strong correlations among sales metrics informed feature reduction

Clustering Suitability

Hopkins Statistic = 0.946

- Indicates strong non-random structure
- Confirms high suitability for clustering

Implication

- Data exhibits clear segmentation potential
- Justifies the use of unsupervised learning methods

The dataset is empirically suitable for meaningful customer segmentation.

DIMENSIONALITY REDUCTION & CLUSTER READINESS

Customer-level features exhibited high correlation and noise, which can distort distance-based clustering. Dimensionality reduction was applied to improve signal quality and computational efficiency.

Principal Component Analysis (PCA)

- PCA applied to standardized customer features
- 7 components retained, explaining ~93% of total variance
- Retained components capture the majority of customer behavioral information while reducing noise

This ensures clustering is driven by dominant behavioral patterns rather than redundant features.

Visualization (Qualitative Validation)

- t-SNE applied to PCA features for 2D visualization only
- Visible groupings observed, providing qualitative confirmation of latent structure

(t-SNE used for interpretation, not for model training)

INTERPRETABILITY, KEY DRIVERS & REPRODUCIBILITY

Notably: demographic and geographic variables showed minimal influence, indicating that behavior and value —not demographics—drive segmentation.

Interpreting Clusters (Unsupervised Context)

Because clustering models do not natively provide feature importance, a surrogate modeling approach was used:

- Provisional clusters generated using K-Means
- A Random Forest model trained to predict cluster membership
- Permutation importance used to identify key drivers of separation

Key Drivers of Segmentation

The most influential features were:

- Customer spend intensity (total, mean, median sales)
- Purchase frequency (number of orders)
- Product diversity (sub-category breadth)
- Sales variability

Reproducibility & Transparency

- Entire pipeline implemented in reproducible Python code
- Outputs saved as reusable analytical artifacts (features, PCA data, importance outputs)

This ensures results are transparent, repeatable, and extensible.

CLUSTERING APPROACH & EVALUATION FRAMEWORK

Model selection prioritized robustness and actionability—not just metric maximization.

Objective

Identify robust, interpretable, and actionable customer segments by implementing and comparing multiple unsupervised clustering algorithms on the PCA-reduced customer feature space.

- Input space: 7 PCA components
- Variance retained: ~93% of customer behavior variance
- Goal: balance cluster quality, stability, and business usability

Models Evaluated

- K-Means (centroid-based, scalable, interpretable)
- Hierarchical Clustering (Ward) (structure-aware, less scalable)
- DBSCAN (density-based, detects noise/outliers)

Evaluation Criteria (Unsupervised Context)

Because no ground truth exists, models were evaluated using:

Cluster Quality

- Silhouette Score (↑ better)
- Davies–Bouldin Index (↓ better)
- Calinski–Harabasz Index (↑ better)

Stability

- Bootstrap resampling
- Mean and variability of silhouette scores

Practical Usability

- Customer coverage
- Interpretability
- Suitability for downstream business use

MODEL COMPARISON & FINAL SELECTION

Summary of Model Performance (Best Configuration)

Model	# Clusters	Noise %	Silhouette	Davies–Bouldin	Calinski–Harabasz	Stability
K-Means	2	0%	0.298	1.375	**363.99**	**High**
Hierarchical	2	0%	0.199	1.801	223.76	Low
DBSCAN	2	~10%	**0.354**	**0.799**	90.64	Not stable

Interpretation

- DBSCAN achieved strong local separation but labeled ~10% of customers as noise, limiting coverage and business usability.
- Hierarchical clustering underperformed across separation and stability metrics.
- K-Means delivered consistently strong separation, high stability, and full customer coverage.

Final Model Selection

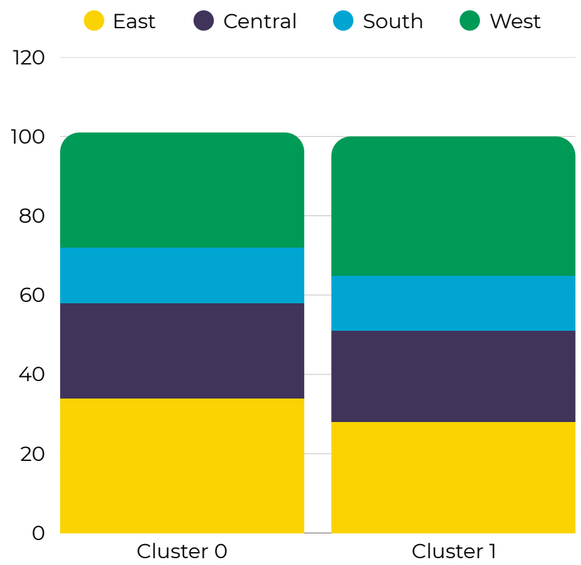
K-Means (k = 2) was selected as the final model because it provides the best balance of:

- Robust internal validation metrics
- High stability under resampling
- Full customer coverage (no noise points)
- Clear interpretability for business use

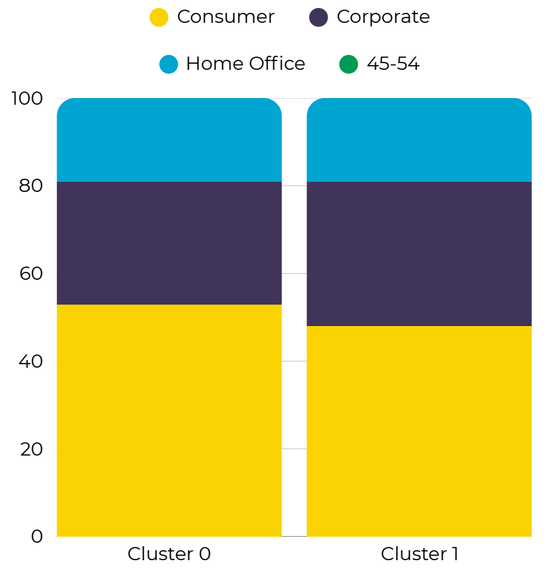
KEY RESULTS - GEOGRAPHIC & CUSTOMER SEGMENT FEATURES

Organizational type and geography are not primary drivers of cluster segmentation.

Region



Customer Segment



KEY RESULTS - BEHAVIORAL FEATURES

Clusters are primarily differentiated by purchasing intensity, spending level, and breadth of engagement.

Cluster	Total Sales	Average Order Value	Typical Purchase Size	Sales Variability	Number of Orders	Product Breadth (Subcategory Diversity)
0	1537.07	140.93	54.76	204.64	5.72	6.91
1	4916.41	353.39	96.54	593.29	7.15	8.63

SUMMARY

Customer segmentation reveals two distinct value-based clusters, differentiated primarily by **spending intensity, purchase frequency, and product breadth**, while remaining largely similar across segment type and region.

Cluster 1: High-Value, High-Engagement Customers

Cluster 1 customers spend more per transaction, purchase more frequently, and engage across a wider range of product categories—indicating deep engagement and strong revenue contribution.

Cluster 0: Lower-Value, Narrower Engagement Customers

Cluster 0 represents lower-intensity buyers, likely driven by occasional or need-based purchasing rather than sustained engagement.

Business Implication

The resulting clusters enable value-based targeting and personalization:

- Cluster 1: retention, loyalty programs, premium offers, cross-sell
- Cluster 0: activation, basket expansion, frequency-building campaigns

This segmentation supports actionable decision-making grounded in customer behavior rather than static attributes.