Cluster Analysis

Theory, Methods, and Practical Application

Juliho Castillo Colmenares

Tec de Monterrey

Today's Agenda

Today's Agenda

1. Customer Segmentation Case Study

What is Cluster Analysis?

Definition: An exploratory technique to discover natural groupings in data **without predefined categories**

What is Cluster Analysis?

Key Characteristics:

- Unsupervised learning method
- No training labels required
- Discovers hidden structure in data
- Groups similar observations together

Goal: Maximize within-cluster similarity and between-cluster dissimilarity

Cluster Analysis vs. Discriminant Analysis

Cluster Analysis	**Discriminant Analysis**
Unsupervised learning	Supervised learning
Discovers unknown groups	Classifies into known groups
No training labels	Requires training labels
Exploratory	Predictive
Groups observations	Creates decision boundaries

Applications: Marketing & Business

Marketing

- Customer segmentation for targeted campaigns
- Market basket analysis

Business

- Fraud detection
- Anomaly identification

Applications: Science & Healthcare

Biology & Medicine

- Disease subtype identification
- Gene expression analysis

Social Sciences

- Community detection in networks
- Document clustering

Distance and Similarity Measures

Why Distance Matters

Clustering depends on measuring how “close” observations are to each other

Common Distance Metrics

1. **Euclidean Distance** (L2 norm) - Most common
2. **Manhattan Distance** (L1 norm) - Robust to outliers
3. **Cosine Similarity** - For high-dimensional data
4. **Correlation Distance** - Pattern similarity

Euclidean Distance

Formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Euclidean Distance

Properties:

- Straight-line distance in n-dimensional space
- Sensitive to scale differences
- Assumes equal importance of all dimensions

Warning: Always standardize variables with different scales!

Manhattan Distance

Formula:

$$d(x, y) = \sum_{i=1}^{p} |x_i - y_i|$$

Manhattan Distance

When to Use:

- Data contains outliers or extreme values
- Variables represent counts
- High-dimensional spaces

Advantage: More robust than Euclidean distance

Why Standardization is Critical

Problem: Variables on different scales dominate distance calculations

Example:

- Age: 20-80 years
- Income: 20,000-200,000 dollars

Without standardization, income dominates!

Z-score Standardization

Solution: Z-score Standardization

$$z_i = \frac{x_i - \mu}{\sigma}$$

Transform to mean = 0, standard deviation = 1

Hierarchical Clustering

Builds a tree-like structure (dendrogram) showing nested clusters

Two Approaches

Agglomerative (Bottom-Up): Most common

- Start: Each observation is its own cluster
- Process: Merge closest clusters iteratively
- End: All observations in one cluster

Two Approaches

Divisive (Top-Down): Less common

- Start: All observations in one cluster
- Process: Split most heterogeneous cluster
- End: Each observation is its own cluster

Linkage Methods

How to Measure Distance Between Clusters?

Single Linkage (Nearest Neighbor)

$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y)$$

Distance between closest points in the two clusters

Complete Linkage (Farthest Neighbor)

$$d(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y)$$

Distance between farthest points in the two clusters

Average Linkage

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y)$$

Average distance between all pairs of points

Ward’s Method

Ward’s Method

Minimizes within-cluster sum of squares

Tends to produce compact, equal-sized clusters

Linkage Methods Comparison

Method	**Outlier Sensitivity**	**Cluster Shape**
Single Linkage	High	Elongated (chaining)
Complete Linkage	Low	Compact, spherical
Average Linkage	Medium	Balanced
Ward's Method	Medium	Compact, equal-sized

Linkage Methods Comparison

Recommendation: Ward’s method often works best in practice

Dendrograms

Visualizing Hierarchical Structure

Reading a Dendrogram

- Horizontal axis: Observations or clusters
- Vertical axis: Distance at which clusters merge
- Height of branches: Dissimilarity between merged clusters

Determining Number of Clusters

- Look for large vertical gaps (jumps in fusion distance)
- Cut dendrogram where there's substantial increase
- Draw horizontal line: number of vertical lines crossed = k clusters

The Chaining Effect

Problem with Single Linkage:

Clusters form long, elongated chains rather than compact groups

The Chaining Effect

Why it Happens:

- Observations connect via intermediate points
- A-B-C-D form chain where each is close to neighbor
- But A and D are far apart

The Chaining Effect

Solution:

- Use complete or average linkage instead
- Or Ward's method for compact clusters

K-Means Clustering

Most popular non-hierarchical method

K-Means Algorithm

1. **Initialize:** Select k random observations as centroids
2. **Assignment:** Assign each point to nearest centroid
3. **Update:** Recalculate centroids as cluster means
4. **Repeat:** Steps 2-3 until convergence

Convergence: When assignments no longer change between iterations

K-Means Objective Function

Goal: Minimize within-cluster sum of squares (WCSS)

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where μ_i is the centroid of cluster C_i

K-Means Properties

- Always converges (finite partitions, monotonically decreasing WCSS)
- Typically converges in 10-30 iterations
- Fast: O(n times k times p times iterations)

K-Means: Advantages

- Fast and scalable to large datasets
- Simple to understand and implement
- Efficient for exploratory analysis

K-Means: Limitations

- Requires specifying k in advance
- Sensitive to initialization (different starts $\longrightarrow$ different results)
- Assumes spherical clusters
- Sensitive to outliers
- Tends to create equal-sized clusters

K-Means++ Initialization

Problem: Random initialization can lead to poor results

K-Means++ Algorithm

1. Choose first centroid randomly
2. For each subsequent centroid:
   - Choose point with probability proportional to squared distance from nearest existing centroid
3. Repeat until k centroids selected

Benefit: Spreads out initial centroids, significantly improves results

K-Medoids (PAM)

Key Difference from K-Means:

- K-means: Centers are computed means (may not be actual points)
- K-medoids: Centers are actual data points (medoids)

K-Medoids (PAM)

Advantages:

- More robust to outliers
- Works with any distance metric
- Interpretable centers (actual observations)

Disadvantage: Slower than k-means (higher computational cost)

How Many Clusters?

The Fundamental Challenge:

No “ground truth” for correct number of clusters

Multiple Approaches

1. **Elbow Method** - Look for bend in WCSS plot
2. **Silhouette Analysis** - Measure cluster quality
3. **Gap Statistic** - Compare to null reference
4. **Davies-Bouldin Index** - Ratio of compactness to separation
5. **Domain Knowledge** - Business requirements

Elbow Method: Procedure

1. Run clustering for k = 1, 2, 3, …, K_max
2. Calculate WCSS for each k
3. Plot WCSS vs. k
4. Look for “elbow” - diminishing returns point

Elbow Method: Interpretation

- WCSS always decreases as k increases
- Elbow indicates where additional clusters don't help much
- Choose k at the elbow point

Limitation: Elbow not always clear - may need other methods

Silhouette Analysis

Measures how well each point fits within its cluster

Silhouette Coefficient

Silhouette Coefficient for observation i:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

- $a(i)$ = avg distance to points in same cluster
- $b(i)$ = avg distance to points in nearest neighboring cluster

Silhouette Interpretation

- $s(i) \approx +1$: Well-matched to cluster
- $s(i) \approx 0$: On border between clusters
- $s(i) \approx -1$: Likely in wrong cluster

Using Silhouette for Optimal k

Average Silhouette Width:

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

Silhouette Procedure

1. Run clustering for different k values
2. Calculate average silhouette width for each k
3. Choose k that maximizes $\bar{s}$

Advantage: Provides both quality measure and optimal k

Cluster Validation

Internal Validation (using data only):

- Within-Cluster Sum of Squares (WCSS) - lower is better
- Silhouette Coefficient - higher is better
- Davies-Bouldin Index - lower is better
- Dunn Index - higher is better

Cluster Validation

External Validation (when true labels available):

- Adjusted Rand Index (ARI)
- Normalized Mutual Information (NMI)

Davies-Bouldin Index

Measures ratio of within-cluster dispersion to between-cluster separation

$$\mathrm{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Davies-Bouldin Index

Interpretation:

- Lower values indicate better clustering
- Compact clusters that are far apart
- Can compare different k values or methods

Curse of Dimensionality

As dimensions (p) increase, problems arise

Curse of Dimensionality: Problems

1. Distance becomes less meaningful (all points appear equidistant)
2. Data becomes sparse (observations spread out)
3. Computational cost increases dramatically

Curse of Dimensionality: Solutions

- Use PCA or feature selection before clustering
- Select only relevant variables
- Use specialized high-dimensional algorithms

Rule: If p is large relative to n, reduce dimensions first

Handling Outliers

Impact by Method:

Method	Sensitivity
K-means	High
Ward’s Method	High
Single Linkage	Medium
K-medoids	Low (Robust)

Handling Outliers: Strategies

- Pre-processing: Detect and remove outliers
- Use robust methods (k-medoids)
- Accept outlier clusters

When to Use Hierarchical Clustering

- Small to medium datasets ($n < 5{,}000$)
- Want to explore different k values
- Need hierarchical structure
- Don't know k in advance

When to Use K-Means

- Large datasets ($n > 5{,}000$)
- Approximately know k
- Need speed and efficiency
- Clusters roughly spherical

Cluster Analysis Workflow

1. **Define objective** - What questions to answer?
2. **Select variables** - Domain knowledge
3. **Preprocess data** - Handle missing values, outliers
4. **Standardize** - If variables on different scales
5. **Choose method** - Based on data characteristics

Cluster Analysis Workflow

6. **Determine k** - Multiple criteria
7. **Run clustering** - Multiple times for k-means
8. **Validate results** - Internal and stability checks
9. **Interpret clusters** - Profile and name clusters
10. **Refine and iterate** - Based on insights

Common Pitfalls to Avoid

1. **Not standardizing** when variables have different scales
2. **Using k-means** with non-spherical clusters
3. **Ignoring outliers** - can severely distort results
4. **Over-interpreting** - clustering always finds structure, even in random data

Common Pitfalls to Avoid

5. **Using too many variables** - curse of dimensionality
6. **Running k-means once** - try multiple initializations
7. **Choosing k without validation** - use multiple methods

Best Practices

1. **Try multiple methods** - Compare hierarchical, k-means, etc.
2. **Validate stability** - Bootstrap samples, different initializations
3. **Visualize extensively** - Scatter plots, dendrograms, parallel coordinates

Best Practices

4. **Use domain knowledge** - Statistical metrics + practical sense
5. **Document decisions** - Why certain methods, parameters chosen
6. **Check interpretability** - Can you explain and use clusters?

Key Takeaways: Fundamental Concepts

- Cluster analysis discovers natural groupings (unsupervised)
- Distance measures are crucial (Euclidean, Manhattan)
- Standardization essential for different scales

Key Takeaways: Methods

- Hierarchical: Creates tree structure, multiple k values
- K-means: Fast, scalable, requires specifying k
- K-medoids: Robust alternative to k-means

Key Takeaways: Validation

- Elbow method and silhouette analysis for optimal k
- Multiple validation measures for quality assessment

Summary: Method Selection Guide

Situation	Recommended Method
Small dataset (n < 1,000)	Hierarchical (Ward's or Average)
Large dataset (n > 10,000)	K-means with k-means++
Outliers present	K-medoids or preprocessing
Non-spherical clusters	DBSCAN or hierarchical

Summary: Method Selection Guide

Situation	Recommended Method
Don't know k	Hierarchical, then elbow/silhouette
High dimensions	PCA first, then k-means
Mixed data types	Gower distance with hierarchical

Advanced Topics (Beyond This Course)

Density-Based Methods:

- DBSCAN - finds arbitrary shapes, identifies outliers

Model-Based:

- Gaussian Mixture Models (GMM) - probabilistic approach

Advanced Topics (Beyond This Course)

Fuzzy Clustering:

- Soft assignment (membership degrees)

Subspace Clustering:

- For high-dimensional data, different subspaces

Real-World Applications: Business

Marketing & Business:

- Customer segmentation for targeted marketing
- Product recommendation systems
- Market basket analysis

Real-World Applications: Healthcare

Healthcare:

- Patient stratification for personalized medicine
- Disease subtype identification
- Medical image segmentation

Real-World Applications: Finance

Finance:

- Fraud detection and anomaly identification
- Credit risk assessment
- Portfolio diversification

Example: Customer Segmentation

Scenario: E-commerce company with 100,000 customers

Variables:

- Purchase frequency
- Average order value
- Product category preferences
- Time since last purchase
- Customer lifetime value

Example: Customer Segmentation Process

1. Standardize variables (different scales)
2. Try k-means for $k = 2$ to 10
3. Use elbow method and silhouette analysis
4. Identify $k = 5$ optimal clusters
5. Profile each segment
6. Develop targeted marketing strategies

Recommended Resources: Books

Textbooks:

- Everitt et al. (2011) - Cluster Analysis (5th ed.)
- James et al. (2021) - Introduction to Statistical Learning

Recommended Resources: Software

Software:

- Python: scikit-learn (KMeans, AgglomerativeClustering)
- R: stats package (kmeans, hclust)

Recommended Resources: Online Videos

StatQuest YouTube Channel:

1. **K-Means Clustering:**
   https://www.youtube.com/watch?v=4b5d3muPQmA
2. **Hierarchical Clustering:**
   https://www.youtube.com/watch?v=7xHsRkOdVwo
3. **Validation Methods (Elbow & Silhouette):**
   https://www.youtube.com/watch?v=DzrvLpxTxJw

Recommended Resources: Additional Online

Additional Online Resources:

- Scikit-learn documentation
- Coursera/edX courses on unsupervised learning
- Interactive clustering visualizations

Part 2: Practical Application

Customer Segmentation Case Study

E-Commerce Cluster Analysis Using Hierarchical and K-Means Methods

Business Context

Business Problem: An e-commerce company seeks to understand their customer base by discovering natural segments based on purchasing behavior, engagement patterns, and browsing habits.

Key Questions:

- How many distinct customer segments exist in our data?
- What behavioral patterns characterize each segment?
- How can we tailor marketing strategies to each discovered segment?

Dataset Overview

Dataset: 2,000 customers with 7 behavioral features

Variables:

- monthly_purchases
- avg_basket_size
- total_spend
- session_duration
- email_clicks
- product_views
- return_rate

Important: No predefined labels (unsupervised learning)

Exploratory Data Analysis

Why EDA Matters

Before attempting to discover customer segments, we must first understand the characteristics and relationships within our data. Unlike supervised learning where we have predefined labels, cluster analysis is exploratory in nature, making this preliminary investigation even more critical.

EDA: Approach

What We Do:

- Visualize distributions using histograms
- Identify skewness, outliers, and typical value ranges
- Compute correlation matrix to understand relationships
- Assess variable redundancy

Why: Understanding data structure before algorithmic analysis helps anticipate which features might drive segmentation.

EDA: Distribution Analysis

```
# Load customer data
df = pd.read_csv("customer_data.csv")

# Summary statistics
print(df.describe())

# Distribution plots for each variable
fig, axes = plt.subplots(3, 3, figsize=(15, 12))
for idx, col in enumerate(df.columns):
    axes[row, col_idx].hist(df[col], bins=30)
```

EDA: Correlation Analysis

```
# Correlation matrix
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True,
            fmt='.2f', cmap='coolwarm')

# Identify strongest correlations
# Example: monthly_purchases <-> total_spend: 0.94
#          avg_basket_size <-> total_spend: 0.89
```

EDA: Key Findings

Outcome:

- Balanced dataset with diverse customer behaviors
- Right-skewed distributions (some extreme high spenders)
- Strong positive correlations between purchase-related variables
- Email clicks correlate moderately with purchase frequency

Implication: Multiple distinct behavioral patterns suggest natural customer segments exist.

Why Standardization is Critical

Data Standardization

Why Standardization is Critical

Cluster analysis algorithms rely on distance metrics to measure similarity. However, our behavioral variables are measured in different units and scales:

- Purchases (counts)
- Spending (dollars: 58 to 7,891)
- Return rate (proportions: 0.0 to 0.5)

The Problem Without Standardization

Without standardization: Variables with larger numeric ranges dominate distance calculations.

Example:

- Customer A: 1 purchase, 100 dollars
- Customer B: 2 purchases, 200 dollars

Distance dominated by dollar difference (100) rather than purchase difference (1).

This leads to biased clustering that reflects scale differences, not true behavioral patterns.

Standardization: Approach

Z-score Standardization:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Properties:

- Transforms to mean = 0, standard deviation = 1
- Preserves distribution shape
- Ensures equal contribution to distance calculations
- Value of 2.0 means “2 std deviations above mean”

Standardization: Implementation

```
from sklearn.preprocessing import StandardScaler

# Standardize features
scaler = StandardScaler()
X_standardized = scaler.fit_transform(df)

# Convert back to DataFrame
df_standardized = pd.DataFrame(
    X_standardized,
    columns=df.columns
)
```

Standardization: Implementation

```
print(df_standardized.describe())
# All means approx 0, all std approx 1
```

Standardization: Outcome

Result: All variables successfully transformed to mean approximately 0 and standard deviation approximately 1.

Impact: Clustering algorithms now treat all behavioral dimensions equally when computing distances between customers.

Example: Total spend (originally 58-7,891 dollars) and return rate (originally 0.0-0.5) now contribute equally to customer similarity.

Hierarchical Clustering

The Fundamental Question

How many customer segments should we look for?

Unlike supervised classification where the number of classes is predetermined, clustering requires us to discover the appropriate number of groups.

Hierarchical clustering provides an elegant solution: Build a complete hierarchy of nested clusters without specifying k in advance.

Hierarchical Clustering: Advantage

Key Benefit: Creates a tree-like structure (dendrogram) showing how customers progressively merge into larger groups.

Result: We can identify the most natural number of segments based on where large jumps in distance occur.

No need to predefine the number of clusters

Hierarchical Clustering: Linkage Methods

How do we measure distance between clusters?

- **Single Linkage:** Minimum distance between any two points
- **Complete Linkage:** Maximum distance between any two points
- **Average Linkage:** Mean distance between all pairs
- **Ward's Method:** Minimizes within-cluster variance

Recommendation: Ward's method typically produces the most balanced clusters for customer segmentation.

Hierarchical Clustering: Implementation

```
from scipy.cluster.hierarchy import linkage, dendrogram

# Compute linkage matrices for different methods
linkage_methods = ['single', 'complete',
                   'average', 'ward']
linkage_matrices = {}

for method in linkage_methods:
    linkage_matrices[method] = linkage(
        X_standardized,
        method=method
    )
```

Dendrogram Interpretation

```
# Create dendrogram for Ward's method
plt.figure(figsize=(14, 7))
dendrogram(linkage_matrices['ward'],
           no_labels=True,
           color_threshold=50)
plt.axhline(y=50, color='r', linestyle='--',
            label='Potential cut (4 clusters)')
```

How to Read:

- Horizontal axis: Customers
- Vertical axis: Distance at which clusters merge
- Large vertical gaps suggest natural cluster boundaries

Hierarchical Clustering: Outcome

Finding: Four dendrograms reveal distinct clustering structures depending on linkage method.

Observations:

- Single linkage: Chaining effects with unbalanced clusters
- Complete linkage: Too many small clusters
- Average linkage: Compromise between extremes
- Ward's method: Clear hierarchical structure, balanced sizes

Decision: Ward's dendrogram suggests 4 clusters as a natural choice.

Extracting Clusters from Dendrogram

```
from scipy.cluster.hierarchy import fcluster

# Extract 4 clusters using Ward's method
n_clusters_hier = 4
hierarchical_labels = fcluster(
    linkage_matrices['ward'],
    n_clusters_hier,
    criterion='maxclust'
)

# Calculate silhouette score
silhouette_hier = silhouette_score(
    X_standardized,
    hierarchical_labels
)
# Result: 0.458 (moderate cluster quality)
```

Hierarchical Clustering: Results

Cluster Sizes:

- Cluster 0: 440 customers (22.0%)
- Cluster 1: 300 customers (15.0%)
- Cluster 2: 560 customers (28.0%)
- Cluster 3: 700 customers (35.0%)

Silhouette Score: 0.458

Interpretation: Customers reasonably well-separated into clusters with balanced sizes. Score above 0.4 is acceptable for customer segmentation where behavioral boundaries are often fuzzy.

K-Means Clustering

Why K-Means?

While hierarchical clustering provided valuable insights through dendrogram visualization, it has computational limitations:

- Complexity: O(n squared) or worse
- Impractical for very large datasets

K-means offers a scalable alternative:

- Works well with larger customer bases
- Computational complexity: O(n times k times p times iterations)

K-Means: The Challenge

Requirement: Must specify the number of clusters (k) in advance.

Solution: Elbow method provides a data-driven approach.

Goal: Identify the point where adding more clusters provides diminishing returns in terms of improved fit.

Elbow Method: Approach

K-means minimizes within-cluster sum of squared distances (inertia):

$$\mathrm{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Procedure:

- Run clustering for k = 2 to 10
- Calculate inertia for each k
- Plot inertia vs. k
- Look for the “elbow” point

Elbow Method: Implementation

```python
from sklearn.cluster import KMeans

# Test different k values
inertias = []
silhouette_scores = []
K_range = range(2, 11)

for k in K_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_standardized)
    inertias.append(kmeans.inertia_)
    silhouette_scores.append(
        silhouette_score(X_standardized, kmeans.labels_)
    )
```

Elbow Method: Results

Plot Analysis:

- Elbow curve shows diminishing returns after k=4
- Silhouette scores peak at k=4
- Both metrics converge on k=4

Decision: Both hierarchical and elbow analyses suggest k=4

K-Means: Final Clustering

```
# Apply k-means with optimal k
optimal_k = 4
kmeans_final = KMeans(
    n_clusters=optimal_k,
    random_state=42,
    n_init=10
)
kmeans_labels = kmeans_final.fit_predict(
    X_standardized
)

# Silhouette score: 0.458 (same as hierarchical)
```

K-Means: Convergence Validation

Key Finding: K-means achieves identical silhouette score to hierarchical clustering (0.458).

Significance: Convergence between two fundamentally different algorithms provides strong evidence that four customer segments represent genuine structure in the data.

Confidence: Clusters reflect real behavioral patterns rather than algorithmic artifacts.

Recommendation: Use k-means for production deployment due to computational efficiency.

Cluster Interpretation and Profiling

From Statistics to Business Value

Successfully identifying clusters is only the first step. The real business value comes from understanding what distinguishes each segment and translating these differences into actionable marketing strategies.

Goal: Transform abstract cluster labels into concrete customer personas.

Interpretation: Approach

```python
# Add cluster labels to original data
df_with_clusters = df.copy()
df_with_clusters['Cluster_KMeans'] = kmeans_labels

# Calculate cluster means (original units)
cluster_profiles = df_with_clusters.groupby(
    'Cluster_KMeans'
)[df.columns].mean()

# Create heatmap
sns.heatmap(cluster_profiles.T, annot=True)
```

Interpretation: Approach

Why original units? Makes profiles interpretable to business stakeholders.

Cluster Characterization: Method

```python
# Compare each cluster to overall means
overall_means = df.mean()

for cluster_id in range(optimal_k):
    cluster_mean = cluster_profiles.loc[cluster_id]

    # Calculate percentage differences
    differences = (
        (cluster_mean - overall_means) /
        overall_means * 100
    )
```

Cluster Characterization: Method

```python
# Identify distinctive features (>10% difference)
high_features = differences.nlargest(3)
low_features = differences.nsmallest(3)
```

Cluster 0: Engaged but Selective Shoppers

Size: 307 customers (15.3%)

Distinctive High Features:

- avg_basket_size: +72.0% vs average
- return_rate: +56.2% vs average
- email_clicks: +51.9% vs average

Distinctive Low Features:

- session_duration: −69.6% vs average
- product_views: −58.0% vs average
- monthly_purchases: −42.1% vs average

Cluster 1: Low-Value Browsers

Size: 707 customers (35.4%)

Distinctive High Features:

- session_duration: +19.0% vs average
- return_rate: +15.2% vs average

Distinctive Low Features:

- total_spend: −83.8% vs average
- email_clicks: −78.3% vs average
- monthly_purchases: −76.2% vs average

Characterization: Spend time browsing but make few purchases.

Cluster 2: Premium High-Value Customers

Size: 432 customers (21.6%)

Distinctive High Features:

- total_spend: +168.3% vs average
- avg_basket_size: +127.5% vs average
- monthly_purchases: +124.1% vs average

Distinctive Low Features:

- return_rate: −24.1% vs average

Characterization: The premium segment with highest spending and purchase frequency.

Cluster 3: Frequent Small-Basket Shoppers

Size: 554 customers (27.7%)

Distinctive High Features:

- monthly_purchases: +23.8% vs average
- product_views: +12.4% vs average

Distinctive Low Features:

- avg_basket_size: −48.8% vs average
- total_spend: −45.1% vs average
- return_rate: −31.8% vs average

Characterization: Regular shoppers with lower average order values.

Cluster Profiling: Outcome

Four Distinct Segments Identified:

Each segment has distinctive characteristics differing by more than 10% from overall average on multiple dimensions.

Confirmation: These represent meaningfully different customer types requiring differentiated marketing strategies.

Next Step: Develop targeted campaigns for each segment.

Cluster Validation with Silhouette Analysis

Beyond Average Scores

While we have identified interpretable customer segments, we should validate the quality of our clustering solution.

Questions:

- Are customers well-matched to their assigned clusters?
- Are some clusters poorly defined?
- Are there misclassifications?

Silhouette Analysis: Approach

Silhouette Coefficient for each customer:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

- $a(i)$ = average distance to points in same cluster
- $b(i)$ = average distance to points in nearest neighboring cluster

Range: −1 to +1

- Near +1: Well-matched to cluster
- Near 0: On border between clusters

Silhouette Analysis: Approach

- Negative: Possible misclassification

Silhouette Plot: Implementation

```
from sklearn.metrics import silhouette_samples

# Calculate silhouette values for each customer
silhouette_vals = silhouette_samples(
    X_standardized,
    kmeans_labels
)

# Create visualization
for i in range(optimal_k):
    cluster_vals = silhouette_vals[kmeans_labels == i]
    cluster_vals.sort()
```

Silhouette Plot: Implementation

```
plt.fill_betweenx(y_range, 0, cluster_vals)
```

Silhouette Analysis: Interpretation

Key Observations:

- Width of each section represents cluster size
- All clusters extend beyond average score (0.458)
- Most customers have positive coefficients
- Few customers near zero (boundary cases)
- Very few negative coefficients (rare misclassifications)

Conclusion: Generally good cluster quality across all four segments. Segmentation is sound.

Visualization in 2D Space

The Challenge

We have worked with seven-dimensional customer data, which is impossible to visualize directly.

Problem: Humans can only perceive 2-3 spatial dimensions effectively.

Solution: Principal Component Analysis (PCA) projects the 7D space onto 2D while preserving as much information as possible.

PCA: Approach

What PCA Does:

- Transforms 7 original variables into uncorrelated components
- Orders components by variance explained
- PC1 captures maximum variance
- PC2 captures maximum remaining variance orthogonal to PC1

Result: 2D visualization retaining most important structure.

PCA Projection: Implementation

```python
from sklearn.decomposition import PCA

# Apply PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_standardized)

# Variance explained
print(f"PC1: {pca.explained_variance_ratio_[0]:.3f}")
print(f"PC2: {pca.explained_variance_ratio_[1]:.3f}")

# Create scatter plot
```

PCA Projection: Implementation

```
plt.scatter(X_pca[:, 0], X_pca[:, 1],
            c=kmeans_labels, cmap='tab10')
```

PCA Visualization: Outcome

Results:

- 2D projection captures moderate proportion of variance (40-60%)
- Reasonably good cluster separation visible
- Some overlap consistent with silhouette analysis
- K-means centroids clearly separated
- Both methods show similar spatial patterns

Interpretation: While 2D view sacrifices some information, it confirms clusters are spatially distinct and well-positioned.

Business Recommendations

From Analysis to Action

Having completed technical analysis and validation, we now translate statistical findings into actionable business strategies.

Goal: Create customer personas and design specific marketing tactics for each segment aligned with their behaviors and needs.

Segment 1: Engaged but Selective Shoppers

Profile: High basket sizes, high returns, engaged with emails but low browsing.

Marketing Strategy:

- Launch VIP loyalty program with exclusive perks
- Offer early access to new products
- Provide free expedited shipping
- Personalized product recommendations
- Improve product information to reduce returns

Segment 2: Low-Value Browsers

Profile: High browsing time but low purchases and engagement.

Marketing Strategy:

- Implement abandoned cart recovery campaigns
- Use retargeting ads to re-engage visitors
- Offer limited-time discounts to incentivize first purchase
- Improve product information and customer reviews
- Create urgency with flash sales

Segment 3: Premium High-Value Customers

Profile: Highest spending, purchase frequency, and basket sizes.

Marketing Strategy:

- Exclusive VIP treatment and recognition
- Premium customer service channel
- Early access to new collections
- Referral program incentives
- Maintain satisfaction to prevent churn

Segment 4: Frequent Small-Basket Shoppers

Profile: Regular purchases with lower average order values.

Marketing Strategy:

- Send targeted discount codes and bundle offers
- Promote free shipping thresholds to increase basket size
- Highlight clearance and sale items
- Create value packs and multi-buy promotions
- Build purchase frequency through regular engagement

Key Findings: Summary

Methodology:

- Analyzed 2,000 customers across 7 behavioral variables
- Standardized data to ensure equal feature weighting
- Applied both hierarchical (Ward's) and k-means clustering
- Used elbow method and silhouette analysis

Result: Both methods converged on 4 distinct customer segments with silhouette score of 0.458.

Key Findings: Segments

Four Distinct Customer Segments:

1. **Engaged Selective (15%)**: High basket, high return
2. **Low-Value Browsers (35%)**: Browse but don't buy
3. **Premium High-Value (22%)**: Highest spenders
4. **Frequent Small-Basket (28%)**: Regular small orders

Each requires tailored marketing strategies.

Methodological Learnings

Hierarchical Clustering:

- Provides hierarchy view
- No need to predefine k
- Best for smaller datasets

K-Means:

- More scalable and faster
- Requires specifying k
- Better for production deployment

Both methods converged: Strong evidence of genuine structure.

Business Value

Enables:

- Targeted marketing campaigns by segment
- Optimized resource allocation to high-value customers
- Opportunities for customer retention and growth
- Data-driven customer understanding

ROI: Improved marketing efficiency and customer lifetime value.

Next Steps

Implementation:

1. Validate cluster assignments with domain experts
2. Implement targeted campaigns for each segment
3. Monitor segment-specific KPIs (conversion, AOV, retention)
4. Re-run clustering periodically to detect behavioral shifts
5. Consider additional features (geographic, demographic) for refinement

Validation in Practice

Important Note: In real-world unsupervised learning, true labels do not exist.

Cluster validation relies on:

- Domain expertise
- Business metrics
- Silhouette analysis
- Stability across different methods

This synthetic dataset included true labels only for educational validation purposes.

Python Code Resources

Complete implementation available in:

- `customer_clustering_analysis.ipynb`
- `fetch_customer_data.py` (data generation)
- `CUSTOMER_DATA_DICTIONARY.md` (variable descriptions)

Key Libraries:

- scikit-learn (KMeans, StandardScaler)
- scipy (hierarchical clustering)
- pandas, numpy (data manipulation)
- matplotlib, seaborn (visualization)

Questions?

Thank you for your attention!

Juliho Castillo Colmenares
julihocc@tec

Office: Tec de Monterrey CCM Office 1540
Office Hours: Monday-Friday, 9:00 AM - 5:00 PM

Next Steps: This Week

For This Week:

- Review lecture notes thoroughly
- Practice with provided examples
- Complete practice questions
- Prepare for E06 quiz

Next Steps: Preparation for Evaluation

Preparation for Evaluation:

- Understand distance measures and when to use each
- Know linkage methods and their properties
- Practice interpreting dendrograms
- Understand k-means algorithm and convergence
- Be able to explain validation methods