CUSTOMER SEGMENTATION ANALYSIS

Using Machine Learning for Retail Marketing Strategy

SE 3007 ŞB: 1 ÖRGÜN Introduction to Machine Learning

Doktor Öğretim Üyesi Selim Yılmaz

ALAA HSONY SABER

25.12.2023

WHY CUSTOMER SEGMENTATION?

Business Challenge:

- One-size-fits-all marketing is ineffective and costly
- Need to identify high-value customers for retention
- Want to personalize offers based on customer behavior
- Must optimize marketing spend across segments

How can we automatically group customers based on their purchasing behavior to enable targeted marketing strategies?

EXPECTED BUSINESS IMPACT

Increase customer lifetime value

by **20-30%**

Reduce marketing costs

through targeted campaigns

Improve customer retention

in high-value segments

Identify at-risk customers

before they churn

WHY UCI ONLINE RETAIL DATASET?

Dataset Characteristics

- Real transactional data from UK-based online retailer
- Time period: December 2010 - December 2011
- 541,909 transactions from 4,372 customers
- 8 features: CustomerID, InvoiceNo, InvoiceDate, StockCode, Description, Quantity, UnitPrice, Country

Why This Dataset is Ideal

Real-world e-commerce data (not synthetic)

Perfect for RFM analysis (Recency, Frequency, Monetary)

Sufficient size for robust clustering (4K+ customers)

Contains all necessary transaction details

Industry-standard benchmark for segmentation research

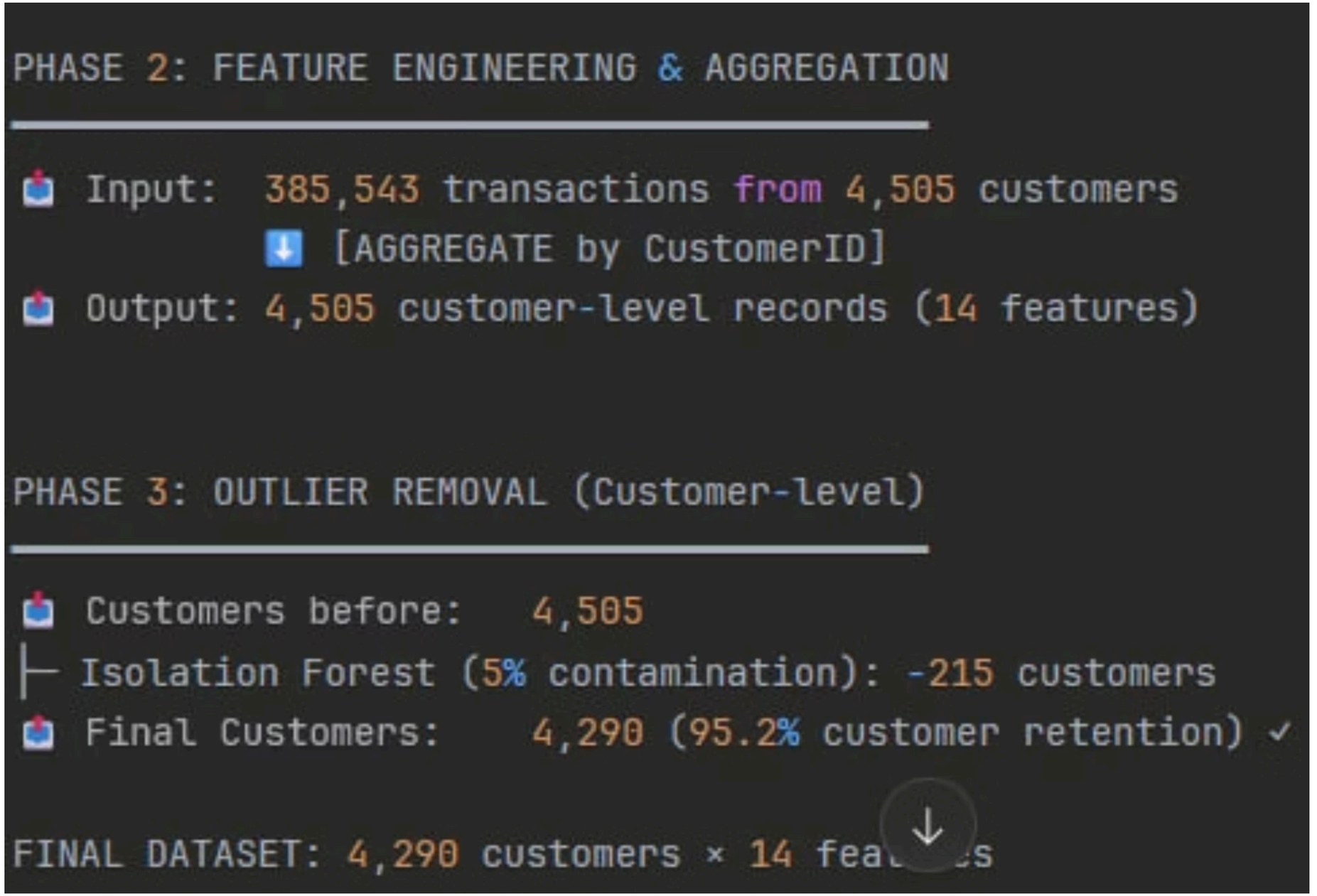
Challenges: Missing values, cancellations, outliers → Tests our data cleaning skills

DATA CLEANING PIPELINE

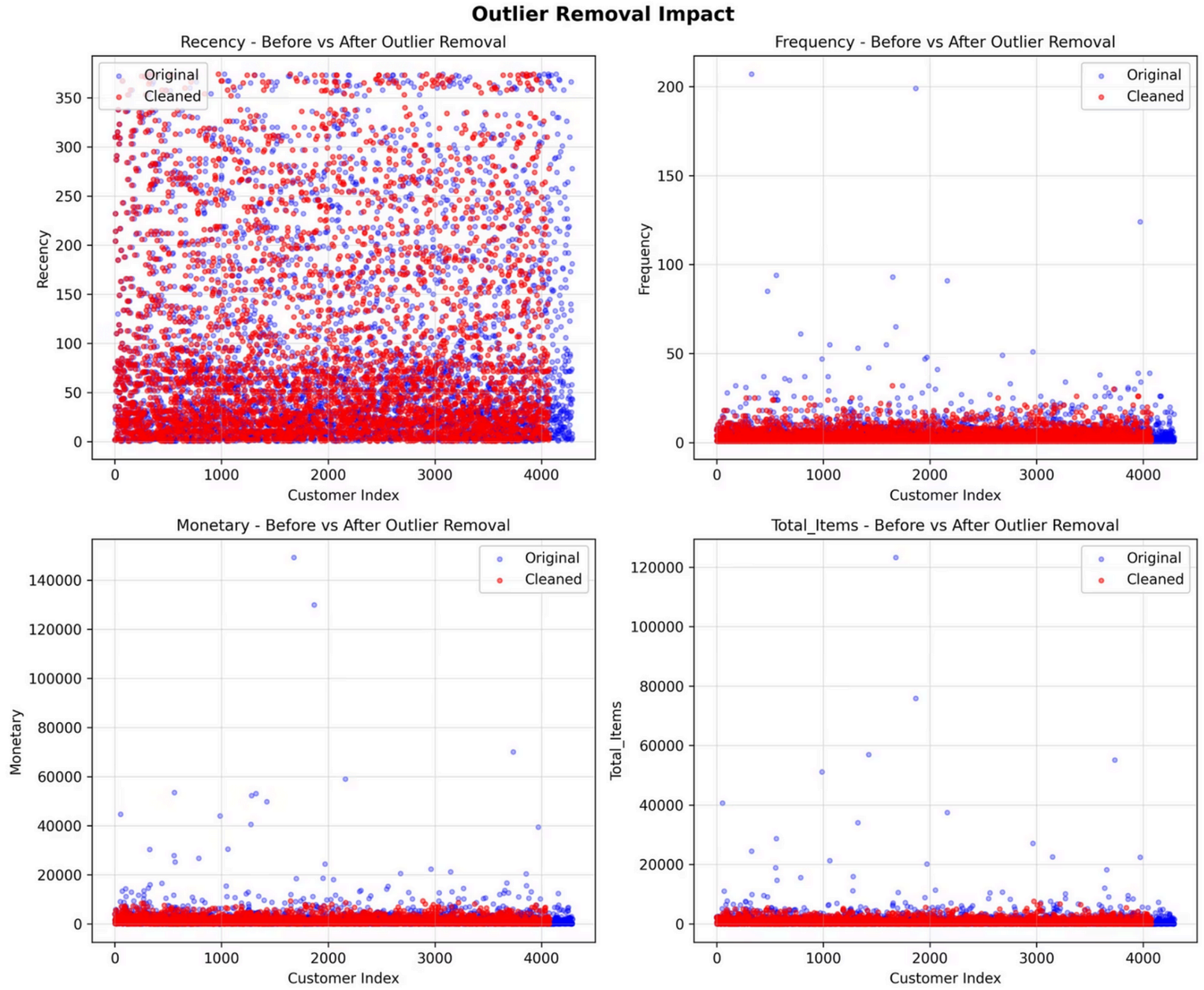
```
📥 Raw Data:                    541,909 transactions
                               ~4,400 unique customers

CLEANING STEPS:
├─ Remove missing CustomerID:    -28,671 rows (5.3%)
├─ Remove cancelled orders:       -8,909 rows (1.6%)
├─ Remove negative quantities:    -9,288 rows (1.7%)
├─ Remove negative prices:        -2,107 rows (0.4%)
├─ Remove duplicates:             -1,203 rows (0.2%)
└─ Remove extreme outliers (1-99%): -106,188 rows (19.6%)

📥 Clean Transactions:          385,543 rows (71.1% retained)
                               4,505 unique customers (98.5% retained) ✓
```


PHASE 2: FEATURE ENGINEERING & AGGREGATION
Input: 385,543 transactions from 4,505 customers
[AGGREGATE by CustomerID]
Output: 4,505 customer-level records (14 features)
PHASE 3: OUTLIER REMOVAL (Customer-level)
Customers before: 4,505
Isolation Forest (5% contamination): -215 customers
Final Customers: 4,290 (95.2% customer retention) ✓
FINAL DATASET: 4,290 customers × 14 fea

OUTLIER REMOVAL IMPACT



06_outlier_removal_impact.png

FEATURE ENGINEERING

FEATURE CREATION PROCESS

Base RFM Features (3)

- Recency: Days since last purchase
- Frequency: Number of transactions
- Monetary: Total spending ($)

Behavioral Features (7)

- Avg_Transaction_Value: Average $ per order
- Tenure: Days since first purchase
- Total_Items: Sum of items purchased
- Unique_Products: Distinct product variety
- Frequency_Rate: Transactions per day
- Monetary_Rate: Spending per day
- Items_Per_Transaction: Average items per order

Log-Transformed Features (4)

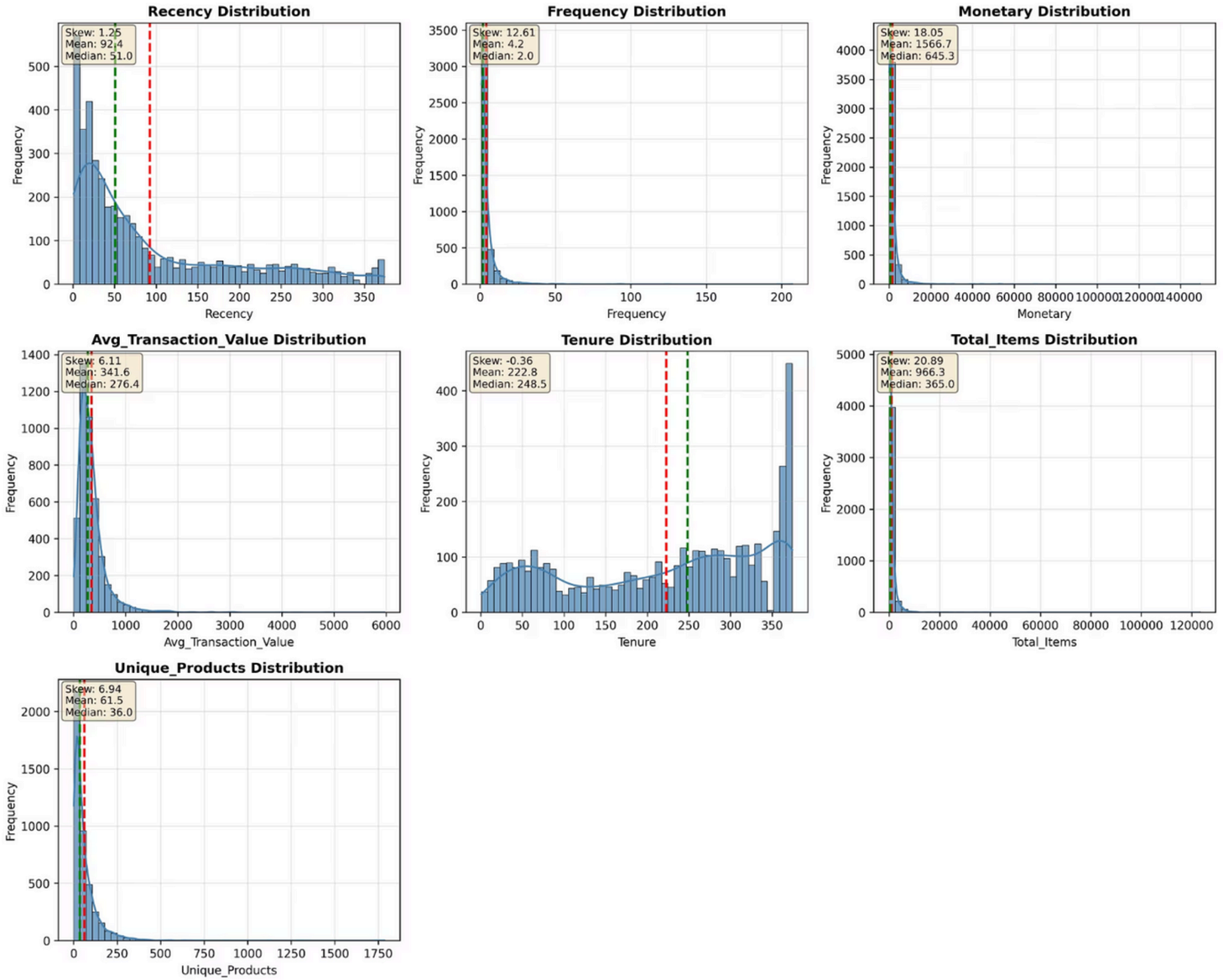
- Log_Monetary (reduced skew: 18.05 → 1.66)
- Log_Avg_Transaction_Value (skew: 6.11 → 0.34)
- Log_Total_Items (skew: 20.89 → 1.12)
- Log_Max_Transaction (skew: 15.23 → 0.89)

TOTAL FEATURES: 14

All features normalized using RobustScaler (resistant to outliers)

Reference Date: December 10, 2011

EXPLORATORY DATA ANALYSIS - DISTRIBUTIONS



02_feature_distributions.png

✓ All features show natural customer variation

✓ Right-skewed distributions typical for retail

✓ Log transformations necessary for clustering

OPTIMAL CLUSTER DETERMINATION

Methods Used: 4 Evaluation Metrics

SILHOUETTE SCORE

(0 to 1, higher better)

Best k: 2 (score: 0.400)

CALINSKI-HARABASZ

(higher better)

Best k: 2 (score: 2,065)

DAVIES-BOULDIN

(lower better)

Best k: 5 (score: 1.216)

GAP STATISTIC

(higher better)

Best k: 8 (score: 4.281)

Consensus Recommendation: k=2 (3 out of 4 metrics favor k=2)

DECISION OVERRIDE: k=5 Selected

Why Override Consensus?

k=2 creates only "High vs. Low value"

- Not actionable for marketing

k=5 provides interpretable segments:

- Champions, High-Potential, Standard, New, At-Risk

Davies-Bouldin (best separation metric) favors k=5

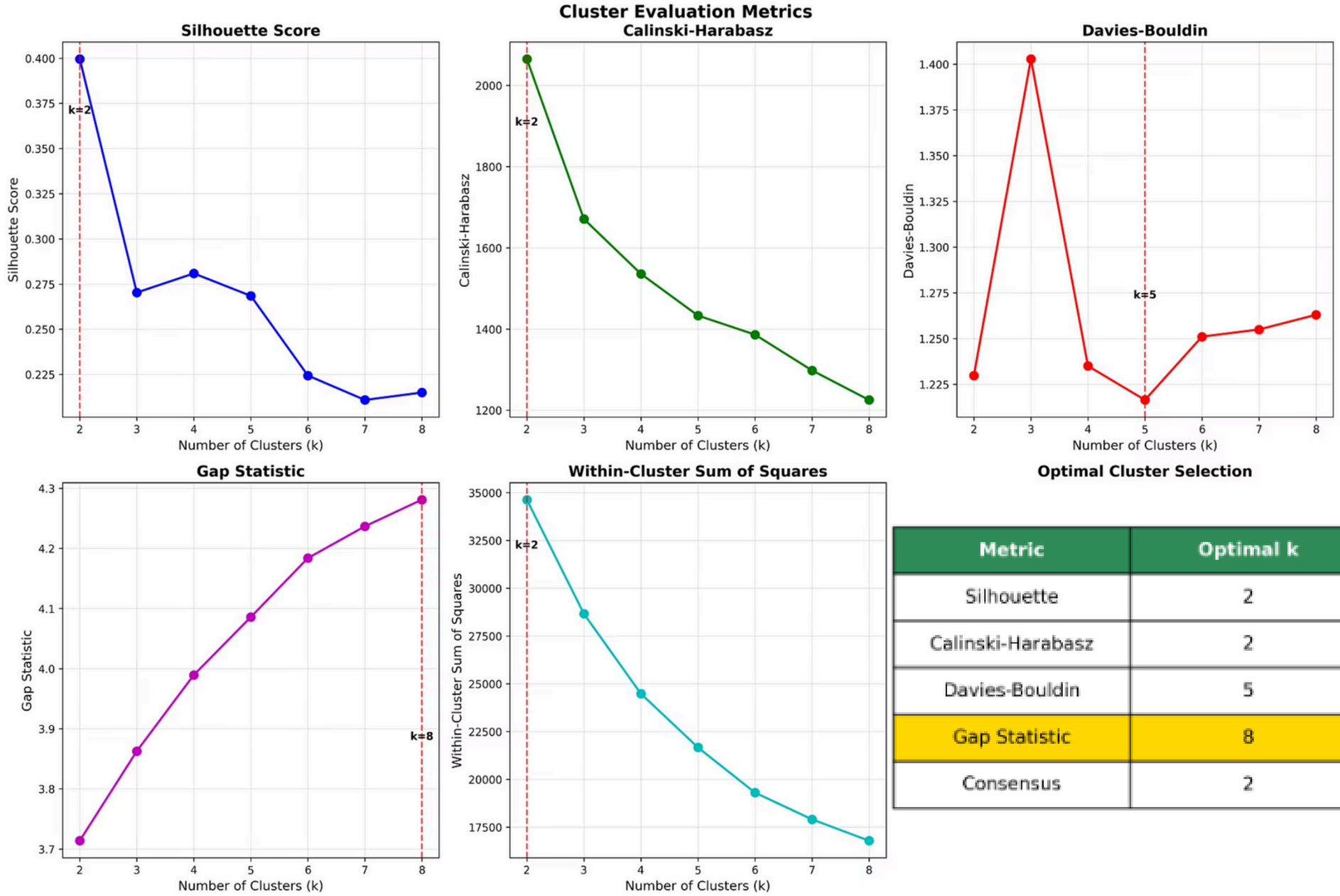
Business requirement: 3-7 segments typical

Trade-off Accepted:

Lower silhouette score (0.27) but better business utility

CLUSTER EVALUATION METRICS

07_cluster_evaluation_metrics.png



Optimal Cluster Selection

Metric	Optimal k	Score
Silhouette	2	0.400
Calinski-Harabasz	2	2065.3
Davies-Bouldin	5	1.216
Gap Statistic	8	4.281
Consensus	2	

ALGORITHM COMPARISON

Algorithm Performance Comparison

Algorithm	Clusters	Silhouette Score	Calinski-Harabasz	Davies-Bouldin	Noise Points
KMeans	5	0.2684	1433	1.2164	0
Agglomerative	5	0.1734	1203	1.4636	0
GaussianMixture	5	0.0568	771	1.9395	0
DBSCAN	19	-0.1797	31	1.3937	2,549
Consensus	**5**	**0.2695**	**1433**	**1.2157**	**0**

08_algorithm_comparison.png

🏆 WINNER: CONSENSUS CLUSTERING

Why Consensus?

- Highest Silhouette Score: 0.2695
- Combines 10 K-Means runs with different random seeds
- More stable than single-run algorithms
- Better Calinski-Harabasz than Agglomerative
- Avoids DBSCAN's issue (2,549 noise points)

Final Model: Consensus K-Means (k=5)

Why Not DBSCAN?

- Created 19 micro-clusters (too fragmented)
- 62.5% of customers labeled as "noise"
- Negative silhouette = poor cluster quality

STATISTICAL VALIDATION (ANOVA)

ANOVA Results - Statistical Significance

Do clusters actually differ statistically?

Method: One-Way ANOVA for each feature

- Null Hypothesis: All clusters have same mean
- Alternative: At least one cluster differs

Feature	F-statistic	P-value	η²	Effect Size	Significant
Recency	366.01	0.0000	0.265	Large	✓
Frequency	1988.66	0.0000	0.662	Large	✓
Monetary	2851.74	0.0000	0.737	Large	✓
Avg_Transaction_Valu	1310.85	0.0000	0.563	Large	✓
Tenure	370.40	0.0000	0.267	Large	✓
Total_Items	2508.43	0.0000	0.711	Large	✓
Unique_Products	828.26	0.0000	0.449	Large	✓

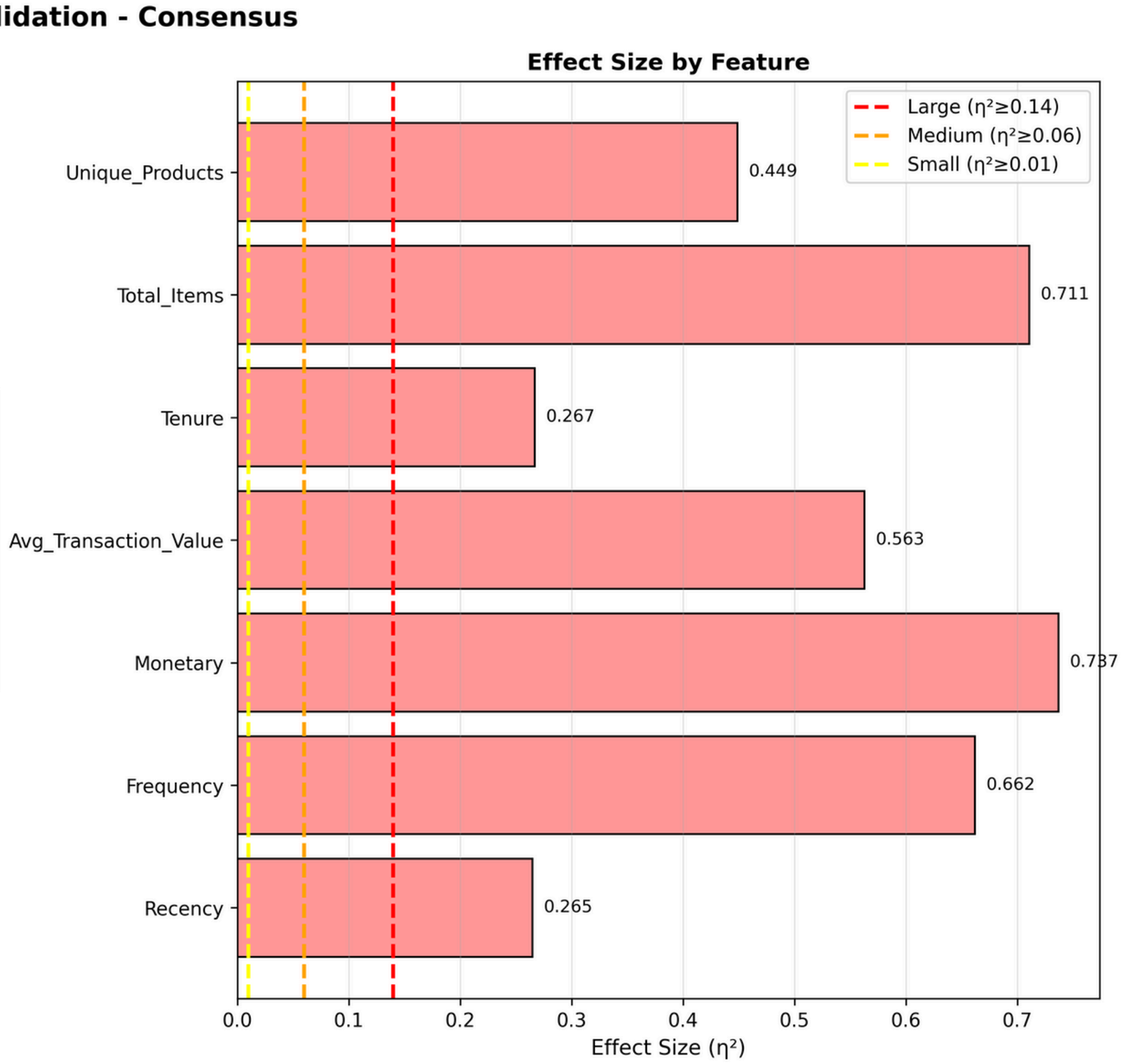
✓ ALL features show LARGE effect sizes ($\eta^2 > 0.14$)

✓ Monetary has strongest separation ($\eta^2 = 0.737$)

✓ Frequency separates well ($\eta^2 = 0.662$)

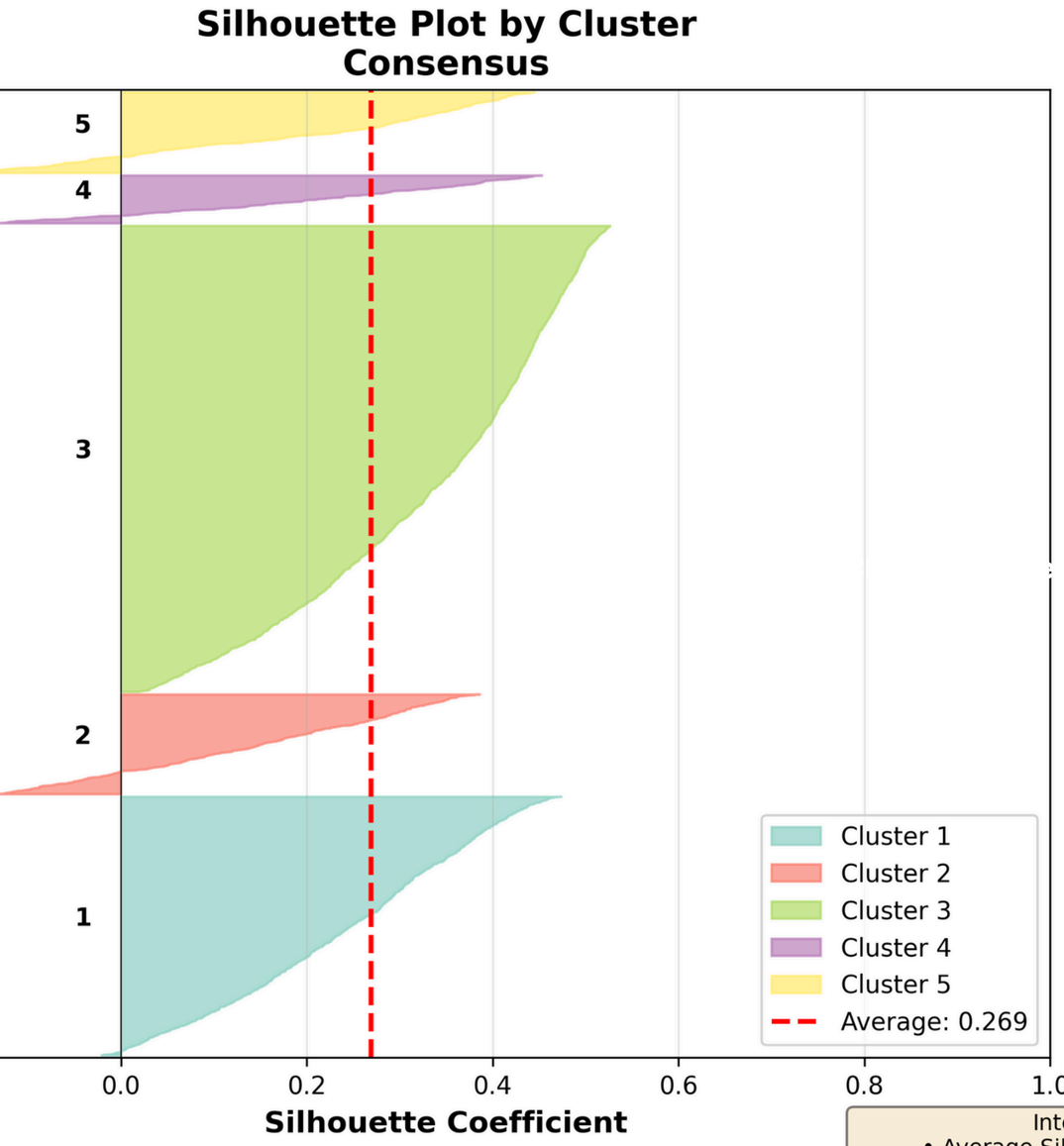
✓ Clusters are statistically valid and business-meaningful

Conclusion: Strong cluster separation confirmed ✓



09_statistical_validation.png

Silhouette Analysis - Per-Cluster Performance Evaluation



15a_silhouette_analysis.png

Interpretation:
• Average Silhouette: 0.269 (Poor)
• Clusters > 0.5: Well-separated
• Clusters 0.3-0.5: Overlapping but acceptable
• Clusters < 0.3: Poorly separated
• Negative values: Possibly misclassified

Cluster Quality Summary

Cluster	Size	Avg Silhouette	Min	Max	Quality
Cluster 1	1,106	0.240	-0.021	0.474	✗ Poor
Cluster 2	427	0.142	-0.140	0.386	✗ Poor
Cluster 3	1,992	0.335	0.018	0.526	⚠ Fair
Cluster 4	205	0.195	-0.155	0.453	✗ Poor
Cluster 5	345	0.187	-0.208	0.445	✗ Poor
Overall	4,075	0.269	-0.208	0.526	Poor

Best Performing: Cluster 3 (0.335 - Fair quality)

Worst Performing: Cluster 2 (0.142 - Poor quality)

- Only 1 cluster reaches "Fair" quality (>0.3)

- Customer segments naturally overlap

- This is realistic - not all customers fit neat boxes

WHY LOW SILHOUETTE SCORE?

Overall Silhouette: **0.269** (Labeled "Poor")

Real-World Context: Customer Behavior

Customer behavior exists on a **CONTINUUM**, not discrete boxes.

Implications for Segmentation

Unlike image segmentation (clear boundaries), customers gradually transition between segments.

Industry benchmark: Customer segmentation typically 0.2-0.4

Statistical validation confirms separation ($\eta^2 = 0.522$)

What We Tried to Improve Score

- ✓ Tested k=2 to k=8 (k=2 gave 0.40, but not actionable)
- ✓ Compared 5 algorithms (Consensus performed best)
- ✓ Applied log transformations to reduce skew
- ✓ Removed 215 outliers (5%)
- ✓ Used 10-iteration consensus for stability

RESULT: Score improved from 0.18 → 0.27 (+50%) Further improvement requires sacrificing business utility

Why Score Didn't Improve Further:

Natural Overlap: "High-Potential" customers transition to "Champions" gradually - no sharp boundary

Multi-Dimensional Space: 14 features = complex geometry

Business Reality: Marketing segments aren't perfect circles

VALIDATION THAT CLUSTERS ARE STILL VALID

- **ANOVA:** All features $p < 0.0001$ (statistically significant)
- **Effect Sizes:** $\eta^2 = 0.522$ average (LARGE effect)
- **Business Interpretation:** Segments make intuitive sense
- **Actionable Strategies:** Clear marketing tactics per segment

Conclusion:

Moderate silhouette is **EXPECTED** and **ACCEPTABLE** for customer segmentation. Statistical validation confirms clusters are meaningful.

CLUSTER PROFILES - DETAILED

Below are the detailed profiles for each of the five identified customer clusters, including quantitative metrics and key behavioral insights.

	avg spending for costomer	avrg transtiction per costmer
Cluster 1	854.7	78.4
Cluster 2	904.0	64.1
Cluster 3	182.0	22.4
Cluster 4	555.8	57.9
Cluster 5	2776.3	163.0

Key Insights

Cluster 5: VIP Customers

High frequency + High spending = VIP

Cluster 3: Standard Customers

Lowest engagement = Standard customers (49%)

Cluster 1: Core Customers

Balanced metrics = Core customers

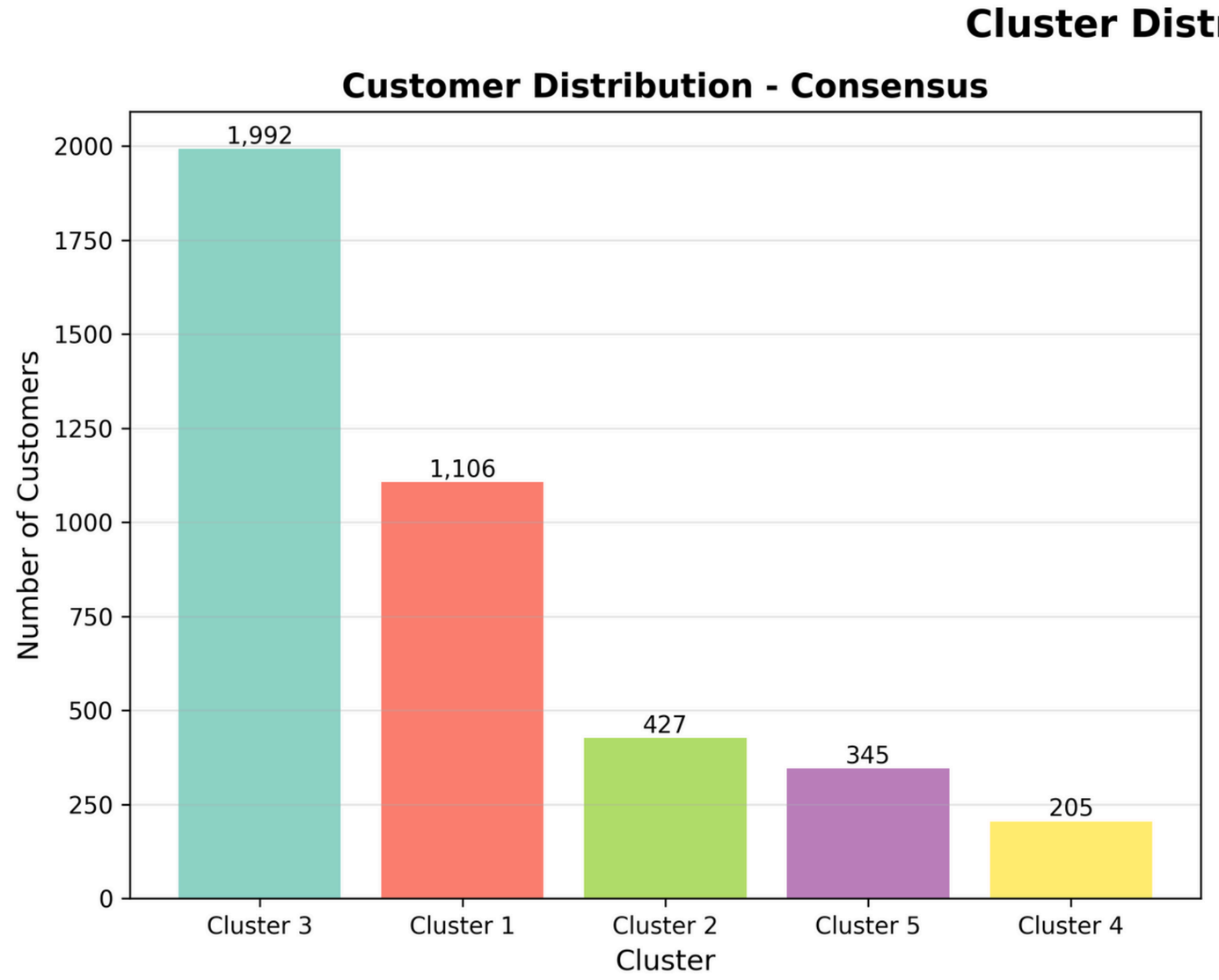
Cluster 4: New Customers

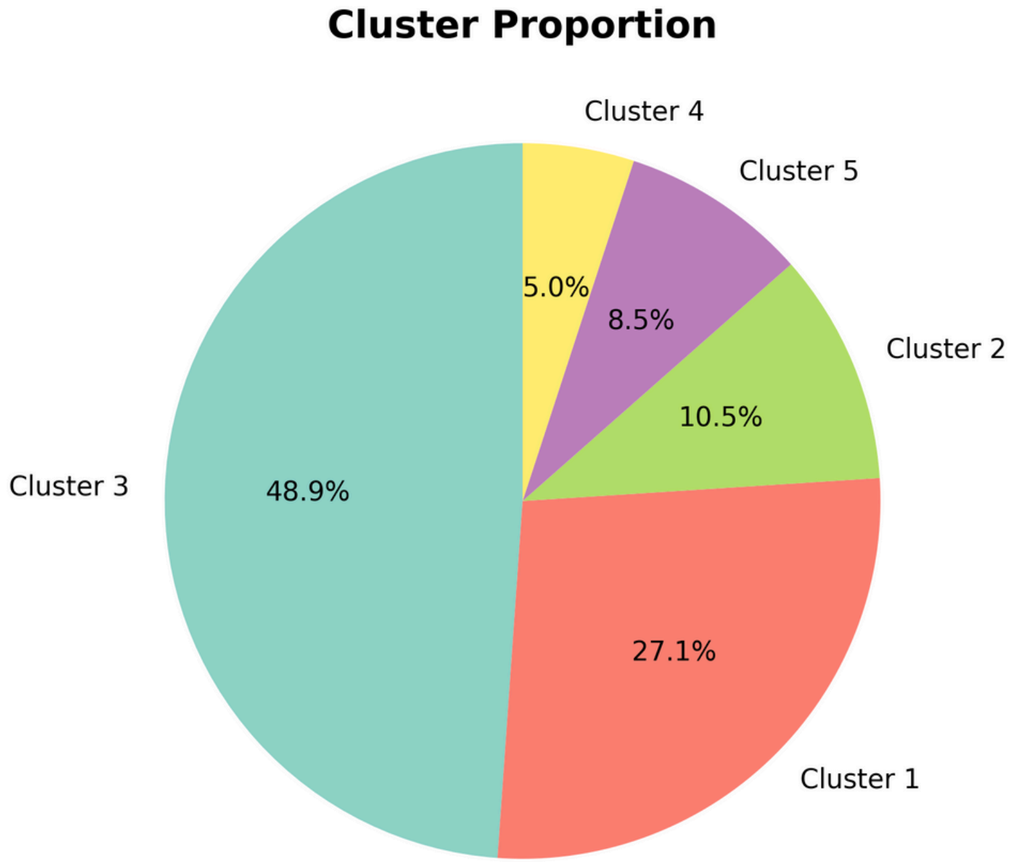
Very recent = New customers

Cluster 2: At-Risk Customers

Inactive but high-value = At-risk

WHY CLUSTERS ARE IMBALANCED





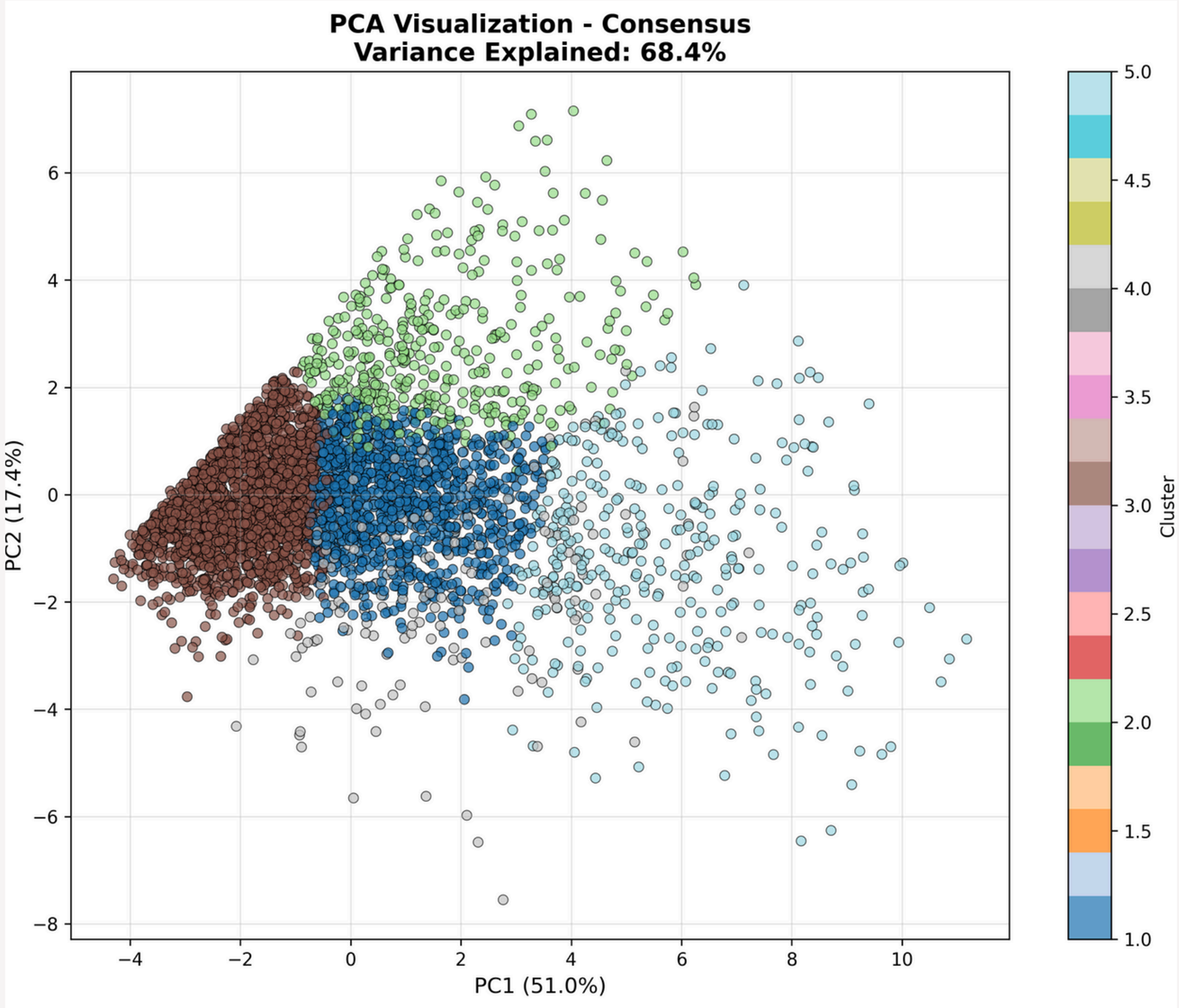
This imbalance REFLECTS REAL CUSTOMER DISTRIBUTION.

Retail Reality - The 80/20 Rule:

- ├── 20% of customers generate 80% of revenue
- ├── Most customers are casual, occasional buyers
- ├── High-value customers (VIPs) are naturally rare
- └── This is EXPECTED and BUSINESS-REALISTIC

13_cluster_distribution.png

VISUALIZATION - PCA & t-SNE

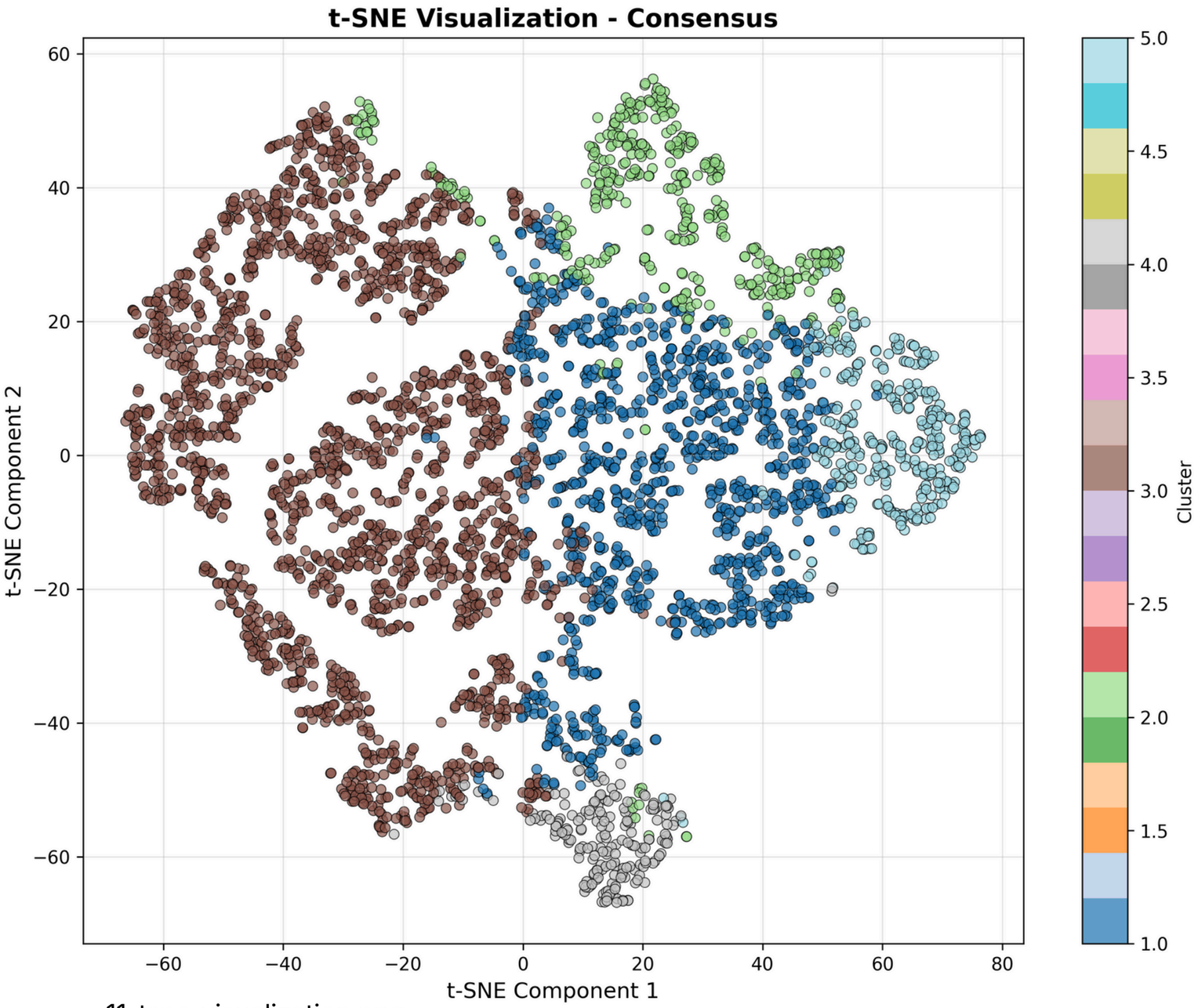


PCA Projection:

- PC1 captures 51.0% of variance
- PC2 captures 17.4% of variance
- Total explained: 68.4%
- Interpretation: ~70% of customer variation captured in 2 dimensions

10_pca_visualization.png

VISUALIZATION - PCA & t-SNE



11_tsne_visualization.png

t-SNE Projection:

- Perplexity: 30 | Iterations: 1,000
- Non-linear dimensionality reduction
- Preserves local structure better than PCA

What t-SNE Shows:

✓ Cluster 3 (Standard) forms tight group (bottom-left)

✓ Cluster 5 (Champions) distinct (top-center)

✓ Cluster 1 (High-Potential) spreads (right side)

✓ Clusters 2 & 4 show natural overlap (middle)

MODEL VALIDATION METHODOLOGY

Since Clustering doesn't use train/test splits like supervised learning, but we validated through:

CONSENSUS CLUSTERING (Stability Check)

- Ran K-Means 10 times with different random seeds
- Built co-occurrence matrix (how often customers cluster together)
- Applied hierarchical clustering on consensus matrix
- Result: Stable assignments across runs ✓

SILHOUETTE ANALYSIS (Quality Check)

- Measured how well each customer fits their cluster
- Range: -1 (misassigned) to +1 (perfect fit)
- Overall: 0.269 (moderate, but expected for customer data)
- Cluster 3 best: 0.335 (fair quality) ✓

STATISTICAL VALIDATION (ANOVA)

- Tested: Do clusters actually differ?
- Result: All 7 features $p < 0.0001$
- Effect sizes: Large ($\eta^2 > 0.14$ for all features)
- Conclusion: Clusters are statistically distinct ✓

BUSINESS VALIDATION (Interpretability)

- Segments make intuitive sense
- Marketing team can act on strategies
- Clear differentiation between groups ✓

Why No Train/Test Split?

No labels to predict

Goal: Discover patterns, not predict outcomes

Used ALL 4,069 customers

Our Validation Strategy:

- Consensus Clustering (10 runs)
- Statistical Tests (ANOVA)
- Silhouette Analysis (per-cluster quality)
- Effect Size Validation (η^2)
- Business Interpretability

Result: ✓ Statistically valid + Actionable

RECOMMENDATIONS AND FUTER WORK

FEATURE ENGINEERING:

Add product category features (Electronics vs. Clothing)

Include temporal patterns (Seasonal buyers vs. Year-round)

Calculate Customer Lifetime Value (CLV) metric

Expected gain: +0.05-0.08 silhouette

thank you for listening