esade

Clustering Algorithms
Online Retail Case Study

Timo Sachs | AI2 – Clustering Assignment
Barcelona, 31.03.2024

Do Good. Do Better.

esade

esade

Smart customer segmentation provides opportunity to retain high-value client revenues and (re-)engage dormant customers

Project Objective

Optimize marketing and customer retention strategies using clustering algorithms to segment existing customer base and make actionable recommendations to improve campaign effectiveness for groups with similar characteristics and purchasing behaviors.

Customer insights

- **4 customer segments**: Dormant Majority (68%), Occasional Spenders (24%), VIP Customers (8%), and High-Volume Anomalies (<1%) with specific marketing strategies available
- Top **7.2% of customers** ('VIP Customers' segment) generate **47.4% of total revenue**
- Approximately 34.4% of customers generate 80% of revenue
- Overall >90% of customers from the UK, while profitable segments are more internationally diverse
- Heavy reliance on a single market presents risk and opportunity

Model findings

- **K-Means performed slightly better** than hierarchical clustering (Silhouette Score: 0.344 vs. 0.300)
- DBSCAN **not effective** with the given dataset and feature engineering approach (only 1-2 clusters)
- Log-transformation and robust scaling absolutely crucial for meaningful cluster results
- **Remaining anomalies (0.02-0.07%)** indicate potential B2B customers or data issue

Analysis process is based on **step-by-step processing pipeline** fed into 3 clustering algorithms to derive customer segments

Data Source → **Explore** → **Clean & process** → **Compare** → **Analyze**

Purchases

Online Retail.xlsx

Missing values
- Impute description
- Assign new ID per invoice (but drop for clustering)

Drop
- Duplicates (5226)
- Specific Outliers
- Negative quantity (Returns)
- Errors (negative price)

Engineer
- Recency
- Frequency
- Monetary
- Avg. Basket Size
- Avg. Basket Volume
- Distinct Items
- IsUK

Transform
- Log-scaling
- Robust Scaling

Determine Parameters
- Elbow
- Silhouette
- Epsilon / min_samples simulation

k-Means

Hierarchical

DBSCAN

Customer Segments
- Method overlap
- Top 3 products
- Recommendations

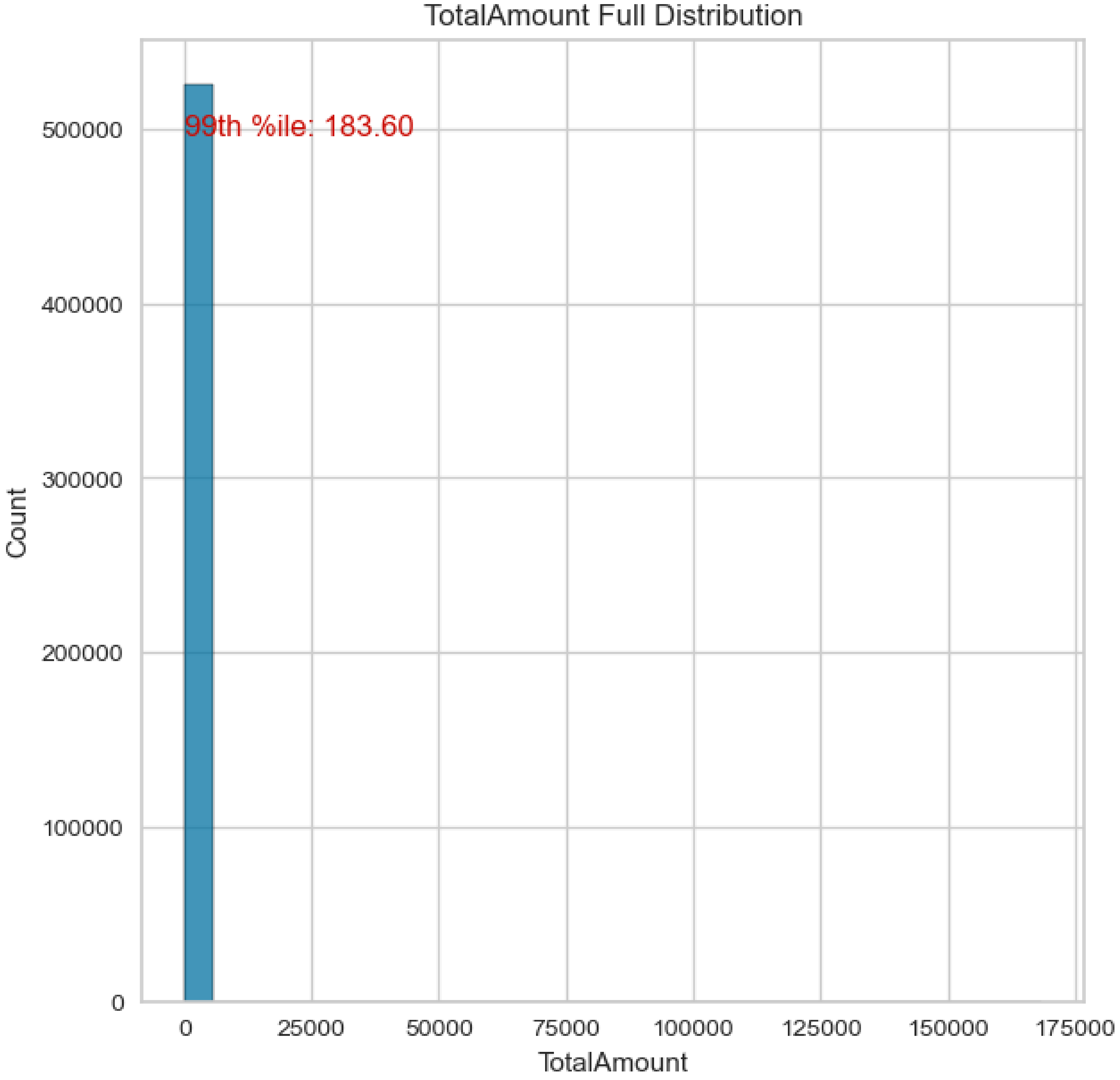
03

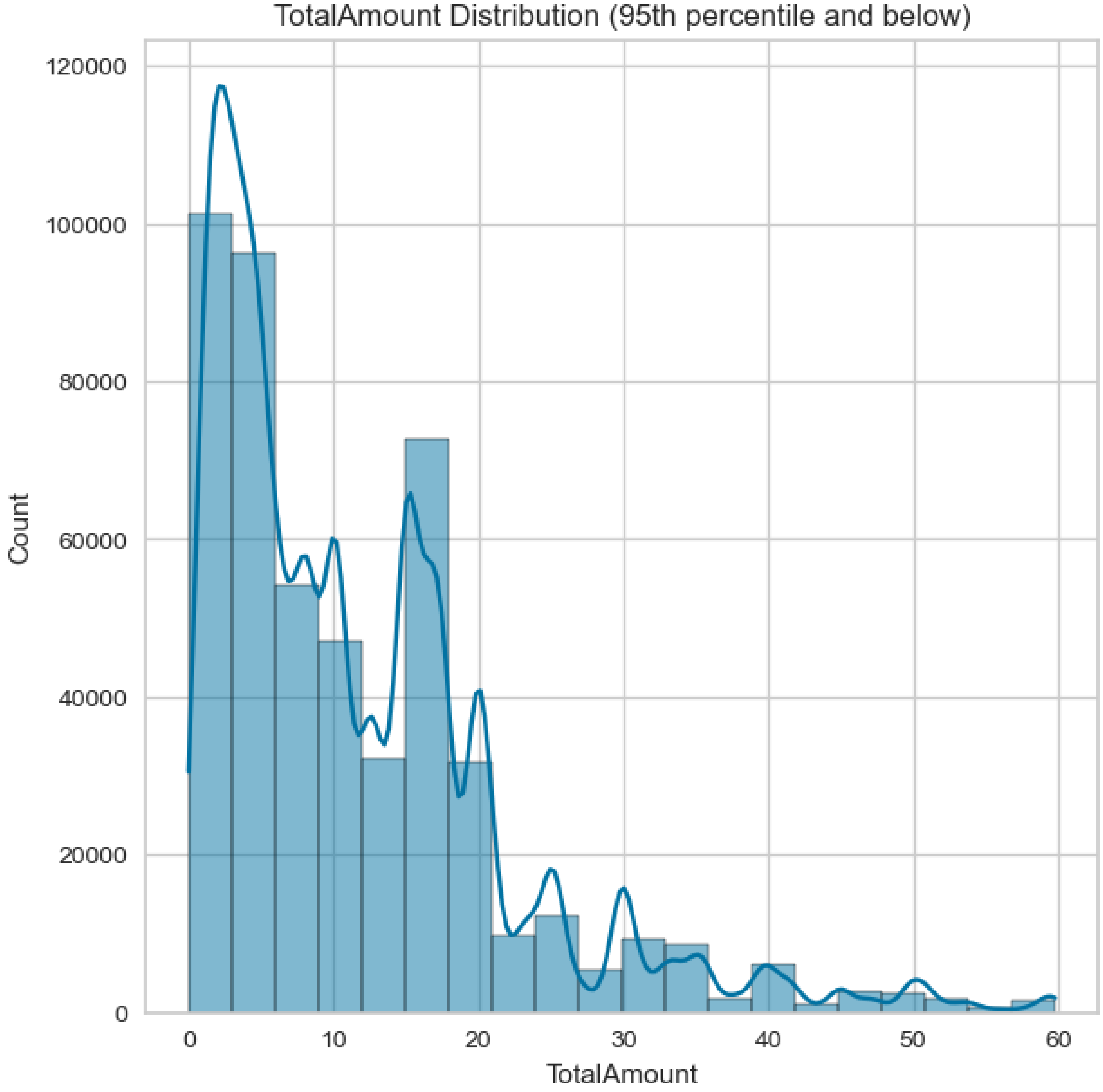
Feature Cleaning, Selection & Processing

Feature selection and preprocessing is based on a mix of exploration, imputation and dropping features

	Contains	Missing values	Processing
InvoiceNo	Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.	No	Used for CustomerID imputation
StockCode	Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.	No	No
Description	Product (item) name.	Yes	Yes, impute with “unknown”
Quantity	The quantities of each product (item) per transaction. Numeric.	No	Calculate total amount
InvoiceDate	Invoice Date and time. Numeric, the day and time when each transaction was generated.	No	Derive monthly activity
UnitPrice	Unit price. Numeric, Product price per unit.	No	Calculate total amount
CustomerID	Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.	Yes	Group per customer id
Country	Country name. Nominal, the name of the country where each customer resides.	No	Create IsUK flag

Total amounts (derived from quantity and unit price) are highly skewed and thus log-transformed after outlier removal





Outlier removal is **not done based on percentiles**, but through in-depth analysis of top/bottom values with **hand-picked removals**

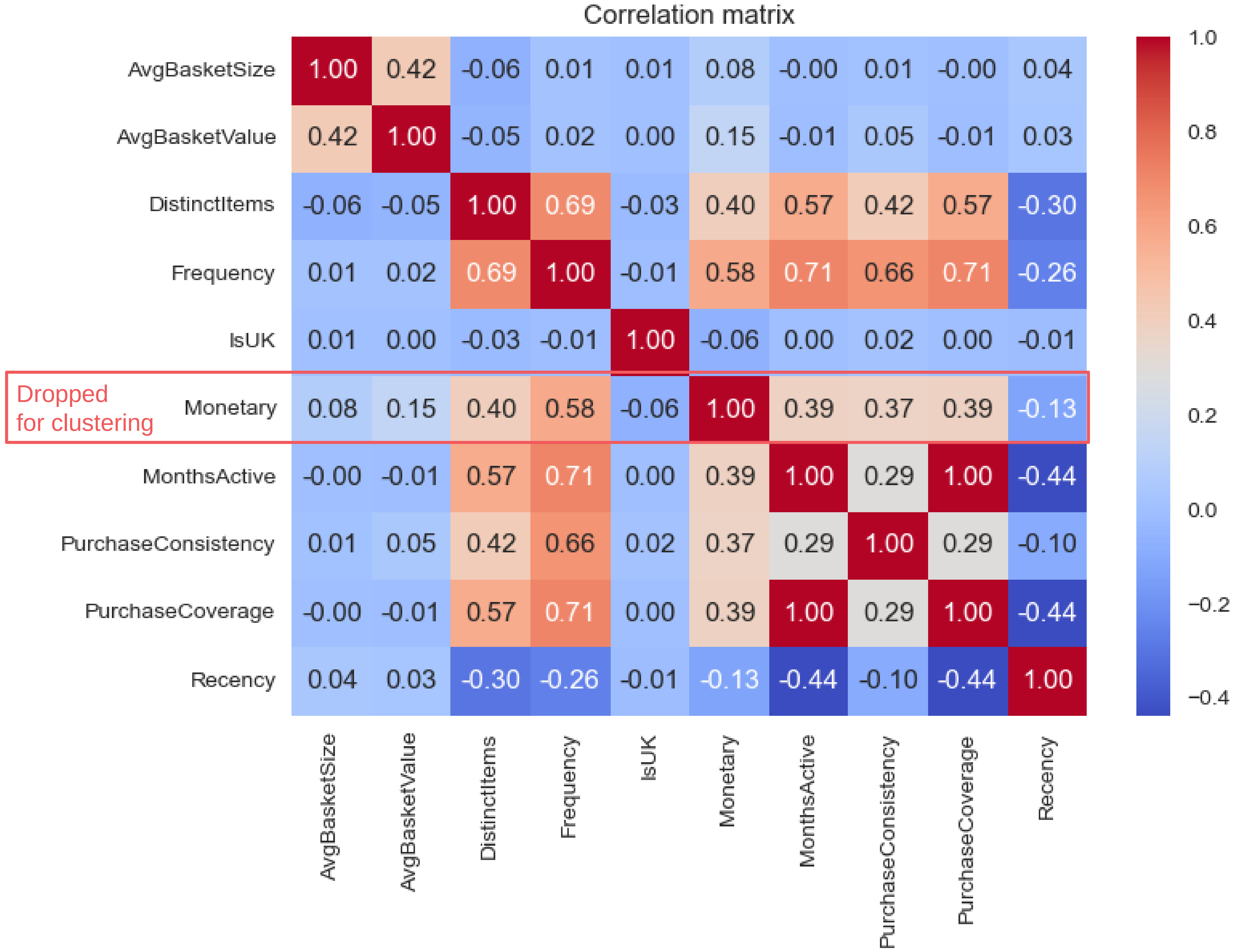
	Stock Code / Invoice No.	Rows removed	Rationale
Manual Bookings	M	316	No insights / information can be derived; Might delute results
Postage Fees	DOT and POST	706 + 1126	Postage is not a product to be bought analyzed
Padding	PADS	3	Padding distorts unit prices due to ultra low price; Also not a real product
Bank Charges	BANK CHARGES	12	Bank charges are part of the profit calculation, but are not counted as product / item in this analysis
Wrong Quantity	556444	.	Fixed quantity based on comparable transaction (reset quantity to 1 instead of 60)
Amazon Fee	AMAZONFEE	2	Excluded since not a real product

Common RFM metrics are enhanced with **additional basket metrics** and seasonality

	Metric / Calculation	Insight goal
Recency	Time since last purchase	Active vs. passive customers
Frequency	Count of purchases	Count of purchases
Monetary	Sum of total spend per customer	Focus on highest-revenue customers
Avg. Basket size	Avg. total spend per invoice	Identify bulk buyers vs. small purchases
Distinct items	Number of unique products bought	Buying behavior (broad vs. specific)
PurchaseConsistency	Number of purchases within active month	Identify high engagement in active month
Purchase Coverage	Months with activity / all months	Identify consistent buyers over time
IsUK	IsUK	Region-specific analysis with less sparsity

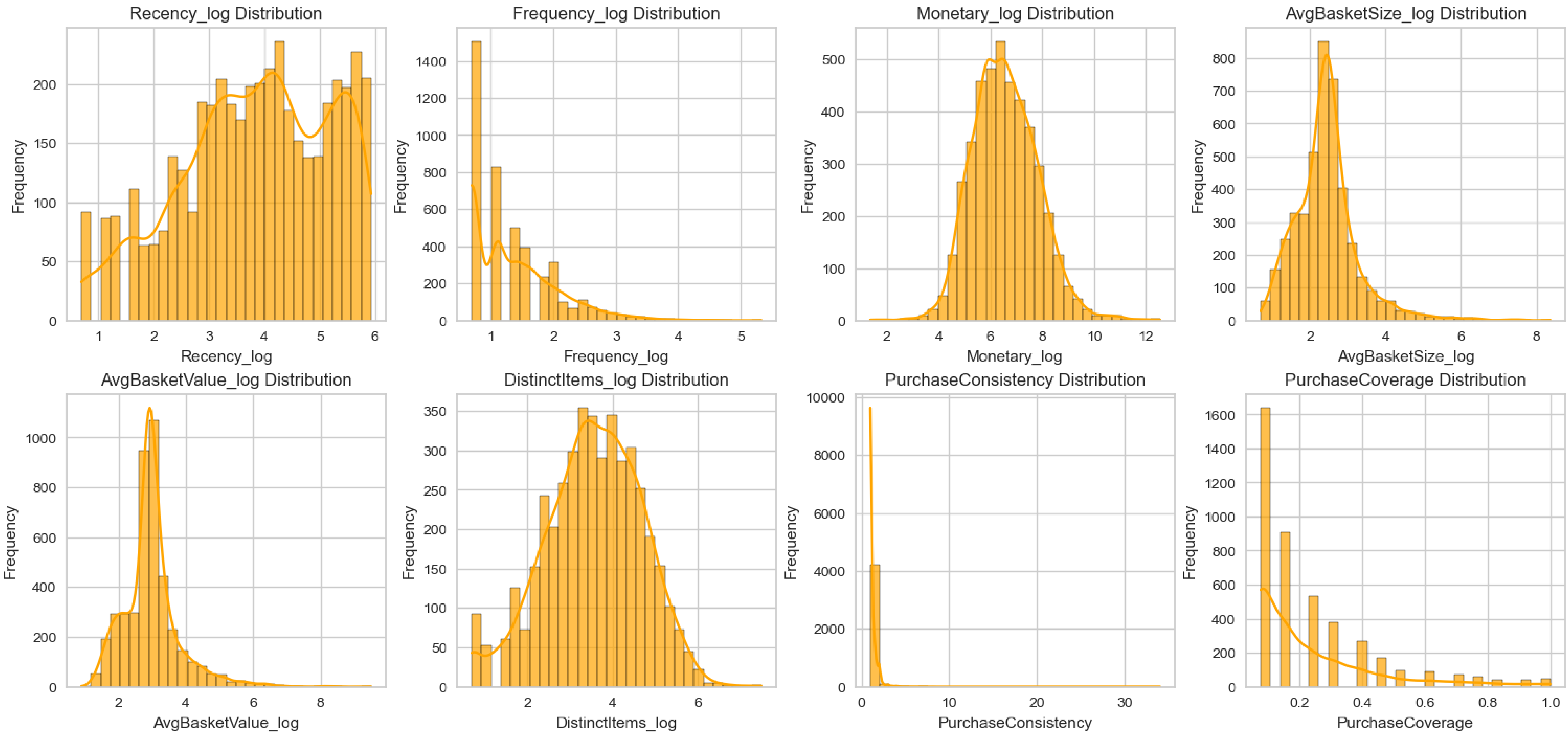
Newly engineered **feature correlation** already provides **insights into behavioral trends** on global customer base level

- **Frequency** and **Distinct Items** have a strong positive correlation (+0.69), meaning that customers who buy frequently also tend to purchase a wider variety of items.
- **Frequency** and **Months Active** are highly correlated, which makes sense—customers who have been around longer tend to buy more often.
- **Purchase Coverage** and **Frequency** are also strongly linked, suggesting that frequent shoppers engage with multiple product categories.
- **Recency** has a negative correlation with **Months Active**, meaning that longer-term customers tend to have less recent purchases—potential sign of disengagement.
- **Monetary** value is most related to **Frequency**, confirming that the more often someone buys, the more they spend overall.



esade

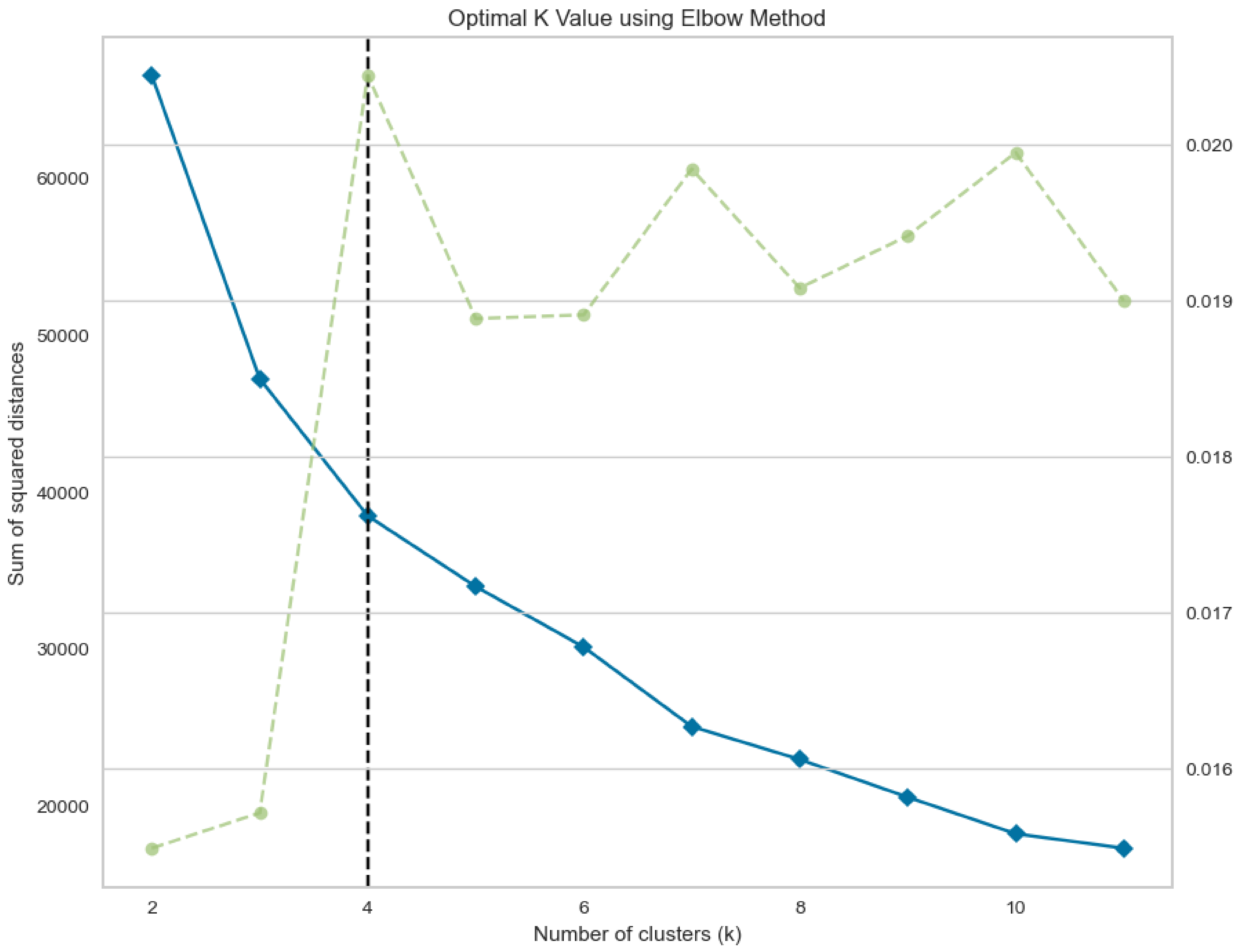
Log-transformation leads to normalized distribution of quantitative features which can be fed into the clustering algorithms

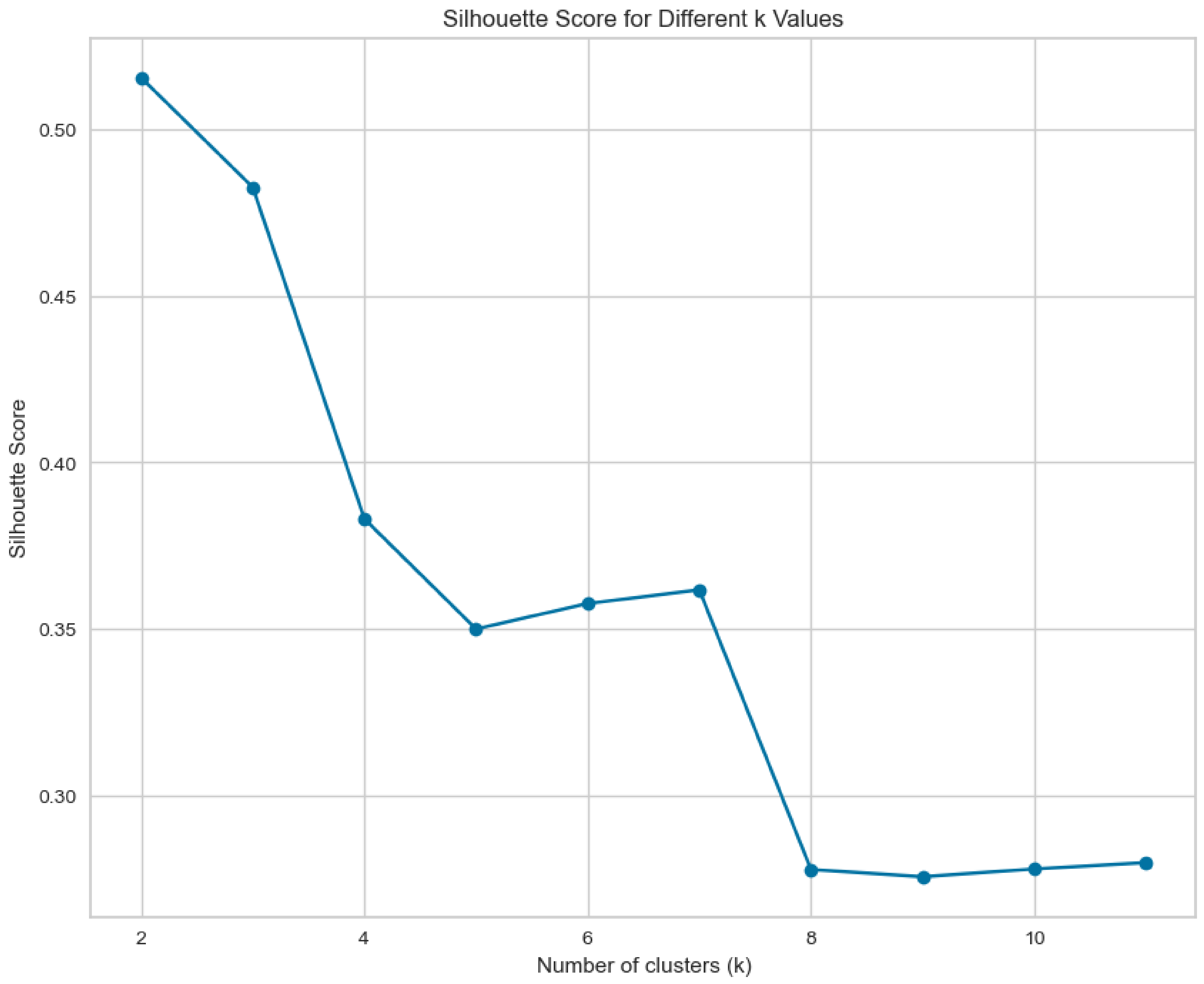


04

Parameter selection

Elbow method suggests an **optimal number of 4 clusters**; Silhouette score quite low due to complex difficult-to-separate customer groups





esade

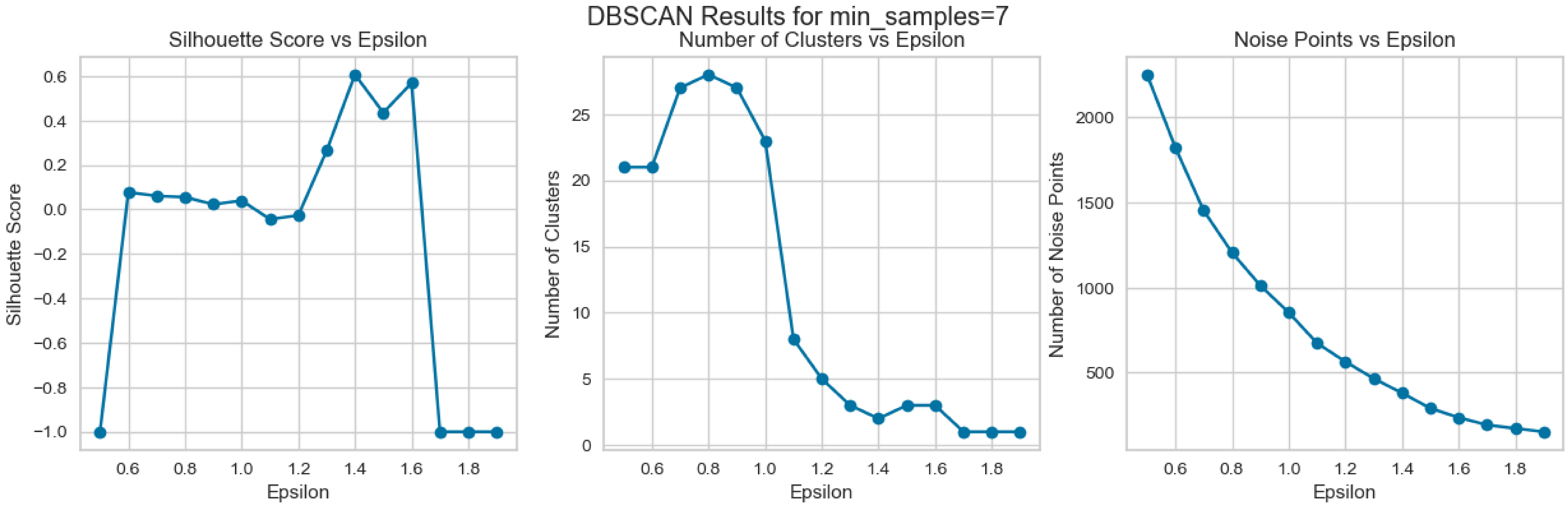
Dendrogram reveals clear customer segments despite heavy branching towards lower distance levels; Same cluster cut is applied as for k-Means

esade

Despite multiple parameter simulations, overall DBSCAN performance was insufficient to derive meaningful customer segments

Best DBSCAN parameters

- Epsilon: 1.4
- Min Samples: 7.0
- Number of clusters: 2.0
- Noise points: 381.0 (8.79%)
- Silhouette Score: 0.6071



05

Clustering Results & Model Comparison

esade

Clustering results show distinct customer segment characteristics with 3 customer groups and 1 outlier / anomaly cluster

	Recency	Frequency	Monetary	Avg Basket Size	Avg Basket Value	Distinct Items	Purchase Consistency	Purchase Coverage	IsUK	# of Custom.	% of Custom.
Dormant Majority	124.22	1.66	583.74	23.06	40.11	31.92	1.09	0.12	90.45	2838	**65.50**
Occasional Spenders	38.13	5.85	2238.66	18.82	27.15	98.96	1.29	0.36	90.25	1179	**27.21**
VIP Customers	12.06	20.27	12858.93	29.19	49.40	177.01	1.97	0.78	90.10	313	**7.22**
High-Value Anomalies	124.67	146.33	59126.27	8.56	16.70	1191.00	21.72	0.69	66.67	3	**0.07**

Detailed distributions in appendix

esade

k-Means and hierarchical clustering produced **similar results**; DBSCAN **struggled to identify customer segments** despite parameter tuning

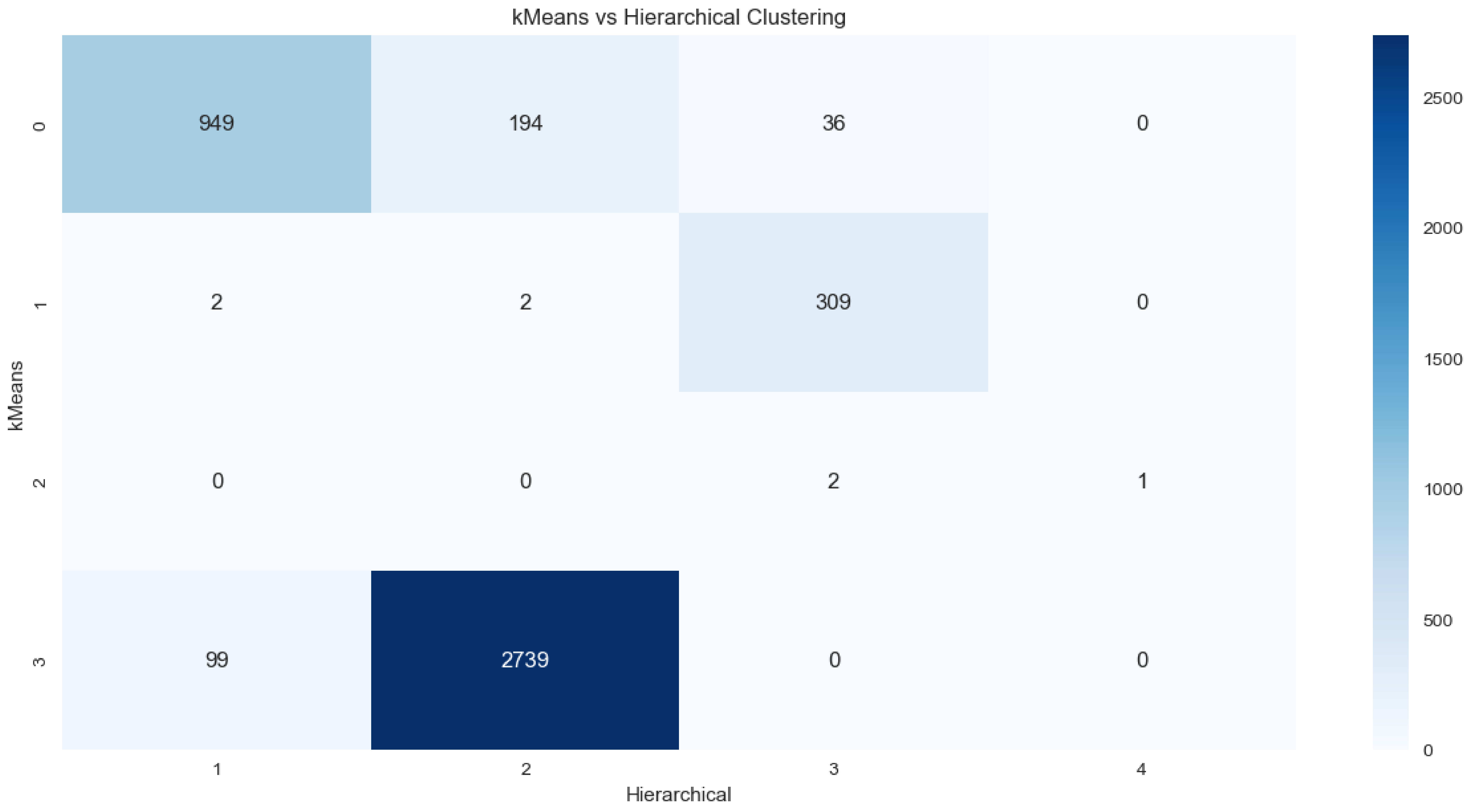
	K-Means	Hierarchical	DBSCAN
+ Strengths	▪ Well-balanced customer segments (8%, 24%, 68%) ▪ Clear, actionable segments with distinct RFM profiles ▪ Good separation between high-/low-value customers ▪ Results immediately applicable to marketing strategy	▪ Visualization (dendrogram) of customer relationships ▪ Flexibility in choosing segment count after analysis ▪ Less affected by outliers than K-means	▪ Automatically detected outliers (high 13.5% of customers) ▪ Did not require pre-specifying number of segments ▪ Found clusters of varying shapes and densities
– Weak points	▪ May not capture complex relationships ▪ Requires choosing k (=4) ▪ Sensitive to outliers (high-spent customers)	▪ Computationally more intensive ▪ Some smaller segments with unclear business value ▪ More difficult to explain to business stakeholders	▪ Separate parameter tuning ▪ Classified many potentially valuable customers as "noise" ▪ Did not identify meaningful clusters

Compare results (K-Means, Hierarchical)

Discard for this analysis (DBSCAN)

esade

Confusion matrix of k-means and hierarchical clustering reveal similar customer attribution despite different labeling…



… which is supported by the overall similar comparison metrics across customer segments

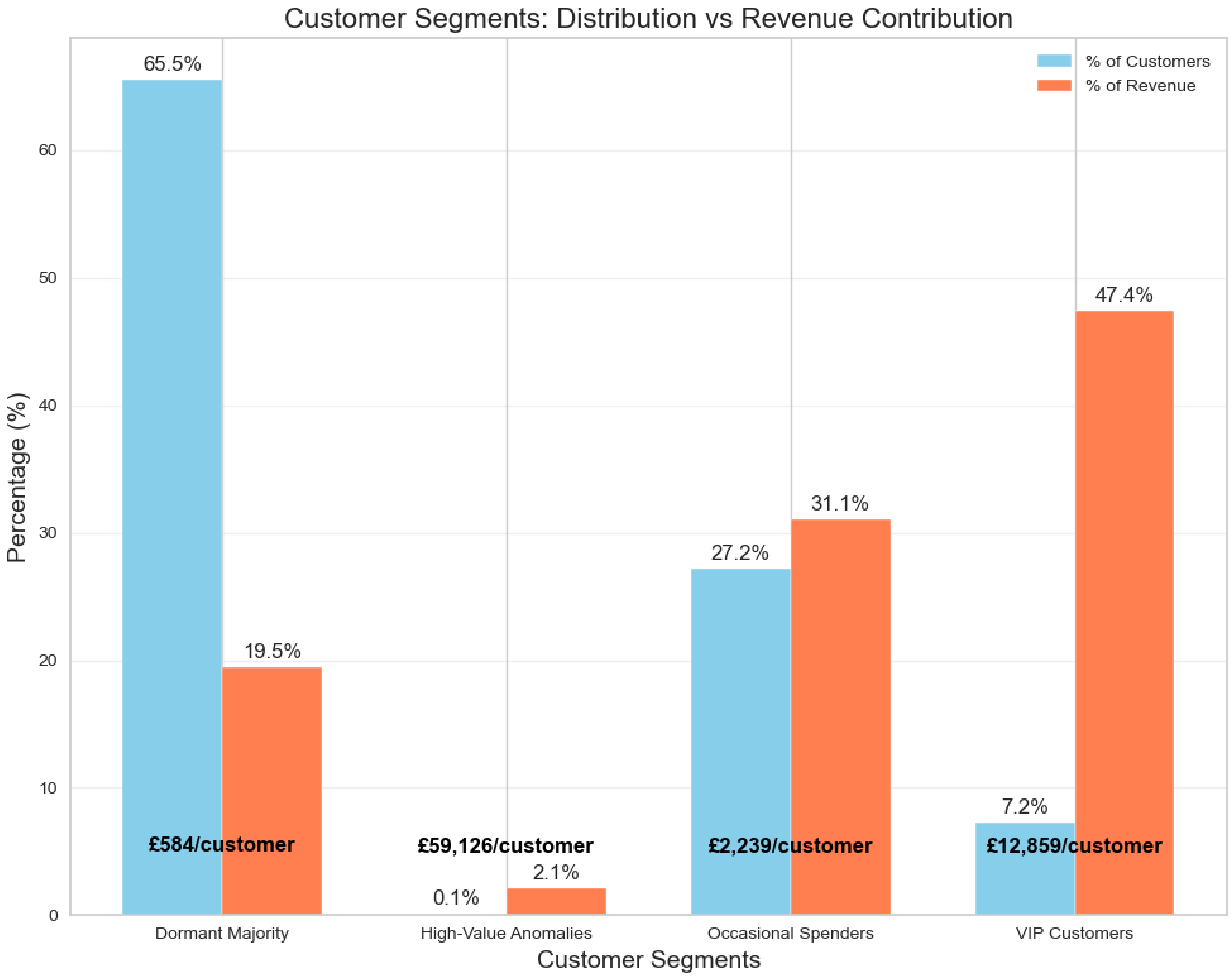
Robust Segmentation	The **remarkable consistency between K-means and hierarchical clustering** validates the segmentation approach. Both methods identified similar customer groups with consistent behavioral patterns, increasing the confidence in these segments as a foundation for strategic planning.
Segment Proportions	Hierarchical clustering assigned slightly **more customers to the dormant segment** (67.74% vs. 65.50%) and fewer to the middle segment (24.23% vs. 27.21%), suggesting some borderline customers may display characteristics that could place them in either group.
Different Outlier Detection	The most significant difference appears in the outlier segment, where **hierarchical clustering (only) identified a single customer** with very different behavioral patterns than the three anomalies found in K-means. This suggests these methods might have different sensitivities to unusual purchasing patterns.
Value Distribution	The distribution of value across segments remains consistent, with a small percentage of customers (8.01% in hierarchical clustering) accounting for disproportionately high spending, frequency, and engagement, while the majority (67.74%) remain relatively disengaged with minimal spending.

06

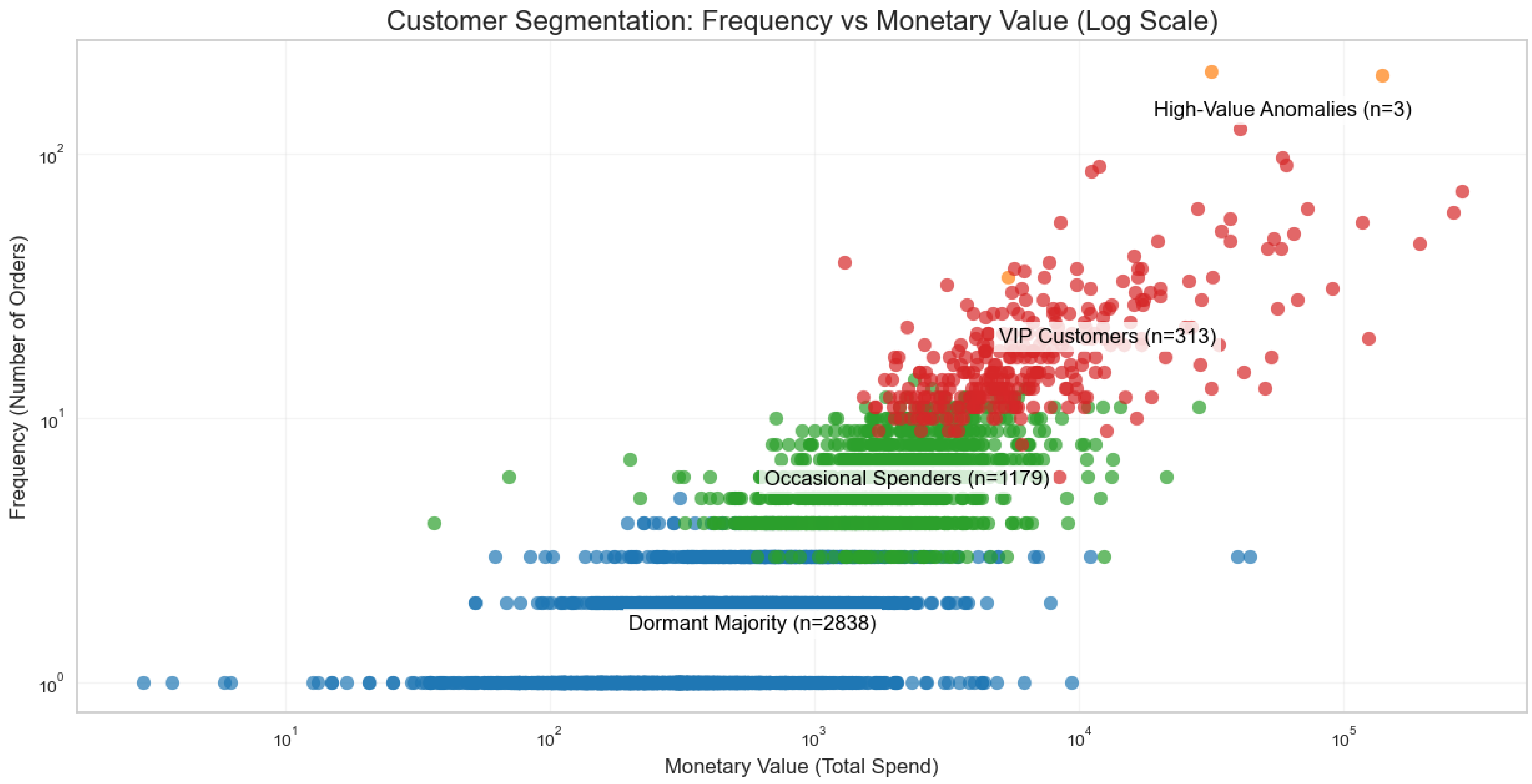
Segment Deep-dive and Marketing Strategies

Revenue contribution by customer segments reveals strong pareto distribution for VIP customers and dormant majority

- **Dormant majority** makes up 65% of customer based but only **19.5% of all revenues**
- **VIP customers** contribute **47.4% of revenues** (~2.5 of dormant but with only ~1/10 of customers demonstrating strong powerlaw distribution and need to retain those customers
- Occasional spenders might tip either way (dormant or VIP)
- High-value anomalies need to be revisited to confirm outlier removal approach or identify fraud / big B2B buyers



Customer segment clusters show relationship between number of orders and total spend, whereas broader base is separated from outliers



The **dormant majority** should be targeted using **reactivation campaigns** to turn them into returning customers

Segment	**Dormant Majority** | 66 %
Key Characteristics	▪ 124 days since last purchase ▪ Only 1.7 purchases on average ▪ $584 total spend per customer
Top 3 Products	▪ World War 2 Gliders ▪ White Hanging Heart ▪ Fairy Cake Flannel
Opportunity	Reactivation could unlock significant revenue

Marketing Recommendation

- **Reactivation Campaign:** "We Miss You" emails with personalized product recommendations
- **Win-Back Incentives:** First-time reorder discount or free shipping
- **Reminder Strategy:** Show previously purchased items with complementary suggestions
- **Seasonal Triggers:** Reach out during key seasonal moments based on past purchase timing

Increase average spending of **occasional spenders** to turn them into VIP customers

Segment	**Occasional Spenders** | 27 %
Key Characteristics	▪ More recent activity (38 days) ▪ Moderate frequency (5.8 purchases) ▪ $2,239 lifetime value ▪ 98 distinct items vs. 31 for dormant
Top 3 Products	▪ World War 2 Gliders ▪ Pack of 72 Retrospot ▪ Small Popcorn Holder
Opportunity	Most likely to convert to VIP status with right engagement

Marketing Recommendation

- **Frequency Program:** Reward increased purchase frequency with escalating benefits
- **Category Expansion:** Introduce related product categories based on past purchases
- **Mid-tier Loyalty:** Create "rising star" tier with visible pathway to premium benefits
- **Personalized Bundles:** Offer curated collections based on previous basket analysis

esade

Maintain **high engagement of VIP clients** by offering premium experience and exclusive benefits

Segment	**VIP Customers** | 8 %
Key Characteristics	▪ Very recent activity (12 days) ▪ High frequency (20.3 purchases) ▪ $12,858 lifetime value (22× dormant) ▪ Highest basket value ($49.4)
Top 3 Products	▪ Jumbo Bag Retros ▪ Small Popcorn Holder ▪ World War 2 Gliders
Opportunity	**Retention is critical** as they drive significant revenue

Marketing Recommendation

- **Premium Experience:** White-glove customer service and early access to new products
- **Relationship Building:** Personal shopping assistance and product customization
- **Exclusive Benefits:** VIP-only events and substantial loyalty rewards
- **Referral Program:** Incentivize bringing in similar high-value customers

esade

Based on the cluster analysis, clear **next steps** can be derived to improve customer retention and increase overall revenues

Customer management

- **Develop VIP retention program:** Early product access, personalized service, exclusive benefits
- **Create conversion path for Occasional Spenders:** Incentivize increased purchase frequency
- **Design reactivation campaigns:** Target dormant customers with personalized recommendations

Analytics Enhancement

- Implement **real-time segment scoring**: Dynamically assign customers as behaviors change
- Conduct A/B testing: Compare segment-specific marketing approaches against generic campaigns
- Longitudinal analysis: Track customer movement between segments over time
- **Integrate customer service data:** Enhance segmentation with support interactions and satisfaction ratings

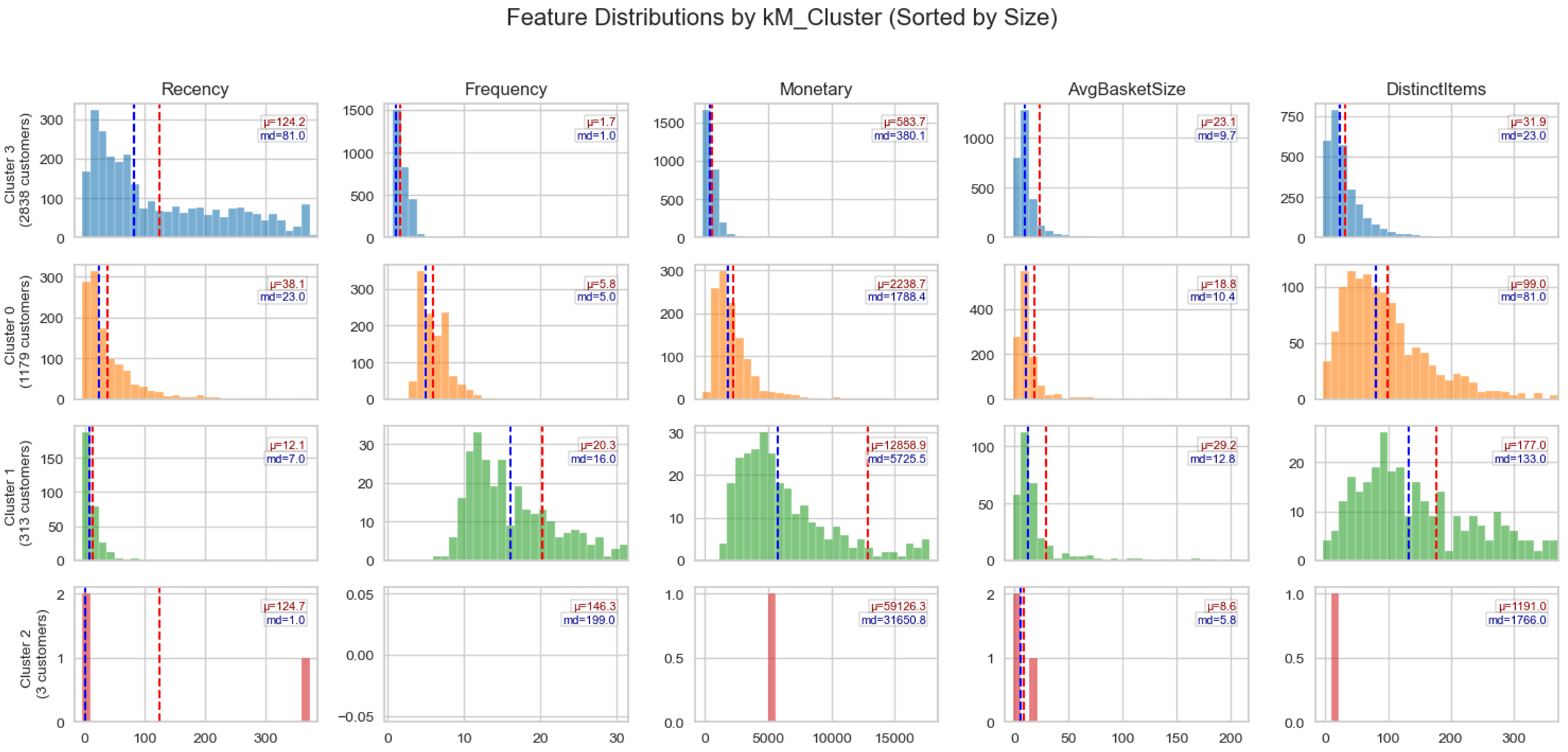
Timeline

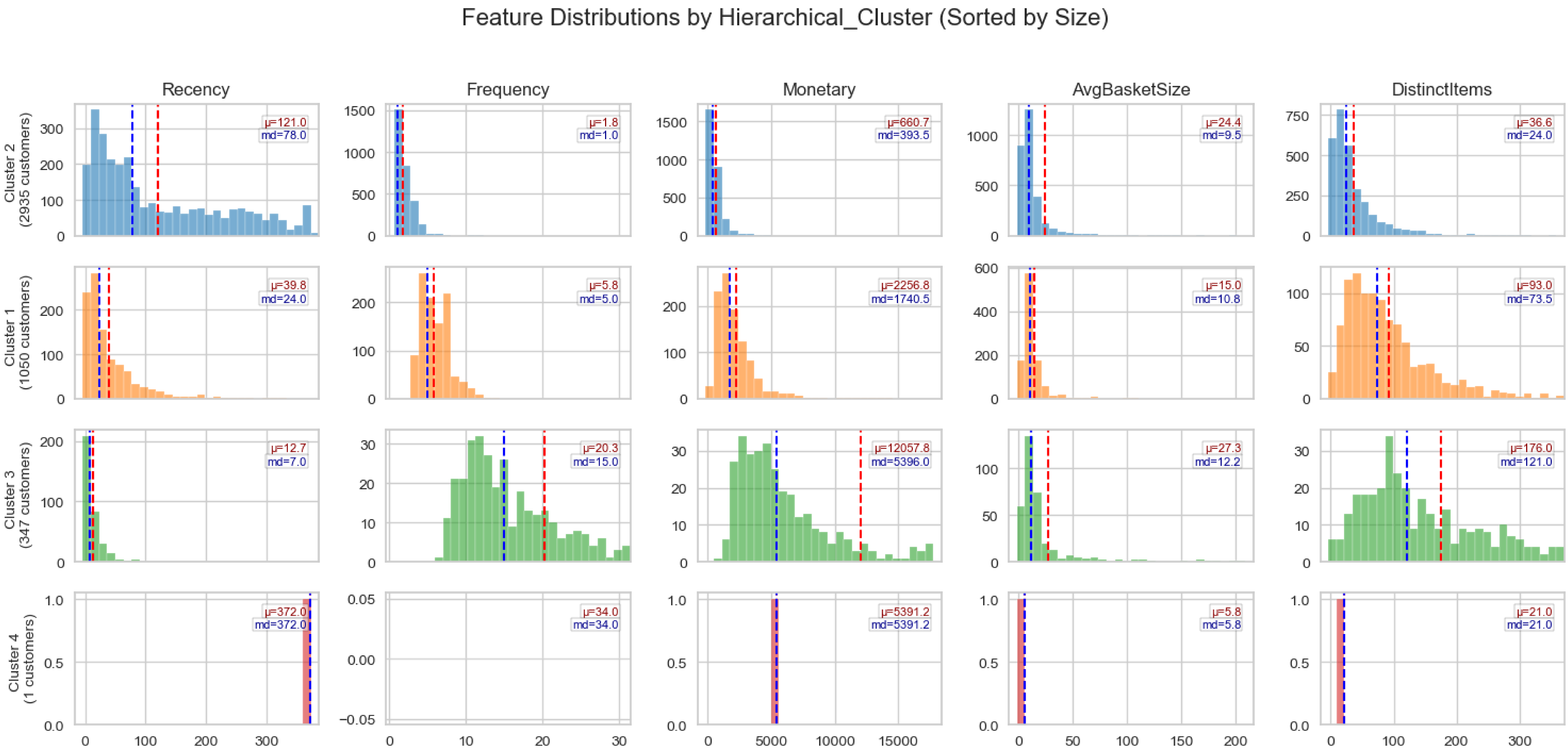
- **Month 1:** Set up segment-based targeting in marketing platforms
- **Month 2-3:** Launch pilot campaigns for each segment with distinct messaging
- **Month 3-6:** Measure segment-specific conversion rates and adjust strategies
- **Month 6+:** Implement predictive analytics to identify at-risk customers and growth opportunities

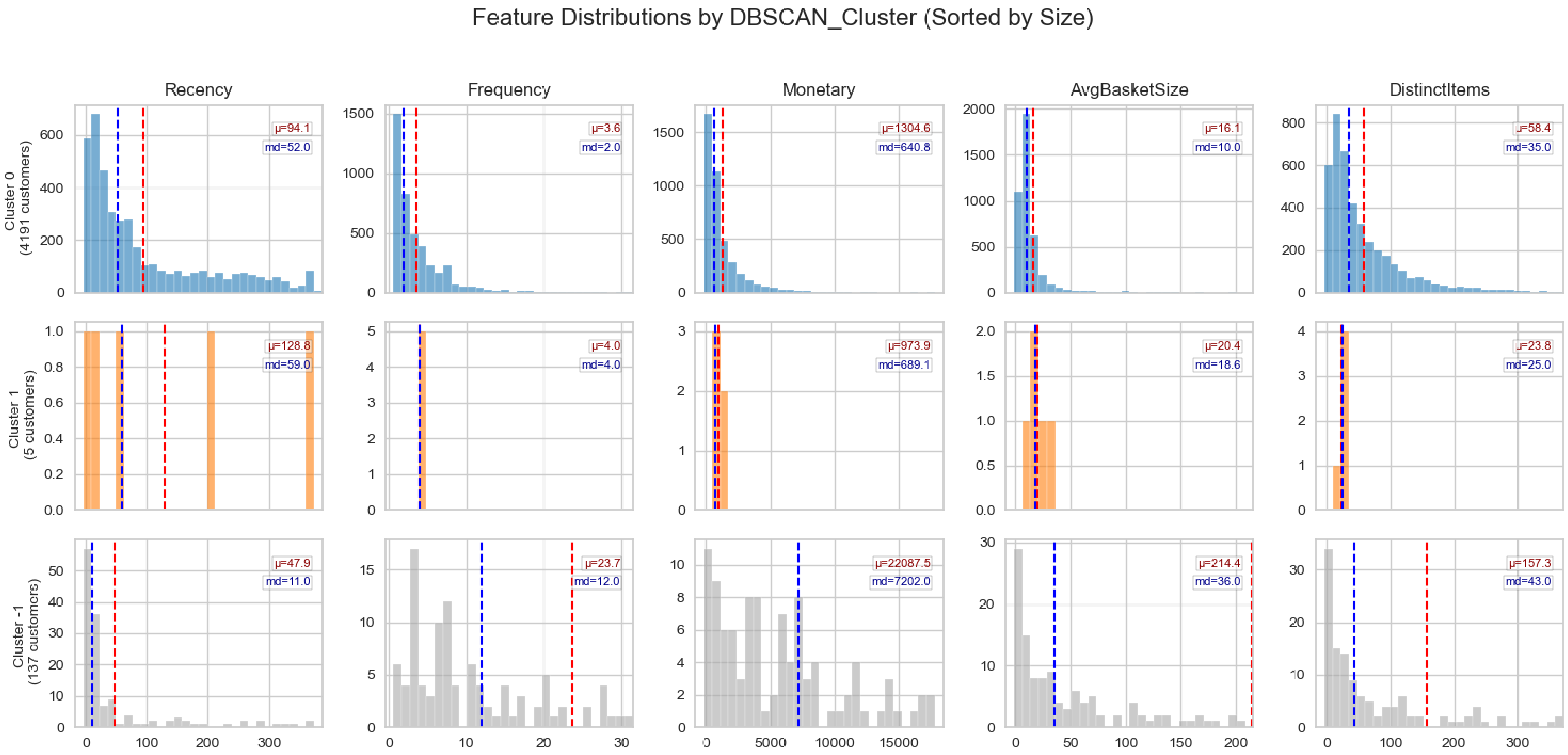
esade

Appendix

Do Good. Do Better.


Feature Distributions by kM_Cluster (Sorted by Size)
Recency
Frequency
Monetary
AvgBasketSize
DistinctItems
Cluster 3
(2838 customers)
μ=124.2
md=81.0
μ=1.7
md=1.0
μ=583.7
md=380.1
μ=23.1
md=9.7
μ=31.9
md=23.0
Cluster 0
(1179 customers)
μ=38.1
md=23.0
μ=5.8
md=5.0
μ=2238.7
md=1788.4
μ=18.8
md=10.4
μ=99.0
md=81.0
Cluster 1
(313 customers)
μ=12.1
md=7.0
μ=20.3
md=16.0
μ=12858.9
md=5725.5
μ=29.2
md=12.8
μ=177.0
md=133.0
Cluster 2
(3 customers)
μ=124.7
md=1.0
μ=146.3
md=199.0
μ=59126.3
md=31650.8
μ=8.6
md=5.8
μ=1191.0
md=1766.0


Feature Distributions by Hierarchical_Cluster (Sorted by Size)
Recency
Frequency
Monetary
AvgBasketSize
DistinctItems
Cluster 2
(2935 customers)
Cluster 1
(1050 customers)
Cluster 3
(347 customers)
Cluster 4
(1 customers)
μ=121.0
md=78.0
μ=1.8
md=1.0
μ=660.7
md=393.5
μ=24.4
md=9.5
μ=36.6
md=24.0
μ=39.8
md=24.0
μ=5.8
md=5.0
μ=2256.8
md=1740.5
μ=15.0
md=10.8
μ=93.0
md=73.5
μ=12.7
md=7.0
μ=20.3
md=15.0
μ=12057.8
md=5396.0
μ=27.3
md=12.2
μ=176.0
md=121.0
μ=372.0
md=372.0
μ=34.0
md=34.0
μ=5391.2
md=5391.2
μ=5.8
md=5.8
μ=21.0
md=21.0


Feature Distributions by DBSCAN_Cluster (Sorted by Size)
Recency
Frequency
Monetary
AvgBasketSize
DistinctItems
Cluster 0
(4191 customers)
Cluster 1
(5 customers)
Cluster -1
(137 customers)
μ=94.1
md=52.0
μ=3.6
md=2.0
μ=1304.6
md=640.8
μ=16.1
md=10.0
μ=58.4
md=35.0
μ=128.8
md=59.0
μ=4.0
md=4.0
μ=973.9
md=689.1
μ=20.4
md=18.6
μ=23.8
md=25.0
μ=47.9
md=11.0
μ=23.7
md=12.0
μ=22087.5
md=7202.0
μ=214.4
md=36.0
μ=157.3
md=43.0

esade

Analysis of top 3 product reveals commonalities across customer segments as well as deviations for VIP and anomalies