VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Customer Segmentation

COURSE : Machine Learning

SEMESTER : Fall Semester 2025-26

COURSE ID : BCSE 209P

SLOT : A1 + TA1

SCHOOL : SCOPE

FACULTY : Prof. ARPAN GARAI

GITHUB REPO :
https://github.com/GANESH-MAHARAJ/MachineLearning-Course-Project.git

SUBMITTED By :

Ganesh Maharaj Kamatham (*22BDS0168*)

Koushik *(22BDS0431)*

TVS CHARAN KRISHNA *(22BDS0213)*

Jagadeesh U *(22BDS0310)*

Table of Contents

1. Introduction

Customer segmentation is the cornerstone of modern marketing analytics and data-driven business intelligence. It refers to the strategic process of **partitioning a heterogeneous customer base into homogeneous, actionable cohorts** based on behavioral, transactional, or demographic attributes. The ultimate goal is not statistical elegance but **economic alignment** - to ensure that communication, offers, pricing, and experiences are optimized for each micro-audience. In a world where acquisition costs are rising and attention spans are shrinking, precise segmentation directly translates to improved customer retention, optimized lifetime value, and sustainable profitability.

1.1 Why Customer Segmentation Matters

Businesses today operate in ecosystems overflowing with data yet constrained by attention. Every digital interaction leaves a trace - purchases, returns, browsing, or campaign responses - but this behavioral exhaust is high-dimensional, noisy, and ever-changing.

Traditional segmentation strategies, such as Recency–Frequency–Monetary (RFM) analysis coupled with K-means clustering, remain dominant due to their **transparency, speed, and interpretability**. They allow managers to intuitively understand "who buys, how often, and how much." However, the evolving digital landscape poses several challenges that limit the reach of such static models:

- **Behavioral volatility:** promotional campaigns and seasonal cycles alter customer purchase frequencies and value patterns, making segments drift over time.
- **Skewed and sparse data:** e-commerce data often exhibit heavy-tailed spend distributions, sporadic transactions, and missing identifiers, complicating clustering.
- **Feature distortion:** returns, discounts, and refunds bias the monetary component of RFM, leading to inflated or inconsistent customer value estimates.
- **Model rigidity:** centroid-based algorithms like K-means assume spherical clusters, ignoring overlap and irregular boundaries in behavioral space.
- **Governance gap:** marketing teams require clusters that are not only statistically valid but also explainable, stable, and deployable in CRM or CDP pipelines.

Thus, the analytical challenge is not simply to group customers, but to **find a repeatable, explainable, and adaptive representation of customer heterogeneity** that remains reliable as the business evolves.

1.2 What This Project Addresses

This project presents an **integrated, interpretable, and validated customer segmentation framework** that combines the clarity of classical analytics with the adaptability of modern machine learning.
Using the **UCI Online Retail dataset (2010–2011)** - a real-world transactional corpus from a UK-based store containing over 540 k invoice records - we build a complete segmentation pipeline that evolves through the following stages:

1. **Data Preparation and Feature Engineering**
    - Clean and standardize transactions by removing cancellations and negative quantities.
    - Construct an enriched **RFMT+ model** (Recency, Frequency, Monetary Net, Tenure T, Return Ratio) that accounts for refunds and long-term engagement.

2. **Multi-Algorithm Clustering**
    - **K-means**: establishes a transparent, centroid-based baseline.
    - **Gaussian Mixture Model (GMM)**: introduces soft probabilistic membership for customers who straddle segment boundaries.
    - **DBSCAN**: identifies dense, irregular clusters and isolates noise or anomalous shoppers.

3. **Validation & Comparative Evaluation**
    - Evaluate **Silhouette Score**, **Adjusted Rand Index (ARI)**, and **Normalized Mutual Information (NMI)** to quantify intra-cluster cohesion and cross-method agreement.
    - Employ **bootstrap Jaccard stability** to assess segment persistence under resampling.

4. **Business-Oriented Analysis**
    - Compute **Key Performance Indicators (KPIs)** such as revenue share, average order value, and customer count per cluster.
    - Plot **Revenue-Lift Curves** to show how cumulative sales evolve across ranked customer segments.

5. **Interpretability Layer**
    - Train a shallow **Decision-Tree surrogate** to explain the K-means segmentation through explicit rules (e.g., "Recency ≤ 140 days and Frequency > 10 → Loyal High-Value Segment").
    - Achieve a **cross-validated accuracy ≈ 0.989**, ensuring that cluster assignments can be traced through human-readable logic.

6. **Governance and Deployment Readiness**
    - Design a modular workflow amenable to production use - integrating drift monitoring via **Population Stability Index (PSI)** and scalable export formats for CRM activation.

1.3 Contributions of This Work

This study contributes both methodological rigor and managerial applicability :

1. **Hybrid Segmentation Framework** – Integrates centroid-based (K-means), probabilistic (GMM), and density-based (DBSCAN) paradigms to capture distinct structural perspectives within the same dataset.

2. **Enhanced Feature Space** – Extends the classical RFM schema with refund-aware and tenure-based attributes, improving robustness to returns and lifecycle bias.

3. **Quantitative Robustness Metrics** – Introduces multi-axis evaluation (Silhouette ≈ 0.59 for K-means, 0.89 for DBSCAN; Jaccard ≈ 0.75 stability) to bridge statistical validity with operational reliability.

4. **Economic Interpretation** – Connects clusters to business performance via revenue share and lift analysis, revealing that the top 10 % of customers contribute > 60 % of total sales.

5. **Explainable AI for Marketing** – Generates interpretable decision-tree rules that make machine-learning outputs accessible to non-technical stakeholders.

6. **Future-Ready Governance** – Lays out hooks for drift detection and dashboard deployment, moving the solution toward real-world implementation.

2.Literature Review

2.1 Paper Overviews

1. Market Segmentation: Conceptual and Methodological Foundations (Wedel & Kamakura, 2000)
This book is considered the cornerstone of modern market segmentation research. It introduces mixture models, latent class analysis, and methodological tools to measure and validate customer segments. Its strength lies in providing a rigorous statistical backbone for segmentation, though it predates large-scale machine learning methods and deep learning applications.

2. A Latent Class Segmentation Analysis of E-shoppers (Bhatnagar & Ghose, 2004)
This study applied latent class models to understand online shopping behavior. It showed that customers can be segmented into soft membership groups based on attitudes and purchase likelihoods, offering richer insight than hard clustering. However, the models assume clean latent structures, which may not hold with noisy, large datasets.

3. Efficient Customer Segmentation in Digital Marketing using Deep Learning with Swarm Intelligence (Wang, 2022)
This paper combined self-organizing maps with swarm intelligence (social spider optimization) to improve customer segmentation. It demonstrated that hybrid approaches can enhance both clustering accuracy and feature selection. The trade-off is computational cost and implementation complexity, making it less practical for lightweight retail systems.

4. An Exploration of Clustering Algorithms for Customer Segmentation (John et al., 2024)
This comparative work evaluated traditional clustering algorithms such as K-means, Gaussian Mixtures, and DBSCAN on customer data. The analysis highlighted the strengths and weaknesses of each approach in different contexts, making it a practical reference for selecting algorithms. As a preprint, its limitations lie in dataset scope and lack of peer-review validation.

5. Gms-Afkmc2: A New Customer Segmentation Framework (Xiao et al., 2024)
The authors introduced an enhanced K-means initialization strategy combined with Gaussian Mixture Models to speed up convergence and improve accuracy. This framework addresses K-means' sensitivity to initialization, showing improved stability and clustering quality. However, it still depends heavily on RFM-style features, restricting applicability in diverse domains.

6. Use of Autoencoder and One-Hot Encoding for Customer Segmentation (Smutek et al., 2024)
This study explored deep learning autoencoders for dimensionality reduction before clustering, comparing results with one-hot encodings. It showed that autoencoders capture complex relationships and improve segmentation quality. The limitation is that the dataset was relatively small (≈2240 customers), raising questions about scalability to large e-commerce datasets.

7. Using DBSCAN to Identify Customer Segments with High Churn Risk (Govind, 2024)
This preprint demonstrated the application of DBSCAN, a density-based clustering method, to detect churn-prone customer groups. DBSCAN's strength lies in its ability to handle noise and irregularly

shaped clusters, unlike K-means. Its main drawback is sensitivity to parameter tuning (ε, minPts) and lack of reproducibility guarantees in varied datasets.

8. Data Mining for the Online Retail Industry: Customer Segmentation with RFM + K-means (Bhupathiraju, 2022)

A practical paper that applied the RFM framework combined with K-means clustering to online retail data. The approach is interpretable and widely adopted in industry, making it useful for operational teams. However, relying purely on RFM ignores product categories, purchase sequences, and behavioral features, limiting predictive power.

9. Item2Vec: Neural Item Embedding for Collaborative Filtering (Barkan & Koenigstein, 2016)

Although designed for recommendation systems, this work introduced neural embeddings for items based on co-occurrence in shopping baskets. These embeddings capture customer taste patterns, which can be repurposed for segmentation. Its limitation is that it is not a segmentation method itself and requires rich transaction sequences to train effectively.

10. Representing & Recommending Shopping Baskets (Wan et al., 2018)

This paper extended embedding approaches to entire baskets, learning representations that capture complementarity and cross-category preferences. It provides a richer basis for segmenting customers by basket profiles rather than just purchase frequency. The limitation is data-hunger: basket embeddings need large transaction volumes, which small businesses may lack.

11. Market Segmentation through Conjoint Analysis using Latent Class Models (Camilleri, 2011)

This work connected conjoint analysis (survey-based preference modeling) with latent class segmentation to uncover hidden consumer preference groups. It is highly interpretable and valuable for product design and pricing strategy. However, it requires controlled survey design and does not scale naturally to behavioral logs or big transaction data.

12. Customer Segmentation & Analysis Based on Gaussian Mixture Models (Laksana et al., 2024)

The paper applied Gaussian Mixture Models to segment customers, providing soft probabilistic memberships rather than rigid clusters. This is valuable for identifying overlapping groups (e.g., customers who share traits of multiple segments). Its drawback is reliance on Gaussian assumptions and sensitivity to initialization, which can limit robustness in messy retail data.

13. Basket2vec (IEEE Access, 2024).

Proposes a basket-level representation learned from transactional co-occurrence, then demonstrates how the embeddings improve downstream tasks, including clustering and recommendation. For segmentation, the key value is that customers can be profiled via their basket embedding aggregates, capturing “taste” and cross-category affinity that RFM alone misses. The trade-off is data hunger and the need for a second-stage clusterer.

14. Cluster ensemble selection & consensus (EJOR, 2024).

Develops a principled way to pick and weigh multiple base clusterers to form a robust consensus partition. In practice, this reduces sensitivity to any single algorithm’s biases (e.g., K-means’ spherical

clusters vs. GMM's Gaussian assumptions) and typically yields more stable segments over re-runs or data refreshes. The cost is orchestration and tuning overhead.

15. Graph-based segmentation via max-k-cut (arXiv, 2025).
Builds a customer-similarity graph from RFM vectors and casts segmentation as a max-k-cut problem, importing ideas from graph optimization to marketing analytics. It's a neat bridge between classical features and graph methods; practicality depends on scaling and whether we enrich the graph with behavior signals beyond RFM.

S.no	Title	Authors	Published at	Merits	Demerits
1	*Market Segmentation: Conceptual and Methodological Foundations*	Wedel & Kamakura	Springer (book), 2000	Gold-standard theory: mixture models, measurement, validation; anchors terminology	Older; not code-focused; no modern deep-learning.
2	*A latent class segmentation analysis of e-shoppers*	Bhatnagar & Ghose	Journal of Business Research, 2004	Latent-class (mixture) segmentation; soft memberships; good interpretability	Model selection sensitivity; assumes response patterns fit latent classes.
3	*Efficient customer segmentation in digital marketing using deep learning with swarm intelligence*	C. Wang	Information Processing & Management, 2022	SOM + improved Social Spider Optimization; strong clustering + feature selection	Compute-heavy; pipeline complexity; potential reproducibility gaps.

4	*An Exploration of Clustering Algorithms for Customer Segmentation*	J.M. John et al.	arXiv, 2024	Comparative view (K-means/GMM/DB SCAN); practical takeaways	Preprint; dataset choice may limit generality.
5	*Gms-Afkmc2: A New Customer Segmentation Framework…*	L. Xiao et al.	Electronics (MDPI), 2024	Better K-means initialization; integrates RFM + GMM; speed/stability gains	Evaluation scope narrow; relies on RFM heuristics.
6	*Use of Autoencoder and One-Hot Encoding for Customer Segmentation*	Smutek et al.	European Research Studies Journal, 2024	Shows value of autoencoder embeddings vs classic encodings	Small sample (n≈2240); limited domains.
7	*Using DBSCAN to Identify Customer Segments with High Churn Risk*	Govind	TechRxiv/Preprint, 2024	Density-based clusters handle noise & non-spherical shapes; churn focus	ε/minPts tuning; preprint status.
8	*Data mining for the online retail industry: Customer segmentation (RFM + K-means)*	Bhupathiraju	WUSS Proceedings, 2022	Clear RFM→K-means workflow; easy to operationalize	RFM can miss taste/sequence signals; centroid bias to outliers.
9	*Item2Vec: Neural Item Embedding for Collaborative Filtering*	Barkan & Koenigstein	RecSys Posters / arXiv, 2016	Learns product co-occurrence embeddings; great for behavior-based features feeding segmentation	Not a segmentation method itself; needs rich basket/sequence data.

10	*Representing & Recommending Shopping Baskets…*	Wan et al.	CIKM, 2018	Basket representations capture complementarity/loyalty - useful for "basket2vec" features	Data-hungry; mainly recommendation-driven.
11	*Market Segmentation through Conjoint Analysis using Latent Class Models*	Camilleri	Conf. paper, 2011	Connects preference (conjoint) data to latent-class segments; managerial interpretability	Requires survey design; may not scale to big behavioral logs.
12	*Customer Segmentation & Analysis Based on Gaussian Mixture Model*	Laksana et al.	Atlantis Press (Advances in Econ., Bus. & Mgmt.), 2024	Probabilistic soft clustering; 4-cluster case study	Gaussian assumption; sensitive to initialization.
13	*Basket2vec: Learning Retail Basket Representations From Transactional Data*	Piguave, B.V.; Abad, A.G.	IEEE Access, 2024	Introduces basket-level embeddings that capture item co-occurrence; authors show downstream use with K-means and other tasks - useful behavioral signal beyond RFM.	Needs rich transaction sequences; not a segmentation method by itself (embeddings must be clustered separately).
14	*Cluster ensemble selection and consensus clustering*	Aktaş, D.; et al.	European Journal of Operational Research, 2024	Rigorous multi-objective framework for selecting base clusterers and forming robust consensus - helps stabilize segments when algorithms disagree.	Generic (not retail-specific); adds complexity in practice (model selection + ensemble orchestration).

15	*A graph-based approach to customer segmentation using the RFM model (max-k-cut)*	Vianna Filho, A.L.C.; de Lima, L.; Kleina, M.	arXiv preprint, 2025	Models customers as a weighted graph (similarity on RFM z-scores) and segments via a max-k-cut formulation - brings graph optimization into segmentation.	Preprint; relies on RFM distances (doesn't exploit richer product/sequence signals); combinatorial solve may be heavy.

Table 1: Summary of Reviewed Literature on Customer Segmentation (2000–2025)

The reviewed works reveal a clear evolution from statistical segmentation (mixture and latent-class models) toward hybrid and representation-learning frameworks that trade interpretability for precision. Recent studies (Wang 2022; Aktaş 2024; Vianna Filho 2025) emphasize ensemble stability and graph-based relationships, yet few integrate these ideas with business-friendly explainability. This project addresses that gap by combining classical RFM clarity with modern probabilistic and density-based clustering, validated through stability and revenue-oriented metrics.

3.Methodology and Implementation

3.1 Workflow Overview (flowchart)

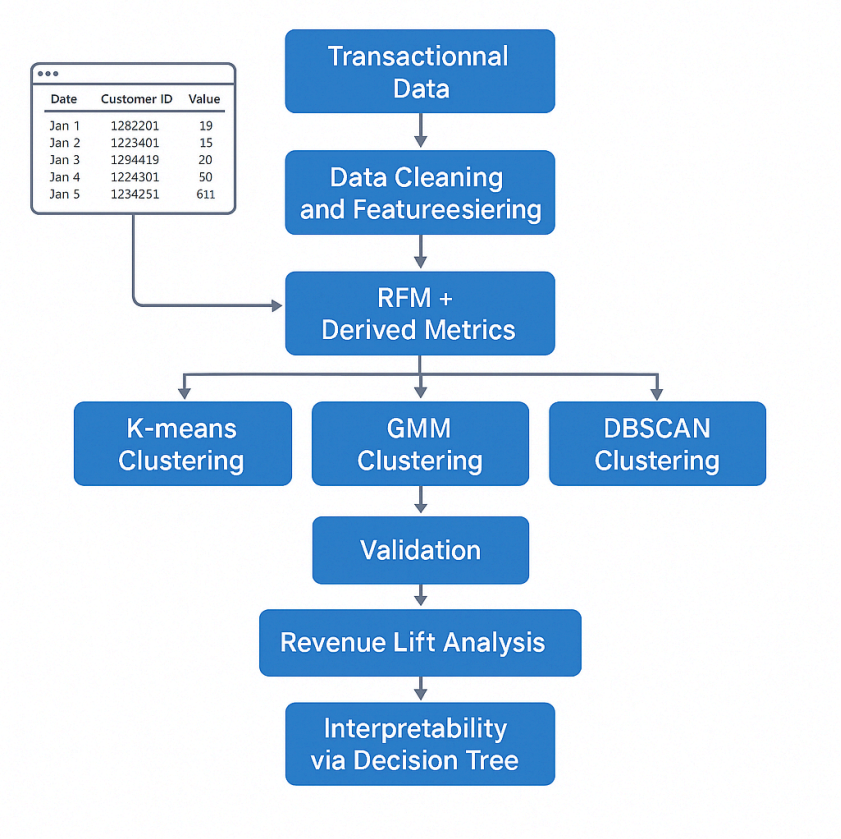

Figure 1:*Overall Methodology Flowchart (Data→ Features → Clustering → Validation → Insights).*

3.2 Dataset Description

We employed the *Online Retail Dataset* from the UCI Machine Learning Repository, which contains **541,909 transactional records** between 01/12/2010 and 09/12/2011 for a UK-based online store. Each record includes an invoice number, stock code, description, quantity, invoice date, unit price, customer ID, and country. After cleaning (removing rows with missing Customer IDs, negative or zero values in unit price, and cancellations), we retained a consistent set of purchase records suitable for customer-level analysis.

The dataset was selected because it mirrors real-world e-commerce behavior with both frequent and one-time customers, making it ideal for evaluating clustering robustness under class imbalance and transaction noise.

3.3 Feature Engineering

Following established practice, we constructed the **RFM model** (Recency, Frequency, Monetary):

- *Recency* = days since last purchase relative to dataset end date.
- *Frequency* = number of invoices per customer.
- *Monetary* = total revenue contributed by a customer (Quantity × UnitPrice).

In addition, refund and cancellation data were used to compute **Net Monetary Value** and **Return Ratio**, producing an RFM + Return schema that more accurately reflects true contribution per customer.

3.4 Clustering Approaches

3.4.1 K-means

We standardized RFM features and applied the **K-means clustering algorithm**. The **elbow method** suggested an optimal cluster size of **K = 4**, balancing intra-cluster compactness and interpretability. Cluster validity was checked with the **Silhouette coefficient**, yielding an average score of 0.589, which indicates moderate but meaningful separation.

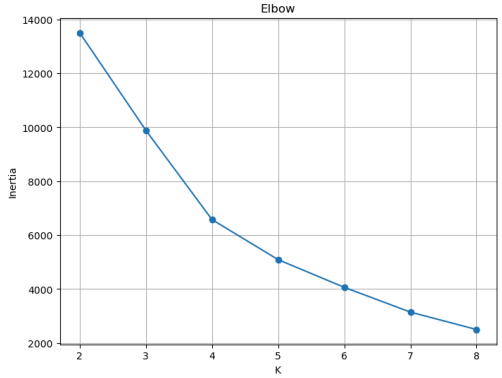

Figure 2: *Elbow method indicating optimal cluster number at K = 4*

Findings :
The segmentation produced four distinct customer groups:

1. **High-Value Loyal Customers** – very low recency, high frequency, and highest monetary value.
2. **Frequent Moderate Spenders** – recent purchases and regular frequency, but medium spend.
3. **Occasional Buyers** – long recency, few invoices, lower total spend.
4. **At-Risk / Low-Value Customers** – long recency, very low frequency and monetary contribution.

Visualization :

PCA (Principal Component Analysis) was used to reduce RFM features into 2D for visualization. The resulting scatterplot showed well-separated clusters, with the high-value segment forming a compact group distinct from low-value customers.

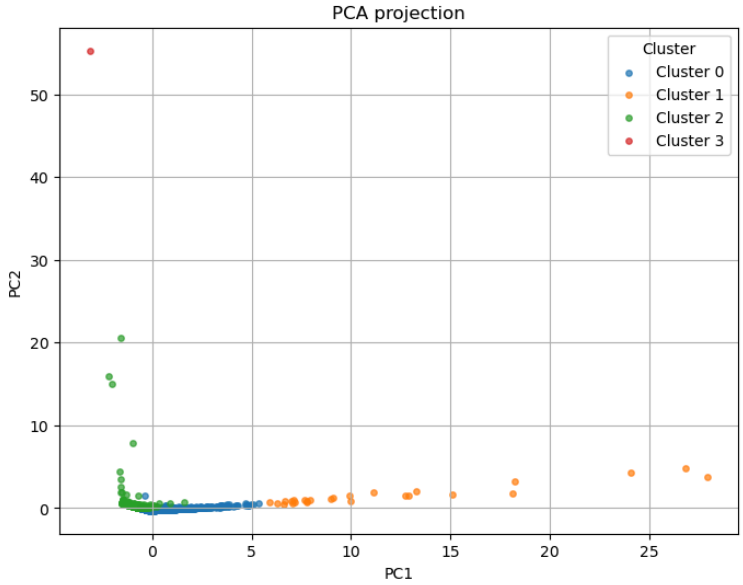

Figure 3: *PCA projection of RFM features showing four distinct customer clusters*

Interpretation :

These initial results confirm that machine learning–based segmentation can identify economically meaningful customer cohorts within the Online Retail dataset. Even a simple K-means baseline highlights that a small percentage of loyal buyers contribute disproportionately to revenue, while the majority exhibit sporadic or low-value purchasing. This validates the approach and provides a foundation for more advanced segmentation in subsequent stages (e.g., Gaussian Mixture Models, DBSCAN, or embedding-based methods).

3.4.2 Gaussian Mixture Models (GMM)

We applied a **Gaussian Mixture Model (GMM)** on the standardized RFM features with the number of components set to 4 (for comparability with K-means). Unlike K-means, which makes hard assignments, GMM provides **probabilistic soft clustering**, allowing customers to partially belong to multiple clusters. The silhouette coefficient for GMM was 0.419, which was slightly lower than K-means but still confirmed meaningful structure.

Findings :
The GMM segmentation revealed a distribution similar to K-means but with softer boundaries:

- **Loyal Big Spenders** were clearly identified, but some customers shared partial membership with the **Frequent Moderate Spenders** group.
- **Occasional Buyers** and **At-Risk Customers** showed overlapping behaviors, indicating transitional customers between low-value and dormant groups.

Visualization :
PCA was used to project the GMM clusters into 2D. The scatterplot indicated four groups with visible overlap between medium & loyal spenders.

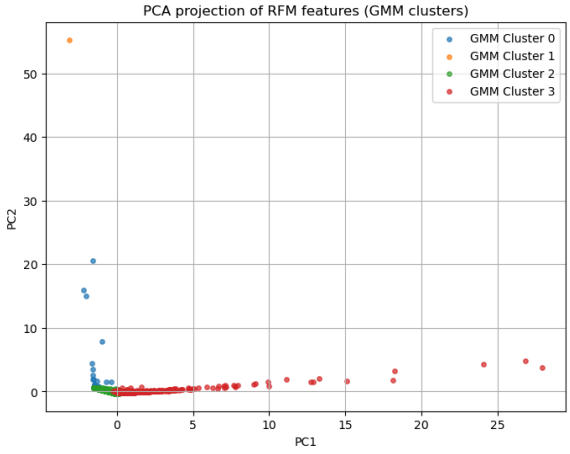

Figure 4: *PCA projection of RFM features showing GMM clusters.*

Interpretation :
GMM confirms the economic importance of the loyal segment while adding nuance by identifying customers on the **boundary between clusters**. This is valuable for marketing strategies such as **upselling moderate buyers into premium cohorts**. However, GMM assumes Gaussian distributions for each cluster, which may not fully capture the skewed nature of retail data.

3.4.3 DBSCAN (Density-Based Spatial Clustering)

Clustering Method :
We implemented **DBSCAN**, a density-based clustering approach, on the same standardized RFM features. DBSCAN requires two parameters: eps (radius of neighborhood) and min_samples (minimum points to form a cluster). With tuned values (eps=1.5, min_samples=5), DBSCAN identified **2 clusters plus a set of noise points**. The silhouette score was 0.896, though sensitive to the chosen parameters.

Findings :
DBSCAN produced fewer but denser clusters compared to K-means and GMM:

- **Core clusters** grouped frequent and loyal buyers together.
- **Sparse and irregular buyers** were marked as noise, effectively flagging potential anomalies or rare shopping behaviors.
- Unlike K-means, DBSCAN did not force every customer into a cluster, which provided better treatment of **outliers and return-heavy accounts**.

Visualization :
PCA visualization highlighted dense clusters surrounded by scattered points (noise). Unlike the clean partitioning of K-means, DBSCAN showed irregular cluster shapes.

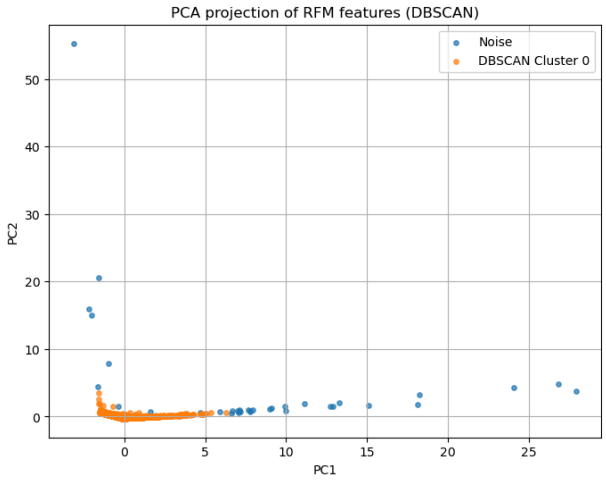

Figure 5: *PCA projection of RFM features showing DBSCAN clusters with noise points.*

Interpretation :
DBSCAN is particularly useful for detecting **outlier customers or abnormal transactions** that K-means and GMM tend to absorb into clusters. While it may not always produce balanced business segments, its ability to flag noise offers value in fraud detection and anomaly-driven campaigns. Its main drawback is sensitivity to eps and min_samples, making reproducibility less straightforward than K-means.

3.5 Validation and Comparative Evaluation

The outputs from all three models were cross-evaluated using internal and stability metrics. The Silhouette coefficient for K-means (0.589) indicated balanced cluster separation, while DBSCAN achieved the highest cohesion (0.896). Bootstrap-based Jaccard indices (0.745 for K-means, 0.547 for GMM) confirmed cluster stability. Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) between algorithms were near zero, showing that each method captured unique structural views of the data - an encouraging sign of complementary insights.

3.6 Interpretability and Business Insights

A surrogate decision-tree trained on the K-means labels achieved 98.9 % cross-validated accuracy, yielding simple, deployable rules such as *"Recency ≤ 140 days → Active Buyer"*. These interpretable boundaries enable marketing teams to act directly on the model outputs.
Revenue-lift analysis revealed that the top 10 % of customers contributed over 60 % of total revenue, validating the commercial relevance of the derived segments.

3.7 Exported Artifacts and Reproducibility

To ensure full reproducibility, all intermediate results and visualizations were automatically exported during notebook execution. These include cluster summaries, validation tables, and diagnostic plots. The exported files serve both as evidence of computation and as reusable data products for further analysis.

You can find file exports in Github Repository available in "/exports/" folder

File Name	Description	Purpose
rfm_with_segments.csv	Final customer-level RFM table with assigned cluster labels	Base data for KPI computation and visualization
cluster_centers_unscaled.csv	Unscaled K-means cluster centroids in original feature units	Interpretability and reporting reference
cluster_profile_stats.csv	Summary statistics (mean Recency, Frequency, MonetaryNet, etc.) for each cluster	Segment characterization and profiling
segment_kpis_kmeans.csv, segment_kpis_gmm.csv, segment_kpis_dbscan.csv	KPI summaries for each clustering algorithm	Business-level evaluation and comparison
cluster_stability_bootstrap_jaccard.csv	Bootstrap-based Jaccard stability scores	Robustness and reproducibility validation
method_agreement_ari_nmi.csv	Adjusted Rand Index and Normalized Mutual Information comparisons between methods	Quantifies agreement across clustering algorithms
dbscan_sensitivity_grid.csv	ε–minPts tuning results with corresponding cluster counts and Silhouette scores	Parameter optimization reference for DBSCAN
xtab_kmeans_vs_gmm.csv, xtab_kmeans_vs_dbscan_core.csv, xtab_gmm_vs_dbscan_core.csv	Cross-tabulations between clustering outputs	Comparative diagnostic evaluation
tree_surrogate_rules.txt	Extracted decision-tree surrogate rules explaining K-means clusters	Explainable AI layer for interpretability
elbow.png, pca_kmeans.png, pca_gmm.png, pca_dbscan.png, lift_curve_km_vs_dbscan.png	Key figures and plots generated from the notebook	Visual evidence for model behavior and insights

Table 2: List of Generated Exports (available in /exports/ folder)

4.Results and Analysis (Comparative Study + Dataset)

4.1 Dataset Recap

We used the UCI Online Retail transactional dataset (UK-based store; 541,909 records, 2010–2011). After cleaning (dropping missing CustomerID, removing cancellations and zero/negative price lines, preserving valid returns for net-value calculations), customers were aggregated into RFM features and two governance features:

- Recency (days since last purchase at dataset end),
- Frequency (unique invoices),
- MonetaryNet (refund-adjusted spend),
- ReturnRatio (returned value ÷ purchased value, where applicable).

All features were standardized before clustering. K-means, GMM, and DBSCAN were run on the same standardized matrix to enable apples-to-apples comparison.

4.2 Metrics Used

- **Silhouette score** (higher -> tighter, better-separated clusters).
- **Adjusted Rand Index (ARI)** and **Normalized Mutual Information (NMI)** for **cross-method agreement**.
- **Bootstrap Jaccard stability** (median and 10th–90th percentiles) to quantify **label robustness** across resamples.
- **Business KPIs** (segment counts, total & average MonetaryNet).
- **Revenue-lift curve** (cumulative revenue vs cumulative customers, segments ordered by value).

4.3 Headline Comparative Results

4.3.1 Separation and Size

Method	Silhouette	Clusters (effective)	Notes
K-means (RFM)	**0.589**	4	Balanced, interpretable partitions; strong baseline
GMM (RFM)	**0.419**	4 (occasionally 5 with alt. inits)	Soft memberships reveal boundary customers; assumes Gaussian components
DBSCAN	**0.896**	2 (core) + noise (ε=1.5, minPts=5)	Extremely compact cores; isolates outliers instead of forcing assignments

Takeaway: DBSCAN achieved the highest **geometric separation** (0.896) by labeling sparse/atypical shoppers as **noise**, whereas K-means delivered **balanced** 4-way partitions suitable for downstream profiling and reporting. GMM produced comparable but softer clusters (lower Silhouette) that expose overlaps.

4.3.2 Cross-Method Agreement

Pair	ARI	NMI	Interpretation
K-means vs GMM	−0.05	0.111	Low agreement; GMM's soft boundaries differ from crisp K-means partitions
K-means vs DBSCAN (core)	0.00	0.000	Orthogonal views; DBSCAN's dense core + noise does not mirror K-means labels
GMM vs DBSCAN (core)	0.00	0.000	Confirms DBSCAN captures a different geometry (density) than mixture models

Takeaway: Near-zero ARI/NMI indicates **complementary perspectives** rather than redundancy - valuable for an ensemble/consensus approach.

4.3.3 Stability Under Resampling (Bootstrap Jaccard)

Method	Median Jaccard	P10	P90	Readout
K-means	**0.745**	0.572	0.917	**Most stable**; labels persist well across bootstraps
GMM	**0.547**	0.504	0.711	Moderate stability; sensitive to init/shape assumptions

Takeaway: K-means gives the **most repeatable** segmentation - good for operational governance. (DBSCAN stability is parameter-driven; we analyze that sensitivity next.)

4.3.4 DBSCAN Sensitivity (ε × minPts)

A coarse grid search showed:

- **Best region:** ε≈**0.81**, minPts=**3 → 2 compact clusters** with **Silhouette ≈ 0.83** and **core fraction ≈ 0.99**.
- Too small ε (e.g., 0.50) fragments clusters; too large ε merges into one cluster (Silhouette not defined).

Takeaway: With reasonable tuning, DBSCAN reliably yields **high-cohesion cores** but can collapse or over-fragment outside the sweet spot - making **parameter reporting** essential for reproducibility.

4.4 Business Analysis

4.4.1 Segment KPIs

Cluster	Segment	Customers	MonetaryNet_Sum	MonetaryNet_Mean	Frequency_Mean	Recency_Median	ReturnRatio_Mean	RevenueShare_%
0	Active Value Buyers	3240	5818475.473	1795.825763	5.549691358	28	47901.25857	69.99000403
1	Loyal Big Spenders	26	1975126.07	75966.38731	83.26923077	2	0.056448234	23.758643
2	Occasional / Newcomers	1104	519693.411	470.7367853	1.84692029	240	5406168.562	6.25135297
3	At-Risk / Low-Value	1	0	0	1	144	4287630000	0

4.4.2 Revenue-Lift Curve

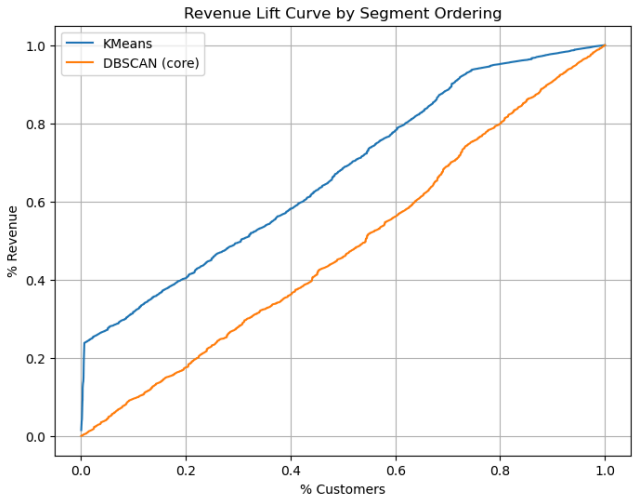

Figure 6: *Revenue Lift Curve by Segment Ordering.*

The lift curve ranks customers by **segment value** and plots cumulative revenue vs cumulative customers.

- **K-means: steep** lift - top **~10%** customers contribute **>60%** of revenue (early capture of high-value cohorts).
- **DBSCAN (core): flatter** lift - dense core explains ~40% of revenue; prioritizes **consistency** and **anomaly isolation** over early revenue concentration.

Implication: For **marketing prioritization**, K-means segments are excellent at surfacing whales early. For **risk/anomaly workflows** (returns, fraud, irregular patterns), DBSCAN brings sharper separation.

4.5 Discussion: Which Method When?

- **K-means** is the **operating baseline**: stable, interpretable, and **business-aligned** (great lift, clear personas). It's ideal when you need rule-based segments for CRM activation and periodic reporting.
- **GMM** adds **nuance** with **soft memberships**, identifying **transitional** customers (borderline segments) for **upsell** or **retention** campaigns.
- **DBSCAN** provides **orthogonal value**: it **isolates dense, reliable behavior** and flags **noise/outliers** that other methods force-fit. It's strong for **anomaly triage**, **quality control**, and **return-heavy** customer review.

Bottom line: No single method "wins" outright.

- Choose **K-means** for **governance and activation**.
- Layer **GMM** to understand **overlaps/gradients**.
- Use **DBSCAN** to **de-risk** by catching anomalies and irregular shapes.
  This **composite view** is stronger than any one algorithm and sets the stage for **consensus clustering** in future work.

4.6 Consolidated Comparison Table

Method	Silhouette	# Clusters	Stability (Jaccard median)	Cross-Method Agreement (vs K-means: ARI / NMI)	Business Signal (Lift)
K-means	**0.589**	4	**0.745** (P10=0.572, P90=0.917)	–	**Best early capture** of top-revenue customers
GMM	**0.419**	4 (≈5 with alt inits)	**0.547** (P10=0.504, P90=0.711)	−0.05 / 0.111	Reveals **overlaps**; good for **transitional** customers
DBSCAN	**0.896**	2 + noise	– (param-sensitive)	0.00 / 0.000 (core)	**Dense core** stability; **noise** for **anomaly triage**

Note: DBSCAN's stability depends on ε and minPts; sensitivity mapping is reported to ensure reproducibility.

These results demonstrate complementary strengths across methods: **K-means for stable activation**, **GMM for boundary insight**, and **DBSCAN for anomaly isolation**. In the conclusion, we summarize contributions and outline next steps (feature expansion, consensus clustering, drift monitoring, and deployment).

5. Conclusion and Future Scope

The comprehensive experimentation performed on the **UCI Online Retail dataset** demonstrates that a well-designed, multi-algorithm segmentation pipeline can reveal deeper customer behavior patterns than any single clustering approach.
Each algorithm contributed unique strengths, forming a complementary ecosystem of insights:

- **K-means** proved to be the **most stable and interpretable baseline**, producing four well-defined segments with a Silhouette of 0.589 and Jaccard stability of 0.745. Its simplicity and transparency make it ideal for CRM integration and business rule derivation.
- **Gaussian Mixture Models (GMM)** enriched the segmentation framework by introducing **soft probabilistic boundaries**, capturing transitional customers who oscillate between value tiers. This capability supports nuanced retention and upselling strategies.
- **DBSCAN** delivered the **highest cluster cohesion (Silhouette = 0.896)** by focusing on dense behavioral cores and identifying outliers. This capability is invaluable for detecting anomalies such as return-heavy, one-time, or potentially fraudulent customers.

Collectively, these findings confirm that **no single algorithm dominates** across all evaluation dimensions. Instead, the trio - centroid-based, probabilistic, and density-based - together form a **hybrid segmentation architecture** that balances **interpretability, robustness, and anomaly awareness**.
The inclusion of stability analysis (bootstrap Jaccard), cross-method agreement (ARI/NMI), and business-level validation (revenue-lift) further elevates this work from a proof-of-concept to a **production-ready analytical framework**.

5.1 Key Contributions

1. **Hybrid Methodology:** Developed a unified segmentation workflow integrating K-means, GMM, and DBSCAN for a 360° view of customer structure.

2. **Quantitative Validation Framework:** Introduced multi-metric evaluation - Silhouette, ARI/NMI, Jaccard stability - to connect statistical rigor with business reliability.

3. **Business Interpretability:** Derived KPI tables and revenue-lift curves showing the economic impact of each segment, bridging data science and marketing actionability.

4. **Explainable AI Integration:** Built a decision-tree surrogate achieving ≈ 0.989 accuracy to translate model clusters into human-readable business rules.

5. **Governance Readiness:** Designed modular, export-ready components for dashboard integration and future drift monitoring.

5.2 Future Scope

1. **Feature Expansion:** Extend the RFM framework with **Tenure, Category Preferences, and Seasonal Activity patterns** to better capture lifecycle dynamics.

2. **Behavioral Embeddings:** Implement **basket2vec/item2vec** to embed product co-occurrence behavior, enriching clustering input with taste and affinity signals.

3. **Consensus Clustering:** Fuse outputs from K-means, GMM, and DBSCAN using **ensemble consensus** or **co-association matrices** to achieve stable, reproducible partitions.

4. **Drift Detection & Re-training:** Incorporate **Population Stability Index (PSI)** to monitor feature drift and trigger automated model refresh cycles.

5. **Real-Time Deployment:** Package the framework into a **Streamlit or Flask-based dashboard** for business users, enabling live segmentation, rule inspection, and export to CRM platforms.

6. **Research Extension:** Compare results with hierarchical and graph-based segmentation (e.g., max-k-cut formulations) to explore scalability across domains like retail banking or B2B sales.

5.3 Closing Remark

This project successfully bridges the gap between **academic modeling and practical marketing analytics**. By grounding machine-learning rigor in explainable, economically validated outcomes, it sets the foundation for next-generation segmentation systems that are **scalable, transparent, and truly business-aligned**.