Customer Segmentation with RFM & Unsupervised Learning

Team: Simeon Musa Nyakeh Vibbi / Joe A.D Allie

Problem Statement (What problem am I trying to solve? And why does it matter?)

Many retailers and SMEs struggle to target customers effectively because they lack data-driven segmentation. This leads to generic campaigns, higher churn, and wasted spend. This project builds a transparent, reproducible customer segmentation pipeline that groups customers by purchasing behaviour to power personalized marketing, retention, and LTV growth.

Existing Solutions

- Marketing suites (CRM/automation tools) offer segmentation but are black-box, subscription-based, and often not tailored to local contexts or smaller datasets.
- Our approach is open and explainable (RFM), includes validation (Silhouette, Elbow, stability checks), and outputs actionable playbooks per segment. All code is reproducible; an optional lightweight app lets non-technical users explore segments.

Objectives (What do I want to achieve with this project?)

- Ingest & clean a recent transactions dataset (Retail Sales Dataset on Kaggle).
- Engineer RFM features (Recency, Frequency, Monetary) and optional enrichments (tenure, AOV).
- Train and compare unsupervised models (K-Means, GMM; optional hierarchical).
- Select optimal k via Elbow + Silhouette score; profile segments.
- Produce segment narratives + recommended actions (business-facing).
- (Optional) Deploy a Streamlit app for interactive segmentation and CSV export.
- Deliver a readable README, figures, and final SEGMENTS.csv (customer_id, segment, R/F/M).

Dataset Information

Dataset: Retail Sales Dataset (Kaggle), maintained by Mohammad Talib.

Link: https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset

Columns present (mapped to this project):

- Customer ID → customer_id
- Transaction ID → order_id
- Date → order_date
- Quantity → quantity
- Price per Unit → unit_price
- Total Amount → amount (used when available)

Why this dataset: It provides recent, transaction-level retail data with the exact fields required for RFM, allowing fast, reliable segmentation and clear business interpretability. The structure aligns with our pipeline and supports deployment for non-technical users.

License & usage: Please refer to the Kaggle dataset page for licensing and usage terms; I will cite the source and respect its terms in our README.

Modeling Plan (What Models do I intend to use and why?)

- Feature Engineering: Recency (days since last purchase), Frequency (distinct orders), Monetary (total spend). Optional: Tenure (first→last purchase span), AOV, product diversity. Transformations: log1p(Monetary) + RobustScaler.
- Clustering Models: K-Means (fast baseline, n_init=20, k∈[2..8]); Gaussian Mixture Models (captures non-spherical clusters, soft probabilities); Optional: Agglomerative/Hierarchical for structure insights.
- Model Selection & Validation: Elbow (inertia) and Silhouette; check cluster sizes and stability. Business validation via segment interpretability (e.g., Champions, At-Risk, Hibernating). Outputs: segment profiles (means/medians of R/F/M + size %), charts, and recommended actions.

Deployment plan

Yes (optional). A Streamlit app where users upload CSV → compute RFM → select k → view segment sizes, profiles, and action tips; download SEGMENTS.csv with assigned labels. Hosted on Streamlit Community Cloud or Hugging Face Spaces.

Community Impact – What is the significance of the project to the community/real world?)

Provides a free, explainable playbook for SMEs to start personalized marketing without expensive tools; encourages evidence-based decisions (who to retain, re-engage, upsell); reusable template for local businesses, startups, and NGOs; builds local capacity in applied data science with clear documentation and an accessible UI.

Acknowledgement & References

- Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn, streamlit.
- Methodology: RFM framework; clustering validation via Silhouette/Elbow.
- Cohort: AI Saturdays Lagos – ML Flipped Cohort (mentorship & structure).
- Dataset: Retail Sales Dataset (Kaggle) — Mohammad Talib; include link and license in README.