Churn Analysis Report

1. Executive Summary

Scope

End-to-end churn analysis using the Customer Purchase & Churn Pipeline: Python ETL, feature engineering (RFM), baseline churn model (logistic regression), and a Power BI dashboard.

Top KPIs (from dashboard cards)

- **Total Revenue:** from transactions_enriched[revenue].
- **Churn Rate:** Churned Customers / Total Customers (DAX provided earlier).
- **Active Customers:** customers with recency_days ≤ 180.

Key trends & insights

- **Recency is the strongest churn signal.** Customers inactive for >180 days are disproportionately churned; recent purchasers show the lowest churn.
- **Low frequency drives risk.** First-time and infrequent buyers (≤2 orders lifetime) make up the bulk of churners.
- **Value concentration.** A small cohort of high-monetary customers (large baskets, frequent orders) contributes a large share of revenue and has the lowest churn.
- **Discounts have a threshold effect.** Moderate discounts can lift order count, but heavy discounting erodes average revenue per order and likely margins; impact varies by category.
- **Geography & income matter.** Churn rate differs by **region** and **income_band**; price-sensitive bands and a few regions show higher churn.

Actions at a glance

- Tighten **win-back** for 90–180 day inactive customers.
- Build **second-purchase nudges** (the biggest step-change in retention).
- Apply **targeted** (not blanket) discounts where elasticity is proven.

---

2. Data & Methodology

Sources & window

- **Customers, Products, Transactions** from the synthetic data in this project.
- Analysis window: **2022-01-01 to 2025-07-31** (transaction dates).

Synthetic data note

- Data is generated to mimic realistic retail patterns. Use results for portfolio demonstration and methodology; not for operational decisions.

ETL steps (Python)

1. Load raw CSVs → clean price/discount fields → recompute line **revenue**.
2. **Enrich** transactions with customer and product attributes.
3. Aggregate to customer level to create RFM and behavioural features:
    - **Recency (days)** since last purchase (snapshot: 2025-08-01)
    - **Frequency** (orders)
    - **Monetary** (total revenue) and **avg_basket**
    - **days_since_signup**
4. Output: transactions_enriched.csv, customer_features.csv.

Feature engineering (RFM & CLV proxy).

- **RFM** used for segmentation and as model inputs.
- Simple **CLV proxy** for dashboarding: monetary and avg_basket × expected_orders_next_period (expected orders derived from frequency segment).

Modelling approach.

- **Logistic Regression** (class_weight='balanced') in a pipeline with one-hot encoding for categorical features.
- Train/test split (80/20, stratified). Evaluate via **AUC**, precision/recall, confusion matrix.
- Goal: interpretability to rank churn drivers (recency, frequency, region/income).

---

3. Customer Behaviour Insights

RFM segments (rules you can show on the dashboard).

- **Champions:** recency_days ≤ 30 and frequency ≥ 6.
- **Loyal:** recency_days ≤ 90 and frequency 3–5.
- **Promising:** recency_days 31–120 and frequency 2–3.
- **At Risk:** recency_days 121–180 or frequency ≤ 2 with declining spend.
- **Lost/Churned:** recency_days > 180.

What we see.

- Revenue skews to **Champions/Loyal**; these groups have the **highest AOV** and stable purchase cadence.
- **Promising** customers are the best upgrade target (respond to bundles and reminders).
- **At Risk** customers show shrinking basket size and longer inter-purchase gaps before churn.
- **By Region/Income:** Certain regions and the <3L/3L-6L bands display **higher churn** and stronger **discount sensitivity**.

Repeat rate & AOV patterns.

- **Repeat rate** increases sharply after the **second purchase** (critical milestone).
- **AOV** (avg revenue per order) declines beyond ~15% discount; category-dependent.

---

4. Churn Drivers

Behavioural thresholds (operational).

- **Recency breakpoints:** 90 days (early risk), 180 days (churn threshold used in labels).
- **Frequency breakpoint:** ≤2 lifetime orders = high risk; ≥6 = low risk.
- **Monetary:** low total spends + high recency is a strong red flag.

Model highlights.

- **Top positive predictors of churn:** higher recency_days, very low frequency, low monetary.
- **Protective factors:** recent purchase, higher frequency, larger avg_basket.
- **Profile effects:** region and income_band terms carry signal; interact with discount response.

Limitations.

- Synthetic data and heuristic churn labelling (recency-based) may not capture contractual churn or service issues.
- No session/web behaviour, marketing exposure, or service ticket data; adding these can improve recall.

---

5. Recommendations

A. Outreach cadence (lifecycle).

- **0–30 days since last purchase:** cross-sell emails/WhatsApp with complementary items.
- **31–90 days:** reminder + value-adds (loyalty points, free shipping thresholds).
- **91–180 days: win-back**: limited-time moderate discount (e.g., 10%) or bundle; emphasize category they last purchased.
- **>180 days:** reactivation sequence (2–3 touches); if no response, suppress for 60–90 days to avoid fatigue.

B. Discount strategy (evidence-led).

- Avoid blanket 20%+ discounts.
- Use the **Discount Impact scatter** (X=discount bucket, Y=Avg order revenue, Size=order count, Legend=category) to:
  - Identify **elastic** categories (discounts lift orders without collapsing AOV).
  - Cap discounts for **inelastic** categories; rely on merchandising/UX and availability.

C. Second-purchase accelerator.

- Dedicated offers to **first-time buyers** within 14 days (free shipping, add-on coupons).
- Personalize based on last category and price point.

D. Cross-sell playbook.

- Use **category affinity** (e.g., Mobiles → Accessories/Audio) to recommend next items; run as email tiles and PDP widgets.

E. Data & model improvements (next steps).

- Capture **marketing touchpoints** (channels, spend, impressions) and **service tickets** to model causal drivers.
- Track **margin** (price – cost) to optimize **profit-focused** discounts, not just revenue.
- Add **model monitoring**: weekly AUC/precision, drift checks on recency/frequency distributions.
- Experiment with **tree-based models** (XGBoost/LightGBM) and **calibration** for better probability estimates.
- Build a **propensity-to-buy** and **uplift** model for truly targeted offers.