Retail Loyalty Lens: End-to-End Data Processing and Analytics Pipeline

A Comprehensive Report on Data Ingestion, Quality Validation, Loyalty Point Engineering, RFM Segmentation and CLV Calculation

Prepared by: Yash Choudhery

November 15, 2025

Abstract

This report presents a complete data engineering and analytics workflow for building a customer-centric loyalty analytics pipeline from raw retail transaction data. It describes the ingestion, cleaning, quality validation, enrichment, loyalty point computation, segmentation and Customer Lifetime Value (CLV) modelling required to support a modern retail loyalty ecosystem. The methodologies, mathematical formulas, threshold selection, and final analytics table designs are provided in detail to ensure reproducibility and operational readiness.

Contents

1 Introduction

Customer loyalty programs play a crucial role in increasing customer retention, driving repeat purchases and understanding consumer behaviour. Modern loyalty systems rely on robust data pipelines that transform raw transactional datasets into enriched analytical datasets.

This report outlines the transformation of a single raw retail table into a complete analytics-ready schema, including:

- Data ingestion and quality validation
- Cleaning and normalization
- Loyalty points calculation
- Customer segmentation using RFM
- Customer Lifetime Value (CLV) modelling
- Analytical table creation (customer, product, loyalty and transaction tables)

The objective is to create a reusable, scalable, and industry-standard pipeline for loyalty analytics.

2 Input Dataset Description

The raw retail dataset contains the following columns:

`['id', 'Transaction_ID', 'Customer_ID', 'Name', 'Email', 'Phone', 'Address', 'City', 'State', 'Zipcode', 'Country', 'Age', 'Gender', 'Income', 'Customer_Segment', 'Date', 'Year', 'Month', 'Time', 'Total_Purchases', 'Amount', 'Total_Amount', 'Product_Category', 'Product_Brand', 'Product_Type', 'Feedback', 'Shipping_Method', 'Payment_Method', 'Order_Status', 'Ratings', 'products']`

This dataset contains both customer-level attributes and transaction-level attributes. It is used to derive multiple tables in the final schema.

3 Data Ingestion and Quality Validation

3.1 Loading Strategy

The data is loaded using `pandas.read_csv()`. To ensure robustness:

- Multiple encodings are attempted (UTF-8, Latin-1).
- Columns are normalized by stripping whitespace and enforcing lower-case naming conventions.

3.2 Date and Time Parsing

Date and time fields are combined into a unified timestamp:

$$\text{timestamp} = \text{pd.to\_datetime}(\text{Date} + \quad \text{“ ”} + \text{Time})$$

Missing or unparsable timestamps are logged in a separate quality report.

3.3 Amount Normalization

Amounts may include currency symbols, commas, or non-numeric characters. Cleaned using:

$$\text{clean\_amount} = \text{regex\_replace}(\text{Amount}, \text{“}[^0-9.-]\text{”}, \text{“”})\text{regex\_replace}(\text{Amount}, \text{“}[^0-9.-]\text{”}, \text{“”})_{\text{regex\_replace(Amoun}}$$

Values failing numeric conversion are marked as unparsable and exported for review.

3.4 Fake Null Replacement

The following common representations are mapped to `NA`:

$$\{\text{“}, \text{“}NA\text{”}, \text{“}N/A\text{”}, \text{“}-\text{”}, \text{“}null\text{”}, \text{“}None\text{”}\} \rightarrow \text{pd.NA}$$

3.5 Duplicate Handling

Transaction duplicates are removed using:

$$\text{unique\_key} = \{Transaction\_ID\}$$

If missing, a composite key is used:

$$(Customer\_ID,\ Date,\ Amount,\ Product\_Type)$$

4 Feature Engineering

Feature engineering converts raw transaction rows into analytical insights.

4.1 Loyalty Points Calculation

The loyalty points system is designed to be intuitive yet incentivise larger purchases.

4.1.1 Simple Rule

$$\text{points} = \left\lfloor \frac{\text{amount}}{100} \right\rfloor$$

This ensures 1 point per 100 spent.

4.1.2 Tiered Rule

For more differentiated rewards:

$$\text{points} = \begin{cases} \left\lfloor \frac{\text{amount}}{100} \right\rfloor & \text{if amount } < 500, \\ \left\lfloor 1.5 \times \frac{\text{amount}}{100} \right\rfloor & 500 \leq \text{amount} < 2000, \\ \left\lfloor 2 \times \frac{\text{amount}}{100} \right\rfloor & \text{if amount } \geq 2000 \end{cases}$$

Business Rationale

- 500 and 2000 correspond to natural spending clusters.
- Higher tiers reward high-value shoppers and encourage larger cart sizes.

4.1.3 Bonus Points

Bonus points are applied for:

- Special promotions
- High-value categories (e.g., electronics)
- Customer birthdays or anniversaries

4.2 RFM Segmentation

RFM (Recency, Frequency, Monetary) is calculated per customer.

$$\text{Recency} = (\text{reference date}) - \max(\text{purchase date})$$

$$\text{Frequency} = \text{count of unique transactions}$$

$$\text{Monetary} = \sum(\text{amount})$$

Quantile-based scoring:

$$R\_rank = \text{qcut(recency, 4)}$$

$$F\_rank = \text{qcut(frequency, 4)}$$

$$M\_rank = \text{qcut(monetary, 4)}$$

$$\text{RFM Score} = R\_rank + F\_rank + M\_rank$$

This creates balanced segmentation even in skewed datasets.

4.3 Customer Lifetime Value (CLV)

Two modelling strategies are used.

4.3.1 Heuristic CLV

$$\text{CLV} = \text{AOV} \times \text{Purchase Frequency} \times \text{Margin} \times \text{Lifetime}$$

Where:

- AOV (Average Order Value) = total spend / number of transactions
- Margin is typically 25–35
- Lifetime may be 12–24 months

4.3.2 Probabilistic CLV (BG/NBD + Gamma-Gamma)

- BG/NBD models transaction frequency over time
- Gamma-Gamma models monetary value

This provides superior forecasting accuracy.

5 Analytical Table Construction

5.1 Customer Master

Contains long-term attributes:

- Demographics
- Loyalty membership status
- Total loyalty points
- Last purchase date

5.2 Sales Transactions

A cleaned transaction-level table including:

- transaction_id
- customer_id
- timestamp
- category, brand, type
- shipping, payment, ratings

5.3 Customer Analytics

Contains:

- RFM metrics
- Product diversity
- Average rating
- CLV or CLV score

5.4 Product Master

Unique product-level attributes.

5.5 Loyalty Transactions

One row per loyalty event (earn/redeem):

- Loyalty transaction ID
- Points earned
- Points redeemed
- Balance after transaction
- Event type & timestamp

6 Threshold Design Principles

Thresholds in retail analytics must be:

- **Data-driven**: using percentiles ensures consistency across categories
- **Interpretable**: values such as 500 and 2000 are intuitive
- **Business-aligned**: segmentation supports marketing objectives

Examples:

- Top 10% by monetary value $\rightarrow$ High-Spenders
- Recency $>$ 30 days $\rightarrow$ At-Risk

7 Evaluation and Monitoring Recommendations

7.1 Evaluation Metrics

- RMSE, MAE for CLV predictions
- Revenue uplift for targeted segments
- Repeat purchase rate after loyalty changes

7.2 Monitoring

- Daily ingestion checks for schema drift
- Anomaly detection on transaction volume
- Loyalty point audit logs

8 Conclusion

This report outlines a complete and scalable approach for converting raw retail data into analytical features used for loyalty programs, segmentation and predictive modelling. The workflow supports both business-driven decisions and advanced data science modelling techniques. This forms the foundation of a production-ready loyalty analytics system.