Data Growth Analyst - SQL

Environment Setup

Google Colab starter template - link Dataset to

upload into Google Colab - link

Dataset Overview

You'll be working with an e-commerce orders dataset with the following schema:

- invoice_id: Unique identifier for each transaction
- line_item_id: Unique identifier for each item in an order
- user_id: Customer identifier
- item_id: Product identifier
- item_name: Product name
- item_category: Product category
- price: Item price in USD
- created_at: Order creation timestamp
- paid_at: Payment completion timestamp

Instructions

- Write SQL queries using pandas SQL (pandasql) syntax
- Provide clear, well-commented code
- Include your analytical reasoning for each question
- Suggested time allocation: 2-3 hours
- **Complete at least 2 questions.**

Output

- Google Colab notebook with code and comments
- Excel / Python for questions, results visualization/presentation

Question 1: Cohort Retention Analysis

Create a comprehensive monthly cohort retention analysis that includes:

1. **Standard cohort table**: Cohort month, cohort size, and retention rates for months 1-12
2. **Resurrection analysis**: Identify customers who return after being inactive for 2+ months and calculate "resurrection rates" by cohort
3. **Quality retention**: Calculate retention rates excluding customers who only made single low-value purchases (<$50 total)

Expected Output:

- Main cohort retention table with monthly percentages
- Resurrection rate table showing what % of "lost" customers return each month
- Comparison of standard vs. quality retention rates

Business Context: Growth team needs to understand true retention patterns to set realistic customer acquisition targets and identify opportunities for win-back campaigns.

Question 2: Customer Lifetime Value s Acquisition Efficiency

Build a CLV model that informs acquisition strategy:

1. **Customer segmentation**: Classify customers based on first 90 days behavior (single vs. repeat purchaser, high vs. low value)
2. **CLV calculation**: For each segment, calculate predicted CLV using:
    a. Average Order Value
    b. Purchase frequency (orders per month)
    c. Estimated lifespan (based on similar customers)
3. **Acquisition ROI**: Determine maximum allowable Customer Acquisition Cost (CAC) for each segment assuming 3:1 LTV:CAC ratio
4. **Validation**: Compare predicted vs. actual CLV for customers with 12+ months history

Expected Output: Table showing segment characteristics, predicted CLV, and recommended max CAC by segment.

Business Context: Marketing needs data-driven CAC limits by customer type to optimize ad spend across different channels and audiences.