Title: Understanding RetailSmart's Customer and Sales Behavior through Data Cleaning and Exploration

Business Context

RetailSmart Analytics Pvt. Ltd. is a mid-sized e-commerce retailer that sells products across multiple categories in both online and offline channels.
The company's leadership suspects that **declining repeat purchases and uneven sales performance** are linked to **data quality issues** and **limited visibility** into customer behavior.

They have provided you with five interconnected datasets derived from the company's operational databases (customers, products, sales, marketing, and reviews).

As a **Data Analyst / Junior Data Scientist**, your first goal is to **explore, clean, and validate the data** to ensure it is ready for modeling and decision-making.

Phase-1 Objective

Perform SQL-based data extraction, cleaning, and exploratory analysis to answer key business questions such as:

1. What does RetailSmart's customer base look like in terms of geography, demographics, and engagement?
2. Which product categories and channels drive the highest revenue and frequency of purchase?
3. Are there anomalies or data quality issues in orders, prices, or payments?
4. What trends are visible in order volumes, revenues, and customer churn over time?
5. How should we prepare this data for predictive modeling in later phases?

Datasets Used

Dataset	**Description**	**Key Columns**
customers.csv	Customer profiles and derived churn flag	customer_id, city, state, total_orders, total_spent, days_since_last_order, churn_flag
sales.csv	Transaction-level data (joined from orders, items, payments)	order_id, customer_id, product_id, category_english, price, payment_type, order_purchase_timestamp, total_price
products.csv	Product attributes	product_id, category_english, product_name_lenght, product_description_lenght, product_photos_qty
marketing.csv	Marketing campaign info	campaign_id, customer_id, channel, spend, conversions, response_rate

Dataset	Description	Key Columns
reviews.csv	Customer review text and scores	review_id, customer_id, review_score, review_comment_message

Tasks Overview

A. SQL & Data Extraction (Conceptual or via SQLite)

1. Create tables for all five datasets.
2. Run basic SQL operations:
   - Retrieve top 10 customers by total spend.
   - Identify the top 5 product categories by revenue.
   - Find the average order value per city/state.
   - Determine the percentage of customers who have churned (churn_flag = 1).
   - Join sales and marketing tables to find conversion rate by channel.
3. Use SQL constraints or queries to detect invalid/missing values.

B. Data Cleaning

1. Handle missing values in category_english, payment_type, review_score, and days_since_last_order.
2. Detect and treat outliers in price, total_price, and spend.
3. Standardize categorical columns (e.g., title-case cities, consistent channel names).
4. Convert timestamps to datetime and derive year, month, weekday fields.
5. Validate referential integrity across datasets (customer_id and product_id consistency).

C. Exploratory Data Analysis (EDA)

1. **Univariate analysis** — distribution of order values, customer spend, churn flag.
2. **Bivariate analysis** — relationship between category and revenue, payment type vs. spend.
3. **Time-series trends** — monthly orders and total revenue over time.
4. **Customer segmentation insights** — RFM scatter plots or boxplots by churn flag.

   RFM stands for **Recency, Frequency, and Monetary value**, which are three key behavioral indicators used to segment customers:

- **Recency (R):** How recently a customer made their last purchase.
  (Smaller = more recent = more engaged)

- **Frequency (F):** How often they purchase.
  (Higher = loyal or repeat customers)

- **Monetary (M):** How much they spend in total.
  (Higher = more valuable customers)

  Together, these dimensions help identify **customer segments** such as "High-Value Loyalists," "At-Risk," or "Churned."

I. **Derive RFM Metrics**
Calculate Recency, Frequency, and Monetary values for each customer_id using the sales data.

   a. Recency = Days since the customer's last purchase

   b. Frequency = Total number of orders placed

   c. Monetary = Sum of total spending

II. **Merge Churn Information**
Join the RFM summary with the customers dataset to include the churn_flag column for each customer.

III. **Visualize with Boxplots**
Create boxplots of Recency, Frequency, and Monetary against churn_flag.
This will help visualize differences between active and churned customers.

IV. **Create Scatter Plot**
Plot Frequency vs. Monetary, using color to represent churn_flag.
This helps identify clusters or behavioral patterns visually.

V. **Interpret the Results**
Observe patterns such as churned customers having higher Recency (long time since last purchase) and lower Frequency or Monetary values.

VI. **Document Key Insights**
Summarize your findings — for example,
"Active customers purchase more frequently and spend more, while churned customers show higher Recency values."

5. **Marketing insights** — response rates by channel and spend bands.

   - **Aggregate by Channel**
   From the `marketing` dataset, calculate the average response rate and average spend for each marketing channel (Email, SMS, Social Media, Affiliate, etc.).

   - **Visualize Channel Performance**
   Create a bar plot showing average response rate by channel to identify which channels are performing best.

   - **Create Spend Bands**
   Group campaign spend into defined ranges, such as 0–2K, 2–4K, 4–6K, 6–8K, and 8K+.
   Use these as spend bands to segment marketing efforts.

- **Analyze Response Rate by Spend Band**
For each spend band, calculate the mean response rate.
This will help identify whether higher spending actually leads to better responses or if there’s a plateau.

- **Visualize Spend Band Results**
Plot response rate against spend bands to highlight trends or diminishing returns.

- **Interpret and Summarize Insights**
Conclude your analysis by noting observations such as,
“Response rates peak between ₹4K–₹6K spend range,” or
“Email campaigns achieve the highest average response rate.”

s