以下是使用 Python 的 `pandas` 和 `faker` 库生成的 10 万条电商销售模拟数据，包含您指定的所有字段，且符合业务逻辑。数据将分为 用户表 和 促销表，并通过 `promotion\_id` 关联。

\*\*数据字段说明\*\*

\*\*用户表 (`users.csv`)\*\*

| 字段名 | 说明 |

|---------------------------|----------------------------------------------------------------------|

| `user\_id` | 用户唯一标识（格式：U000001） |

| `is\_new\_user` | 是否新用户（0=老用户，1=新用户） |

| `registration\_date` | 注册日期（过去2年内随机） |

| `historical\_purchase\_freq`| 历史消费频率（泊松分布，λ=3） |

| `avg\_order\_value` | 客单价（正态分布，均值200元，标准差50元） |

| `ltv` | 用户生命周期价值（历史消费频率 × 客单价 × 随机系数） |

| `promotion\_participation\_count` | 参与促销次数（0~5次） |

\*\*促销表 (`promotions.csv`)\*\*

| 字段名 | 说明 |

|---------------------|----------------------------------------------------------------------|

| `promotion\_id` | 促销活动ID（格式：P0001） |

| `promotion\_date` | 促销日期（过去1年内随机） |

| `promotion\_type` | 促销类型（60%折扣，40%满减） |

| `discount\_rate` | 优惠力度（折扣：5%~30%；满减：50/100/150/200元） |

| `target\_user\_group` | 覆盖用户群体（新用户/老用户/高价值用户） |

\*\*行为表 (`behavior.csv`)\*\*

| 字段名 | 说明 |

|---------------------------|----------------------------------------------------------------------|

| `user\_id` | 用户ID |

| `promotion\_id` | 参与的促销活动ID |

| `purchase\_amount` | 单次促销中的购买金额（受促销类型和用户属性影响） |

| `days\_since\_last\_purchase`| 距离上次购买的天数（1~90天随机） |

---

\*\*数据合理性设计\*\*

1. 用户属性：

• 新老用户比例：30% 新用户，70% 老用户。

• 消费频率：泊松分布模拟低频到高频用户。

• 客单价：正态分布避免极端值，均值200元符合常见电商场景。

• LTV：基于历史消费和随机波动，反映用户价值差异。

2. 促销活动：

• 促销类型：60% 折扣（如8折）、40% 满减（如满200减50）。

• 优惠力度：折扣率5%~30%，满减金额50/100/150/200元。

• 目标用户：促销定向新用户、老用户或高价值用户。

3. 用户行为：

• 购买金额：折扣促销按比例减少金额，满减促销按阈值减少金额。

• 购买间隔：1~90天随机，模拟不同用户的复购周期。

• 促销参与次数：0~5次，高频用户参与更多促销。

\*\*使用建议\*\*

1. 数据分析：

• 探索促销类型对购买金额的影响（如折扣 vs. 满减）

• 研究购买间隔时间与消费频率的关系。

• 分析不同用户群体（新用户/老用户）的促销参与率和 LTV 差异。

2. 模型构建：

• 预测用户参与促销的概率（分类模型）。

• 预测用户购买金额（回归模型）。

• 用户分群（聚类分析）。

3. 数据验证：

• 检查字段范围（如 `discount\_rate` 是否在合理区间）。

• 验证促销参与次数与行为记录的一致性。

如果需要调整参数（如客单价分布、促销频率等），可修改代码中的随机分布参数。