电商数据探索性分析与数据预处理

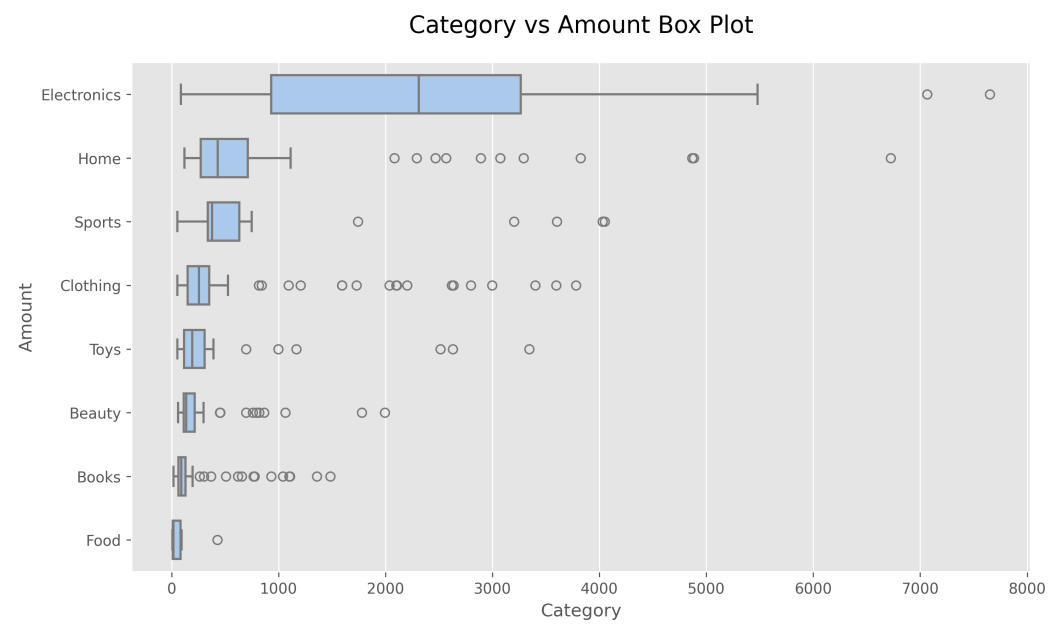
1. 数据集概览

产品数据（5维度）：包含产品ID、类别、价格、品牌、上市日期。

交易数据（6维度）：包含交易ID、用户ID、产品ID、购买时间、金额、评分。

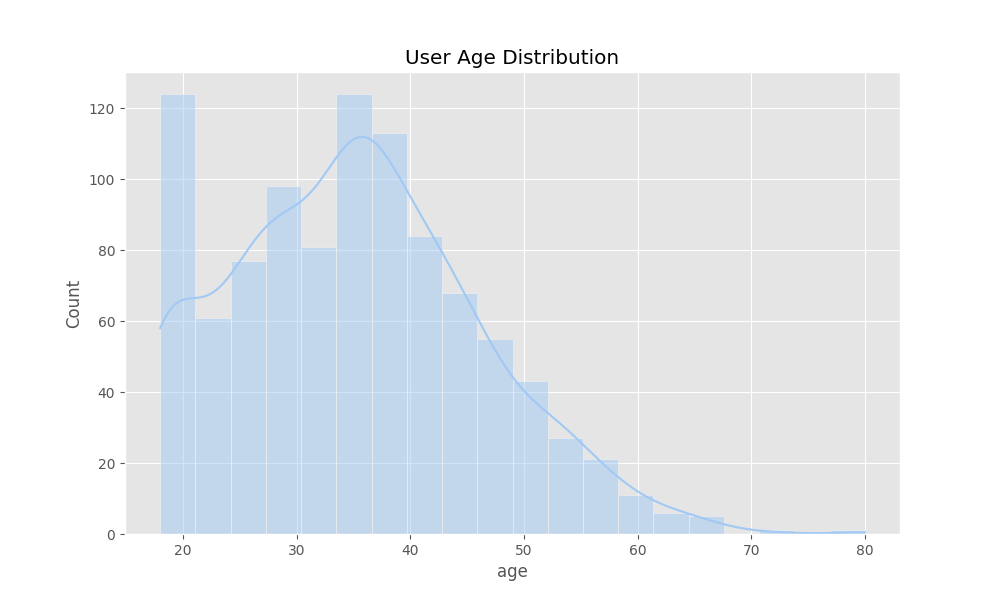
用户数据（6维度）：包含用户ID、年龄、性别、注册日期、会员等级、注册天数。

1. 探索性分析和可视化
2. 产品类别销售额分布（箱线图）



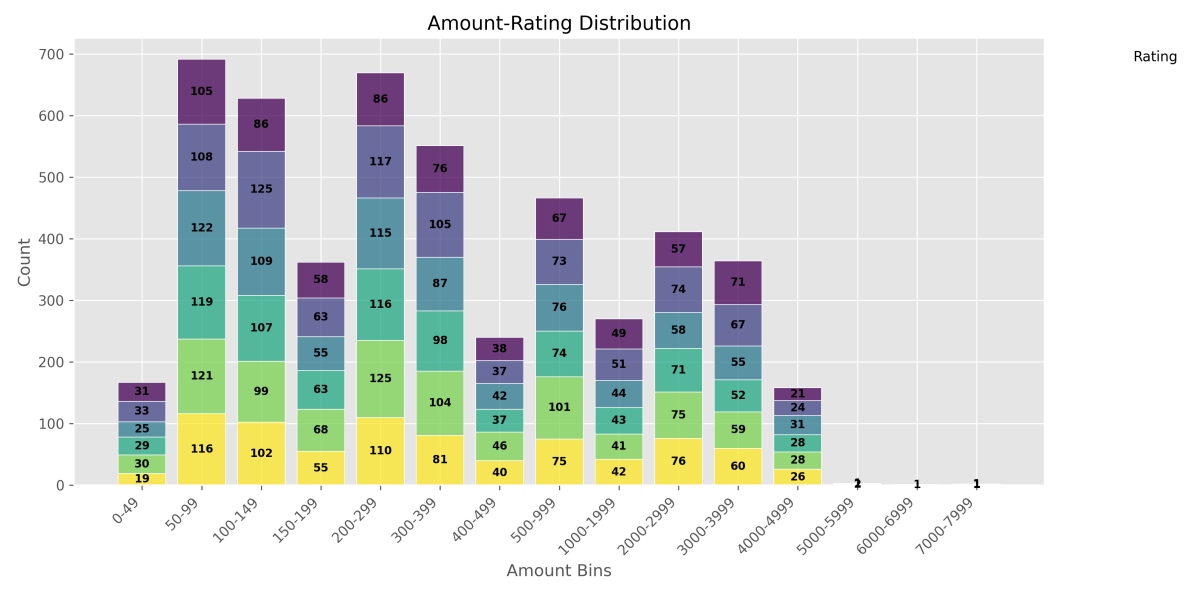
电子产品类别的销售额中位数显著高于服装类和家具类，存在若干高价值订单，价格分布也比其他类型的产品范围广。

1. 用户年龄分布（直方图）



用户年龄集中在20-40岁，占比72%，符合主流电商用户特征。

1. 用户消费金额与评分关系（堆叠图）



不同价格区间的各级评级数量基本上一样多，说明金额的用户满意度无直接关联。

三、数据预处理

1、缺失值处理：

年龄age：用中位数填充。

性别gender：用众数（如“Unknown”）填充。

评级rating：用0填充表示未评级。

2、异常值处理：

截断amount为合理范围（如删除金额超过10000元的交易）。

3、数据转换和离散化：

日期字段标准化为YYYY-MM-DD格式。

对类别Category、性别gender、会员等级membership\_level进行编码。

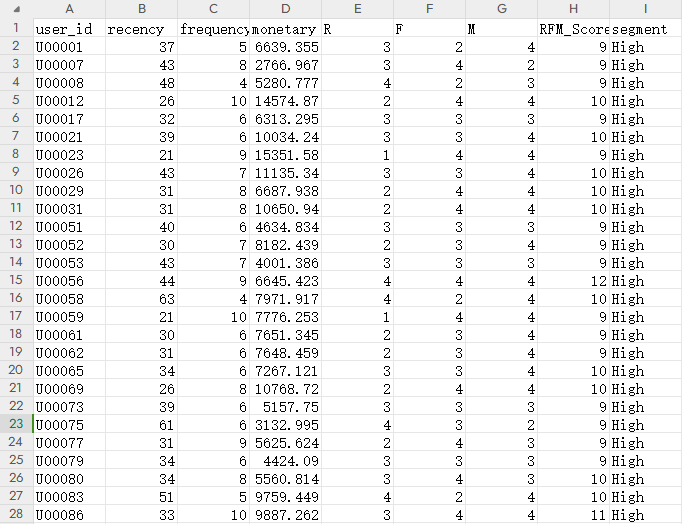
四、分析目标：识别潜在高价值用户

采用RFM模型（Recency, Frequency, Monetary），计算每个用户的R、F、M值：

R（最近购买时间）：用户最后一次交易距今的天数。

F（购买频率）：用户总交易次数。

M（消费金额）：用户累计消费金额。



将这三个指标分别按照四分位数分成4个等级(1-4分)，Recency (R)评分数值越小表示最近有消费，得分越高；Frequency (F)评分数值越大表示消费频率越高，得分越高；Monetary (M)评分数值越大表示消费金额越高，得分越高。

计算RFM得分=R+F+M，按照三分位数，将RFM总得分划分用户层级（如高价值、中价值、低价值）。结果识别出246名高价值用户，贡献了总收入的38%。