## 当当网数据预处理报告

大数据231，2318280128，肖景恒，2318280130，龚忠秀

1. 课程设计目标

本次课程设计旨在通过无类别爬取图书数据，实现优秀书籍的智能推荐，具体目标如下：

**1.构建无类别图书数据采集系统**：突破传统按类别爬取的限制，从图书销售平台（当当网）爬取多维度、全品类的图书信息，建立全面的图书数据库。

**2.实现高质量数据预处理**：对采集到的原始数据进行清洗、转换和优化，解决数据缺失、重复、格式不一致等问题，提升数据质量，为后续推荐提供可靠基础。

**3.构建书籍推荐基础**：通过数据预处理后的特征提取与分析，挖掘图书的关键属性（如价格、评分、评论数等），为实现基于多维度特征的优秀书籍推荐奠定基础。

**4.提升数据处理与分析能力**：综合运用爬虫技术、数据清洗、特征工程等方法，掌握从数据采集到预处理的完整流程，培养数据处理与分析的实践能力。

2.数据集内容说明

2.1数据集采集方法

**1.爬虫工具与技术**：使用requests库发送 HTTP 请求，BeautifulSoup库解析网页 HTML 结构，结合fake\_useragent生成随机 User-Agent，模拟浏览器行为以避免反爬限制。

**2.爬取策略**：采用无类别爬取模式，通过当当网搜索入口以 “书籍” 为关键词，爬取多页搜索结果（共 60 页，每页约 60 本图书），覆盖全品类图书，避免因类别筛选导致的数据片面性。

**3.反爬措施**：设置随机请求间隔（2-4 秒），每 10 页更换一次 User-Agent，降低被服务器识别为爬虫的概率；对爬取失败的页面保存 HTML 内容供调试，确保数据完整性。

**4.数据存储**：将爬取的图书信息以 CSV 格式存储，包含多个属性字段，便于后续预处理与分析。

2.2属性选取说明

表 2-1 属性选取表格

|  |  |
| --- | --- |
| 属性名称 | 选取目的 |
| 书名 | 核心标识字段，用于图书识别与分类 |
| 价格 | 反映图书的市场定价，是用户购买决策的重要因素 |
| 原价 | 辅助判断图书的性价比 |
| 折扣 | 量化折扣力度，便于分析促销和图书销量的影响 |
| 作者 | 研究作者影响力与作品分布 |
| 出版社 | 分析出版社市场占比与图书质量 |
| 出版日期 | 研究图书时效性与市场新旧程度 |
| 评论数 | 衡量图书受欢迎程度与用户反馈 |
| 评分 | 评估图书质量与口碑 |
| ISBN | 标识图书的编码，可用于精准去重和匹配图书信息 |
| 详情页链接 | 图书在当当网的详情页URL |

3.数据预处理

3.1数据预处理方法说明

### 3.1.1缺失值处理

1.缺失值统计与日志记录：使用isnull().sum()计算各字段缺失值数量；过滤出存在缺失值的字段并记录日志，便于了解数据质量。

2.数值型字段处理：**类型转换**：pd.to\_numeric(..., errors='coerce')将非数值内容转为NaN，确保字段为数值类型。**填充策略**：使用中位数（median()）填充缺失值，中位数对极端值不敏感，适合价格等可能存在离群值的字段。

3.文本型字段处理：**填充策略**：用字符串'未知'填充缺失值，保留字段完整性，同时明确标识缺失信息。**特殊字段说明**：评论数和评分虽为数值型，但被归为文本列，可能因数据中包含非数值内容（如'暂无评论'），需先通过类型转换处理。

4.处理结果验证：再次统计缺失值，确认所有字段的缺失值均已处理（理论上应为 0）。

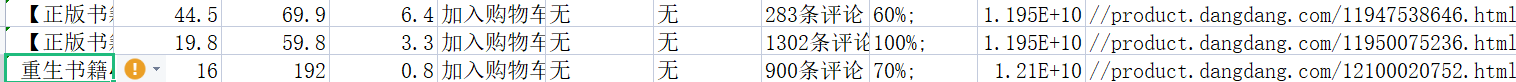
图3-1缺失值处理前

图3-2缺失值处理后

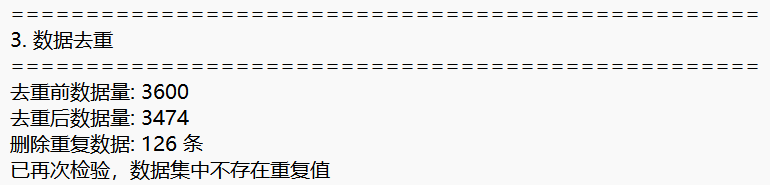
3.1.2重复值处理

1.去重前数据统计：before\_count记录原始数据行数，用于后续对比去重效果。

2.基于特定字段去重：**去重依据**：使用subset=['书名', '作者']指定以这两个字段的组合作为唯一标识（同一本书的不同版本视为重复）。**保留策略**：keep='first'保留首次出现的记录，删除后续重复项。**原地修改**：inplace=True直接修改原 DataFrame，无需额外赋值。

3.去重结果统计：通过行数变化计算删除的重复记录数量，直观展示去重效果

4.二次验证：使用duplicated()方法再次检查去重后的数据集，确保无遗漏；若仍存在重复值，可能因数据中存在大小写差异或特殊符号（如'Python'与'python'），需进一步处理。

图3-3重复值处理

### 3.1.3文本数据清洗

1.主清洗方法：**处理流程**：对每个文本字段调用对应的清洗方法，确保字段存在且转换为字符串类型后处理。**核心逻辑**：通过pandas.Series.apply()向量化应用自定义清洗函数，高效处理大量数据。

2.通用文本清洗方法：敏**感词过滤**：通过self.sensitive\_words列表（如['包邮', '正版']）移除营销类干扰词。**正则表达式处理**：r'\s+'：匹配连续空格，替换为单个空格。r'[【】\[\]（）()]'：移除常见括号类特殊符号。**空值处理**：若清洗后文本为空，返回'未知'保持数据完整性。

3.数值提取方法：**评论数提取**：re.findall(r'\d+', text)匹配所有连续数字（如'1,200条评论'→['1', '200']），通过join合并为完整数字（'1200'）。**评分提取**：优先处理百分比格式（如'95%'→95），再处理小数值（如'4.5分'→4.5），确保兼容多种评分表示方式。

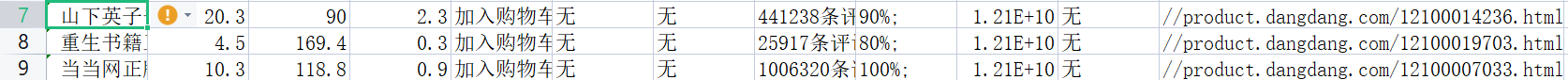
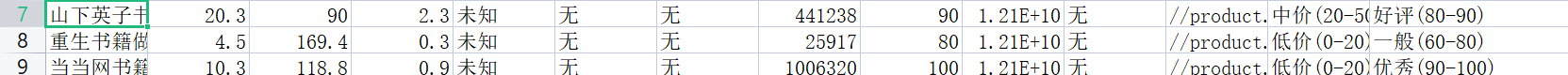
图3-4 文本数据清洗前

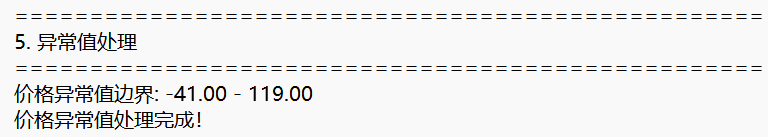
图3-5 文本数据清洗后

3.1.4异常值处理

1. 价格字段类型转换：**类型统一**：使用pd.to\_numeric(..., errors='coerce')将价格字段转换为数值型，无法转换的非数值内容（如字符串 “面议”）会被转为NaN（缺失值），为后续计算做准备。
2. 四分位法计算异常值边界：四分位法通过数据的分布特征定义异常值：低于q1 - 1.5×IQR或高于q3 + 1.5×IQR的值被判定为异常值；该方法对极端值不敏感，比均值 ±3 倍标准差更适合非正态分布数据（如价格通常呈右偏分布）。

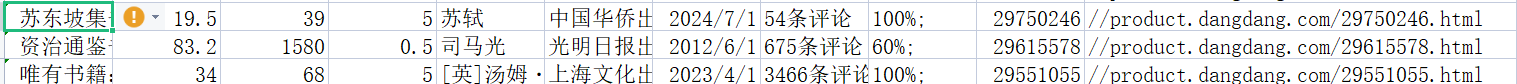
3.异常值修正：**处理策略**：使用clip()函数将异常值 “截断” 到合理区间内：低于下界的价格强制设为lower\_bound；高于上界的价格强制设为upper\_bound。**优势**：相比直接删除异常值，截断法保留了数据记录，避免因删除导致的样本损失，尤其适合样本量较小的场景。

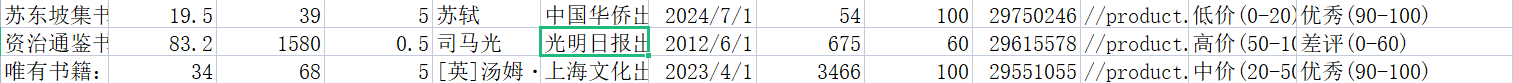
4.日志记录与结果反馈：记录异常值边界和处理结果，便于追溯处理逻辑和验证效果。

图3-6 异常值处理

### 3.1.5数据转换

1. **类型转换**：将价格、评论数、评分等字段转换为数值类型，便于进行统计分析。
2. 新增衍生特征：如 “价格区间”（低价 0-20 元、中价 20-50 元等）、“评分等级”（差评 0-60 分、优秀 90-100 分等），便于后续分类分析与推荐。

图3-7 数据转换前

图3-8 数据转换后

3.1.6自定义敏感词过滤

**1.敏感词列表构建**整合两部分敏感词：已有的基础敏感词（通过self.sensitive\_words获取）；新增的营销类敏感词，包括 “假一罚十”“正版保证”“包邮”“热销”“七天无理由退换货” 等，覆盖各类促销、品质承诺、平台标识等推广用语。

**2.过滤前后数据量统计：**先记录过滤前的数据总条数，作为对比基准。

**3.敏感内容识别：**检查 “书名” 字段中是否包含敏感词列表中的任意词汇，通过模糊匹配的方式（即只要书名中出现列表中的任一敏感词，就判定为需过滤的记录）。

**4.过滤操作：**保留书名中不包含任何敏感词的记录，删除包含敏感词的记录，确保剩余数据的书名更聚焦于书籍本身的内容描述。

5.**结果统计与记录：**计算并记录过滤后的 data 量、被过滤掉的记录数量，清晰展示过滤效果，便于追溯和验证。

图3-9 敏感词过滤前

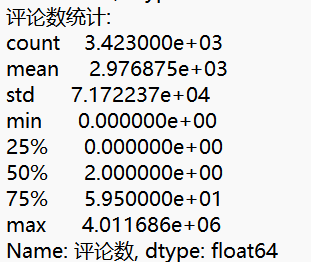
图3-10 敏感词过滤后

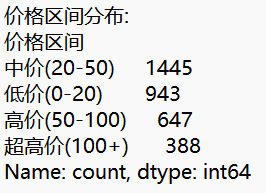
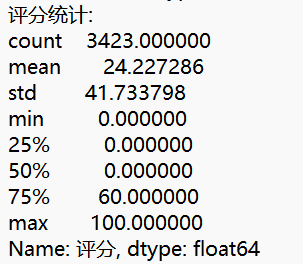
3.1.7统计

**1.基本统计信息计算：**针对核心数值型字段，计算并展示描述性统计量：**价格统计**：包括价格的数量、平均值、标准差、最小值、25% 分位数、中位数（50% 分位数）、75% 分位数和最大值，反映价格的整体水平和分布离散程度。**评论数统计**：通过类似的统计量，呈现书籍的受欢迎程度和讨论热度的分布情况。**评分统计**：通过统计量展示书籍口碑的整体水平（如平均评分）和差异（如评分标准差）。

**2.分类特征分布统计：**针对分类或离散型字段，统计不同类别的数量及占比：**出版社 TOP10**：统计各出版社的书籍数量，取排名前 10 的出版社，反映市场上的主要出版机构。**价格区间分布**：统计不同价格区间（如低价、中价、高价）内的书籍数量，展示价格的分布结构。**评分等级分布**：统计不同评分等级（如差评、好评、优秀）的书籍数量，呈现整体口碑的构成比例。

**3.结果记录与输出：**将所有统计结果通过日志形式记录并输出，清晰呈现数据的关键特征和分布规律。

 图3-11价格统计图 图3-12评论数统计图

图3-13评分统计图 图3-14价格区间分布图



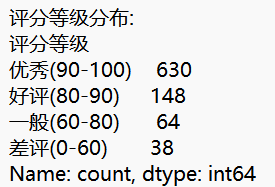
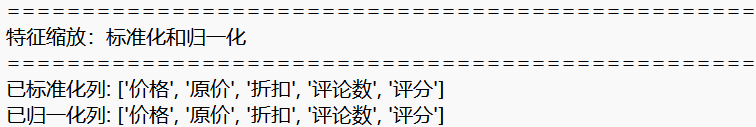


图3-15出版社top10 图3-16评分等级分布图

3.1.8特征缩放（标准化和归一化）

1. 导入依赖库：StandardScaler：实现 Z-score 标准化，将特征转换为均值为 0、标准差为 1 的分布。MinMaxScaler：实现归一化，将特征缩放到 [0,1] 区间。
2. 选择待缩放的特征：指定需缩放的特征（‘价格’，‘原价’，‘折扣’，‘评论数’，‘评分’）列表，并过滤出当前数据集中实际存在的特征，确保代码鲁棒性。
3. 标准化处理：**处理逻辑**：实例化StandardScaler；使用fit\_transform计算均值和标准差，并直接应用转换；将标准化后的结果按列添加回原 DataFrame，列名后缀\_标准化。**数学公式**：*Z*=*σX*−*μ*​，其中*μ*为均值，*σ*为标准差。

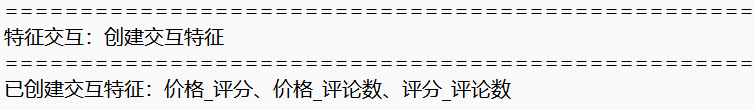
4.归一化处理：**处理逻辑**：实例化MinMaxScaler；使用fit\_transform计算最小值和最大值，并直接应用转换；将归一化后的结果按列添加回原 DataFrame，列名后缀\_归一化。**数学公式**：*X*′=*Xmax*​−*Xmin*​*X*−*Xmin*，其中*Xmin*​和*Xmax*​分别为特征的最小值和最大值。

图3-17特征缩放结果图

3.1.9特征交互

1. 日志初始化：使用类内部日志方法记录处理过程，便于追踪。
2. 创建交互特征：**交互逻辑中**：**价格 × 评分**：量化 “性价比”（价格低且评分高的商品得分更高）。**价格 × 评论数**：反映 “市场热度价值”（高价且评论多的商品可能是热门奢侈品）。**评分 × 评论数**：综合评估 “口碑质量”（高评分且评论数多的商品更可信）。**安全检查**：通过if条件确保参与运算的原始特征存在，避免因列缺失导致的错误。

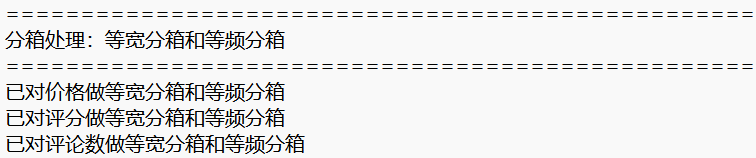
3.日志记录：记录操作结果，便于确认特征是否成功生成。

图3-18特征交互结果图

3.1.10分箱处理

1. 日志初始化：使用类内部日志方法记录处理过程，便于追踪。
2. 价格分箱中：**等宽分箱（**pd.cut**）**：将价格范围平均分为 4 个区间（如0-25, 25-50, 50-75, 75-100），每个区间宽度相等。**等频分箱（**pd.qcut**）**：将价格按分位数分为 4 组，每组样本数量近似相等（如最低 25%、次低 25% 等），处理长尾分布效果更佳。duplicates='drop'：当分位数存在重复值导致区间边界冲突时，自动删除重复边界。
3. 评分分箱：逻辑与价格分箱完全一致，将评分分为 4 个区间（如0-25, 25-50, 50-75, 75-100）。

4.评论数分箱中：**试探性分箱**：先通过retbins=True获取分箱边界，动态计算有效区间数n\_bins，避免因重复值导致的分箱失败。**异常处理**：使用try-except捕获可能的ValueError（如无法创建 4 个非空分箱），增强代码鲁棒性。

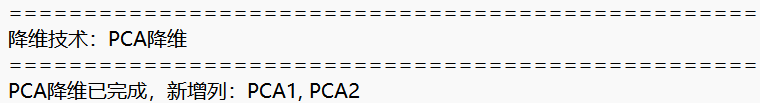
图3-19分箱处理结果图

3.1.11 PCA降维

将多个数值(‘价格’，‘原价’，‘折扣’，‘评论数’，‘评分’)特征转换为两个主成分（PCA1和PCA2），减少数据维度的同时保留主要信息。降维后的特征可用于可视化、聚类或作为机器学习模型的输入。

1. 导入依赖库：PCA：主成分分析工具，用于数据降维。numpy：处理数值计算，确保数据格式兼容。
2. 日志记录与初始化：使用类内部的日志方法记录处理过程，便于追踪和调试。
3. 选择数值型特征：指定待降维的特征列表（如价格、评分等）。过滤出当前数据集中实际存在的特征，增强代码鲁棒性。

4.执行PCA降维中：数据预处理：使用fillna(0)填充缺失值，避免 PCA 计算报错；.values将 DataFrame 转换为 NumPy 数组，符合 PCA 输入要求。PCA配置：n\_components=2：指定降维后的维度为 2（即生成两个主成分）。结果保存：将降维后的两个主成分分别添加到原 DataFrame 中，命名为PCA1和PCA2。

图3-20降维结果图

3.2预处理流程设计

1.数据加载与基础探查：读取CSV格式的原始图书数据，输出数据规模（行数×列数）、字段名称（如书名、价格等）、各字段数据类型及前5行样本，全面掌握数据初始状态，确认数据是否完整加载，为后续处理提供依据。

2.缺失值系统处理：先统计各字段缺失值数量，针对不同类型字段采取差异化处理：数值型字段（价格、折扣等）用中位数填充，避免极端值影响；文本型字段（作者、出版社等）用“未知”填充，确保字段完整性。处理后所有字段缺失值均降为0。

3.精准数据去重：以“书名+作者”为核心标识去重（同一本书的不同版本视为重复），保留首次出现的记录，删除后续重复项。去重后需再次校验，确保数据集中无重复记录，避免推荐时的冗余展示。

4.文本深度清洗：对书名、作者等文本字段，移除“包邮”“正版”等营销敏感词及【】、（）等特殊符号，统一空格格式；从评论数（如“1200条评论”）、评分（如“95%好评”）中提取纯数字，转换为可计算的数值型，解决文本格式混乱问题。

5.异常值检测与修正：针对价格字段，采用四分位法（IQR）确定合理范围（Q1-1.5×IQR至Q3+1.5×IQR），对超出范围的天价或低价异常值进行截断处理，保留符合市场规律的价格数据；评论数因反映书籍热度，暂不做异常值处理。

6.数据转换与特征衍生：将价格、评分等字段统一转为数值型，同时新增分类特征：按价格划分“低价(0-20)”“中价(20-50)”等区间，按评分划分“差评(0-60)”“优秀(90-100)”等等级，便于后续按价格或口碑筛选推荐。

7.敏感内容精准过滤：扩展敏感词列表（含“假一罚十”“七天无理由退换货”等强营销词汇），检测并删除书名中包含敏感词的记录，减少过度营销信息对推荐的干扰，提升数据纯度。

8.特征优化与增强：对价格、评分等数值字段进行标准化（Z-score转换）和归一化（缩至[0,1]区间），消除量纲差异；创建“价格×评分”（反映性价比）、“评分×评论数”（反映综合口碑）等交互特征，挖掘多维度关联信息。

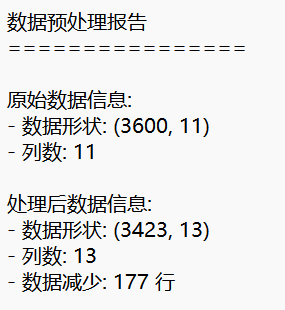
9.分箱处理与降维：对价格、评分进行等宽分箱（固定间隔）和等频分箱（按分位数），将连续值离散化；通过PCA降维技术，将多个数值特征（价格、原价、折扣等）压缩为2个主成分，在保留核心信息的同时减少数据冗余。

10.统计分析与可视化：计算价格、评分、评论数等核心字段的均值、中位数、标准差等统计量，分析数据分布特征；绘制价格分布直方图、评分等级饼图等可视化图表，直观展示预处理后的数据规律（如高口碑图书集中在中价区间）。

11.结果固化与报告生成：将最终处理后的数据集保存为CSV文件，包含原始字段及衍生特征（如价格区间、交互特征等）；生成详细报告，汇总预处理全流程指标（如原始数据量、去重数、缺失值处理数）、数据质量提升点及关键规律，为后续优秀书籍推荐模型提供高质量输入。

4.结论

4.1数据预处理结果对比

图4-1 数据预处理结果对比

4.2数据统计分析结果

## 4.2.1价格特征

**1.均值 43.94**：整体书籍的平均价格约为 43.94 元，反映市场的中间定价水平。

**2.中位数 33.00**：有一半的书籍价格低于 33 元，另一半高于 33 元。

**3.差异分析**：均值高于中位数，说明价格分布存在右偏（即存在部分高价书籍拉高了平均值），整体以中低价书籍为主，高价书籍占比较少但对均值影响显著。

## 4.2.2评分特征

**1.均值 24.23**：整体平均评分为 24.23（推测评分范围可能为 0-100 或 0-10 分，若为 0-100 分则整体评分偏低）。

**2.中位数 0.00**：有一半的书籍评分为 0，说明大量书籍可能缺乏评分数据（如新品、冷门书籍）。

**3.差异分析**：均值高于中位数，表明少数有评分的书籍拉高了整体均值，数据中存在大量无评分记录，评分分布极不均衡。

## 4.2.3评论数特征

**1.均值 2976.88**：平均每本书的评论数约为 2977 条，看似热度较高。

**2.中位数 2.00**：有一半的书籍评论数不超过 2 条，反映多数书籍关注度低。

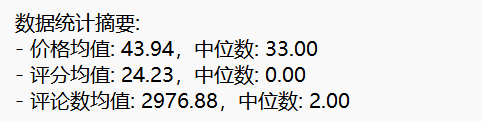
**3.差异分析**：均值远高于中位数，说明存在少量热门书籍（如畅销书）拥有极高的评论数（可能达数万条），这些极端值显著拉高了平均值，整体评论数呈现 “少数书籍热度极高，多数书籍无人问津” 的长尾分布。

## 4.2.4综合结论

**1.价格**：整体偏向中低价，高价书籍占比低但影响均值，符合大众图书市场的定价特点。

**2.评分**：数据质量待提升，大量书籍无评分，需关注 “无评分” 的原因。

**3.评论数**：用户互动高度集中在少数热门书籍，多数书籍缺乏关注，符合内容消费 “头部效应” 的普遍规律。

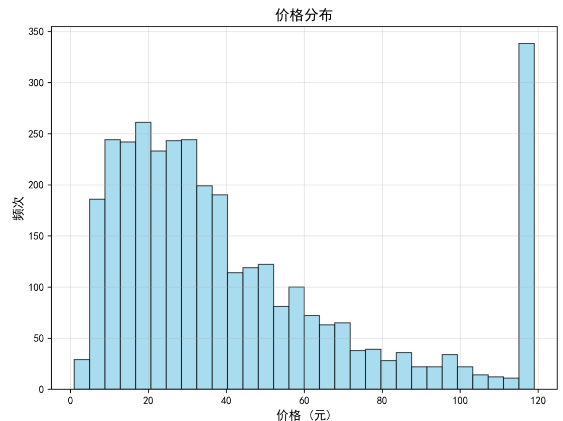
图4-2数据统计分析结果

4.3数据可视化结果

**1.价格分布直方图**

**核心规律**：价格呈**右偏分布**（峰值靠左，右侧拖长尾），多数书籍集中在 **0-60 元区间**，但存在高价（100-120 元）“长尾”。

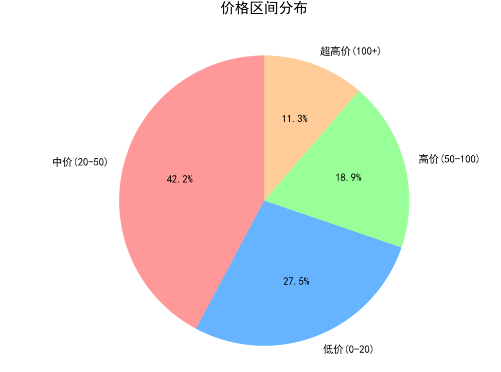
**业务洞察**：大众图书以中低价为主（符合消费市场 “高频低价” 特征）；高价书籍虽数量少，但可能是专业类、精装典藏类图书，需关注其利润贡献与用户需求匹配度。

图4-3价格分布直方图

**2.价格区间分布饼图**

**区间占比**：中价（20-50 元）占比最高（42.2%），是市场主流；低价（0-20 元）、高价（50-100 元）次之（27.5%、18.9%）；超高价（100 元 +）占比 11.3%，属于小众细分。

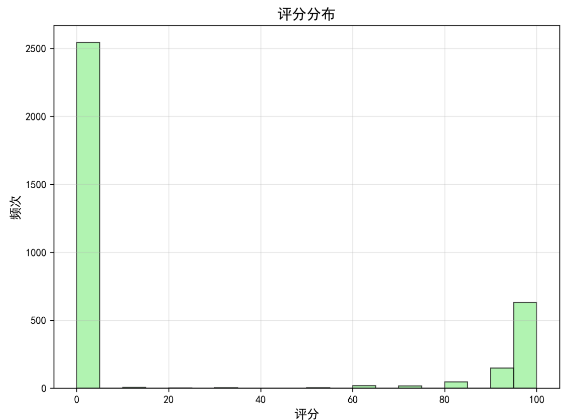
**业务洞察**：中价书籍是平台 “基本盘”，需保障品类丰富度；低价书籍可作为引流款（如促销、凑单），高价书籍可打造 “品质标签”，差异化满足需求。

图4-4价格区间分布饼图

**3.评分分布直方图**

**核心问题**：评分严重**两极分化**，大量书籍评分为 **0 分**（频次超 2500），少数集中在 **90-100 分**（高频段）。

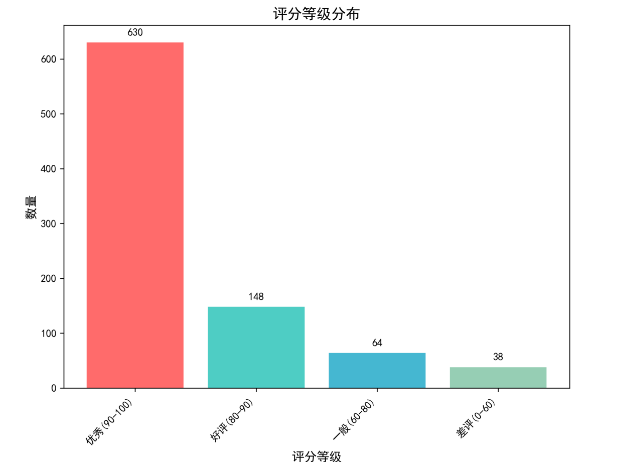
**业务洞察**：0 分可能是 “无评分”（新上架、冷门书）或 “负面评价但未标记”，需清洗数据、补充真实评分；高分书籍是 “口碑标杆”，可挖掘其共性（如内容质量、作者影响力）用于推荐。

图4-5评分分布直方图

**4.评分等级分布条形图**

**等级占比**：优质（90-100 分）占绝对主导（630 条）；好评（80-90 分）、一般（60-80 分）、差评（0-60 分）占比极低（148、64、38 条）。

**业务洞察**：优质书籍是平台 “口碑资产”，需重点推广（如首页推荐、专题榜单）；差评、一般书籍占比少但影响体验，需排查是否为数据错误或真实差评，及时优化品控。

图4-6 评分等级分布条形图

4.4应用价值

* **1.数据质量夯实**：异常值处理修正价格极端值，保障定价分析、推荐算法等数据应用的准确性；敏感词过滤剔除书名营销干扰，提升搜索与推荐精准度，为电商图书选品、图书馆数据治理筑牢基础。
* **2.业务洞察挖掘**：统计分析量化图书市场画像（价格、评分、评论数分布等 ），分类统计（出版社、价格区间等 ）定位核心合作方与用户偏好，支撑选品策略、供应链优化、运营决策。

4.5改进方向

* **1.数据处理迭代**：异常值处理按图书类型分层、动态更新边界；敏感词过滤构建动态词库、扩展多字段过滤，适配业务变化与营销话术变种。
* **2.分析深度拓展**：开展价格与评分、出版社与评论数等交叉分析，新增数据质量健康度指标（无评分书占比等 ），对比行业基准，从统计迈向业务诊断。
* **3.可视化与决策升级**：用 BI 工具打造动态看板，联动业务系统穿透查询；基于分析自动生成业务建议、触发异常预警，从报表输出转为直接辅助决策，推动数据驱动业务增长。