BDA 数据分析师实践考核

报告名称：当当网 2020-2023 年图书榜单销售分析

考核级别：初级

作　　者：王文銮

单　　位：无

联系方式：18067322536

提交时间：2024 年 9 月 24 日

目 录

当当网 2020-2023 年图书榜单销售分析

摘 要

本论文主要研究当当网 2020 至 2023 年的畅销书排行榜的销售情况，运用 Excel 对数据进行单变量分析、多变量分析和相关性分析，探索得出影响图书销量的主要因素，并根据分析结果对图书卖方提出了一些相关建议。

关键词：图书销售、单变量分析、多变量分析、相关性分析

引 言

随着我国经济的发展,文化产业面临着难得的发展机遇，其中，我国出版业市场化程度不断加深,特别是加入 WTO 后,政府开放了图书的批发零售权限,来自内部和外部的各种力量使得中国出版业的市场竞争越来越激烈。与此相应,出版体制改革也在持续推进,经营性的出版社将从传统的事业单位向现代企业转制。随着我国出版业市场化程度的加深和出版体制改革的推进,构建我国市场化的图书营销模式必不可少。这要求图书销售方的经营更加深入、细致，以提高对图书市场资源的可控程度。因此，了解读者的行为和需求对于制定有效的营销策略至关重要。本文将对影响图书销量的因素进行多维度的相关性分析，解释导致图书销量变动的主要因素，旨在帮助销售者了解经济现象，为其提供一些有价值的售卖建议。

1.文献综述与研究目标

1.1 文献综述

中国出版产业正面临着出版全球化所带来的机遇和挑战。对于中国的出版集团而言，如何按照比较竞争优势理论，介入到全球出版产业链和价值链中就成为一个非常重要的问题。我国加入 WTO 的以后，图书出版业加快了产业化发展的步伐。产业化要求出版企业尽快走向市场，加入 WTO 意味着市场开放与准入，其核心问题都是竞争。市场竞争已不可避免地摆在所有出版企业面前。

从趋势看，出版业面临着新进入者和替代品的巨大威胁，社会资本和国外资本进入将导致业内竞争激烈化，出版企业竞争对手转换，竞争核心将表现在与国

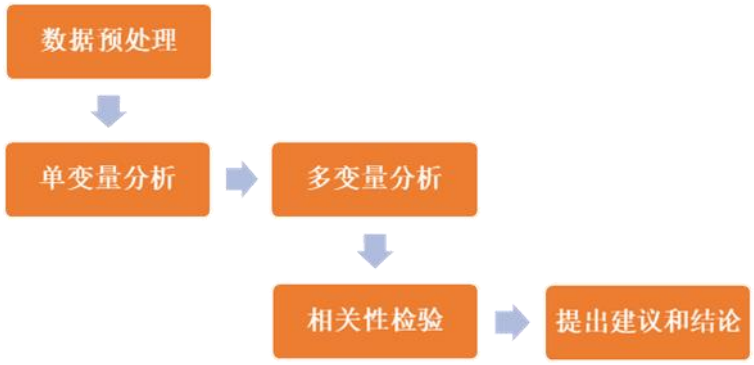
外资本的对抗上，出版业发展将呈现出很大的不确定性。高度重视高质量优秀图书的出版，扩大图书发行覆盖面，已成为我国出版业在全面建设小康社会的新时期一项紧迫任务。

本次研究将从出版年份、售价、折扣比例、推荐值、上榜次数、平均排名、图书类别等几个维度探索与图书销量的相关性（由于所运用的数据集不包含销量这一字段，而通常来说评论数越多，销量也就越高，因此使用能够量化“销量”的“评论数”作为代理指标），把握读者图书需求的产生及其规律，并对出版社和网站提出相关建议。

1.2 研究目标

对 2020 至 2023 年上榜图书销售数据进行分析，从而把握读者的需求和偏好，以便出版社和网站根据需求调整决策，制定更有效的、“以读者为核心”的市场策略，增强读者黏性，对于如何提高图书出版的销售量提供参考价值。

2.数据分析流程



3.分析工具

本报告主要采用 MySQL 数据库、Navicat、Excel 等数据分析工具，对进行数据清洗、查询和分析，数据可视化主要通过 Excel 进行实现。

4.数据概述与数据预处理

4.1 数据概述

本数据来源于和鲸社区数据集中的当当网图书畅销榜单（2020-2023）。原始数据共 2000 条记录、12 个字段，包含书名、作者、出版日期、出版社、原价、售价、折扣比例、排序、排行榜类型、推荐值、电子书价格、评论数等字段。

4.2 数据预处理

（1）数据处理：

利用 Excel 对数据进行清洗，无重复值。在推荐值字段发现有 4 个异常值，用 1.000 填充。作者字段中有 3 个缺失值和 1 个出版日期缺失值，用网络查询书籍得到数据并填充；评论数有 4 个缺失值，用平均数进行填充；考虑到电子书价格缺失值为 1307 占比过大，分析此字段很难得出准确的结论，且电子书价格不是我们本次分析的主要方向，故删除该字段。

将原价、售价、折扣比例转换为浮点数，将评论数转换为整数型；调整推荐值格式均保留 3 位小数。

去掉折扣比例中的“折”字符；排行榜类型中只保留年份的数据。

（2）数据加工：

为了便于计算，将排序、排行榜类型替换为 2020 排名、2021 排名、2022 排名、2023 排名等四个字段。

增加了上榜次数、平均排名（四年的平均排名）字段，调整平均排名格式保留 2 位小数。

通过网络数据采集，增添了图书类别这一字段。

由于同一书名不同年份的排名合并为一行数据，该数据出现重复值，故删除 901 条重复值。

清理结束，剩余 1099 条记录、16 个字段。

5.数据分析

5.1 单变量分析

5.1.1 不同出版年份的上榜情况分析

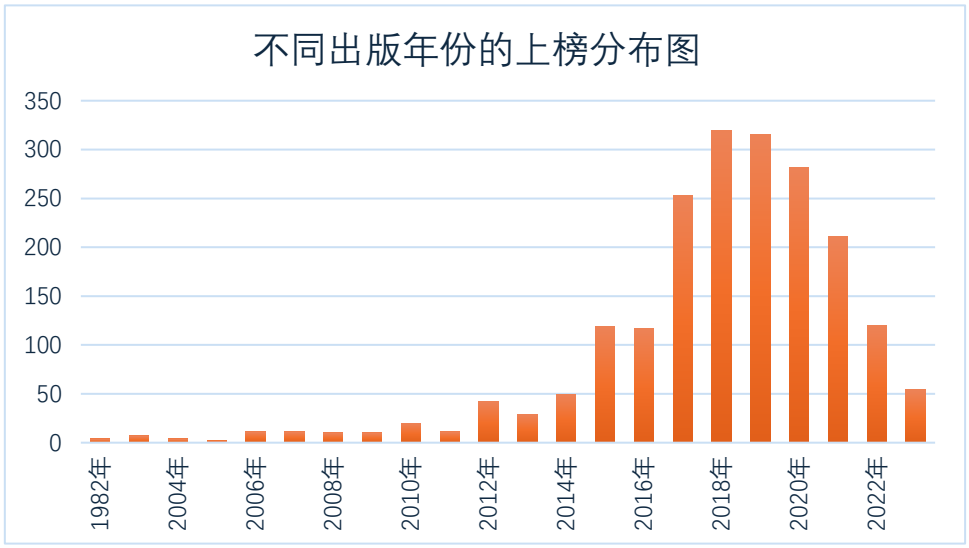


图 1 不同出版年份的上榜分布图

从图 1 中我们可以看出，上榜书籍的出版年份出现频数最高的是 2018 年，从整体上看，出版年份集中在 2017-2021 年。

5.1.2 不同出版社的上榜情况分析

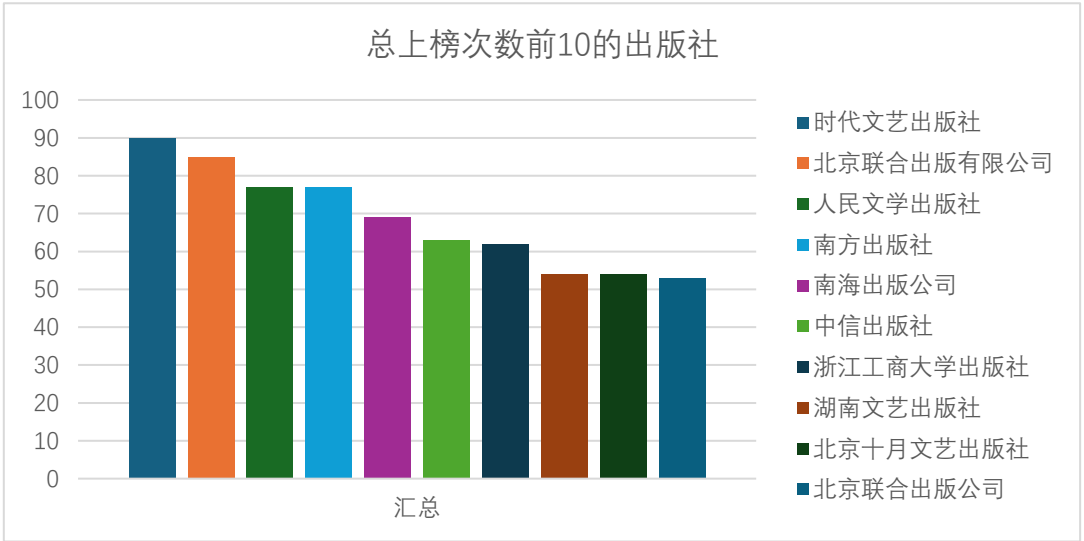


图 2 总上榜次数前 10 的出版社情况图

观察图 2 我们发现：

上榜最常见的出版社是时代文艺出版社，上榜共有 90 本书籍，由此可以看出，在数量的角度上，时代文艺出版社出版的书籍在榜单上具有较高的占有量。

5.1.3 不同作者的上榜情况分析

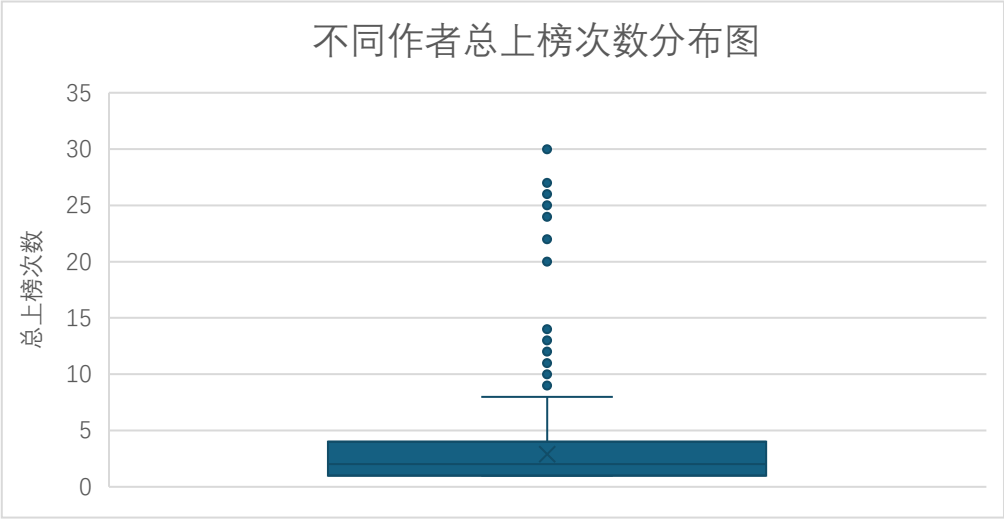


图 3 作者总上榜次数分布图

从图 3 可以看出：

大多数作者总上榜次数在 5 次以下，少部分作者达到了惊人的 20 次以上。

为进一步探究上榜次数偏高的作者具体是谁，因此绘制了总上榜次数前 10 的作者图。

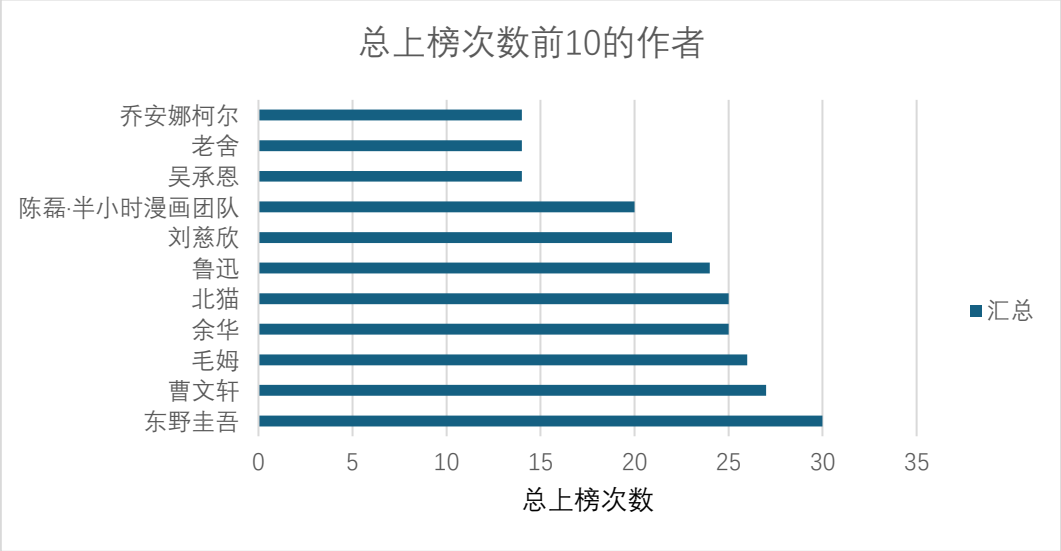


图 4 总上榜次数前 10 的作者情况图

我们分析总上榜次数前 10 的作者可以看到：

东野圭吾的上榜次数达到 30 次，位居最高，此外曹文轩、毛姆、余华、北猫等的上榜次数也达到了 25 次及以上，意味着这些作者的作品受到读者的广泛

欢迎。对于这些频繁上榜的作者，出版商可以考虑建立长期合作关系，为他们提供更多的资源和支持，以保持其作品的持续热度和市场竞争力。

为了寻找这些上榜次数多的作者的作品趋势，以帮助出版商预测和引导市场潮流，接下来对他们上榜的图书类型分布情况进行分析：

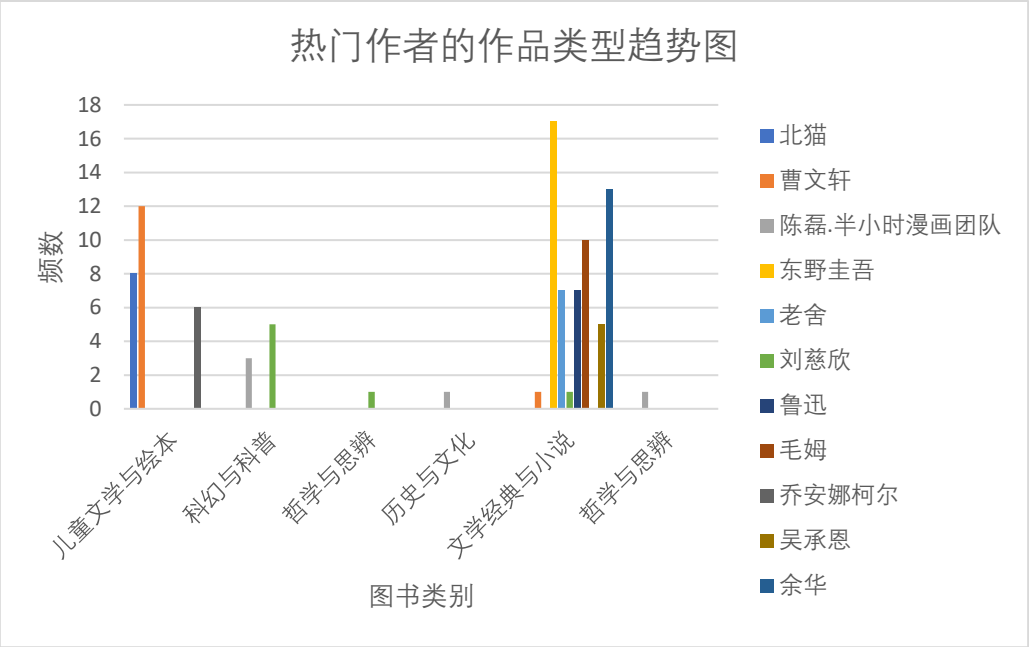


图 5 热门作者的作品类型趋势图

从图 5 这张图可以发现：

整体上看，这些较为热门的作者的图书类型大多趋向于文学经典与小说，其次则是儿童文学与绘本。意味着这两类书籍在市场上有较强的吸引力，因此出版商应该更关注文学经典与小说与儿童文学与绘本这两类书籍的供应。

东野圭吾作为上榜次数最高的作者，其作品类型也在文学经典与小说中显著分布。而东野圭吾作为悬疑小说的代表作家，它的作品类型指向了特定读者群。出版商可以强化悬疑、推理小说的市场定位。

此外，总上榜次数位居第 2 的作者曹文轩的作品在儿童文学与绘本的分布表现突出。而曹文轩作为著名的儿童文学作家，他的作品通常具有深厚的文学价值。出版商可以扩展儿童文学与绘本系列，满足市场需求，并且推广其作品的国际版本，扩大全球影响力。

5.1.4 书籍的价格描述分析

对书籍的原价、售价、折扣比例进行描述分析：

	最小值	最大值	平均值
原价	14.8	828	61.71
售价	2.1	455.4	32.42
折扣比例	0.8	9.9	5.26

表 1 上榜书籍的原价、售价、折扣比例

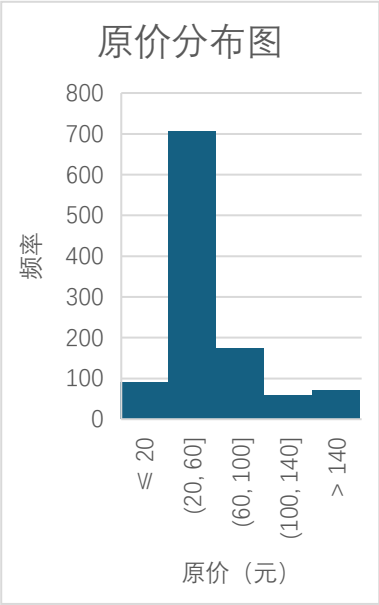


图 6 原价分布图

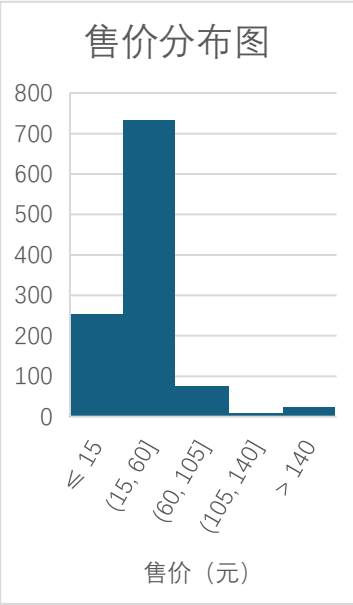


图 7 售价分布图

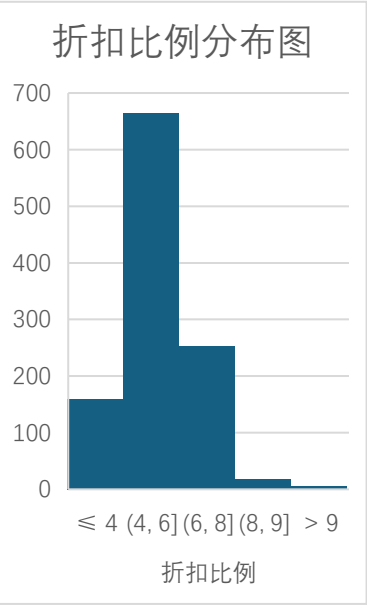


图 8 折扣比例分布图

利用 Excel 的数据透视表以及数据分析功能得出以上图表。

（1）从原价上看，图书的原价范围是 14.8 到 828 元，平均值约为 61.71 元；集中在 20-100 元之间，有少数书籍的原价较高，超过 200 元。

（2）从售价上看，图书的售价范围是 2.1 到 455.4 元，平均值约为 32.42 元；主要集中在 15-80 元之间，分布与原价相似，但整体偏低。

（3）从折扣比例上看，折扣比例的范围是 0.8 到 9.9 折，平均值约为 5.26 折；集中在 4-6 折之间，少数书籍的折扣比例较低（折扣高）。

5.1.5 书籍的评论数描述分析

	最小值	最大值	平均值
评论数	49	3427401	377124.7

表 2 上榜书籍的评论数

从表 4 中我们可以看出：

数据透视表显示评论数的范围从 49 到 3427401 条，平均值约为 377124.7 条。

5.1.6 书籍的推荐值描述分析

	最小值	最大值	平均值
推荐值	0.935	1	0.999

表 3 上榜书籍的推荐值

观察数据得出：

数据透视表显示推荐值的范围从 0.935 到 1，平均值约为 0.999，整体差异不大，均较高。

5.1.7 书籍的上榜次数描述分析

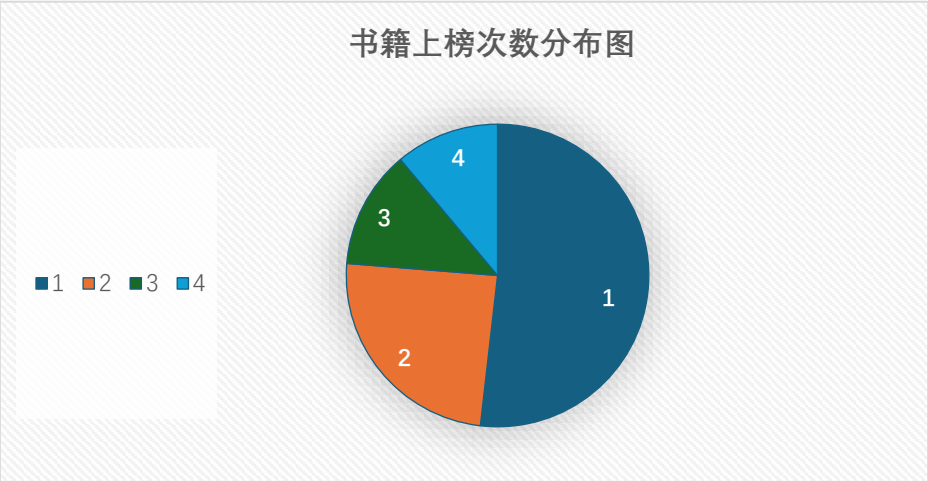


图 9 书籍上榜次数分布图

我们通过观察图 9 可以看出：

只上榜过 1 次的书籍占比较高，可能是有些书籍出版时间在 2020 年以后，但也反映了能够保持 3 次及以上上榜的书籍，其风格特点是受大众青睐的对象。

图中显示，上榜次数大于等于 3 的书籍也占有不小的比重，说明存在一部分书籍常年占据在榜单上，而这类书籍则是大众愿以消费的对象。

为了使出版商可以更好地满足市场需求，接下来将探究上榜次数偏高的书籍的具体类型是什么：

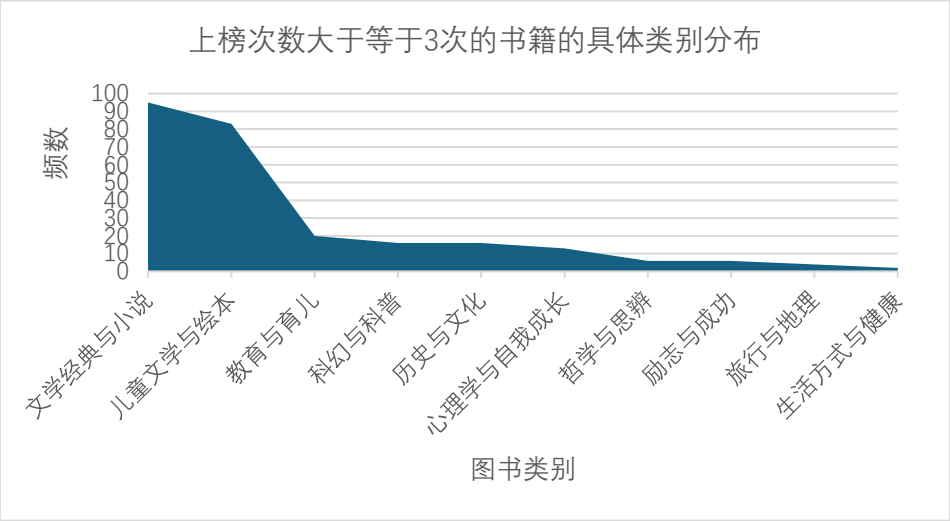


图 10 上榜次数大于等于 3 次的书籍的具体类别分布图

观察图 10 可以得到：

上榜次数偏高的书籍的具体类型大多为文学经典与小说和儿童文学与绘本。说明这两类书籍的市场热度表现较好。

为进一步分析上榜次数偏高的图书类别的销售情况，找出不同图书类别在上榜次数与平均销量中，综合表现较好的图书种类，我绘制了如下这张图：

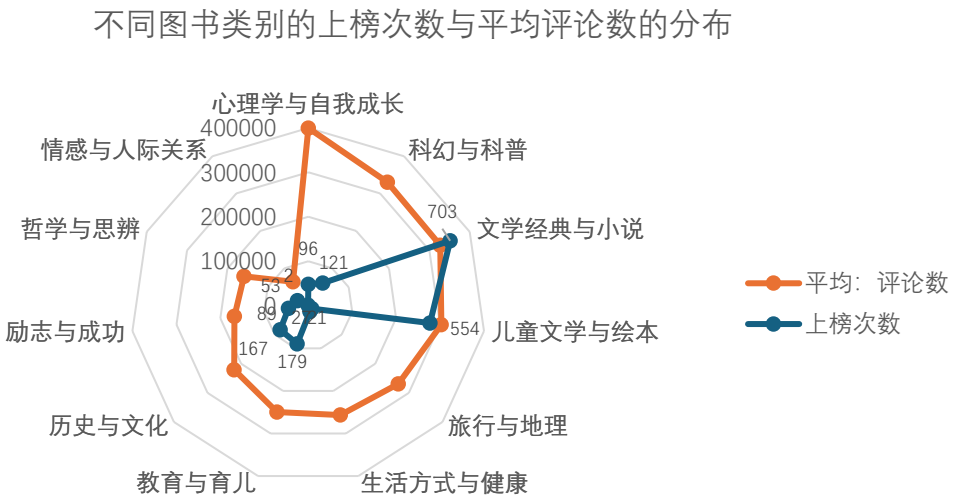


图 11 不同图书类别的上榜次数与平均评论数的分布

观察图 11 我们可以发现：

心理学与自我成长这类书籍的平均评论数最多，可能因其对个人发展的重要性而受到广泛关注，但上榜次数偏低，其综合代表性低，但可能具有一定的发展

潜力，出版商可以多多探究。

而文学经典与小说、儿童文学与绘本这两类书籍的平均评论数和上榜次数都比较高，综合代表性强，意味着文学作品和儿童文学这两类书籍在市场上具有较高的流行度。文学作品往往具有较深的社会和文化影响力，能够吸引大量读者参与评论和讨论。而儿童文学反映了家长和教育者对儿童阅读材料的重视。

因此，出版商可以持续发掘和出版高质量的文学作品和儿童图书；利用社交媒体、读书俱乐部、学校合作等渠道进行营销推广，增加书籍的曝光率和读者的参与度；可以考虑将优秀的国内作品翻译成外语，推向国际市场，同时也可以引进国外的优秀儿童文学和文学作品。

5.2 多变量分析

利用 Excel，探究读者的消费偏好及其影响因素。

为了分析书籍的出版社、作者、图书类别、售价、折扣比例、推荐值、上榜次数、平均排名和销量之间的关系，由于所运用的数据集不包含销量这一字段，而通常来说评论数越多，销量也就越高。因此使用能够量化“销量”的“评论数”作为代理指标。

5.2.1 探究出版社与销量（评论数）的关系

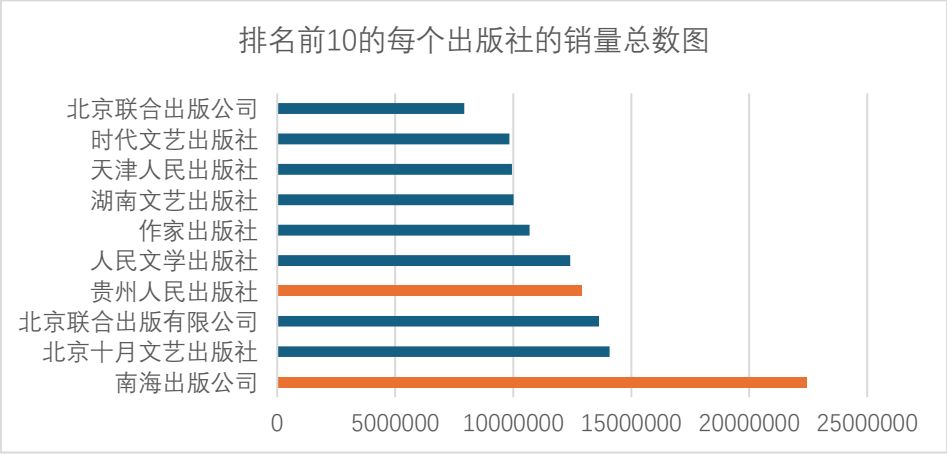


图 12 排名前 10 的每个出版社的销量总数图

从上面的条形图中，可以观察到不同出版社的图书销售总数情况（通过评论数来估计）：

（1） 出版社销售排名：位于底部的出版社拥有最高的总评论数，意味着它们的图书销量可能较高。比如说“南海出版公司”和“人民文学出版社”等在评论总数上表现突出。

（2） 出版社销售差异：不同出版社的图书销售情况存在显著差异，部分出版社表现突出。

以上反映了排名前 10 的每个出版社对整体市场的占有情况，为了探究每个出版社出书的质量，即靠量还是靠质，或者两者兼具，进行了以下分析：

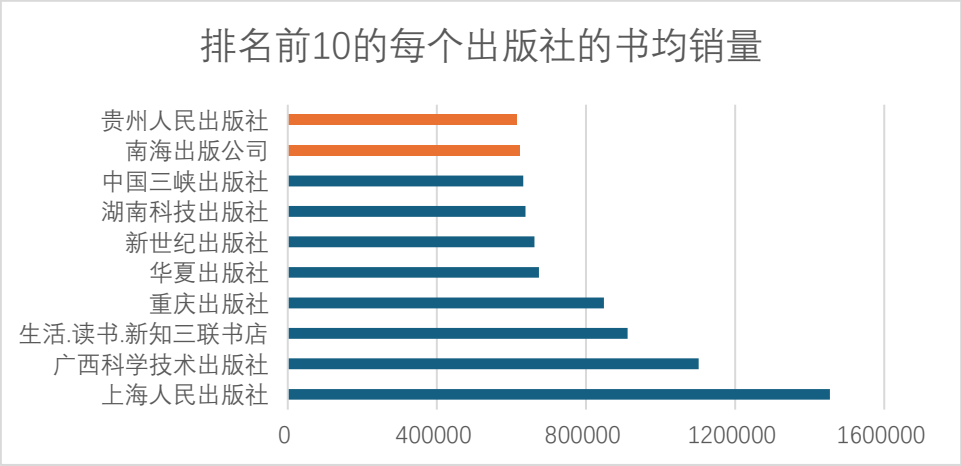


图 13 排名前 10 的每个出版社的书均销量

如图 13，通过对每个出版社的书均销量进行分析，发现上述中销量总数排名前 10 的南海出版公司和贵州人民出版社这两个出版社，仍然出现出版社书均销量前 10 的榜单上，说明了这两个出版社，无论是对整体市场的占有情况，还是其出书的质量，都是相当不错的。

因此销售者在进行图书筛选步骤时，可以优先考虑南海出版公司和贵州人民出版社这两个出版社出版的图书。

5.2.2 探究作者与销量（评论数）的关系

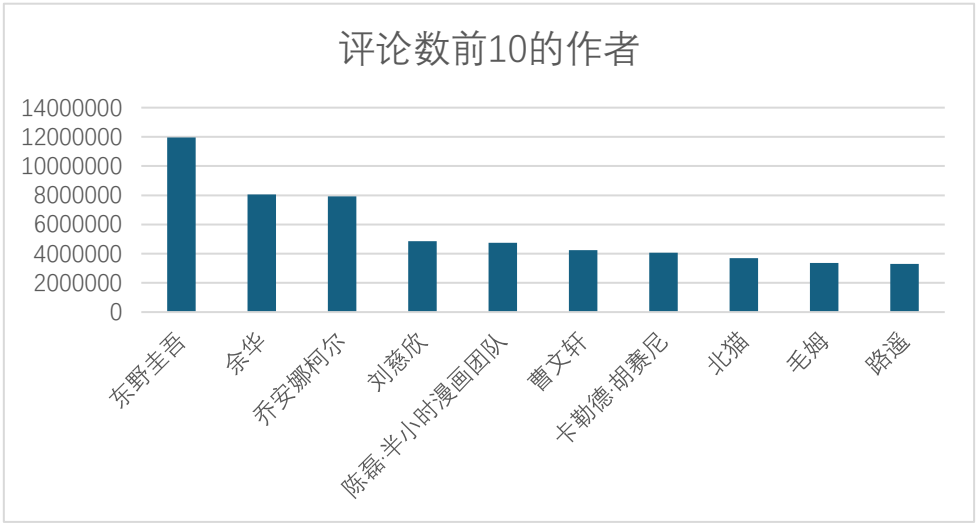


图 14 作者与评论数的关系图

从上面的条形图中，可以观察到评论数前 10 的不同作者的图书销售情况（通过评论数来估计）：

（1） 作者销售排名：位于左侧的作者拥有最高的总评论数，这表明他们的图书销量可能较高。例如：东野圭吾和钱钟书等在评论总数上表现突出。

（2） 作者销售差异：不同作者的图书销售情况存在显著差异，部分作者表现突出。

结合上述单变量分析中作者上榜情况的分析，建议出版商与作者保持良好的关系，确保其作品的稳定输出和长期合作；利用作者的知名度开发多种媒介形式的产品，如电子书、有声书等；出版商可以最大化地利用作者的知名度和图书的销量优势，提高市场占有率和销售业绩。

5.2.3 探究不同图书类别的平均售价、平均折扣比例与评论数的关系

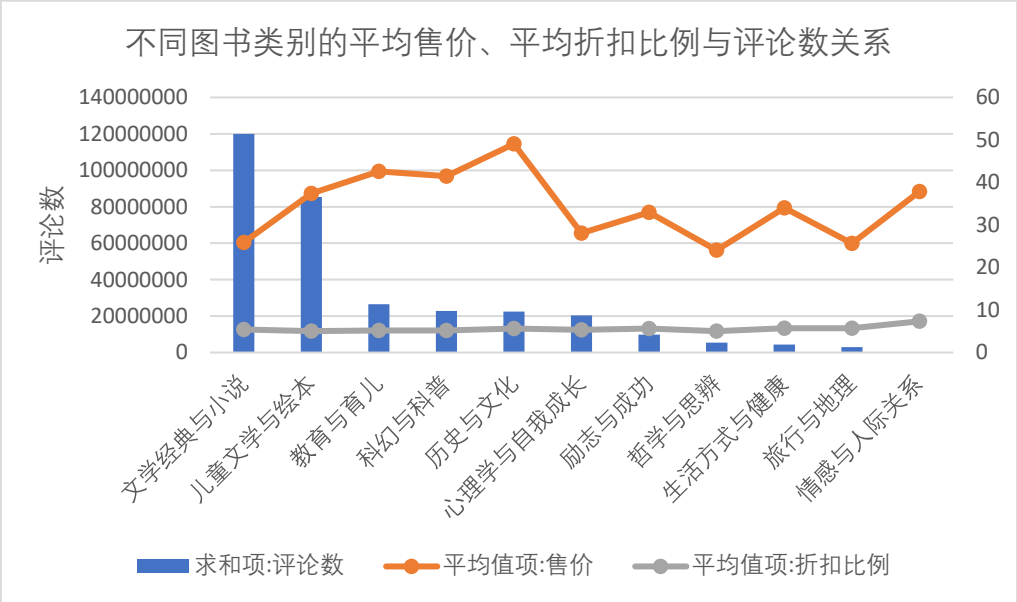


图 15 不同图书类别的平均售价、平均折扣比例与评论数关系图

我们从以上这张图可以发现：

销量最高的两大图书类别分别是：文学经典与小说、儿童文学与绘本。且儿童文学与绘本的平均售价高于文学经典与小说。从整体上看，不同图书类别的平均售价存在差异；不同图书类别的平均折扣比例相似，主要集中在 5 折左右，整体差异不大。

5.2.4 探究售价、折扣比例与销量（评论数）的关系

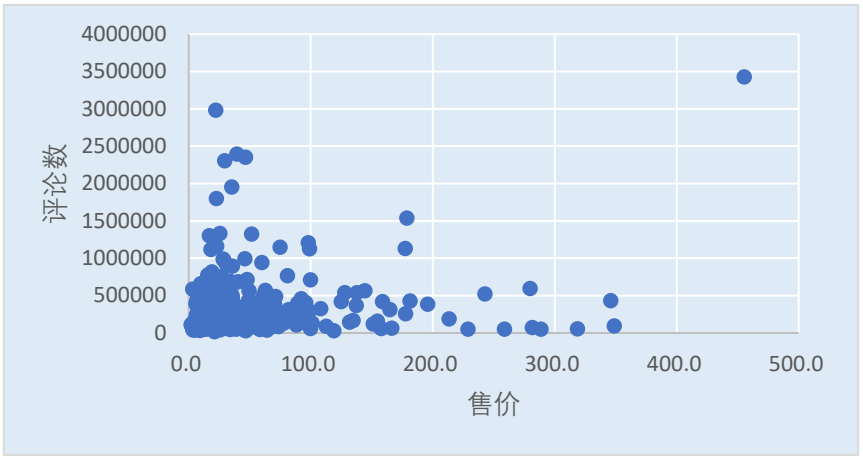


图 16 售价与评论数的关系

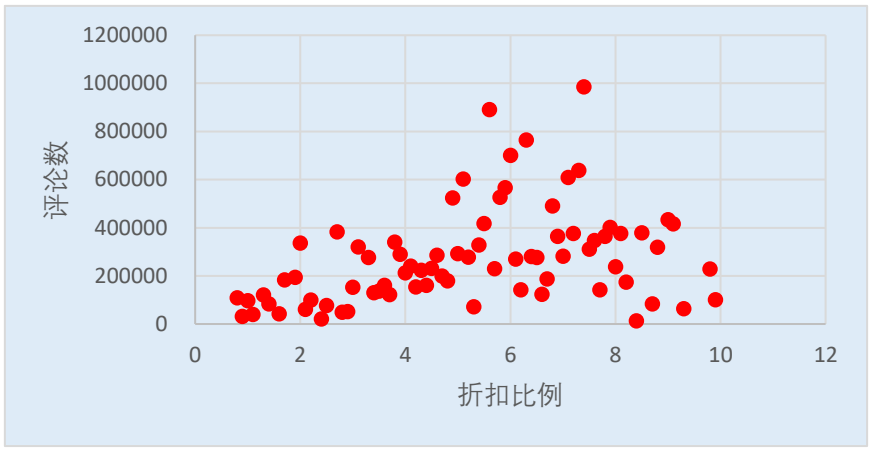


图 17 折扣比例与评论数的关系

从上面的散点图中，可以观察到以下关系：

(1)售价与评论数的关系：

图 11 显示，大部分图书的售价集中在较低的区域。售价与评论数没有明显的相关关系。则说明售价可能不是决定销量的主要因素。

(2)折扣比例与评论数的关系：

图 12 显示，折扣比例主要集中在 4 到 7 折之间。与售价不同，折扣比例与评论数有一定的相关性。通常折扣越低（即折扣比例越高），评论数也越多，这可能意味着销量也越高，说明读者对折扣有一定的偏好。

结合上述分析，销售者可以在特定时期或对特定图书实施限时折扣促销活动，以增加销量和评论数。例如，可以为新书发布或节假日提供特别折扣。

5.2.5 探究推荐值与销量（评论数）的关系

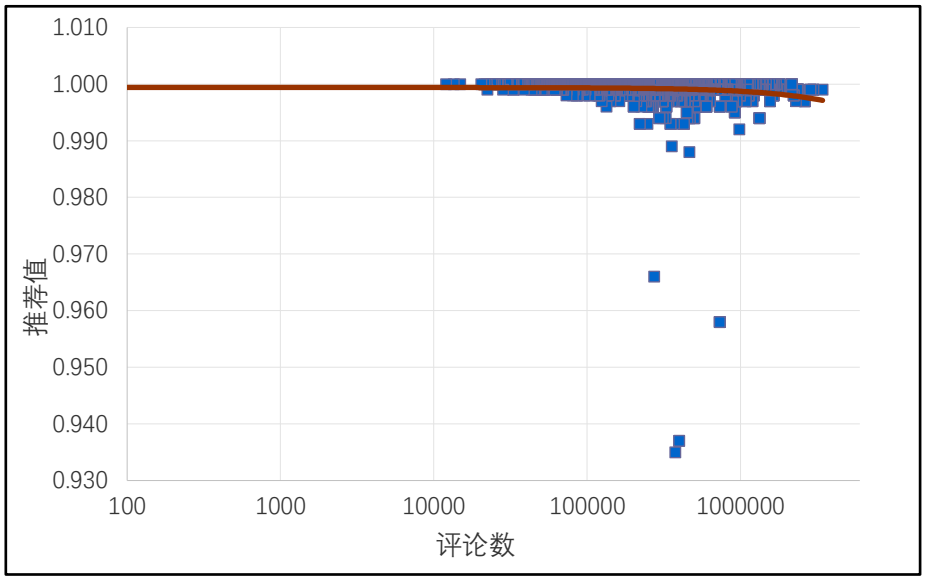


图 18 推荐值与评论数的关系图

从上面的散点图中，可以观察到以下关系：

(1)推荐值与评论数的关系：

大部分图书的推荐值集中在接近 1.000 的区域，推荐值与评论数（销量）之间没有明显的正相关或负相关关系,甚至当评论增多，推荐值有呈下降的趋势。这可能表明推荐值不是决定销量的主要因素。

(2)其他因素的影响：

如出版社、作者知名度、图书类别等，这些因素可能与推荐值共同影响图书销量。

5.2.6 探究平均排名与销量（评论数）的关系

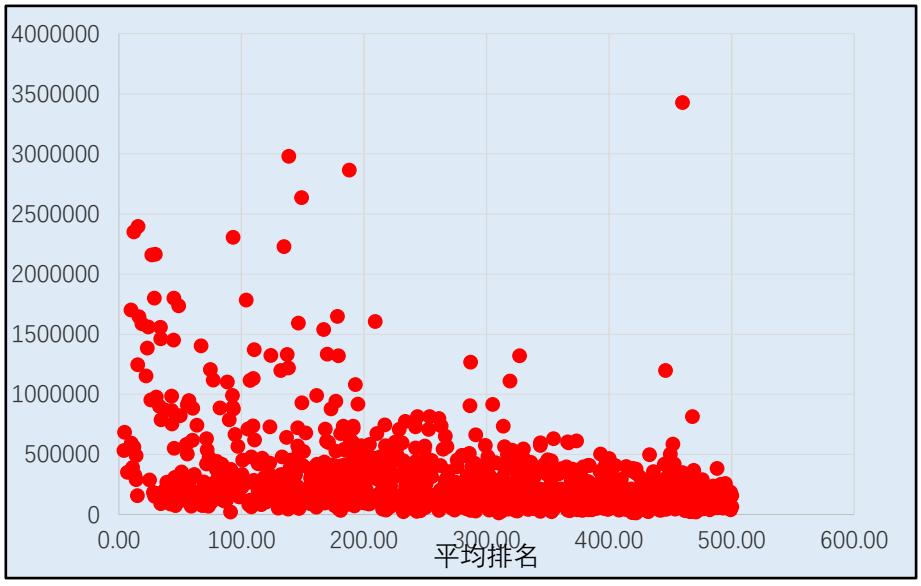


图 19 平均排名与评论数的关系图

从上面的散点图中，可以观察到：

(1)平均排名与评论数的关系：

随着平均排名越靠前，评论数通常会增加，平均排名与评论数之间存在一定的负相关关系，则表明平均排名较低的图书通常销量更高。

(2)排行榜的重要性：

排行榜在图书销售中起着较为重要的作用，排名较高的图书往往能吸引更多的读者和评论。

意味着销售者可以对排名靠前的图书，增加其在销售渠道的曝光度，如在书店的显眼位置展示、在线书店的首页推荐等；将排名靠前的图书与其他图书进行捆绑销售，以提高整体销量。

5.3 相关性分析

通过之前的分析，初步得知评论数与折扣比例、平均排名等因素存在关系。本节将通过热力图的相关系数来分析评论数与其他因素之间的相关关系。

	出版年份	上榜次数	平均排名	评论数	推荐值	原价	售价	折扣比例
出版年份	1.00	-0.27	0.19	-0.25	0.15	0.07	0.07	-0.01
上榜次数	-0.27	1.00	-0.83	0.43	-0.03	-0.07	-0.09	-0.08
平均排名	0.19	-0.83	1.00	-0.58	-0.01	0.06	0.06	0.03
评论数	-0.25	0.43	-0.58	1.00	-0.35	0.07	0.10	0.16
推荐值	0.15	-0.03	-0.01	-0.35	1.00	-0.01	-0.01	0.00
原价	0.07	-0.07	0.06	0.07	-0.01	1.00	0.96	0.00
售价	0.07	-0.09	0.06	0.10	-0.01	0.96	1.00	0.47
折扣比例	-0.01	-0.08	0.03	0.16	0.00	0.00	0.47	1.00

图 20 斯皮尔曼相关性矩阵

我们从热力图中可以看出：

a) 出版年份： 较新的书籍通常有较高的推荐值，但评论数和上榜次数较少，而

且原价和售价也稍微较高；出版年份与评论数存在弱的负相关关系。

b) 上榜次数： 上榜次数与评论数之间有中等强度的正相关关系，这意味着上榜次数越多的书籍的评论数越多。

c) 平均排名：评论数与平均排名有中等强度的负相关关系，意味着排名越靠前的书籍评论数越多。

d) 推荐值： 推荐值与评论数有中等强度的负相关关系，意味着评论数较多的书籍推荐值较低。

e) 折扣比例：折扣比例与评论数之间存在弱的正相关关系。

6.结论与建议

综合所有的分析结果，报告在读者消费行为和偏好方面获得了洞悉。

6.1 结论

（1）读者消费偏好：

a.出版年份与评论数存在弱的负相关关系，意味着较新的书籍的评论数较少，需要借助广告等营销方式，打开新书的销售市场。

b.评论数与平均排名有中等强度的负相关关系，意味着排名越靠前的图书能吸引更多的读者和评论。可能因为是读者偏好在排行榜上搜索并购买热门书籍，也可能是排名靠前的书比较符合大众喜好。

c.评论数与折扣比例之间存在弱的正相关关系，通常折扣越低（即折扣比例越高），评论数也越多。说明读者对折扣有一定的偏好。

（2）固定消费群众：

上榜次数大于等于 2 的书籍比重较大，说明存在一部分书籍常年坐稳在榜单上，其消费群众数量时常较多，且可能拥有相同的画像特征。若能找出该消费群众的画像特征，对图书销量定能提供不少帮助。

（3）热门消费：

a.南海出版公司和贵州人民出版社这两个出版社，无论是对整体市场的占有情况，还是其出书的质量，都是相当不错的。

b.东野圭吾、乔安娜柯尔、刘慈欣、余华等作者的评论总数位居前列。东野圭吾、北猫、余华、毛姆、曹文轩等的上榜次数达到了 25 次及以上，表明这些作者创作了很多优秀的作品，并深受读者的欢迎。

c.文学经典与小说、儿童文学与绘本这两类书籍的平均评论数和上榜次数都比较高，综合代表性强，意味着这两类书籍在市场上具有较高的流行度。而这两

类书籍中最具有代表性的作者分别为悬疑小说的代表作家：东野圭吾，以及著名的儿童文学作家：曹文轩。

d.售价集中在 15-80 元之间。

6.2 建议

a)确保采购图书的质量，对较新的书籍通过广告等营销方式，加大宣传力度，增加访问量，尽可能地提升排名并上榜，增加评论数，打开新书的销售市场。

b)对排名靠前的图书，增加其在销售渠道的曝光度，如在书店的显眼位置展示、在线书店的首页推荐等；将排名靠前的图书与其他图书进行捆绑销售，以提高整体销量。而对于排名靠后的图书，可以考虑适当的折扣策略来吸引更多的购买者，但要注意保持利润率。

c)鼓励满意的顾客留下评论，可以通过提供小额折扣或奖励来实现。这样做不仅可以增加评论数量，而且高质量的评论可能会吸引更多的购买者。

d)合理调整图书市场的供应结构。购书网站应根据读者的需求去采购图书，比如像东野圭吾、余华、北猫等具有知名度的作者的书籍是比较受读者欢迎的；也可根据图书类别按销量的比重去采购，像文学经典与小说、儿童文学与绘本这两类书籍的市场占有比重是较高的。

e)可对读者画像做进一步调研。根据画像对读者进行分层，找到潜在/特定客户，进而采取措施提升图书消费量。

f)销售者在进行图书筛选步骤时，可以优先考虑南海出版公司和贵州人民出版社这两个出版社出版的图书。

g)对于东野圭吾、余华、曹文轩、毛姆、刘慈欣、乔安娜柯尔、北猫等热度较高的作者，建议出版商与作者保持良好的关系，确保其作品的稳定输出和长期合作，为他们提供更多的资源和支持，以保持其作品的持续热度和市场竞争力。

h)出版商应持续发掘和出版高质量的文学作品和儿童图书。一方面强化悬疑、推理小说的市场定位，以抓住稳定客户群体；另一方面扩展儿童文学与绘本系列，满足市场需求，并且推广其作品的国际版本，扩大全球影响力。

7.局限性

一方面，由于本次数据与有关销售量情况的仅有 2020 排名、2021 排名、2022 排名、2023 排名、上榜次数、平均排名、推荐值、评论数等字段数据，不存在销售量这一数据，根据实际情况用评论数替代，因此对这块没有做到很精确的分析。

另一方面，数据未能提供购买者的个人信息特征，无法从消费群体活跃度角度判断是否存在特定或潜在消费人群。

以上是本报告无法避免的局限性。

参考文献

[1]王冰.产业链演化下图书出版企业投资转型研究[D].中南大学,2013.

[2]王文民.中国图书市场分析[D].西南交通大学,2007.

[3]郭友安.图书营销渠道的管理与整合[J].出版科学,2003,(01):38-40.

[4]程小东.我国出版企业图书市场营销策略研究[D].哈尔滨工程大学,2008.

[5]洪玉华.透过数据看上半年图书市场[N].中国新闻出版广电报,2024-07-15(005).DOI:10.28907/n.cnki.nxwcb.2024.001598.

[6]唐薇.当前图书市场现状以及改进措施探究[J].群文天地,2012,(13):15.

附 录

【数据来源】和鲸社区，当当网图书畅销榜单（2020-2023）数据集 https://www.heywhale.com/mw/dataset/66613ba63f6df7924cddb3b