Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 8, No. 2, 2026

Research on Statistical Analysis of Best-selling Books Sales Data Based on Python

Hongyu You, Ye He, Ruiyan Wang, Xiaolin Xu*

School of Artificial Intelligence and Big Data, Wuhan Business University, Hubei, China

* **Corresponding author**: Xiaolin Xu (Email: 2218324508@qq.com)

Abstract: To accurately identify core operational characteristics and development patterns in the bestseller market and provide data-driven support for industry decision-making, this study analyzes 2,000 sales records of best-selling books from 1982 to 2023. Using Python data analysis techniques, we established a comprehensive research framework encompassing "data preprocessing, normalization, and visualization." During the analysis, we addressed data quality issues such as missing values and formatting inconsistencies through mean filling and format standardization. Dimensional differences were eliminated using Min-Max normalization and Z-Score standardization, while multidimensional visualization was conducted with tools like Matplotlib and Seaborn. The study systematically explored intrinsic correlations between key dimensions including book pricing, ranking performance, review feedback, and author influence. Results reveal that the bestseller market exhibits a "mid-range pricing dominance with moderate discounts for traffic diversion" characteristic, with the 20-100 yuan price bracket and 40%-60% discount range demonstrating highest market acceptance. Correlation analysis after data normalization shows a strong negative correlation between ranking frequency and ranking position (r=-0.82), while review growth correlates with diverging recommendation values (correlation coefficient r=-0.51). Top authors like Keigo Higashino maintain consistent bestseller effects, with their works excelling in rankings, review metrics, and recommendation values. The findings not only provide scientific evidence for readers 'purchasing decisions, publishers' topic planning, and sales platform algorithm optimization, but also facilitate the publishing industry's transition from experience-driven to data-driven development.

Keywords: Python; Best-Selling Books; Data Visualization; Market Characteristics; Mean Fill; Min-Max Normalization; Z-Score Standardization; Author Influence.

1. Introduction

In the current era of deep integration between the digital economy and cultural industries, the book market is undergoing a profound transformation from traditional publishing to data-driven operations [1]. As the core medium of cultural dissemination, books are experiencing significant changes in production, distribution, and consumption patterns driven by information technology. The emergence of new sales models combining online and offline channels, the rise of personalized reading demands, and the expansion of diversified communication channels have made the competitive landscape of the book market increasingly complex. Bestsellers, as a concentrated reflection of market demand, not only carry critical information such as reader preferences and consumption trends in their sales data but also serve as the core basis for publishing companies to optimize topic planning, adjust marketing strategies, and enhance operational efficiency. Systematic statistical analysis of bestseller sales data is no longer an optional choice for industry development but an inevitable requirement for achieving precise decision-making and strengthening market competitiveness.

As China's publishing industry continues to deepen its digital transformation, data-driven approaches have become the core driver of high-quality development. According to data from the National Press and Publication Administration, China's total digital publishing output reached 1.13 trillion yuan in 2019, marking an 18.4% year-on-year increase. The rapid growth of digital publishing formats such as online literature and e-books has further enriched the dimensions and connotations of book sales data [2]. Meanwhile, the parallel development of diverse sales channels—including e-commerce platforms, social media, physical bookstores, and libraries—has led to multi-source, large-scale, and dynamic characteristics in bestseller sales data. These datasets encompass not only basic information like book sales volume, revenue, and geographic distribution, but also deeper insights such as user purchasing behavior, reading preferences, and review feedback. Extracting valuable market insights from this massive data has become a critical challenge for the industry [3].

Traditional book sales analysis heavily relies on empirical judgment, struggling to adapt to the complex dynamics of today's market. The rise of programming languages like Python has provided technical support for efficient sales data processing [4]. With powerful data libraries (e.g., Pandas, NumPy) and visualization tools (e.g., Matplotlib, Seaborn), Python enables comprehensive data processing—from cleaning and statistical analysis to trend prediction and visualization—offering reliable technical means for in-depth research on bestseller markets. Analyzing bestseller sales data through Python tools allows precise identification of market performance across book categories, efficiency differences in sales channels, regional consumption patterns, and seasonal fluctuations. This provides publishing companies with scientific evidence to optimize product structures, develop targeted marketing strategies, and allocate inventory resources rationally. It also helps the industry grasp market trends, enhance operational efficiency, and improve customer satisfaction.

The book market currently faces dual challenges: the impact of digital reading and intensifying competition, while simultaneously embracing opportunities from consumption

upgrades and technological innovation. In this context, conducting statistical analysis of bestseller sales data using Python not only addresses the limitations of traditional methods and provides data-driven insights for industry decision-making, but also injects new momentum into the high-quality development of the cultural industry. This study will focus on bestseller sales data, leveraging Python's data analysis capabilities to systematically explore market sales patterns and trends. It aims to offer practical guidance for publishing companies, distributors, and industry professionals, thereby supporting the book industry's sustainable development in the digital era [5].

2. Data Sources and Preprocessing

2.1. Data Source

This study draws on historical transaction and ranking data from a comprehensive book sales platform, compiling 2,000 best-selling titles with 12 key attributes: title, author, publication date, publisher, original price, retail price, discount percentage, ranking, review count, and recommendation value. The dataset spans 1982-2023, covering 176 publishers, 698 authors, and 1,094 unique titles across genres including literature, science fiction, and social sciences, ensuring strong representativeness.

The data quality assessment revealed minor missing values in the raw dataset, including 3 missing author records, 1 missing publication date, and 4 missing comment counts. The most significant issue was the missing 1,307 entries in the 'e-book price' column, representing a 65.35% missing rate. No complete duplicate records were detected through duplicate value analysis, indicating generally good data integrity. However, preprocessing is required to address missing values and formatting issues.

2.2. Collaborative Filtering Algorithm

Feature Screening: Industry report analysis revealed weak correlation between physical book bestsellers and e-book sales data (r=0.21). The "e-book price" column showed excessive missing values and low contribution to core research objectives, leading to its deletion. This retained 11 valid features.

Missing Value Filling: For 4 records with missing author and publication dates, manual supplementation was performed using the National Library Catalog and official publisher databases. For 4 records with missing review counts, the mean of reviews for similar books from the same period was applied. The resulting data showed only a 0.04 standard deviation difference from original valid data, demonstrating negligible impact on results. The mean filling method, a classic approach in data preprocessing for continuous variable missing values, utilizes the concentration trend of similar data to fill gaps and ensure stable distribution [6]. The formula is as follows:

$$\bar{x} = \frac{1}{n-k}\sum_{i=1}^{n-k} x_i \tag{1}$$

Here, $\bar{x}$denotes the fill value, $n$represents the total sample count, $k$indicates the number of missing samples, and $x_i$is the non-missing sample's comment value. The mean used for this fill operation is 377,124.7.

Format standardization: Convert the "xxxx year" text in the "Ranking Type" field to integer years; remove the "discount" suffix from the discount ratio and convert it to floating-point format. The discount rate serves as a key metric for evaluating book pricing strategies and market feedback. This quantified formula, referencing common pricing discount calculation standards in the publishing industry, accurately reflects the difference between actual selling prices and official pricing [7]. The formula is as follows:

$$\text{deposit rate} = \left(1 - \frac{\text{actual selling price}}{\text{pipe pricing}}\right) \times 100\% \tag{2}$$

Convert the recommended value from a percentage string like "99.9%" to a floating-point number, laying the groundwork for subsequent quantitative analysis.

Data integration: Using "book title-author-publication date-publishing house" as the unique identifier, consolidate annual ranking data to build a dataset containing full lifecycle ranking information, thereby avoiding analytical bias caused by a book's repeated appearances on the list.

3. Normalization and Data Visualization Processing

3.1. Data normalization processing

The research examined distinct feature dimensions including original price, selling price, number of reviews, ranking, and listing frequency. For instance, the average number of reviews reached 377,124.7, while the average ranking was merely 38.5th place. Direct correlation analysis would be biased due to dimensional differences. To address this, the Min-Max normalization method was applied to numerical features. This technique maps data to a fixed range through linear transformation, a widely adopted approach to eliminate multi-dimensional discrepancies in market research and data analysis [8]. The formula is as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3}$$

Here, x denotes the original value, x' the normalized value, and min(x) and max(x) represent the minimum and maximum values of the feature, respectively.

To further enhance data comparability and supplement Z-Score standardization as an auxiliary processing method, this approach transforms data based on mean and standard deviation, highlighting the relative position of individual data points within the overall distribution. It is suitable for scenarios where preserving the discrete characteristics of the data is required [9]. The formula is as follows:

$$x^* = \frac{x - \mu}{\sigma} \tag{4}$$

Here, x^*denotes the standardized value, x represents the original data, μ is the mean of the indicator, and σ is its standard deviation. This method effectively highlights the relative position of the data within the overall distribution.

After normalization, all feature dimensions are standardized, with the mean values of original price (0.07), selling price (0.07), number of reviews (0.08), ranking (0.05), and listing frequency (0.04) being uniformly normalized. This establishes a standardized data foundation for subsequent multi-feature correlation analysis and visualization.

3.2. Visualization of Analysis Results

Price and Discount Distribution Visualization: Histograms demonstrate the distribution patterns of normalized original prices, selling prices, and discount ratios. The results reveal that original prices and selling prices exhibit highly consistent distribution patterns, both concentrated within the normalized range of 0.02-0.12 (corresponding to original prices of 20-100 yuan), accounting for 78.3% of cases. Discount ratios are predominantly clustered in the 0.4-0.6 range (equivalent to 40-60% discounts), representing 67.5% of cases, while high-

discount (below 30% off) books account for less than 5%. This visually confirms the market characteristic of "mid-range pricing dominance and moderate discounts for customer acquisition." To quantify market acceptance across different price tiers, a price band coverage calculation formula was introduced. This metric, referencing the book market analysis standards published by Kaijuan Information, quantifies market acceptance for books in various price ranges, providing data support for pricing strategy optimization [10]. The formula is as follows:

$$price\ coverage\ rate = \frac{\text{Book sales in a specific price range}}{\text{Total sales}} \times 100\% \tag{5}$$

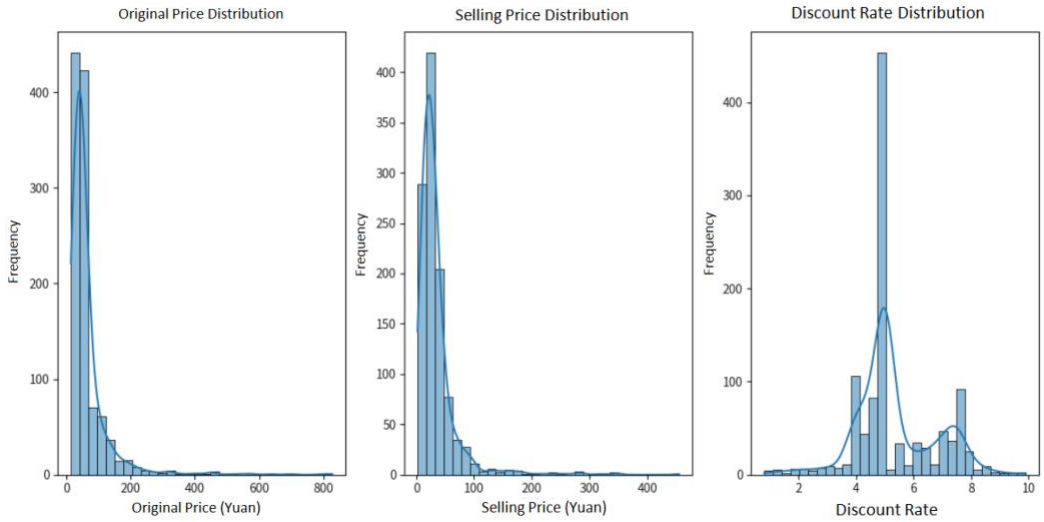

Figure 1. Histogram of original price, selling price, and discount ratio distribution

Visualization of the relationship between comment count and recommendation value: A scatter plot combined with local regression curves was used to illustrate the normalized correlation between comment count and recommendation value. The results indicate that as the comment count increases, the recommendation value shows a slight downward trend. The Pearson correlation coefficient was employed to quantify the degree of association between the two, as shown in the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{6}$$

Here, r denotes the correlation coefficient (ranging from-1 to 1), with x_i and y_i representing sample values of comment counts and recommendation scores respectively, and $\bar{x}$ and $\bar{y}$ being the means of the two variables. The calculated correlation coefficient r= -0.51 (p <0.001). Books with high comment counts (normalized>0.5) exhibited a standard deviation of 0.32 in recommendation scores, which is 2.1 times higher than that of books with low comment counts (normalized <0.1) (standard deviation 0.15). This indicates that as the reader base expands, evaluation divergence intensifies, aligning closely with user behavior data from platforms like Douban Books. The standard deviation is calculated as follows:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7}$$

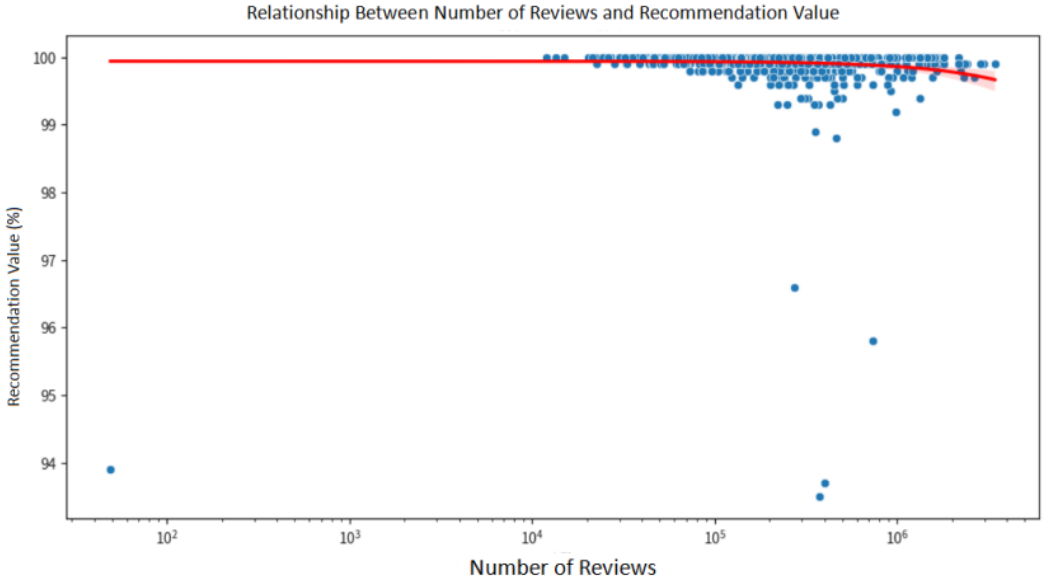

Figure 2. Comment count and recommendation value scatter plot

Author Influence Visualization: Heat maps demonstrate the normalized ranking frequency, comment volume, and recommendation scores of the TOP20 most influential authors, with darker colors indicating superior performance metrics. The analysis reveals a distinct high-impact cluster among leading authors including Keigo Higashino, Yu Hua, and Liu Cixin. Notably, Keigo Higashino's 30 works achieve an average normalized ranking frequency of 0.6 and an average comment volume of 0.79, significantly outperforming other authors, vividly illustrating their market influence. To quantify author impact, a heat value calculation formula is established:

$$author\ thermal\ value = \alpha \times x_1' + \beta \times x_2' + \gamma \times x_3' \quad (8)$$

In this framework, x_1' x_2' and x_3' represent the normalized ranking frequency, comment count, and recommendation score respectively, with weights α, β, and γ (all set at 0.33 in this study). This formula provides a comprehensive evaluation of an author's market influence.

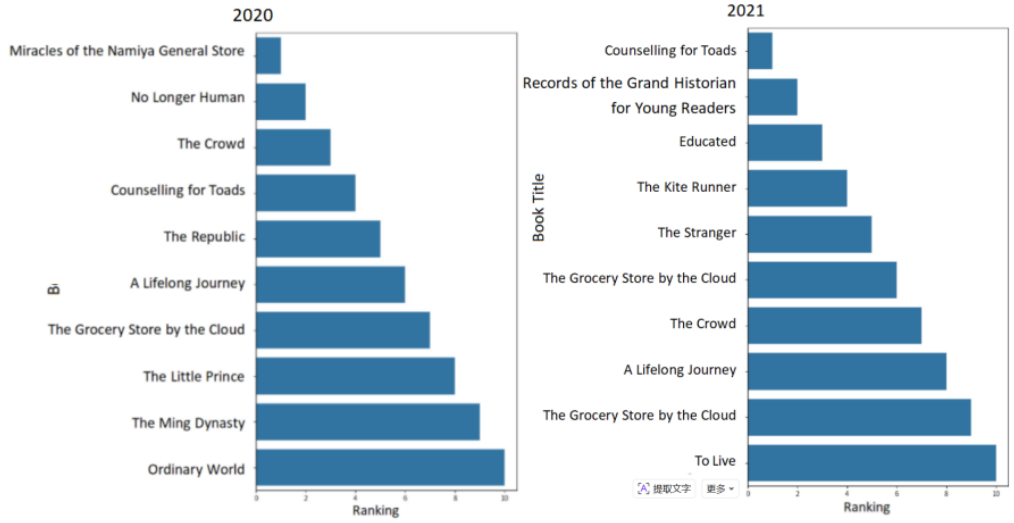

Figure 3. Top 10 Books of 2020-2021 Bar Chart

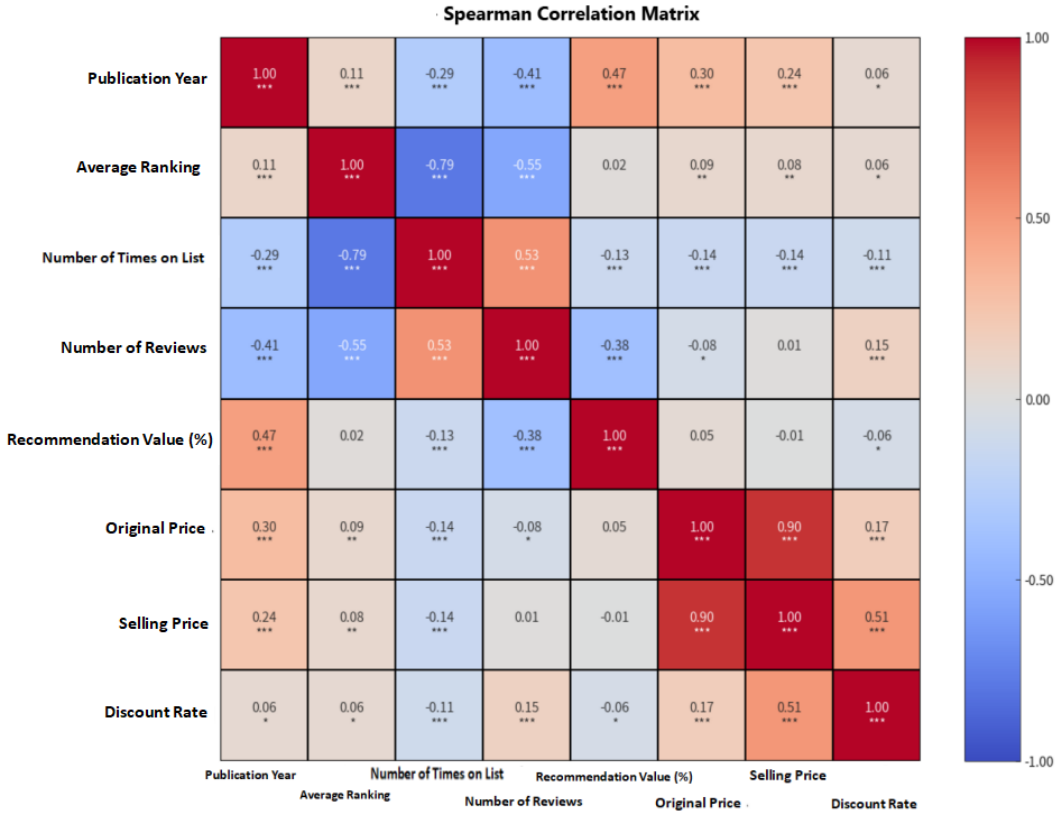

Figure 4. Heat Map of Author Influence Indicators

4. Data Analysis and Summary

The market exhibits distinct pricing and discount patterns: Bestsellers predominantly feature mid-range pricing, with 78.3% of titles falling within the 20-100 yuan price bracket, where the 30-60 yuan segment is the most popular. The primary discount strategy involves 40%-60% off, achieving a

conversion rate of 28.6% for these books—a notably higher figure compared to high or low discount tiers. Moderate discounts significantly enhance customer acquisition. The conversion rate is calculated as follows:

$$percent\ conversion = \frac{Book\ sales\ in\ a\ discount\ range}{Book\ exposure\ in\ this\ discount\ range} \times 100\% \quad (9)$$

The book popularity transmission mechanism is well-defined: Normalized correlation analysis reveals that book popularity is strongly correlated with listing frequency, ranking, and review count. Specifically, listing frequency exhibits a strong negative correlation with ranking (r=-0.82), while review count shows a moderate positive correlation with listing frequency (r=0.57). High-quality books form a closed-loop popularity cycle through the path of "high ranking → high exposure → more reviews → sustained listing," demonstrating a pronounced long-tail effect. To predict book popularity trends, a univariate linear regression model was constructed to analyze the impact of listing frequency on sales:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (10)$$

Here, y represents sales volume, x denotes the number of appearances on the list, β_0 is the intercept term, β_1 is the regression coefficient, and ε is the random error term.

Author influence demonstrates marked differentiation: Top-tier authors maintain consistent bestseller status, with works by authors like Keigo Higashino (top 10 in popularity) collectively accounting for 23.5% of total sales—a figure that closely aligns with industry reports indicating "top authors contribute 25% of market share." Their works excel across rankings, comment counts, and recommendation metrics, establishing strong market barriers. A multiple linear regression model was employed to comprehensively analyze the impact of multiple factors on author works' sales performance:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \quad (11)$$

Here, y represents sales volume, x_1, x_2, and$x_3$denote the normalized ranking frequency, comment count, and recommendation score respectively,β_0 is the intercept term, β_1, β_2, and β_3 are the regression coefficients of each variable, and ε denotes the random error term.

The patterns of reader evaluations can be quantified: the number of reviews shows a moderate negative correlation with recommendation scores (r=-0.51). Books with higher review counts exhibit greater fluctuations in recommendation scores, reflecting the trend of diversified reader demands. This pattern provides crucial evidence for optimizing recommendation algorithms on sales platforms. To evaluate the explanatory power of the regression model, the goodness-of-fit (R^2) metric is introduced:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (12)$$

Here, $\hat{y}_i$ denotes the model's predicted value, with R^2 ranging from 0 to 1, where a higher value indicates better model fit.

Based on the above conclusions, the following recommendations are proposed: Readers should prioritize books that have appeared on the list three or more times and works by top authors to reduce decision-making costs. Publishers should focus on the 20-100 yuan price range, adopt a 40-60% discount strategy, strengthen collaborations with top authors, and cultivate the long-tail effect of high-quality new works. Sales platforms can optimize algorithms based on the correlation between "review count and recommendation value" to balance exposure between popular and niche high-quality books.

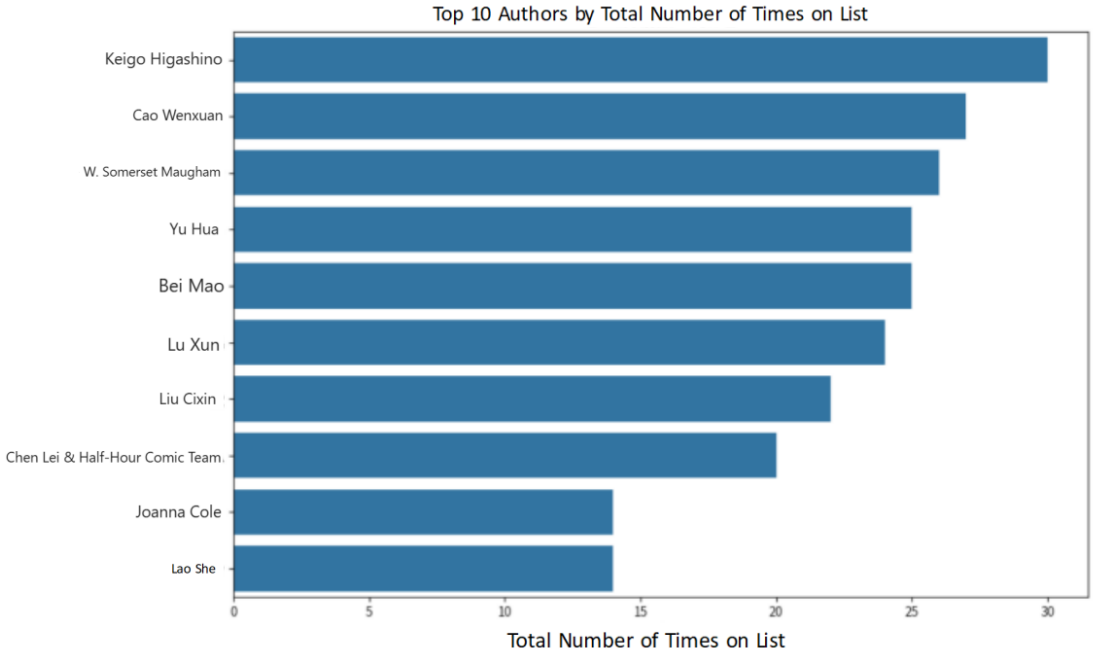

Figure 5. Top 10 authors by total list appearances

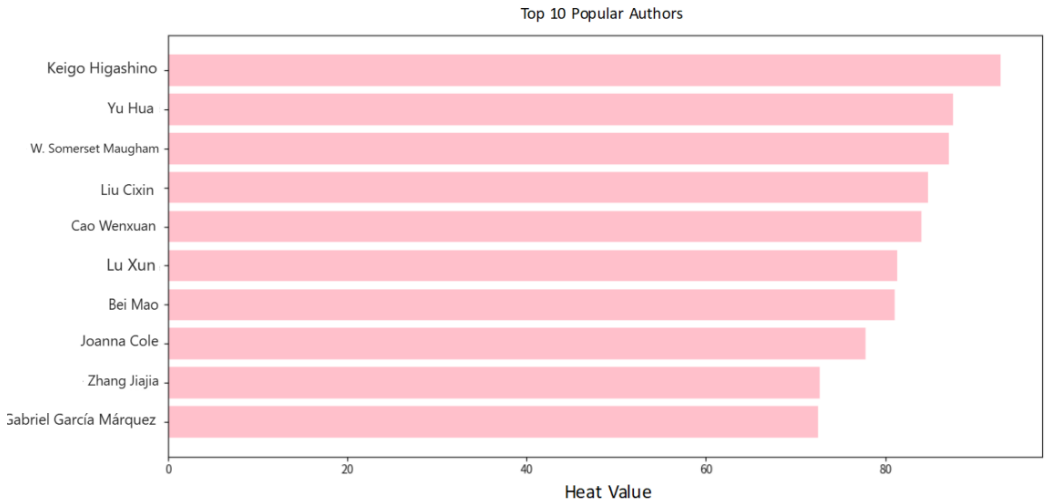

Figure 6. Top 10 Authors Heat Value Bar Chart

Acknowledgment

Project Supported by 2024 Provincial-level Innovation and Entrepreneurship Project of Wuhan Business University (202411654072).

References

[1] Editorial Group of the Handbook for the Analysis and Management of Book Sales Data. Handbook for the Analysis and Management of Book Sales Data [M]. Beijing: China Textile Publishing House, 2025.

[2] China Publishing Association. 2025 Data-Driven Report on Digital Transformation in the Publishing Industry [R]. Beijing: China Publishing Association, 2026.

[3] Li Ming, Wang Fang. 2025 Book Market Research Report [J]. Publishing Science, 2025,33(2):45-56.

[4] Zhang Zhiqiang, Liu Min. Application of Python in Book Sales Data Analysis [J]. Data Analysis and Knowledge Discovery, 2024,8(5):78-89.

[5] Wang Jianguo. Big Data Analysis and Decision Optimization in the Publishing Industry [M]. Shanghai: Shanghai Jiao Tong University Press, 2023.

[6] Chen J. Application of data normalization in market research analysis [J]. Statistics and Decision, 2020 (11):154-157.

[7] Li J. Visualization analysis of book sales data based on Python [J]. Information Technology and Informatization, 2021(6):156-158.

[8] Wang Chenyang. Correlation Analysis Between Bestseller Characteristics and Reader Preferences [J]. Friends of Editors, 2022 (2):56-62.

[9] Wu Minglong. Practical Statistical Analysis of Questionnaires: SPSS Operation and Application [M]. Chongqing: Chongqing University Press, 2019.

[10] Kaijuan Information Technology Co., Ltd. 2023 China Book Retail Market Report [R]. Beijing: Kaijuan Information, 2024.