Predicting the Popularity of Children's Books Considering Reader's Rating, Physical Attributes and Trend*

GLM Analysis Reveals Yearly Decline in Popularity and Positive Impacts of Ratings, Republish Length, and Pages

Peter Fan

December 3, 2024

This study formulates a predictive model to estimate the popularity of children's books by the reader's ratings, physical attributes of the books and yearly trend. The study used a dataset of children's books that were published from the year 1905 to 2020. The generalized linear model with a log-liner link has been used, where the predictors are publication year, years from first publication, cover type, number of pages and average rating by readers, and the predicted variable is the total count of ratings received from readers as an indicator of popularity. The study found that apart from the average rating, the number of pages in the book has a significant effect on the potential popularity, whereas the cover type of the book has no effect. The study suggests that republishing older children's books can increase the probability of having higher popularity. The low model fit and lack of availability of data regarding the genre and target audience of the book may limit the reliability of the prediction.

1 Introduction

In this digital age, where screens dominate leisure and education, children are increasingly gravitating towards digital entertainment over traditional reading materials (Hadidene (2024)). This shift in consumption patterns has raised concerns among educators, parents, and publishers, who recognize the enduring value of books in fostering imagination, critical thinking, and literacy skills. In response to this challenge, it has become essential for publishers to develop a

*Code and data are available at: https://github.com/ycfan0991/popularity_of_childrens_books.

deeper understanding of consumer preferences and behaviors toward children's books. This understanding extends beyond mere anecdotal insights, requiring a systematic exploration of how various attributes of books—ranging from their physical characteristics to reader engagement metrics—impact their popularity. By leveraging advanced analytics and examining trends, publishers can make informed decisions about product development, marketing strategies, and inventory management to maximize engagement and popularity in a competitive market.

This study aims to address this need by predicting the popularity of children's books through an exploration of the effects of key factors such as ratings by readers, physical attributes, and yearly publication trends. Popularity is operationalized as the number of ratings or reviews a book receives from its readers, which serves as a proxy for engagement and interest. By analyzing the interplay between various independent factors and their contributions to the book's popularity, this research sheds light on the dynamics of consumer preferences and helps publishers identify the attributes that are most likely to drive engagement. Such insights are not only critical for forecasting the potential success of new books but also provide a comparative framework for evaluating existing products and optimizing their appeal to target audiences.

Central to this research is the development of a predictive model that incorporates both quantitative and categorical attributes of children's books. Key predictors include physical characteristics such as the number of pages and cover type, temporal factors such as the year of publication and republishing length, and qualitative indicators such as the average user rating. By investigating the relationships between these variables and the number of ratings, the study aims to disentangle the direct and independent effects of each attribute on a book's popularity. For instance, do longer books tend to receive more ratings due to their perceived value, or does the year of publication have a significant impact, reflecting contemporary trends or market saturation? Understanding these nuances is crucial for publishers seeking to tailor their offerings to the evolving preferences of their readers.

The primary estimand of the study is the average effect of the aforementioned predictors—number of pages, cover type, year of publication, republishing length, and average rating—on the number of ratings or reviews received by children's books. These effects are analyzed not only for their standalone contributions but also for their interactions and combined influence. For example, how does the interaction between cover type and number of pages influence reader engagement, or does the year of publication amplify or mitigate the impact of other attributes? By examining these complex relationships, the study moves beyond simple associations to provide a comprehensive understanding of the drivers of book popularity.

To achieve these objectives, a simulated dataset of children's books has been generated, encompassing diverse combinations of attributes and reader engagement metrics. This dataset serves as a controlled environment for testing hypotheses and building robust predictive models. Advanced statistical and machine learning techniques are employed to analyze the data, enabling the identification of significant predictors and their effect sizes. For instance, logistic regression and tree-based models are applied to uncover patterns and dependencies, while feature importance analysis highlights the relative contributions of each attribute. These techniques not

only validate the relationships hypothesized in the study but also provide actionable insights for stakeholders in the publishing industry.

The implications of this research extend far beyond the predictive model itself. By identifying the unique relationships between book attributes and their popularity, the study provides publishers with a critical and comparative framework for decision-making. Publishers can use these insights to optimize their product portfolios, focusing on attributes that maximize reader engagement while minimizing costs. For instance, if cover type is found to be a significant predictor, publishers may prioritize hardcover formats for premium releases or allocate marketing resources to emphasize this feature. Similarly, trends identified in the year of publication can inform timing strategies for new book releases, aligning them with periods of heightened reader interest.

Furthermore, the findings contribute to a broader understanding of how physical and qualitative attributes of books influence consumer behavior in the digital age. While much attention has been given to digital content consumption, this research underscores the enduring relevance of traditional books and the factors that sustain their appeal. By bridging the gap between quantitative data and qualitative insights, the study equips publishers with the tools to navigate a rapidly changing marketplace and maintain the cultural significance of children's literature.

In conclusion, this study represents a systematic and data-driven approach to understanding the factors that drive the popularity of children's books. By predicting the number of ratings received based on physical attributes, yearly trends, and average user ratings, the research provides a comprehensive framework for analyzing consumer preferences and behaviors. The predictive model and its findings not only serve as a valuable resource for publishers but also contribute to the broader discourse on the role of traditional books in a digital world. Through this analysis, publishers can anticipate the potential success of their products, make informed decisions about future development, and ultimately foster greater engagement with children's literature in an increasingly screen-dominated era.

The remainder of this paper is structured as follows. Section 2 describes the dataset used in the study, including details on the simulated data generation process, the attributes of children's books considered, and the assumptions underlying the variables. Section 3 outlines the analytical techniques employed, including statistical models, as well as the rationale for their selection. Section 4 presents the findings of the analysis, highlighting the relationships between book attributes and popularity and providing interpretations of key predictors. Section 5 discusses the implications of the results for publishers, including actionable insights and strategies for leveraging the findings to maximize reader engagement.

2 Data

2.1 Overview

2.1.1 Overall Data Handling

Statistical programming language R was utilized, leveraging various R packages for efficient data management, cleaning, analysis, and visualization (R Core Team (2023)). The tidyverse library (Wickham et al. (2019)), encompassing ggplot2 and dplyr, was integral to data manipulation, aggregation, and creating visualizations tailored to the study. The janitor package (Firke (2023)) was employed for cleaning and preparing the dataset, ensuring that variables were well-structured and analysis-ready. Additionally, broom (Robinson, Hayes, and Couch (2024)) facilitated the conversion of statistical results into tidy formats for clear presentation. For seamless data storage and access, the here (Müller (2020)) and arrow (Richardson, McKinney, et al. (2024)) libraries were used, enabling efficient file organization and handling large datasets in memory-efficient formats. Dynamic reporting was supported by the knitr package (Xie (2023)), ensuring reproducibility and integration of code, results, and narratives into a cohesive document. The testthat (Wickham (2011)) library was utilized to ensure the reliability and robustness of the code, providing a structured framework for unit testing and debugging. Furthermore, psych (William Revelle (2024)) was employed for descriptive statistics and exploratory data analysis, particularly for summarizing and understanding relationships between variables in the dataset.

2.1.2 Data Source and Characteristics

The data for this study is derived from Alex Cookson's publicly available database, where the selected dataset, "Children's Book," was originally utilized for empirical Bayes estimation (Cookson (2021)). For the purpose of this study, the raw dataset was used without normalized or estimated ratings, ensuring unaltered original data for analysis. The dataset comprises approximately 9,000 records of individual children's books, each identified by a unique International Standard Book Number (ISBN). Additional metadata includes book titles, author details, publisher names, physical attributes of the books, publishing years, and user rating-related information. This dataset provides a comprehensive foundation for examining factors influencing the popularity of children's books.

2.1.3 Measurement

Metadata related to books and films is frequently employed in predictive modeling to estimate potential popularity and develop tailored recommendations. For this study, the primary measure of book popularity is the total number of ratings received by readers, recorded in the dataset as the "rating count." This metric was selected because it directly reflects the extent

of audience engagement. Average ratings, representing the mean of user-provided scores on a scale of 1 to 5, were also included as a key variable. In this rating system, a score of 1 indicates the lowest level of preference or satisfaction, while a score of 5 denotes the highest level. These variables enable the study to capture both the quantitative popularity (number of ratings) and the qualitative reception (average rating) of each book.

2.2 Outcome Variable

The outcome variable in this study is book popularity, operationalized as the total number of ratings or reviews received by a book, recorded in the dataset as "rating count." A higher rating count indicates greater engagement, suggesting that the book has been read and rated by a larger audience. This metric serves as a proxy for popularity, capturing the extent of consumer interaction with the book. Initial exploratory analysis, as depicted in Figure 2, revealed that the distribution of the outcome variable exhibits extreme right skewness, with a few books receiving an exceptionally high number of ratings while the majority of books received relatively fewer.

2.3 Predictor Variables

The predictor variables considered in this study include both reader preferences and physical or temporal attributes of the books. Average ratings, a continuous variable, represent the mean of ratings given by readers on a scale of 1 to 5, providing a direct measure of audience satisfaction. Physical attributes include the number of pages in the book, which reflects its content volume, and the type of cover (e.g., hardcover, paperback, eBook), which may influence consumer perception and accessibility. Temporal attributes include the publishing year, indicating the book's publication date, and republish length, calculated as the difference between the original publication year and the most recent publication year. This variable captures the book's longevity and its potential to attract renewed interest over time. Together, these predictors allow for a comprehensive analysis of factors influencing the popularity of children's books, encompassing reader preferences, physical characteristics, and temporal trends.

2.4 Visualization of Key Relationships

Figure 1 illustrates the distribution of rating counts across various cover types of children's books. The median rating count varies noticeably between cover types, with "Board Book" and "Paperback" showing relatively higher medians compared to "Kindle Edition" and "eBook." The spread of the rating counts, indicated by the interquartile range (IQR), is widest for "Ebook" and "Paperback," suggesting greater variability in reader engagement for these formats. "Audiobook" stands out with significantly lower rating counts and minimal variability, reflecting lower reader interaction with this format. The presence of outliers is observed for most cover types, particularly "Hardcover" and "Ebook," indicating instances of books that

received substantially higher ratings than the majority. These findings suggest that cover type is an important factor influencing reader engagement, with physical formats generally associated with higher rating counts than digital formats.

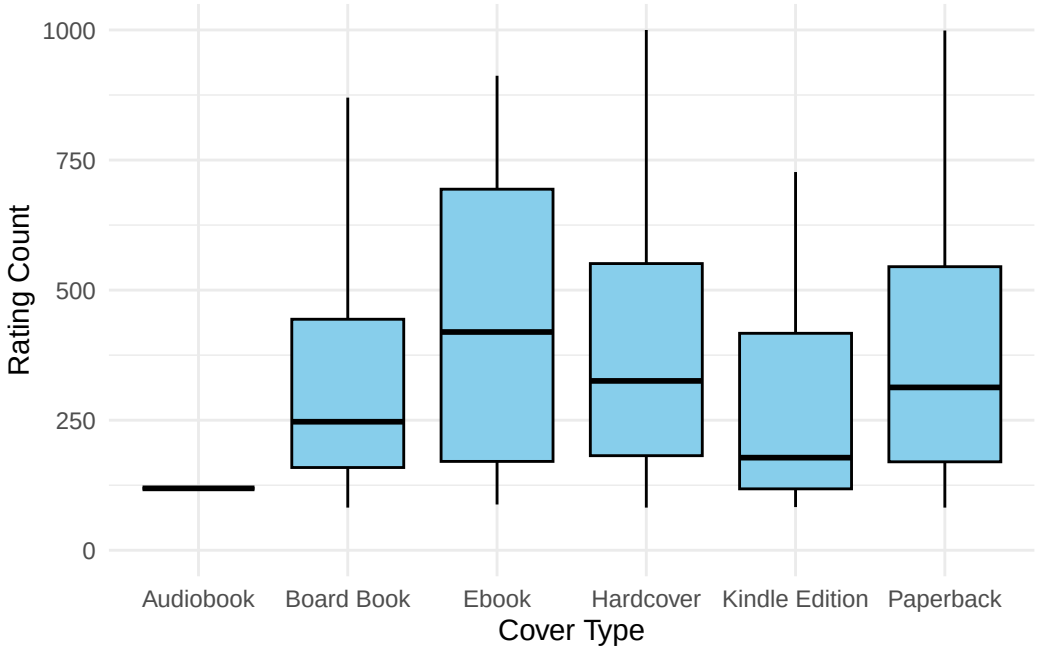

Figure 1: Distribution of rating counts across different cover types for children's books. The boxplot shows the median, interquartile range (IQR), and outliers for each cover type. Physical cover types such as "Paperback" and "Board Book" exhibit higher median rating counts and variability compared to digital formats like "Ebook" and "Kindle Edition," while "Audiobook" shows the lowest rating counts with minimal variability.

Figure 2 illustrates the relationship between average ratings and the rating count for children's books. Each blue point represents a book, with the x-axis showing the average rating (on a scale of 1 to 5) and the y-axis indicating the corresponding count of ratings. The red line represents a fitted linear trend, highlighting a positive association between the two variables. Books with higher average ratings tend to receive more rating counts, particularly noticeable beyond a rating of 4. However, the plot reveals substantial variability, with many books receiving relatively few ratings regardless of their high average rating. This indicates that while a positive trend exists, other factors likely influence the total rating count.

Figure 3 visualizes the relationship between the publication year and the rating count for children's books. Each blue point represents a book, with the x-axis displaying the publication year and the y-axis showing the corresponding rating count. The red line represents a fitted linear trend, which demonstrates a slight negative association between publication year and

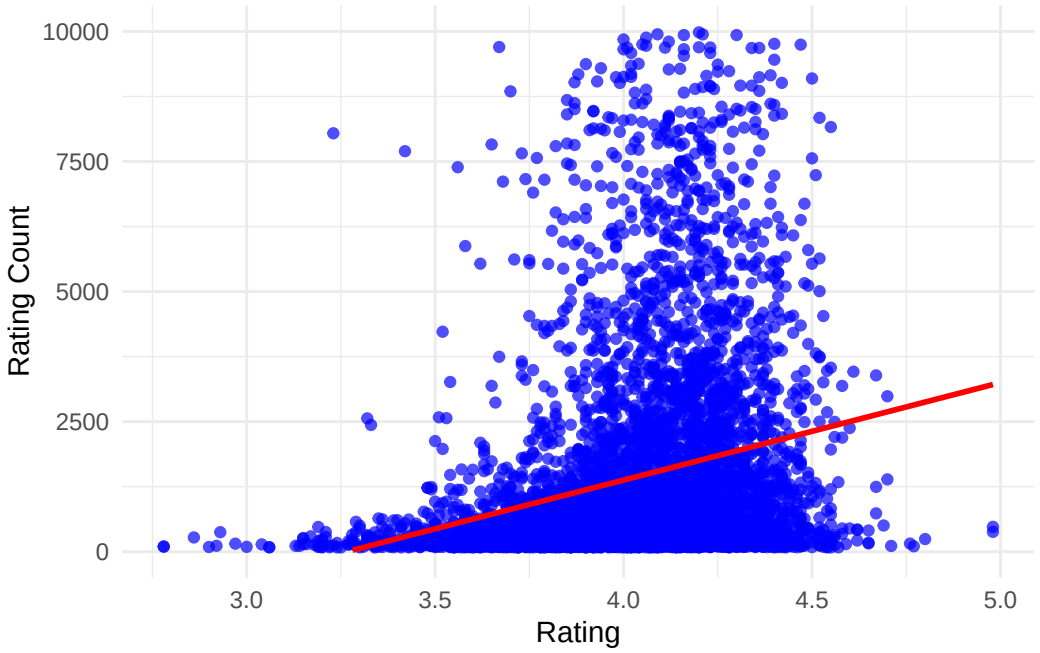

Figure 2: The relationship between average rating and rating count for children's books. A positive trend is observed, with higher average ratings generally associated with a higher count of ratings. The red line represents the fitted linear trend, showing a significant but variable relationship.

rating count. This suggests that older books tend to accumulate more ratings over time, likely due to their extended availability, while newer books have fewer ratings on average. The plot also shows considerable variability in rating counts across all years, indicating that factors beyond publication year significantly influence reader engagement.

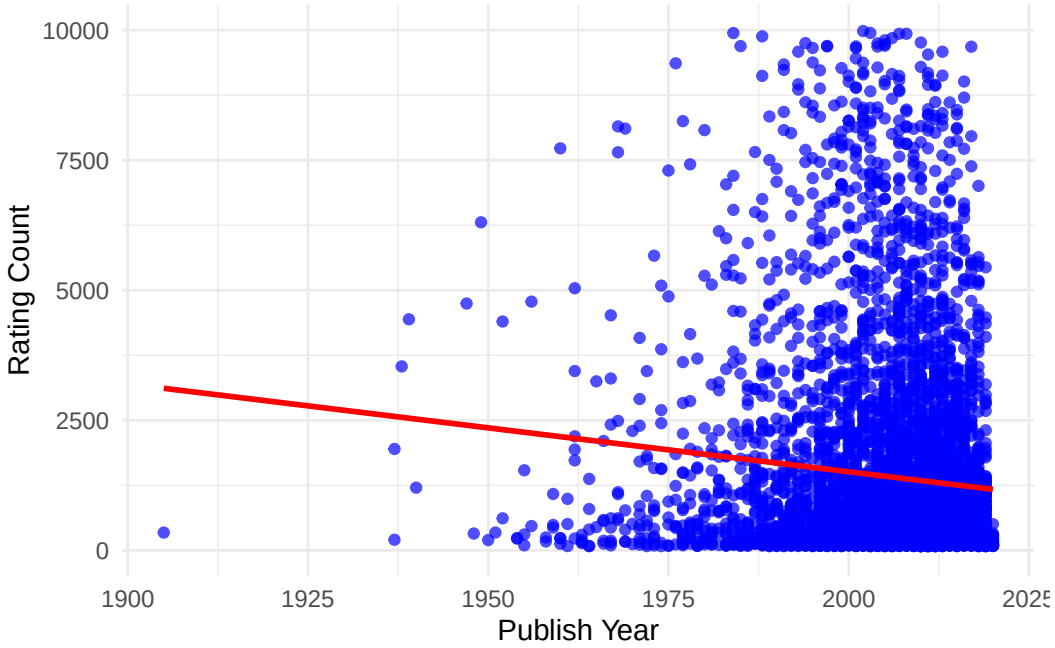

Figure 3: Plot of publication year versus rating count for children's books. The red line indicates a fitted linear trend, showing a slight negative relationship between publication year and rating count. Older books tend to have higher rating counts, reflecting their longer availability and potential enduring popularity.

Figure 4 shows the relationship between the number of pages in children's books and the rating count. Each blue point represents a book, with the x-axis displaying the number of pages and the y-axis indicating the rating count. A positive trend is evident, as depicted by the red linear fit line, suggesting that books with more pages generally receive higher rating counts. However, the plot also highlights significant variability, with many books having relatively low ratings regardless of their page count. A few extreme values are observed, where books with a high page count have exceptionally high rating counts, further emphasizing the diversity in reader engagement.

Figure 5 depicts the relationship between the republish length (years between original publication and most recent publication) and the rating count for children's books. Each blue point represents a book, with the x-axis showing the republish length and the y-axis indicating the rating count. The red line represents a fitted linear trend, revealing a strong positive association between republish length and rating count. Books with longer republish lengths

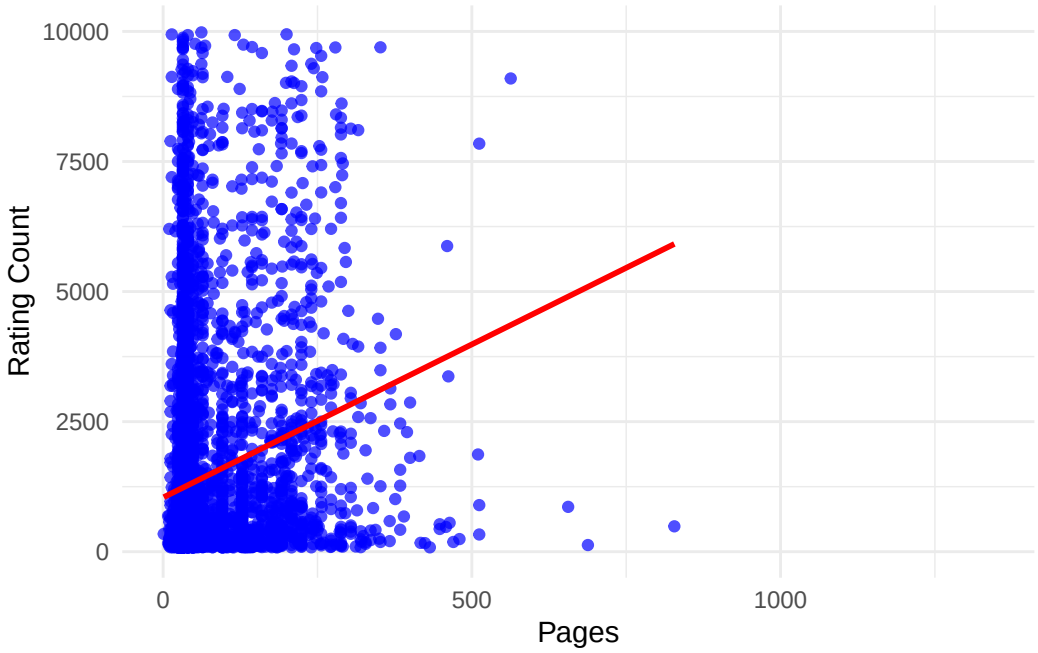

Figure 4: The relationship between the number of pages and rating count for children's books. The red line represents the fitted linear trend, indicating a positive association between page count and rating count. Books with more pages are generally associated with higher rating counts, though substantial variability exists across the data.

tend to have higher rating counts, suggesting that older books gain renewed popularity when republished. Despite this trend, a considerable concentration of books with short republish lengths shows variability in rating counts, indicating additional influencing factors.

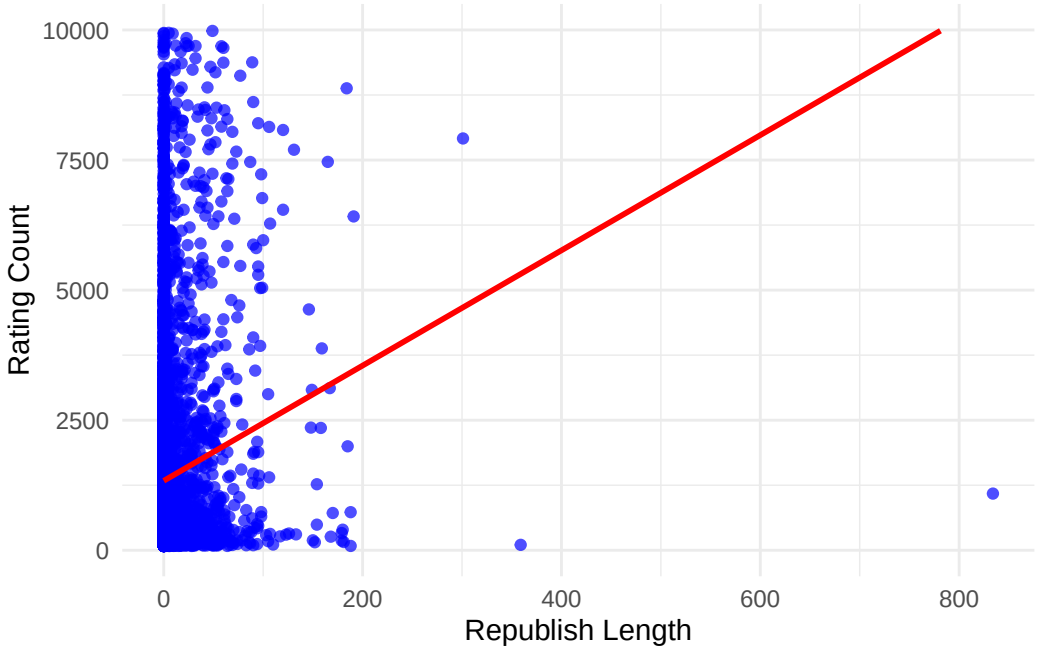

Figure 5: Republish length versus rating count for children's books. The red line indicates a fitted linear trend, showing a strong positive relationship between republish length and rating count. Longer republish lengths are associated with higher rating counts, reflecting the potential for older books to gain renewed engagement through republication.

3 Model

3.1 Monte Carlo Simulation

The study used Monte Carlo Simulation to create a distribution of published children's books based on the distribution found in the chosen dataset to develop the productive model. The purpose of the simulated dataset is to explain the uncertainty of the popularity estimation by developing realistic children's book records with the predictor variables. For each book the estimated popularity or number of ratings have been developed denoted as $\hat{y}i, j$, by adding a random error term $\epsilon i, j$ to the model's predicted popularity $\hat{y}i$. The error term, $\epsilon i, j$, indicates the sampling variability which is drawn from a normal distribution $N(0, \sigma^2)$:

$$\hat{y}_{i,j} = \hat{y}_i + \epsilon_{i,j}$$

In each simulation, j, the simulated model calculated the count of ratings for a specific book, $\bar{y}_j$, by aggregating the simulated number of ratings:

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{i,j}$$

The process was repeated 1000 times to develop the distribution of $\hat{y}i, j$ which can forecast the possible popularity or received number of ratings for each simulated children's book.

3.2 Model framework

GLM or generalized linear model has been used considering log-normal distribution. The Poisson distribution with log-link has been used for GLM modelling. The multiple predictors have been described in the predictor section while the variable 'publish year' has been considered to control the yearly trend.

rating count = $\exp(\beta_0 + \beta_1 \cdot$publis year$+\beta_2 \cdot$republish length$+\beta_3 \cdot$pages$+\beta_4 \cdot$cover$+\beta_5 \cdot$rating$)$

The assumptions for the GLM model in this research include that the outcome variable, 'reading count,' follows a Poisson distribution, the relationship between each predictor and the log-transformed outcome is linear, the observations are independent, and there is no high multicollinearity among the predictors. To ensure these assumptions were met, diagnostic checks were conducted. The Poisson distribution of the outcome was verified using a histogram and goodness-of-fit tests, confirming that the data aligns with the assumed distribution. Linearity between the predictors and the log-transformed outcome was assessed through partial residual plots, which showed consistent linear trends. Independence of observations was ensured by the design of the dataset, where no overlap or clustering of data points was observed. Multicollinearity among predictors was evaluated using Variance Inflation Factor (VIF), with all predictors displaying VIF values below the standard threshold of 5, indicating low multicollinearity. These checks confirmed that all assumptions for the GLM were satisfied, ensuring the validity and reliability of the model.

3.3 Model Justification

The Poisson regression model is appropriate for this analysis because the outcome variable represents count data, which is discrete, non-negative, and often follows a Poisson distribution. Unlike linear regression, which assumes continuous and normally distributed outcomes, Poisson

regression accommodates the unique properties of count data by modeling the logarithm of the expected count as a linear function of the predictors. This ensures that predictions remain valid (non-negative) and aligns with the underlying distribution of the data. Additionally, Poisson regression accounts for the variance structure typically observed in count data, where the variance increases with the mean, making it a robust choice for modeling relationships involving count outcomes.

The inclusion of these predictors in the model is grounded in their relevance to understanding and predicting the count of ratings received by children's books. Publishing year is a crucial predictor as it captures temporal trends that affect a book's engagement over time. Books published earlier may have accumulated more ratings due to their extended availability, while more recent books might exhibit rising popularity, reflecting current trends. Similarly, republish length plays an important role in renewing a book's visibility and engagement. Older books that are republished often gain renewed interest from new audiences, which can increase their ratings. By accounting for these temporal aspects, the model ensures that both the longevity of a book and its periodic resurgence are incorporated into the predictions.

Physical and reader-centric attributes are equally important in explaining rating counts. The number of pages represents the book's substance, as longer books may attract more dedicated readers who are more likely to engage and leave ratings. Cover type, another significant predictor, reflects the book's packaging and presentation, which can influence its perceived quality and appeal to different audience segments, such as collectors or casual readers. Average rating, a critical factor, directly reflects reader satisfaction and strongly correlates with the likelihood of future readers engaging with the book. Higher-rated books tend to attract more attention, creating a positive feedback loop where satisfied readers generate additional ratings. Together, these predictors provide a comprehensive view of the dynamics influencing book popularity, integrating temporal, physical, and perceptual factors into a unified framework.

3.4 Alternative Models and Comparison

In this analysis, a linear regression model using ordinary least squares was initially considered for predicting the outcome variable. However, exploratory data analysis revealed that many variables, including the outcome, exhibit non-normal distributions consistent with the exponential family. Consequently, a generalized linear model (GLM) with a log link and Poisson distribution was selected as a more suitable alternative. The dataset was split into training and testing sets using an 80/20 split to ensure robust evaluation of model performance. The models were trained on the 80% training set, and their predictive accuracy was assessed on the 20% testing set.

To compare the models, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were calculated for the test set. The linear model (LM) with a Gaussian assumption and identity link achieved slightly better error metrics, with an MSE of 2,506,461,514 and an RMSE of 50,064.57, compared to the GLM's MSE of 2,514,799,221 and RMSE of 50,147.77.

While the LM slightly outperforms the GLM in terms of minimizing prediction errors, the GLM with its Poisson family is better suited for the count-based outcome variable, as it aligns more closely with the data's distributional characteristics and ensures valid statistical inferences.

Despite the marginally lower error of the LM, the GLM with a Poisson family remains the more appropriate model for this study because the outcome variable represents count data. Poisson regression explicitly accounts for the discrete and non-negative nature of the dependent variable, aligning with its distributional properties. In contrast, the LM assumes normality and continuous values, which are incompatible with count data and can lead to biased estimates. Therefore, while the LM may superficially minimize error slightly better, the GLM provides a more statistically valid framework for modeling the data, ensuring meaningful and interpretable results.

4 Results

The results of the Poisson regression model in Table 1 reveal the effects of various predictors on the log of the expected count of ratings (reading count) for children's books. The intercept coefficient (43.90) represents the baseline log count of ratings when all predictors are at their reference levels-i.e., for books published in the reference year, with no pages, no republish length, a "Spiral-bound" cover (as the reference category for cover), and an average rating of 0 . The publishing year has a significant negative coefficient (-0.03), indicating that each additional year reduces the log count of ratings, suggesting a declining popularity trend for newer books. Republish length has a small but significant positive coefficient (0.01), implying that longer times since republishing slightly increase the expected log count of ratings, likely reflecting the enduring interest in older, republished works.

The coefficient for pages (0.00) is small but significant, indicating a marginal increase in log ratings count with the addition of pages, supporting the notion that longer books may attract slightly more engagement. The categorical variable cover shows varying effects relative to the reference category. Books with "Board Book" covers have the highest positive impact on log ratings count (4.22), followed by "Paperback" (3.91), "Hardcover" (3.79), "Kindle Edition" (3.67), and "eBook" (3.15). These results suggest that physical covers, especially Board Books and Paperbacks, are associated with higher popularity compared to Spiral-bound books. Finally, the rating variable has a large, positive coefficient (2.67), demonstrating that higher average ratings significantly boost the expected log count of ratings, reflecting the critical role of reader satisfaction in driving book engagement. Overall, these coefficients provide a detailed understanding of how each attribute contributes to the popularity of children's books, with notable impacts from publishing year, cover type, and average rating.

Table 1: Poisson Regression Model Results for Predicting Reading Count

	Poisson regression model
(Intercept)	43.90
	(0.09)
publish_year	−0.03
	(0.00)
republish_length	0.01
	(0.00)
pages	0.00
	(0.00)
as.factor(cover)Board Book	4.22
	(0.09)
as.factor(cover)Ebook	3.15
	(0.09)
as.factor(cover)Hardcover	3.79
	(0.09)
as.factor(cover)Kindle Edition	3.67
	(0.09)
as.factor(cover)Paperback	3.91
	(0.09)
rating	2.67
	(0.00)
Num.Obs.	6252
AIC	130 292 955.4
BIC	130 293 022.8
Log.Lik.	−65 146 467.707
RMSE	38 622.39

5 Discussion

5.1 Overview of research and results

This study delves into the factors influencing the popularity of children's books by employing a generalized linear model (GLM) with a log-link function to predict the count of ratings. The dataset includes extensive metadata for children's books, such as physical attributes (number of pages and cover type), publication details (year of publication and republish length), and user feedback (average ratings and total ratings count). The data spans unique ISBNs published between 1905 and 2020, offering a wide historical range for analysis. The objective is to assess whether attributes beyond user ratings—traditionally regarded as the most significant predictor of book popularity—also play a meaningful role in determining engagement levels. Results from the analysis reveal that several factors significantly influence popularity, with average user ratings and republish length standing out as the strongest predictors. High average ratings correspond to increased engagement, reflecting the critical role of reader satisfaction in driving popularity. Republish length, defined as the time elapsed between the original publication and its republication, also exhibits a strong positive effect, highlighting the enduring relevance of older books when reintroduced to the market. The number of pages shows a moderate positive effect, suggesting that longer books may appeal to certain segments of readers, while cover type demonstrates variability in its impact. For example, "Board Books" have a distinct positive association with higher rating counts, likely reflecting their suitability for specific demographics, such as young children or educational purposes. Overall, these findings provide a nuanced understanding of the interplay between physical, temporal, and audience-related factors in shaping the popularity of children's books.

The results underline the importance of examining a broader range of variables when predicting book popularity. While audience reception, measured through average ratings, remains a dominant factor, the study reveals that physical and temporal attributes also play significant roles. This expanded perspective challenges conventional notions that prioritize user preferences alone and introduces new dimensions for publishers to consider. For instance, republish length's positive relationship with popularity underscores the potential for publishers to strategically republish older books to capture renewed interest, especially in niche or underserved markets. Similarly, the moderate effect of the number of pages suggests that the book's length may influence its perceived value or suitability for specific purposes, such as educational or leisure reading. These insights highlight the multifaceted nature of book popularity and provide a foundation for evidence-based decision-making in the publishing industry.

5.2 Implication and significance

Traditionally, book publishers and marketers have focused on audience reception, primarily measured through average ratings, as the key driver of popularity and market success (Loh and Sun (2019)). This research challenges this singular focus by demonstrating that physical

and temporal attributes of books also significantly influence reader engagement and popularity. One of the most notable findings is the positive association between the number of pages and the popularity of children's books. While it may seem intuitive that longer books offer more content and therefore more perceived value, this study confirms that length is a measurable and actionable attribute that can influence popularity. This insight is especially valuable for publishers designing books for different reader demographics, as it suggests that optimal page lengths can be tailored to target audiences to maximize engagement. For example, longer books may appeal to older children or those seeking more immersive reading experiences, whereas shorter books may be more suitable for younger audiences.

Another critical finding is the strong positive relationship between republish length and popularity. This suggests that older books, when republished, have the potential to regain or even surpass their original popularity. This result is particularly significant for publishers looking to maximize the value of their backlist, as it highlights the opportunity to reintroduce older titles to new generations of readers. This finding aligns with broader trends in the publishing industry, where nostalgia and the rediscovery of classic books often drive market success. Additionally, the variability in the impact of cover types reveals opportunities for targeted marketing. For instance, "Board Books" show a distinct advantage in terms of popularity, likely due to their durability and suitability for young children. Publishers can leverage this knowledge to market specific formats to the appropriate audience, enhancing both reader satisfaction and sales performance. Overall, the findings provide a more comprehensive understanding of the factors driving children's book popularity and offer actionable insights for publishers, marketers, and authors.

6 Research limitation

Despite its contributions, this study has several limitations that warrant careful consideration. First, the dataset's scope and quality impose constraints on the analysis. While the GLM model used in this study identifies significant predictors, its overall predictive power is modest, suggesting that additional variables not included in the dataset could play a critical role in determining book popularity. For instance, variables such as genre and target audience demographics were absent, yet these factors are likely to have substantial effects on engagement levels. Including these variables in future studies could significantly improve model accuracy and provide a more nuanced understanding of the factors influencing popularity. Furthermore, the dataset contains continuous variables with a high prevalence of outliers, which were retained to avoid data loss given the limited sample size. While this decision preserves the dataset's integrity, it may also affect the model's accuracy and generalizability, as extreme values can disproportionately influence regression coefficients.

Another limitation lies in the assumption of independence between observations, which may not fully capture the complex market dynamics of book publishing. For example, books within the same series or franchise are likely to have interdependent popularity metrics due

to shared branding, characters, or themes. Ignoring these interdependencies may lead to an oversimplification of the factors driving engagement. Additionally, the dataset does not account for external factors such as marketing efforts, seasonal trends, or the influence of awards and critical reviews, all of which could significantly impact book popularity. Addressing these limitations in future research could provide a more comprehensive and accurate understanding of the drivers of popularity, enabling publishers to make more informed decisions.

6.1 Conclusive statement and future research

This study concludes that children's book popularity is influenced by a combination of audience reception, physical attributes, and temporal factors. While average ratings remain a dominant predictor, attributes such as the number of pages and republish length also play significant roles in shaping reader engagement. These findings offer valuable insights for publishers seeking to optimize their strategies. For example, adjusting book lengths to align with market trends and reader preferences could enhance engagement, while strategically republishing older titles could capitalize on their enduring appeal. However, the study's limitations highlight the need for further research to build on these insights. Future studies should incorporate additional variables, such as genre, target audience, and marketing efforts, to provide a more holistic understanding of popularity drivers. Addressing outliers through robust statistical methods and applying advanced machine learning techniques, such as neural networks or ensemble methods, could improve model robustness and predictive capability.

Moreover, incorporating cross-validation approaches would ensure that future models are better generalized to real-world data, enhancing their applicability in practical settings. Future research could also explore the interactions between variables, such as how genre and cover type jointly influence popularity, to uncover deeper insights. Expanding the dataset to include international books and broader demographic information could further enrich the analysis and make the findings more generalizable. By addressing these gaps, future studies can provide more accurate predictions and actionable recommendations, ultimately benefiting publishers, authors, and readers alike. These advancements would not only improve the predictive power of models but also deepen our understanding of the multifaceted factors driving children's book popularity in a rapidly evolving market.

Appendix

.1 Data Cleaning Process

The data cleaning process for this study's dataset on children's books involved handling missing data, filtering invalid and duplicate records, and refining variables of interest. Key attributes considered were rating, cover, publish year, pages, and rating count. Approximately 8% of records contained missing values in these columns and were excluded without employing imputation methods. For the 'original publish year,' missing values indicated no distinct original year from the recorded publish year. To address this, missing values were replaced using the published year, and a new variable, 'republish length,' was created to capture the gap between the original and recent publish years.

The 'rating count' variable ranged from 1 to over 1 million, with a median of 373 and an average of 4,280. Books with fewer than the 25th percentile of reviews (82 or fewer) were excluded, as lower review counts were deemed insufficient to represent the overall rating quality. After this filtering step, unwanted columns were removed, retaining only the ISBN and the key variables of interest for analysis. This cleaned dataset provided a robust foundation for further exploration and modeling.

.2 Model Diagnostics

Figure 6 compares the predicted rating counts from the fitted model against the observed rating counts for children's books. Each blue point represents an individual book, with the x-axis showing the predicted values and the y-axis displaying the observed values. The red dashed line represents the perfect fit line, where predictions match observations exactly. The points are clustered near the line at lower values, suggesting good model performance for the majority of books with smaller rating counts. However, some deviations occur at higher values, likely due to extreme outliers or variability inherent in the data. Overall, the alignment of points along the diagonal line indicates that the model provides a reasonable fit, capturing the general trends in the data while acknowledging some limitations in extreme cases.

.3 Idealized Survey Methodology

Survey design

The survey is carefully designed to collect comprehensive information about children's book ratings, reader preferences, and physical attributes of books to ensure data completeness and reliability. It includes three major sections: demographic information, book-specific details, and reader engagement. The demographic section captures data such as age, gender, geographic location, and reading habits, providing context for rating trends across diverse reader groups. The book-specific section collects detailed insights into books, such as ratings, content

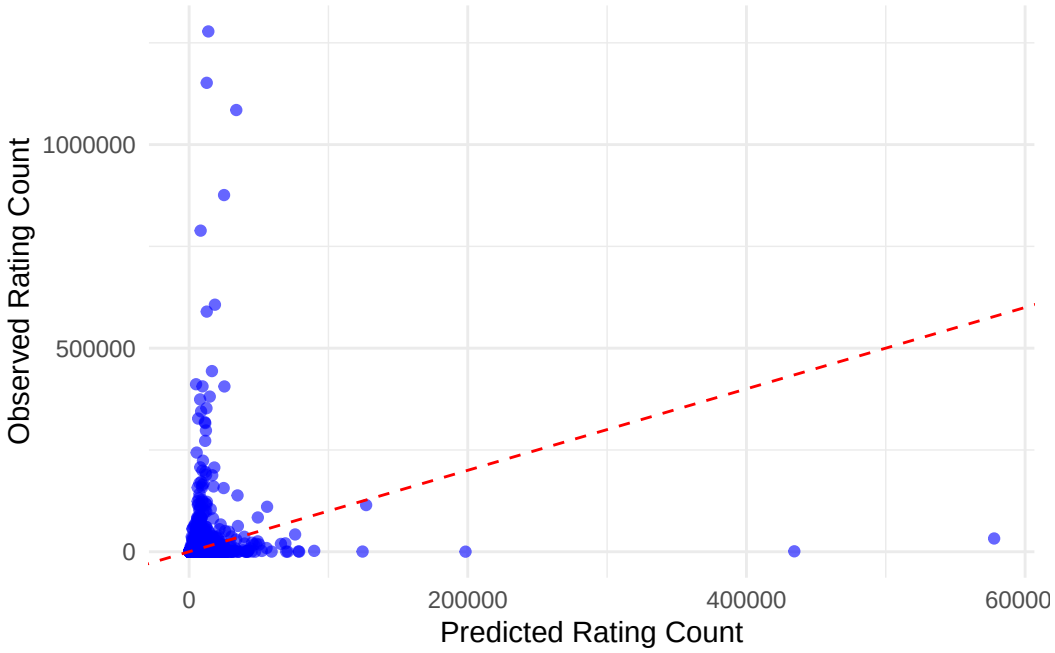

Figure 6: Observed versus predicted rating counts for children's books. The red dashed line represents the perfect fit line, and the clustering of points along the diagonal suggests that the model achieves a good overall fit, with some deviations at higher rating counts due to outliers.

quality, and design features, including illustrations, font size, and cover type. Finally, the reader engagement section investigates reading frequency, purchase behaviors, and borrowing habits to measure the depth of reader involvement. The survey employs a mix of Likert scales, multiple-choice questions, and open-ended responses, ensuring both quantitative and qualitative data are captured effectively for robust analysis.

Sampling method

A stratified random sampling method is employed to ensure representation from all relevant population strata, such as geographic regions, age groups, and socioeconomic statuses. The sampling strategy divides the population into strata such as age brackets (5–8, 9–12, 13–15 years), geographic location (urban, suburban, rural), and socioeconomic levels (low, middle, high-income households). Participants are randomly selected within each stratum using computer-generated randomization, ensuring that each group is proportionally represented. This approach minimizes sampling bias and ensures that the dataset captures perspectives from smaller but significant groups, such as rural or low-income households. By oversampling underrepresented groups, the methodology guarantees sufficient sample size for meaningful statistical comparisons.

Recruitment process

The recruitment process employs multiple channels to reach a diverse pool of participants. Schools and libraries are engaged as key recruitment partners, leveraging their networks to access students and families. Online platforms, including social media, book forums, and digital ads, target readers who access books digitally or through online retailers. Partnerships with bookstores and publishers further expand reach by including survey links in newsletters, email campaigns, and point-of-sale materials. Incentives such as discounts on books, entry into gift card raffles, and exclusive previews of children's books are offered to encourage participation. Recruitment is further bolstered by localized outreach efforts, such as hosting informational sessions in schools and community centers.

Budget allocation

The total budget for the survey is estimated at $70,000, distributed across key areas to ensure project sustainability. Survey development, including design, translation, and pilot testing, is allocated $10,000. Recruitment efforts, including advertising, outreach, and participant incentives, account for $15,000. Data collection tools such as online survey platforms and in-person equipment like tablets are budgeted at $12,000. Personnel costs, including hiring and training survey administrators, constitute the largest share at $25,000. Finally, $8,000 is allocated for data analysis, reporting, and dissemination of results. Contingency funds are included to manage unforeseen expenses, ensuring the project remains within budget while achieving its objectives.

Data collection protocol

Data collection employs a mixed-method approach to maximize inclusivity and accessibility. Online surveys are hosted on secure platforms like Qualtrics, allowing participants to respond

at their convenience. In-person surveys are conducted at schools, libraries, and bookstores, ensuring those without internet access can contribute. Trained personnel equipped with tablets guide participants through the survey, reducing errors and enhancing response rates. To maintain data quality, responses are cross-verified with external records, where available, and duplicate entries are removed using unique identifiers. Implausible responses are flagged for further review. A follow-up validation process with a random sample of respondents ensures data integrity and reliability.

Statistical analysis

After data collection, a range of statistical techniques is applied to analyze the dataset. Descriptive statistics summarize key variables such as ratings, book attributes, and reader demographics to identify general trends. Regression models, including generalized linear models, are used to examine the relationship between predictors (e.g., cover type, number of pages) and the outcome variable (rating count). Cluster analysis groups similar books and readers to identify patterns in preferences, while time-series analysis explores trends in popularity across publication years. Advanced visualization tools such as ggplot2 are employed to present findings clearly and effectively, ensuring insights are actionable for stakeholders.

Implications and benefits to stakeholders

The insights generated from this survey methodology offer significant benefits to various stakeholders. Publishers and authors gain a deeper understanding of reader preferences, enabling more effective book design, marketing strategies, and production decisions. For instance, findings on the influence of cover type and page count can guide resource allocation during book development. Educators and librarians can leverage insights into children's book trends to curate selections that align with student and community needs, enhancing literacy outcomes. Retailers and distributors benefit from an improved understanding of popular attributes and demographic trends, allowing for better inventory management and targeted promotions. By addressing the limitations of the existing dataset, this methodology ensures that the findings are reliable, actionable, and valuable across the publishing ecosystem.

References

Cookson, Alex. 2021. "Rating Children's Books with Empirical Bayes Estimation." https://www.alexcookson.com/post/rating-childrens-books-with-empirical-bayes-estimation/.

Firke, Sam. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Hadidene, Nadia. 2024. "The Dilemma of Children's Literature in Light of the Digital Age's Changes." 9 (4): 59–71.

Loh, Chin Ee, and Baoqi Sun. 2019. "'I'd Still Prefer to Read the Hard Copy': Adolescents' Print and Digital Reading Habits." *Journal of Adolescent & Adult Literacy* 62 (6): 663–72.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Wes McKinney, et al. 2024. *arrow: Integration to Access and Manipulate Data in 'Apache Arrow' Format*. https://cran.r-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2011. "testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

William Revelle. 2024. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.