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 2, 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 has been used considering the `R` packages for data storing, data cleaning and pre-processing (R Core Team (2023)). The library `tidyverse` (Wickham et al. (2019)) which includes `ggplot2`, `dplyr` and other essential packages has been used for data handling and visualization. The library of `janitor` (Firke (2023)), `broom` (Robinson, Hayes, and Couch (2024)) have been used for data cleaning and structuring the results. The library `here` (Müller (2020)) and `arrow` (Richardson, McKinney, et al. (2024)) have been used to efficiently store and retrieve the data.

2.1.2 Data Source and Characteristics

The data has been collected from Alex Cookson's database, where the chosen dataset 'children's book' was previously used for empirical Bayes estimation (Cookson (2021)). The raw data of the study was taken without the normalized and estimated ratings. The raw dataset contains 9000 records of individual children's books with ISBN, title, author details, publisher's name, physical attribute of book, publishing years and different user rating-related details.

2.1.3 Measurement

The meta-data of books and films are often used for predicting possible popularity while developing appropriate recommendations for books and film development to have higher potentiality of popularity. In order to examine the popularity of the book while considering the scope and limitation of the chosen dataset, this research considered the number of ratings given by the readers as the measure of the popularity of the book. The average rating of the book is measured by taking the average of the total ratings received by the readers. The readers have given a rating from 1 to 5, where 1 implies the lowest rating or lowest preference and 5 implies the highest rating or highest level of preference.

2.2 Outcome Variable

The outcome variable is the popularity of the book, which is measured by considering the column 'rating count', which indicates the total number of review ratings received by the consumers of the book. The higher number of ratings indicates the higher number of readers and the higher number of readers indicates the popularity of the book. Figure 2, represents that the outcome variable is distributed with an extremely right skewness.

2.3 Predictor Variables

The predictor variables are average ratings given by the reader with a decimal or float variable. The average rating of the book indicates the readers' preference for the book, which is an obvious predictor of the popularity of the book. The physical attributes of a book include physical attribute includes the number of pages in a book and the physical property of the cover. The publishing year represents the publication year of the book. The republication length indicates how old the original publication of this book is from its recent publication.

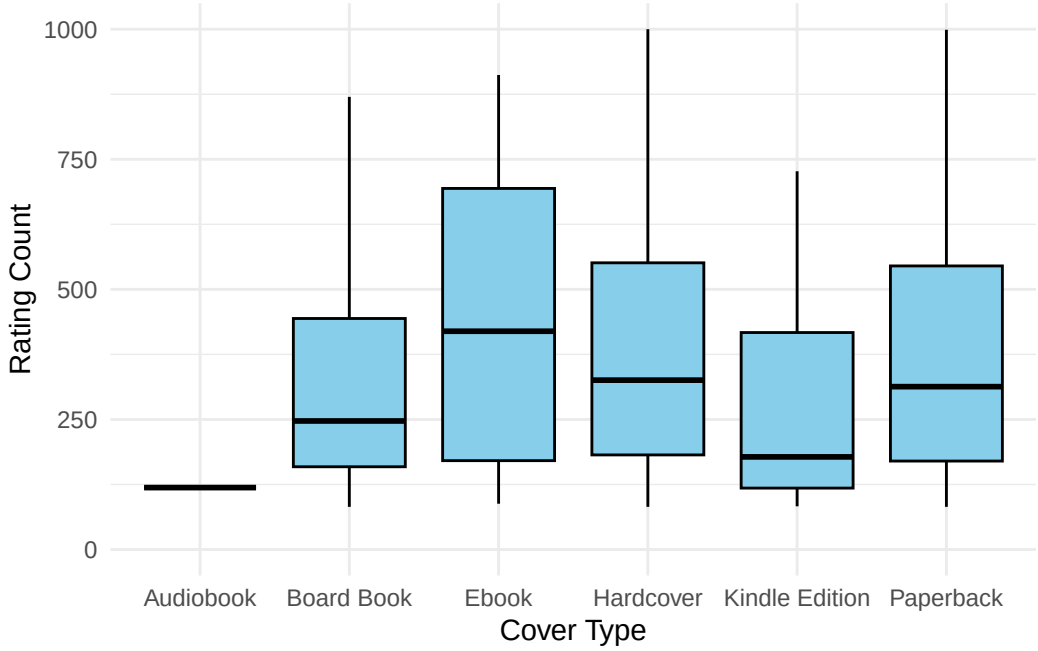


Figure 3: Scatter Plot of Rating vs. Rating Count up to 10000

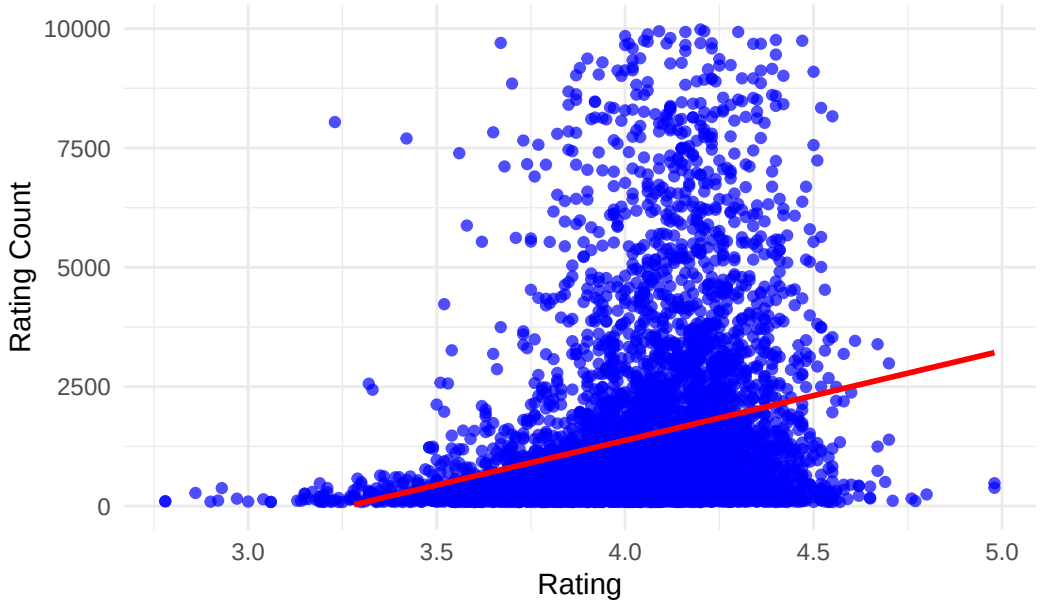


Figure 4: Scatter Plot of Publish year vs. Rating Count up to

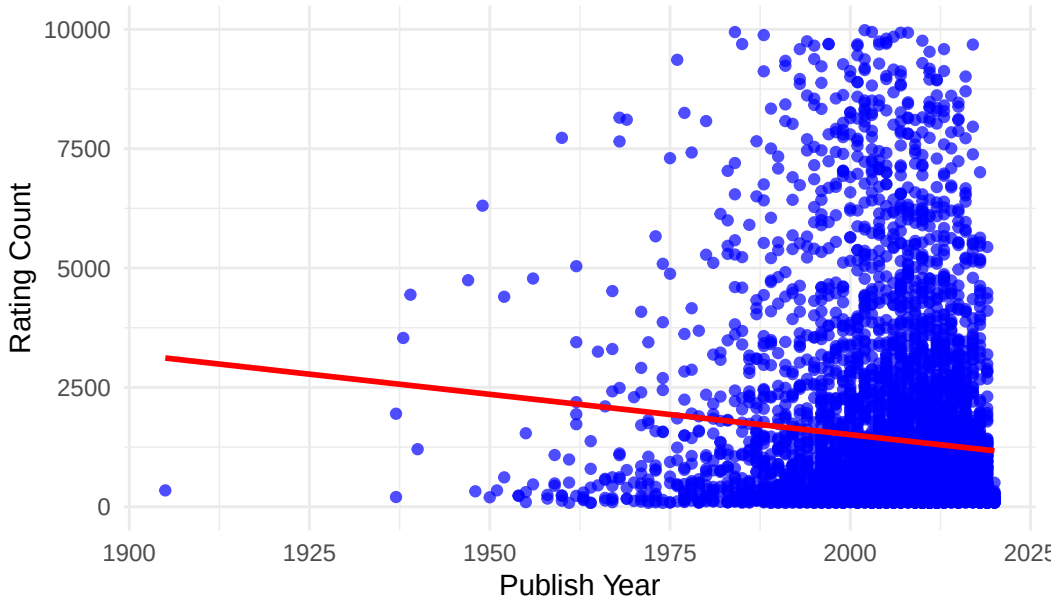


Figure 5: Scatter Plot of Pages vs. Rating Count up to 10000

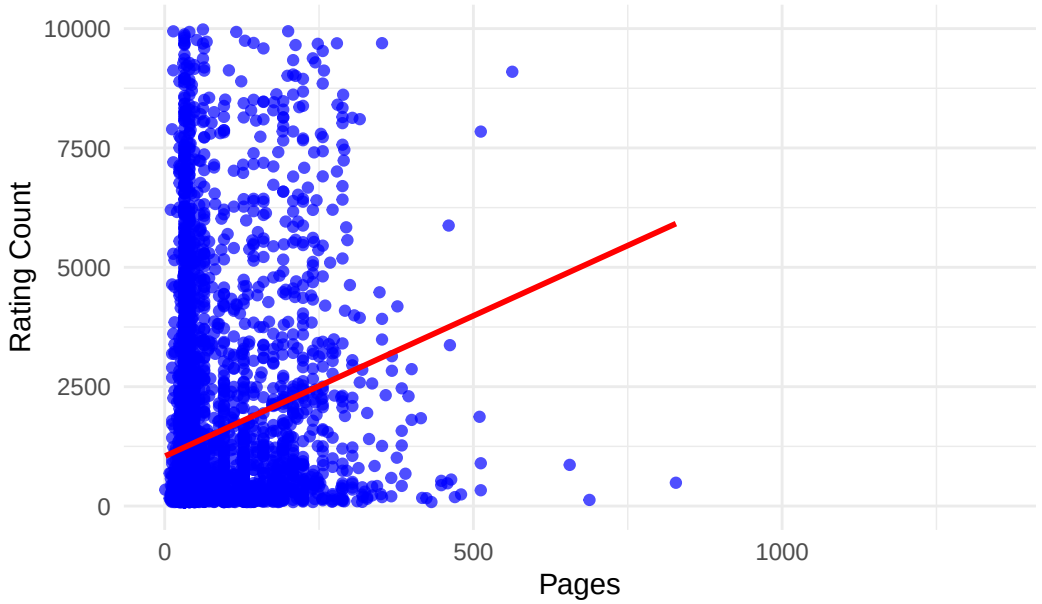
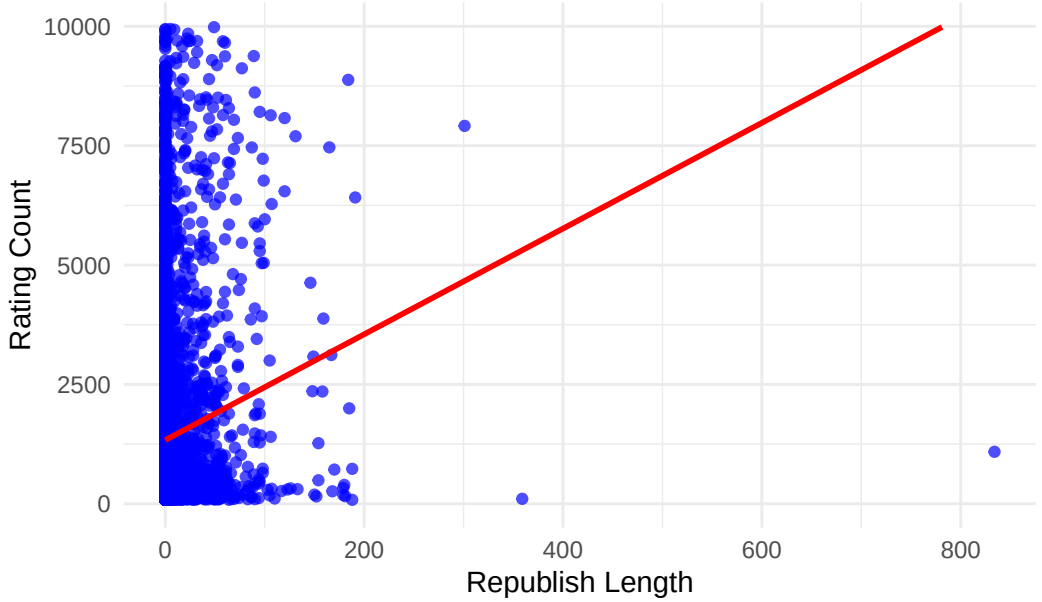


Figure 6: Scatter Plot of Republish Length vs. Rating Count



As per Figure 2, it can be seen that the average rating count and the distribution notably differ based on different cover types of the book, which indicates that the cover type could be a predictor of the received number of ratings. As per Figure 3, a positive correlation has been found from the trend line or linear regression line between Average Rating and Rating

Count. There is a plausible association between the audience preference and the popularity of the book, which is also reflected in this scatter plot. A negative regression line or potentially negative association has been found in Figure 4, between the published year and rating count (popularity), which indicates the published year can be a predictor of popularity. As per Figure 5, a positive association has been found from the trend line or linear regression line between the number of pages in a book and the Rating Count, which makes the 'pages' a possible predictor of popularity. Figure 6: Indicates that Republish Length and Rating Count could have a positive association and linear relationship. Therefore, the Republish Length could be a potential predictor of popularity.

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 Generalized Linear Model

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

GLM Model Assumptions

The assumptions for the GLM model of this research are:

- The outcome variable 'reading count' has a distribution that follows a Poisson distribution
- The relationship between each predictor and log-transformed outcome variable 'reading count' is linear
- The observations are independent of each other
- There is no high multi-collinearity within the predictors

4 Results

	Model 1
(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)

```
------------------------------------------------
AIC                              130292955.41
BIC                              130293022.82
Log Likelihood                   -65146467.71
Deviance                         130238910.48
Num. obs.                             6252
================================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

From the GLM output, it has been found that the publishing year is a significant predictor with a negative effect size. It indicates that with the progression of the year the popularity of books reduced. The strongest predictor of rating count with significant positive effect size is rating. It shows, that having a higher average rating increases the popularity of children's books. Republish length is a significant predictor of the popularity of a book with a positive effect size. It indicates when a publisher republishes older books it increases the probability of having high popularity. The variable 'pages' is also a significant predictor of rating count or popularity with a positive effect size. It indicates that containing a higher number of pages in children's books increases popularity. However, the different types of cover do not have any significant effect on the rating count. It implies there is no direct relationship between the type of cover and the popularity of a children's book.

5 Discussion

5.1 Overview of Research and Results

In this study, the underlying factors that could have effects on the popularity of children's books have been examined using a generalized linear model. The study considered an already existing dataset on children's book ratings that also includes the physical attribute of the book, the publication details and readers' ratings with unique ISBNs published from the year 1905 to 2020. Through developing a predictive model for estimating the popularity of children's books, the study examined whether there are other attributes apart from the obvious factor such as user preference can contribute to higher popularity. The results indicate that the average rating of the books, the number of pages in the book, the publishing year and the length of republishing or the duration of republishing of the book from the original year have significant predictability to estimate the popularity of a children's books.

5.2 Implication and Significance

As per the conventional understanding of the print media, the audience review or audience likability of the product is the major predictor of the potentiality (Loh and Sun (2019)).

This research highlights that not only the liability or average rating of the audience, the physical attributes such as the number of pages also major contributors to the popularity of the book. These findings will enable the publishers to consider the number of pages in a book to estimate its potential popularity. Besides, the study also highlights that older books which were originally published many years ago, can also have larger popularity than comparatively newly published books. The findings will enable publishers to focus on republishing older books to gain more popularity within the market of children's books.

5.3 Research Limitation

The major limitation of this research is associated with the limitation of the dataset of this study. The regression model that has been developed using the GLM method has low fitness and low predictability, which indicates that there are a considerable number of potential predictors and confounders which has not been considered in this study. The dataset of this study does not include information about the genre of the book and the target audiences, which could increase the accuracy of the predictability with a more critical understanding of the underlying factors of popularity. Besides, each continuous variable of this dataset consists of a large number of outliers, which has not been addressed in this dataset in order to avoid any predictive anomaly due to a lower sample size.

5.4 Conclusive Statement and Future Research

As per the findings of this study, it can be concluded that apart from the reader's preference or average of received ratings, the physical attribute namely the number of pages of a book can have a significant incremental effect on the popularity of the book. It can be also concluded that republishing older children's books could lead to higher popularity among readers. However, the predictability of popularity using the publishing details and physical attributes is not adequate, which requires more information regarding the book such as the genre of the book and target audience. In future research, the information regarding the genre of the book and target audience should be included as predicted. Besides, the outliers should be handled by future research while considering the cross-validation method and advanced machine learning methods such as neural networking for predicting the popularity of the book.

Appendix

.1 Data Cleaning Process

The data cleaning process for the chosen raw dataset of this study involves handling missing data, and filtering out the invalid and duplicate records (rows) and attributes (columns) from the dataset. In the selected dataset of children's books, first, the missing data are explored considering each attribute or column of the dataset. The variables of interest in this dataset are rating, cover, publish year, pages and rating count. There are rows only 8% of records that include missing values in any of these columns, and therefore, without using any amputation methods the records have been excluded to develop the filtered dataset. However, the missing values in the column 'original publish year' represent that there is no separate original year from the recorded publish year. Therefore, for this variable, the missing data has been replaced using values from published data and a new variable is created named 'republish length' by subtracting the original year from the published year. The variable signifies the gap between the original publish year and from recent publish year. The 'rating count' includes values from 1 to more than 1 million reviews with an average of 4280 (Q1 = 82, Q2 = 373, Q3 = 1459), whereas the lower number of reviews is not valid enough to represent the overall review rating of any book. Hence, books that have more than 25% quintile reviews (>82) have been selected in the filtered dataset. Then from the filtered dataset, the unwanted columns are excluded while including only ISBN and variables of interest.

.2 Alternative Model

In this research, the linear regression model using ordinary least squares was initially chosen for developing the prediction model. However, from an exploratory analysis of variables it was found that a large number of variables including outcome variables have a non-normal distribution that belongs to the exponential family. Therefore, a generalized linear model with log-normal link has been selected for this study.

.3 Observational Data

The author of the dataset Alex Cookson created the dataset of children's books with given ratings by the readers for empirical Bayes estimation of reader's ratings. The data collection process of the author of the dataset is not mentioned. However, this study collected the data from the GitHub data repository. The dataset includes all details as per the specific ISBN of the books. The dataset represents ratings of the readers of children's books, whereas each instance or record represents a single book with a specific ISBN. Each record contains character and numerical type data to represent the book title, publisher name, ratings, published year, page number and book cover type. The raw data consists of 9240 books with unique ISBN where the children's books that were published from 1905 to 2020 have been included.

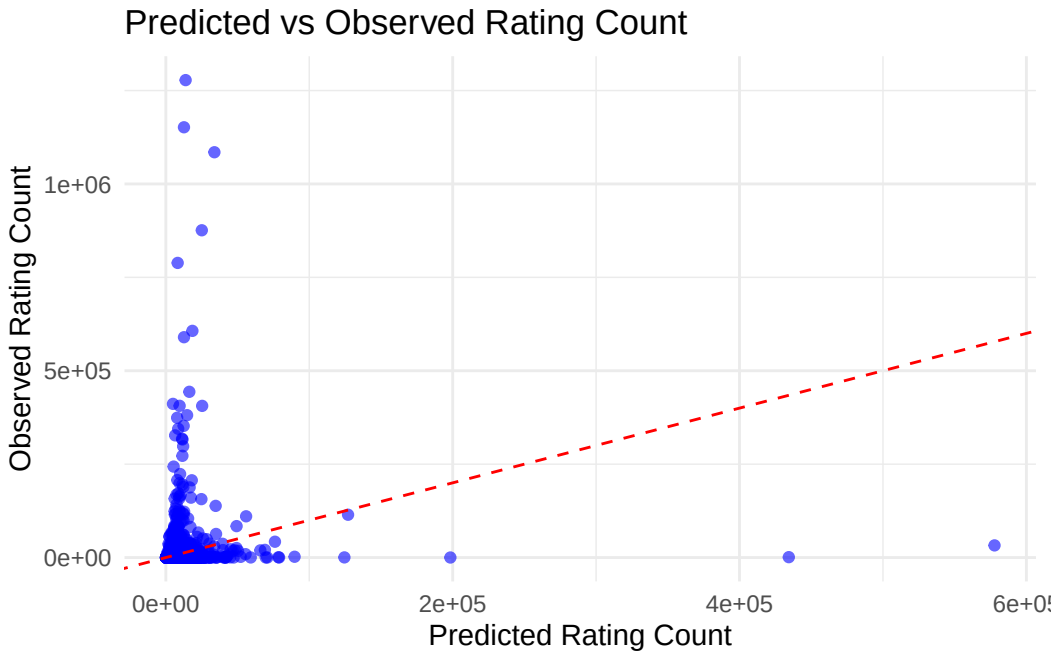
.4 Sampling

All records within the children's book dataset of Alex Cookson have been included in this study using no selection criteria. Convenient sampling has been used for selecting all the available records within the selected dataset.

.5 Model Diagnosis

```
GLM Model MSE: 1491689002
```

```
GLM Model RMSE: 38622.39
```



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, 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.