Data 557
Winter 2022

Goodreads Data Analysis

GitHub Link

Members:

- Anuhya Bhagavatula *(anuhyabs@uw.edu)*
- Juhi Choubey *(choubju1@uw.edu)*
- Eli Corpron *(ecorpron@uw.edu)*
- Aishwarya Singh *(aish25@uw.edu)*
- Hunter Yobei Thompson *(hunteryt@uw.edu)*

Table of Contents

INTRODUCTION

I. PROJECT MOTIVATION

An average publishing company receives thousands of transcripts daily and must prioritize publishing those transcripts that have higher chances of selling to maximize profits. Since data for transcripts are not available, the next viable option is data on published books and authors. The purpose of this project is to provide the publishers with an understanding of factors that influence the profits by considering the rating counts for published books.

Audience: Book Publishing Companies

II. USE CASES

Factors affecting Book Popularity:

When a publishing company receives a book to publish, it is typically provided with the following information -
Author Information: Name, Sex, Age, Previous works, Fanbase
Book Information: Genre, Number of Pages

Publisher ← Book

Information at hand

Sex of Author
Exp of Author
Author Fanbase/Rating
#Pages
Genre, etc.

Goal: Can we utilize historical information available on book & author popularity to gauge what book should be prioritized by the publisher to achieve maximum sales?

Metric to gauge Book Popularity:

Publishing companies are interested in publishing books that will sell comparatively better than the rest. The dataset does not have any information on book price, profits or readers reached by publishers. As a proxy measure, we choose to proceed with 'rating_count' as it provides the best approximation of the reader population compared to other scores.

III. QUESTIONS

The goal of the project is to answer the questions from the point of interest of a publishing company/publisher:

Question 1:

We want to identify if the gender of the author has any role to play in terms of books being sold. Hence the first question that we want to answer; Is the average rating count different for male and female authors?

Question 2:

Authors can be classified based on years of experience or the number of books published. Highly experienced authors can be expected to have larger rating counts. From a publisher's point of view, it would be useful to assess if publishing works of new authors (who typically have 0-3 works published) would look promising, which brings us to the next question; Do authors with fewer published works have a different average rating count than those who have published more?

Question 3:

The number of pages in each book may vary and it takes a significant amount of time for the publisher to read the book and make a decision. Is it even worth it? We decided to categorize the number of pages into three bins and find out whether the average rating count is equal across all bins.

Question 4:

The genre of the book plays a major role in the number of books being sold and we are interested in identifying which genre is more popular. We have grouped the genres into three categories and we attempt to answer the question if fiction has the same average rating count as non-fiction and others.

DATA

DATA SOURCES

The primary source for the two datasets was Kaggle. These datasets can be considered observational real-world data and are web scraped from the Goodreads website:

1. Goodreads Books Data
   This dataset contains information on published books from 1600-2020 and has 50K+ records on books.
2. Goodreads Authors Data
   This dataset contains information on Authors and has 200K+ records on authors.

DATASET DESCRIPTION

Below is the Entity-Relationship Diagram of both the datasets highlighting the features in each. After merging the 2 datasets, we have records on 34k+ books & 12k+ authors.

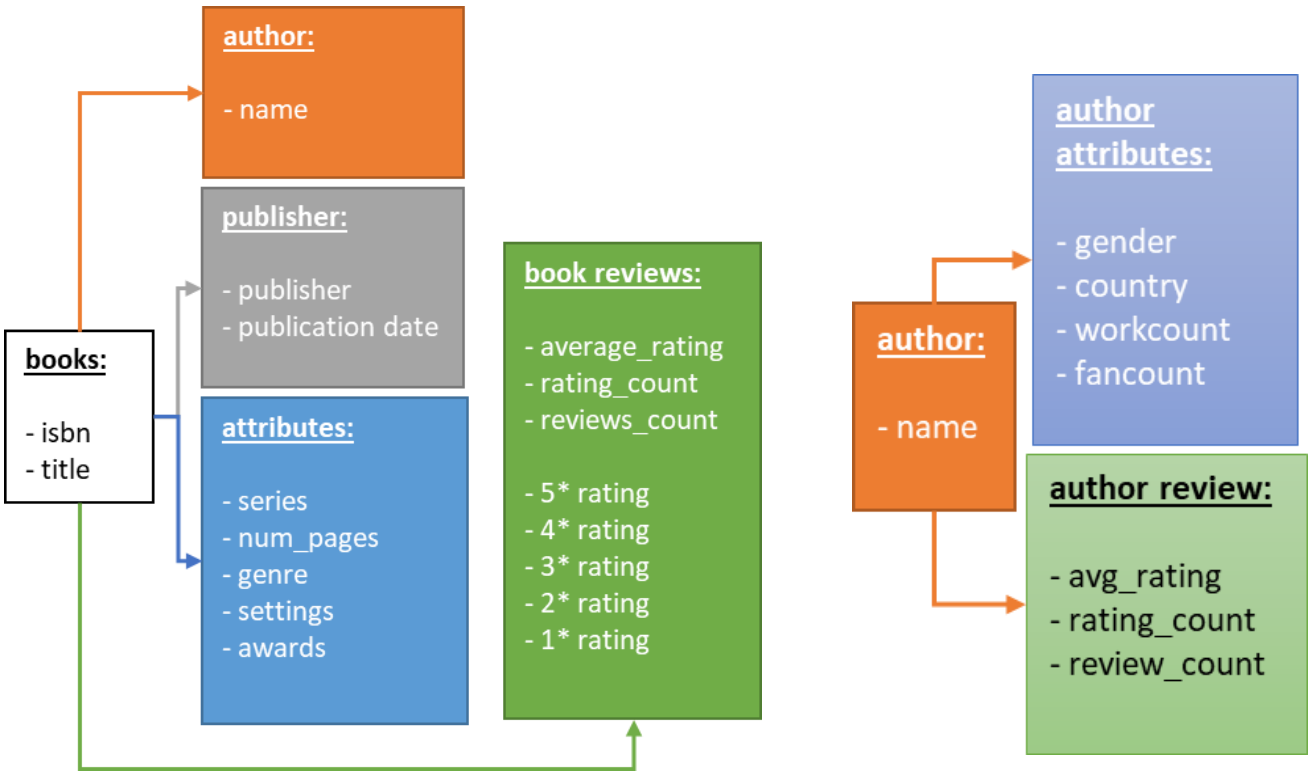

Figure 1: Dataset Description

VARIABLE	DESCRIPTION	INFORMATION
book	Title of the book	~34K+ Book Names
author	Author name	Raw column includes names of all contributors: Primary author, co-authors, etc.
rating_count	The number of ratings for the book	Integer column counting #times the book was rated on Goodreads
page_count	Number of pages in the book	#Pages in the book as mentioned on Goodreads. Includes records of books with 0 page count (audiobooks/CDs)
genre	Category of the book	Raw column: genre_and_votes Includes all genres associated with the book, and #votes received in each genre category. Sorted by #votes descending.
author_sex	Gender of the author	Male/Female authors only
author_work_count	Number of works done by author	#works of the author including individual /collaborated, articles/scripts/short essays, etc.
author_exp	Author_work_count after binning	Derived field after categorizing author_work_count
book_size	Page_count after binning	Derived field after categorizing page_count
genre_category	Genre after binning	Derived field after classifying the genre

Table 1: Variable Descriptions & Associated Information

CHALLENGES

1. The data does not include many metrics to understand how well a book sells from a publisher perspective such as information about the sale price of a book and how many copies have been sold, due to which we had to rely on less direct metrics such as the number of ratings a book has.
2. The dataset also contains entries for things such as audiobooks, CDs, and Plays.
3. Books could also get multiple entries if they were a part of a volume or a collection of novels. This can inflate the published works per author since each entry is counted.
4. Authors' data does not have a unique author ID. The next best option is to merge entries by author name. This fix is an issue if different authors with the same name are treated as one.
5. The majority of the books were also published in the United States, with very few books in the dataset published anywhere else. A large number of the entries also had corrupted names either for the author name or in the book title.

DATA CLEANING

General Data Cleaning Pointers Implemented:

- ❖ Text cleaning all categorical fields:
  - Text was made to be only lowercase
  - Removed unnecessary spaces
  - Removed non-alphanumeric characters
- ❖ Handling missing values:
  - Records with NULL values in relevant fields were discarded. As the data size in the analysis was large enough, removing records with null values did not result in a significant loss of information. We were still able to deploy large sample tests for statistical inferences.

Specialized Data Cleaning Pointers Implemented:

- ❖ Books Data:

	title	series	author	page_count	genre
0	Inner Circle	(Private #5)	Kate Brian, Julian Peploe	220.0	Young Adult 161, Mystery 45, Romance 32
1	A Time to Embrace	(Timeless Love #2)	Karen Kingsbury	400.0	Christian Fiction 114, Christian 45, Fiction 3...
2	Take Two	(Above the Line #2)	Karen Kingsbury	320.0	Christian Fiction 174, Christian 81, Fiction 58
3	Reliquary	(Pendergast #2)	Douglas Preston, Lincoln Child	464.0	Thriller 626, Mystery 493, Horror 432, Fiction...
4	The Millionaire Next Door: The Surprising Secr...	NaN	Thomas J. Stanley, William D. Danko	258.0	Economics-Finance 1162, Nonfiction 910, Busine...

1. Author: There are multiple entries with multiple authors. Only the initial author was kept. For instance, in the first record for the book 'Inner Circle' there are 2 author entries: Kate Brian and Julian Peploe. The first author (Kate Brian) present is considered the primary author and is kept for the analysis.
2. Genre: Books have multiple associated genres, which are voted on by the reader's. Based on the number of votes, only the highest voted genre for each book was considered to be the primary genre. For instance, in the first record for the book 'Inner Circle', there are 3 associated genres. As the first genre 'Young adult' has the highest number of votes, it is considered as the primary genre for the book and the rest are removed.
3. Page_Count: Records with unusual page counts such as 0, 1, 2, and >5K were also present in the dataset. Upon looking up such records on Goodreads, we found out that these entries were of Audiobooks, CDs, single page poetries, and volumes or collections of multiple books. Such records were not considered in the analysis.

- ❖ Authors Data:

	author	work_count	fan_count	sex	author_avg_rating	author_rating_count	author_review_count	author_country
0	Jason Wallace	2	13	male	3.74	1028	175	United Kingdom
1	Rosan Hollak	4	0	unknown	3.73	15	1	NaN
2	Nanna Foss	6	156	female	4.35	1172	205	NaN
3	Terri Savelle Foy	23	125	female	4.56	1054	151	NaN
4	Vishwas Nangare Patil	1	127	unknown	4.15	725	43	NaN

1. Sex: Records of authors with 'unknown' sex were discarded. For our analysis we only considered authors with sex entries of 'Male' and 'Female'.
2. Author_country: The Majority of the data present in the authors' dataset hailed from the USA. Therefore, to maintain the integrity of the data, we focused the scope of our analysis to books published in the USA.

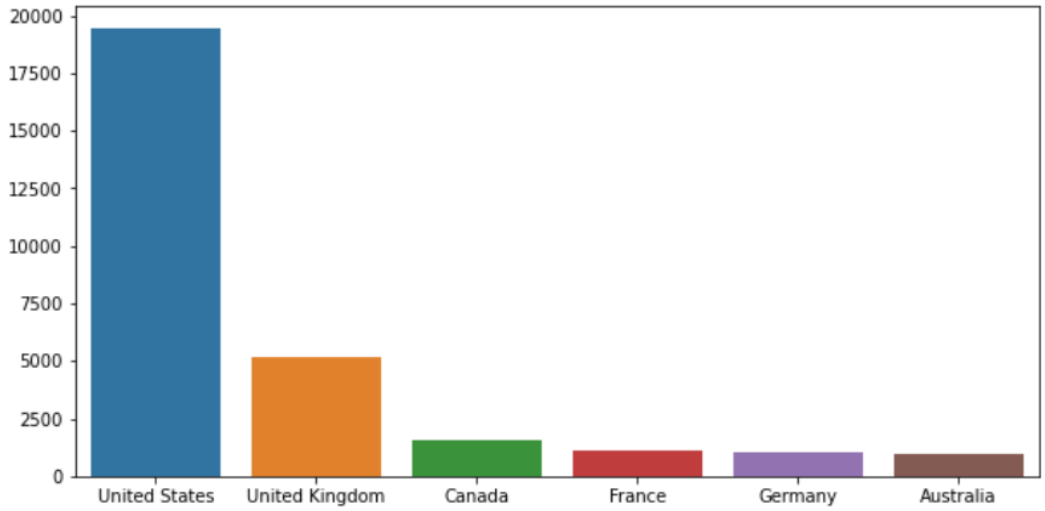

Figure 2: #Authors by Country (Top 6)

3. Duplicate entries:

author	sex	work_count	fan_count	author_avg_rating	author_rating_count	author_review_count
robin hardy	female	36	34	4.22	1691	119
robin hardy	male	23	4	3.77	754	88

There were duplicate entries for some authors in the authors dataset. Due to the absence of any unique author ID, the next viable option for uniquely identifying authors was 'author name'. Therefore, post merging the authors with books dataset, the records with the same author name but different gender were filtered (~6), manually looked up on Goodreads, and updated accordingly in the master dataset. The unique authors were de-duplicated with aggregating the work count, and other numerical entries, for maximum value.

STATISTICAL METHODS

Hypothesis 1: Male vs Female Authors

The publisher is interested to understand if the rating count between male & female authors is significantly different. If so, books from the group that holds larger rating counts would be prioritized over the other.

Null Hypothesis: Average rating counts b/w male & female authors are equal

EDA:

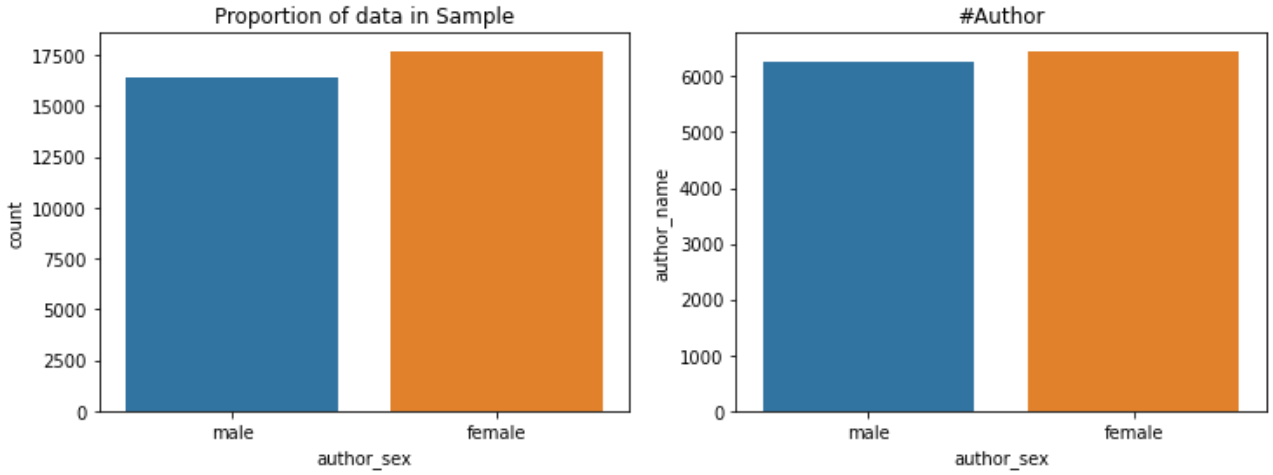

Figure 3: Proportion of author gender

Initial EDA seems to show a large, similarly size (~ 6000) male and female authors

Test:

Independent 2-tailed Z-test to compare mean reader count among the 2 groups. We do not perform ANOVA because the assumption of equal variance is not met.

Assumptions:

- Large Sample Size met to satisfy Central Limit Theorem (around 12000 authors in total)
- Independent samples met through aggregating average rating count by author. Otherwise, authors with multiple would not yield independent samples.

Outcome: p-value of the sample under Null Hypothesis
Considering a 5% level of significance, if the p-value would be less than 5%, then we would reject the null hypothesis.

Hypothesis 2: Across Author Experience

The publisher is interested to know if new authors have a similar rating count as a more established author. This is to provide evidence to publishers to help avoid the newcomer trap (when a publisher discards the book solely reasoning that the author has no experience).

Variable Selection:
Work_count provides the #works (books, articles, revisions, etc) of an author. We classify author experience into three bins based on work_count.

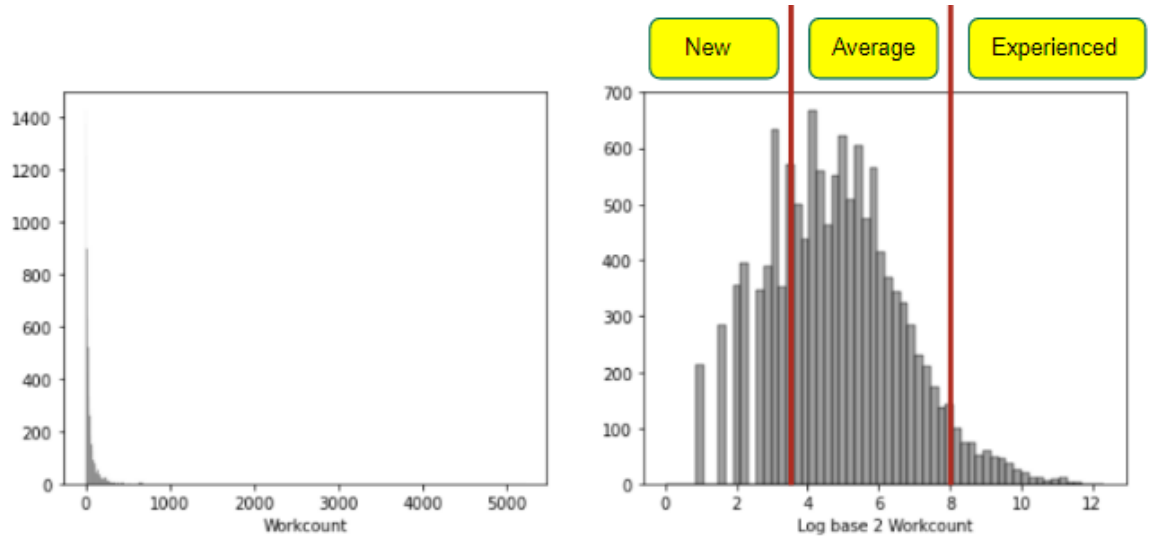

Figure 4a & 4b: Binning of work_count variable

EDA: Plotting authors based on their raw work count shows a distribution that looks roughly exponential and very right-skewed. To accurately bin our data, we instead used a logarithm transformation of work count and grouped data roughly as:

- New: > 1 std. Dev below the mean
- Average: within 1 std. Dev from the mean (both sides)
- Experienced: > 1 std. Dev above mean

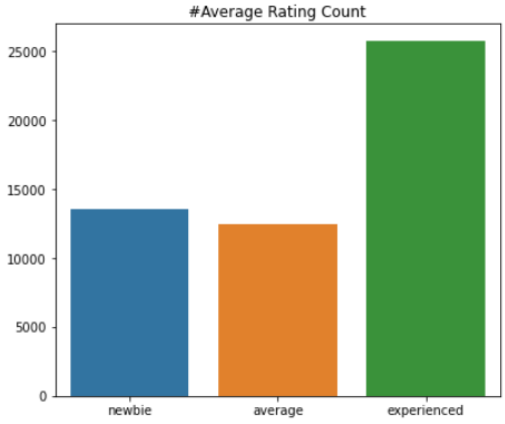

Figure 5: Average Rating Count by Experience

The average rating count between the groups shows that experienced authors seem to outperform new and average authors in terms of rating count, while new and average authors have a relatively comparable average rating count.

Null Hypothesis: Average rating count of the new, average, and legendary authors are equal.

Test:
Pairwise independent 2-tailed Z-test to compare mean rating count between all three groups of authors

Assumptions:
- Large Sample Size met to satisfy Central Limit Theorem (around 12000 authors in total)
- Independent samples met through aggregating average rating count by author. Otherwise, authors with multiple would not yield independent samples.

Outcome:
We are interested in an overall significance level of 0.05; because we are performing three pairwise tests, $\alpha = 0.05/3 = 0.0167$ using Bonferroni correction.

Further Analysis

It can also make sense to view the effect of author gender with author's work experience (combining Hypothesis 1 & 2) on average rating count.

Null Hypothesis: Average rating count of the new, average, and legendary authors is equal for male and female authors.

Test:
Two sets of three pairwise independent 2-tailed Z-tests to compare mean rating count between experience levels. We do not perform ANOVA because the assumption of equal variance is not met.

Assumptions:
- Large Sample Size met to satisfy Central Limit Theorem (around 12000 authors in total)
- Independent samples met through aggregating average rating count by author. Otherwise, authors with multiple entries would not yield independent samples.

Outcome:
We are interested in an overall significance level of 0.05; because we are performing three pairwise tests for each gender, $\alpha = 0.05/3 = 0.0167$ using Bonferroni correction.

Hypothesis 3: Across Book Sizes

Publishers receive books with pages ranging from 1-10K+! Bulky books require a lot of time to be reviewed. If they don't sell more, it's just a waste of effort. Publishers want to see if books with a higher page count have a lower rating count on average as compared to books with fewer pages.

EDA:
Page_count is the variable that tracks the number of pages in a book.

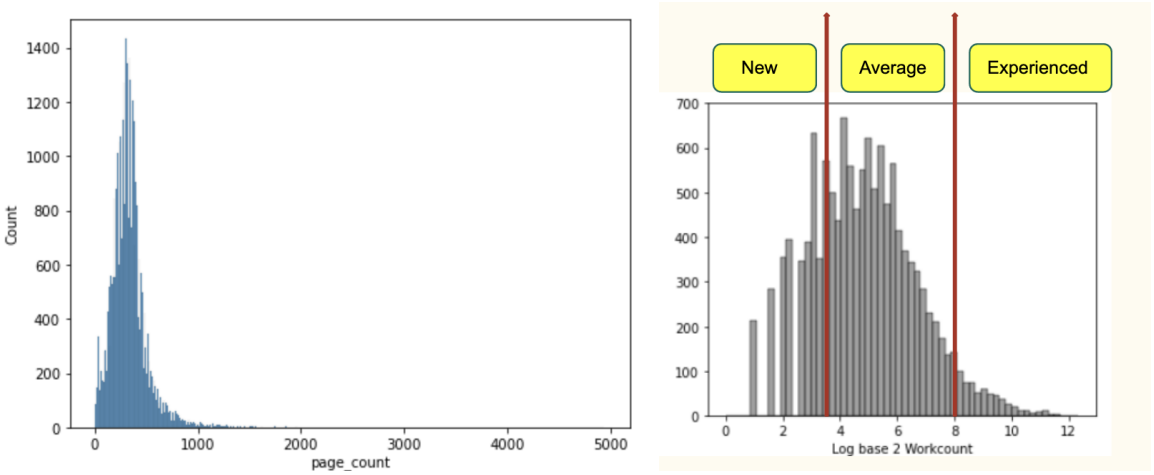

Figure 6a & 6b: Distribution of Page Count

Through some research, we found that the average page length of a book is 200-400 pages. With that we decided on three bins:

- Light: <200 pages
- Average: 200-400 pages
- Bulky: >400 pages.

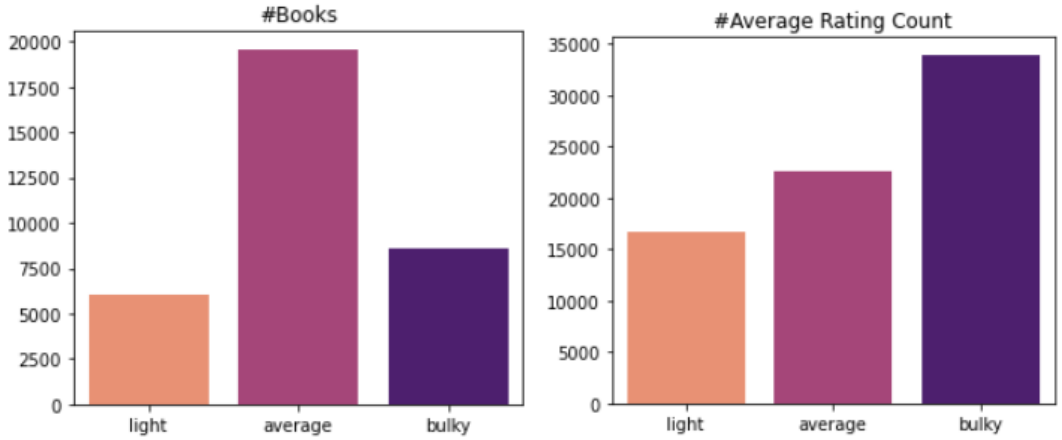

Figure 7a & 7b: Average Rating Count by Book Size

It appears that bulky books have a much higher rating count as compared to light and average books.

Null Hypothesis: Each combination of book sizes has the same average rating count.

Test:
Three independent pairwise 2-tailed Z-tests to compare mean rating count among the 3 groups. We do not perform ANOVA because the assumption of equal variance is not met.

Assumptions:
- Large sample size met to satisfy Central Limit Theorem
- Independent samples met since books can be assumed to be published independently

Outcome:
We are interested in an overall significance level of 0.05; because we are performing three pairwise tests, $\alpha = 0.05/3 = 0.0167$ using Bonferroni correction.

Hypothesis 4: Across Book Genres

The genre of the book is an obvious factor that can have a grossing effect on the rating count. Two of the most popular and general genres are fiction and nonfiction. We prefer reading fiction over non-fiction, and a publisher would want to know if there is a difference in performance between fiction and nonfiction books. If we find out that the rating count of fiction is not greater than non-fiction (or others), then we would know that the bias towards fiction books is ill-formed.

EDA:
We looked at the most popular genres across the entire dataset based on different business values:

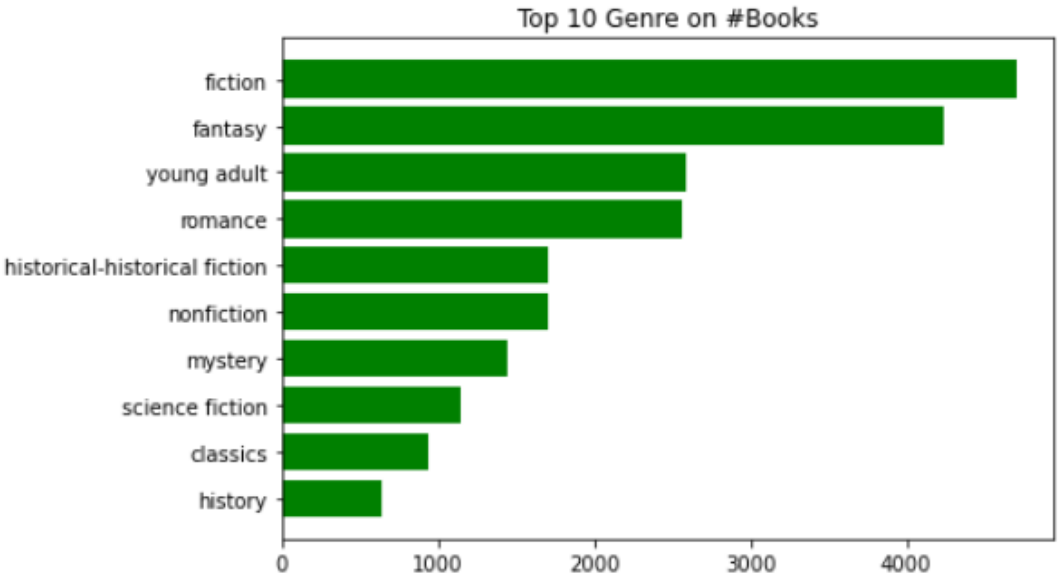

Figure 8: Top 10 Genres based of count of books

Grouping fiction genres under the genre category fiction, nonfiction genres under nonfiction, and other genres under the group 'others':

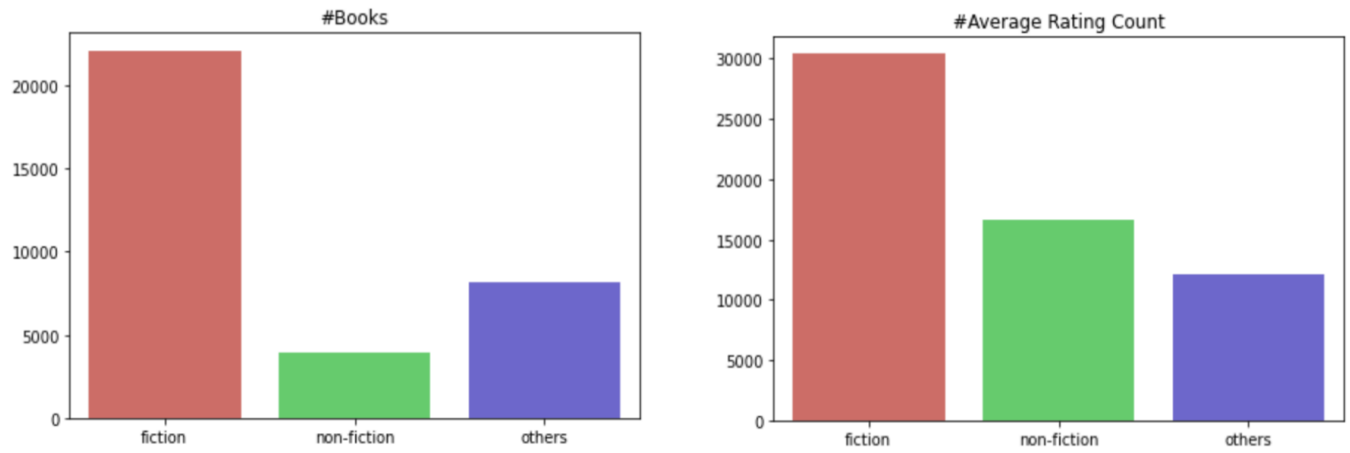

Figure 9a & 9b: Average Rating Count by Books Genre

It appears that our assumptions hold, and that fiction books have a higher average rating count as compared to nonfiction and other genres, but that nonfiction and other genres do not have too different of an average.

Null Hypothesis: Average rating count of Fiction books is greater than or equal to that of non-fiction & others.

Test:

Three pairwise comparisons of means between fiction and nonfiction, fiction and others, and nonfiction and others.

Assumptions:

- Large sample size met to satisfy Central Limit Theorem
- Independent samples met since books can be assumed to be published independently

Outcome:

We are interested in an overall significance level of 0.05; because we are performing three pairwise tests, $\alpha = 0.05/3 = 0.0167$ using Bonferroni correction.

RESULTS

Hypothesis 1: Male vs Female Authors

Result:

Z-test static is -0.2470396
p-value is 0.8049 > 5% level of significance.

Interpretation:
We did not find statistically significant evidence to reject the null hypothesis that male & female authors have equal average rating counts.

Hypothesis 2: Across Author Experience

Result:

We are interested in an overall significance level of 0.05; because we are performing three pairwise tests, $\alpha = 0.05/3 = 0.0167$

Pair	Test Statistic	P value	Conclusion
New v/s Average	1.02	0.31	Do not reject Null Hypothesis
New v/s Exp.	-3.99	6.5e-05	Reject Null Hypothesis
Average v/s Exp.	-6.55	5.6e-11	Reject Null Hypothesis

Interpretation:

- New Vs. Average: Cannot reject the null hypothesis that the two groups are equal

Would recommend a publisher to consider new authors just as much as average authors

- Other two pairs: Reject the null hypothesis that two groups are equal

Would recommend publisher to heavily consider works of experienced authors

Hypothesis 1 & 2: Further Analysis

Result: We are interested in an overall significance level of 0.05; because we are performing three pairwise tests for each gender, $\alpha = 0.05/3 = 0.0167$

Male Authors:

Pair	Test Statistic	P value	Conclusion
New v/s Average	0.17	0.86	Do not reject Null Hypothesis
New v/s Exp.	-3.62	0.0002	Reject Null Hypothesis
Average v/s Exp.	-5.41	6.18e-08	Reject Null Hypothesis

Interpretation (Males):

- New Vs. Average: Cannot reject the null hypothesis that the two groups are equal

Would recommend a publisher to consider new male authors just as much as average male authors based on this result

- Other two pairs: Reject the null hypothesis that two groups are equal

Would recommend publisher to heavily consider works of experienced male authors

Female Authors:

Pair	Test Statistic	P value	Conclusion
New v/s Average	1.06	0.28	Do not reject Null Hypothesis
New v/s Exp.	-2.35	0.02	Do not reject Null Hypothesis
Average v/s Exp.	-3.94	8.08e-05	Reject Null Hypothesis

Interpretation (Females):

- New Vs. Average & New vs Experienced: Cannot reject the null hypothesis that new authors are different from average and experienced authors

Would recommend a publisher to heavily consider the works of new female authors as there is no significant evidence to show a difference between new female authors and experienced or average female authors

- New vs Average: Reject the null hypothesis that two groups are equal

Hypothesis 3: Across Book Sizes

Result:

Pair	Test Statistic	P value	Conclusion
Light v/s Average	-3.11	0.0018	Reject Null Hypothesis
Light v/s Bulky	-7.64	2.1e-14	
Average v/s Bulky	-6.04	1.4e-09	

Interpretation:

We reject the null hypothesis that Avg. rating counts across all three pairs are equal.

$$\mu_{Bulky} > \mu_{Average} > \mu_{Light}$$

Hypothesis 4: Across Book Genres

Result:

Pair	Test Statistic	P value	Conclusion
Fiction v/s Non-Fiction	5.13	2.7e-07	Reject Null Hypothesis
Fiction v/s Others	9.84	7.4e-23	
Non-Fiction v/s Others	4.59	4.2e-06	

Interpretation:
We reject the claim that Avg. rating count across all three pairs of genres is equal.

$$\mu_{fiction} > \mu_{non\text{-}fiction} > \mu_{others}$$

LINEAR MODEL

The publisher would also like to see if we can predict the average rating count of a book based on the features we just tested.

Reader Count ~ Page Count + Volume Flag + Genre + Author Sex + Author Exp + Author Rating

Numerical variables such as Page Count and Author Exp were converted into categorical variables. We plan to retain them as categorical features in the model to maintain consistency between the statistical tests performed and the linear regression model built. Depending on the skewness in the data points, we will look to apply any transformations on any variable too.

Checking assumptions: We observed that the data does not appear to be normally distributed. Moreover, looking at the residual plots (image below) we concluded that the assumption of normality of errors was also not met. Since the sample size is large, we decide to proceed with the large sample size assumption.

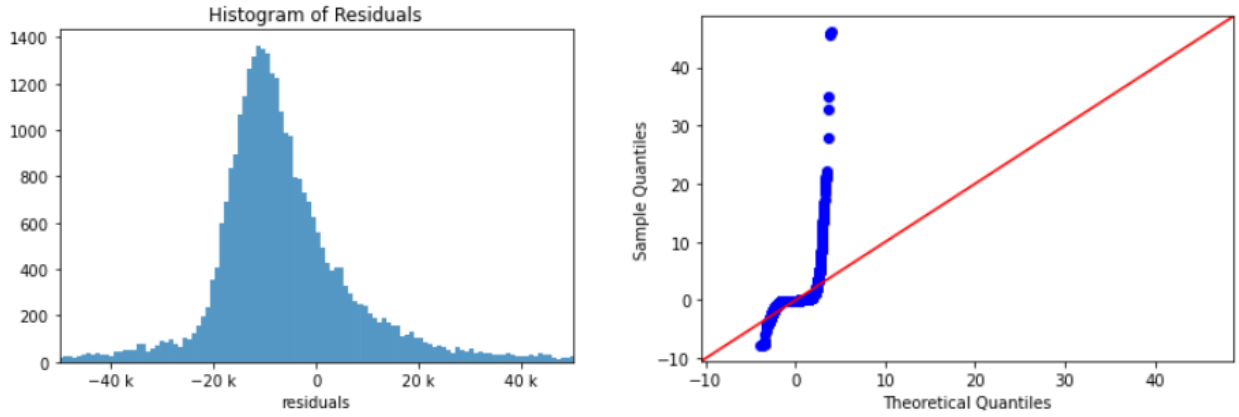

Figure 10a & 10b: Distribution of Residuals

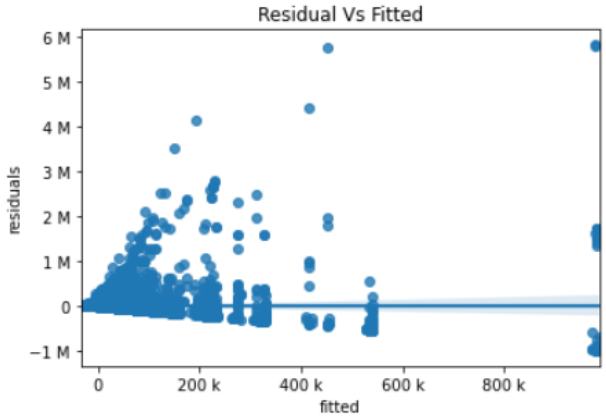

Figure 11: Residual v/s Fitted

Linear Model Assumptions

1. Independence of observation assumption was satisfied
2. The Linearity assumption was **not met**
3. Normality of Residuals was satisfied (considering large sample size)
4. Constant Variance **not met**

Result:

We built a linear regression model and here are the results from the same. We obtained an Adjusted R-squared: 0.145

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.376e+04	1.35e+04	-2.508	0.012	-6.02e+04	-7375.169
fiction	8911.6421	1705.276	5.226	0.000	5569.244	1.23e+04
non_fiction	7047.6607	2570.333	2.742	0.006	2009.721	1.21e+04
light	-5661.8046	1912.096	-2.961	0.003	-9409.576	-1914.033
bulky	1148.7870	1665.099	0.690	0.490	-2114.862	4412.436
is_volume	-6007.2410	1522.866	-3.945	0.000	-8992.109	-3022.373
male_author	-1619.1884	1476.721	-1.096	0.273	-4513.611	1275.234
newbie_author	2846.8157	1910.629	1.490	0.136	-898.082	6591.713
experienced_author	-1.982e+04	2154.718	-9.196	0.000	-2.4e+04	-1.56e+04
author_rating_count	0.0393	0.001	72.741	0.000	0.038	0.040
author_avg_rating	1.051e+04	3355.369	3.133	0.002	3934.493	1.71e+04

Interpretation:

From the above results, we observed that the sex of the author has no effect in predicting the rating count of the book. This observation aligns with our results from hypothesis 1.

DISCUSSION

The following conclusions were obtained from the significance tests performed:

- There is no significant evidence to conclude that the gender of the author affects the average number of ratings.
- Experienced authors outperform new and average experienced authors, but average experience does not outperform new authors.
- We found statistically significant evidence that books with more pages have a higher mean number of ratings.
- We also found statistically significant evidence that fiction has a higher number of ratings than non-fiction or other genre types.

The dataset had certain limitations such as not having information on how many issues of a book were sold and the price of the book. We would like to be able to use these variables for further research.

We would also like to have more information about sales of books outside of the United States. The U.S. is only a small portion of the book-buying population, so it would be helpful to get information from other markets. The linear regression model is a work in progress and we would like to apply log transformation to the rating count variable as well as look for other transformation strategies that could be applied. Lastly, if we could get information from websites and marketplaces like Amazon and Google we may be able to make further insights.

These limitations in the dataset also have an impact on the scope of inference for the results. Since we restricted the books to ones published in the United States, our conclusions can only be made about books published in the United States. Any conclusions also need to be restricted to the number of reviews. As such it is hard to extrapolate the conclusions beyond Goodreads.

APPENDIX

Other Statistical tests: Across Book Types

The publisher receives different types of books daily: some are solo books, while others may be part of a series, or a completed set, or even some random collection of books to be published as one. We are interested in checking if the books that are part of a volume have similar rating counts as standalone books.

Class I	Class II
Standalone	Parts of Volume

Null Hypothesis: Average rating count b/w standalone books and books that are part of a Volume is equal.

Test:

Independent 2-tailed Z-test to compare mean reader count b/w Class I and Class II books.

Assumptions:

- Large Sample Size met to satisfy Central Limit Theorem (around 12000 authors in total)
- Independent samples met since books can be assumed to be separate

Outcome: p-value of the sample under Null Hypothesis
Considering 5% level of significance, if the p-value would be less than 5%, then we would reject the null hypothesis.

Result:

Z-test static is -1.542843
p-value is 0.12286 > 5% level of significance.

Interpretation:
We did not find statistically significant evidence to reject the null hypothesis that standalone books & books that are a part of a volume have equal average rating counts.