Text

Description automatically generated

MS Data Science 2021 Cohort

   BIOST 557 A Winter 2022: Applied Statistics & Experimental Design

   Dated: February 23, 2022

**PROJECT PROPOSAL**

**Title:** Book Popularity Analysis

**Members:**

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

**Audience:** Book Publishing Companies

**Description:**

An average publishing company may receive thousands of transcripts daily. The company must select or atleast prioritize to publish those transcripts that have high chances of selling to maximize profits. The purpose of this project is to provide the publishers an understanding of factors that influence the reader base for a book.

**Data Sources:**

1. Books: <https://www.kaggle.com/austinreese/goodreads-books>
2. Authors: <https://www.kaggle.com/choobani/goodread-authors>

Chart, treemap chart

Description automatically generatedDiagram

Description automatically generated

**Use-Case:**

* Factors affecting Book Popularity:

When a publishing company receives a transcript 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 **←**  TranscriptInformation at hand

| *Sex of Author* |
| --- |
| *Exp of Author* |
| *Author Fanbase/Rating* |
| *#Pages* |
| *Genre* |
| *Any other info on Book* |

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

* Metric to guage Book Popularity:

We have the following 3 score types for each book to assess book popularity -

| Feature | Description | Interpretation |
| --- | --- | --- |
| rating\_count | Number of times the book was rated | Can provide a typical idea of the number of readers |
| review\_count | Number of text reviews for the book | Difficult to interpret |
| average\_rating | Weighted average of all ratings of the book | Gives an idea of whether the book is positively, negatively, or neutrally received by the reader. Does not help to guage the reader population |
| 5\* rating  4\* rating  3\* rating  2\* rating  1\* rating | Number of times the book was rater for the specified star rating | It does give an idea about the number of readers favoring/nont favoring the book; but, does a publisher really need to care for that? |

Publishing companies are interested to publish books that would sell comparatively more than the rest. For this reason, we choose to proceed with ‘rating\_count’ as it provides the best approximation of the reader population compared to other scores.

**Hypothesis:**

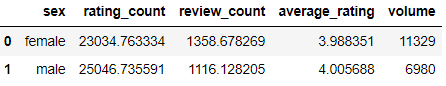
1. Is the average reader base different for male and female authors?
2. Authors can be classified based on years of experience or number of books published. Highly experienced authors can definitely be expected to have larger reader base. From a publisher point of view, it would be useful to assess if publishing works of new authors (who typically have 0-3 works published) would look promising.
3. Categorizing number of pages into ‘n’ bins, is the average reader count equal across all bins?
4. Each book can either be a standalone book or part of a volume. Intuitively, volumes can be expected to have greater reader base than standalone books. Is this true?
5. Are the reader bases for the top published genres reflective of rating counts? No. of books published for a genre vs no. of reader base for the genre

**Hypothesis Testing Detailed:**

1. Question:

* The publisher is interested to understand if the reader base between male & female authors is significantly different. If so, transcripts from the group that holds larger reader base would be prioritized over the other.

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

**

Test:

* \*\*(remove) EDA: If the difference in reader count b/w the 2 groups is *approximately same*, we will conclude that they are significantly equal and may not refer to any statistical test
* Else: Independent 2-tailed Z-test to compare mean reader count among the two groups.

Outcome:

* \*\*(remove) EDA: Bar plot
* Test: 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.

Open Question:

1. We are planning to review the proportion of male & female authors and the proportion of books published by the 2 groups. This is just to present a complete picture - do you think it is too much?

*(This is a common question across all hypotheses; as all the hypotheses involve assessing difference in reader count among 2 or more groups)*

1. Question:

* The publisher is interested to know if new-comer authors (published < 5 books) have a similar reader base as an average author (published average #books). This is to help avoid the new-comer trap! (when a publisher discards the transcript solely reasoning that the author has no experience)
* Call-out: Work\_count provides the #works (books, articles, revisions etc) of an author. It can be classified into 3 bins:

- New-comers: <5 (or <2std dev from mean)

- Average: within 1 std. Dev from mean (both sides)

- Legendary: > 1 std. Dev from mean

*Null Hypothesis: Average reader count of a new-comer author is less than or equal to that of an average author. (null hypothesis should state that there is no differed)*

Test:

* Independent 1-tailed Z-test to compare mean reader count b/w New-comers and Average authors.

Outcome:

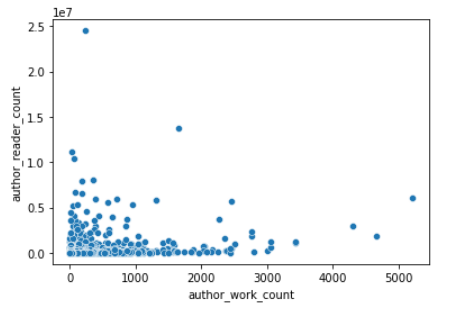
* Metric of Interest: p-value of the sample under Null Hypothesis

Considering 5% level of significance, if the p-value would be less than 5%, then we would reject the null hypothesis.

If it can be proved that new-comers have similar reader base to an average author, it will help addressing the prejudice against them.

Open Question:

1. Handling Outliers: some authors, such as: Jane Austen, Stephen King, and Stephene Meyers, are outliers to the data. They have slightly greater than average number of books published, but have very high reader count. Should they be removed from data to avoid skewing the test? (do not remove/consider log on rating count)

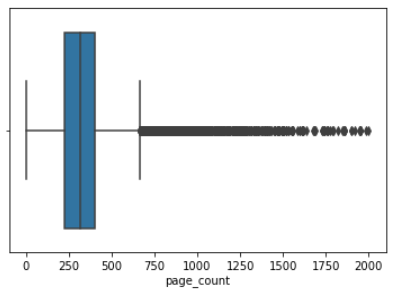


1. It can also make sense to view the effect of author gender with author’s work exp (clubbing Hypothesis 1 & 2) on average reader count. We are not sure how it will play; so if there is insight - we would want to share.
2. Question:

* The publisher receives transcripts with pages ranging from 1-10K+! Bulky transcripts require a lot of time to be reviewed. And if they do not sell more, it’s just waste of effort. The publisher is interested to see if the bulkier transcripts have less reader count on an average as compared to transcripts with average #pages.
* Call-out: books can be classified into 3 bins:

- Light: <100 pages (or <2std dev from mean)

- Average: within 1 std. Dev from mean page count (both sides)

- Legendary: > 1 std. Dev from mean page count

*Null Hypothesis: Average reader count of a bulky book is greater than or equal to that of an average size book.*

Test:

* Independent 1-tailed Z-test to compare mean reader count b/w New-comers and Average authors.

Outcome:

* Metric of Interest: 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.

Open Question:

1. Handling extreme values:

* 0: deleted
* 1,2: audiobooks, poetry
* >1500: Volumes/collections (remove as they are not books at all)

The extreme values of <10 and > 1500 show that the books belong to categories such as plays, poetries, or volume sets. But, poetries and plays are also present as a class in genre (with greater number of pages). If we are removing outliers on page number, should we also consider removing these type of books from genre?

1. Bulkier books may have a different selling pattern across fiction & non-fiction genres. Does it make sense to check effect of #pages independent of genre? We also want to club Hypothesis 3 & 5 and check the effect of book genre & size on average reader count.
2. Question:

* Genre of the book is an obvious factor that can affect the reader count of the book.
* *Either Fiction v/s Non-Fiction (we can either filter for fiction & non-fiction; but that would omit some books that are fiction but have other genre name present such as fantasy/thriller etc. This would require binning remaining ~400 genres)*
* *Or Top 5 Genre (no)*
* *(All other genres: fiction, non-fiction, others)*

1. Question:

* The publisher receives different types of transcripts 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 reader base as standalone books.

| Class I | Class II | Class III | Class IV |
| --- | --- | --- | --- |
| Standalone | Parts of Volume | Volume set | Collections |
| Pride & Prejudice | HP: #3 POA | HP Set of 7 | Set of 6 Jane Austen Books |

Test:

* Independent 2-tailed Z-test for mean reader count b/w Class I and Class II books

Outcome:

* Metric of Interest: 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.

**Open Question:**

1. Can we build a Linear regression model to predict reader count for a particular transcript based on the above features?

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

*(Most of the features are categorical. Does it make sense to think about a linear regression? Is assumption of Independence violated?) - maintain consistency in column type.*

*Thoughts:*

* *Transcripts vs next best thing: published books*
* *Outcome variable: Rater count: maintain consistency in terminology (remove reader term)*
* *We do not have book price or money made: so next best thing or proxy measure is rating\_count.*
* *Deal with 0 reader count (if taking log) - replace instead of remove (+1 to entire column)*