PROJECT PROPOSAL

Title: Book Popularity Analysis

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)*

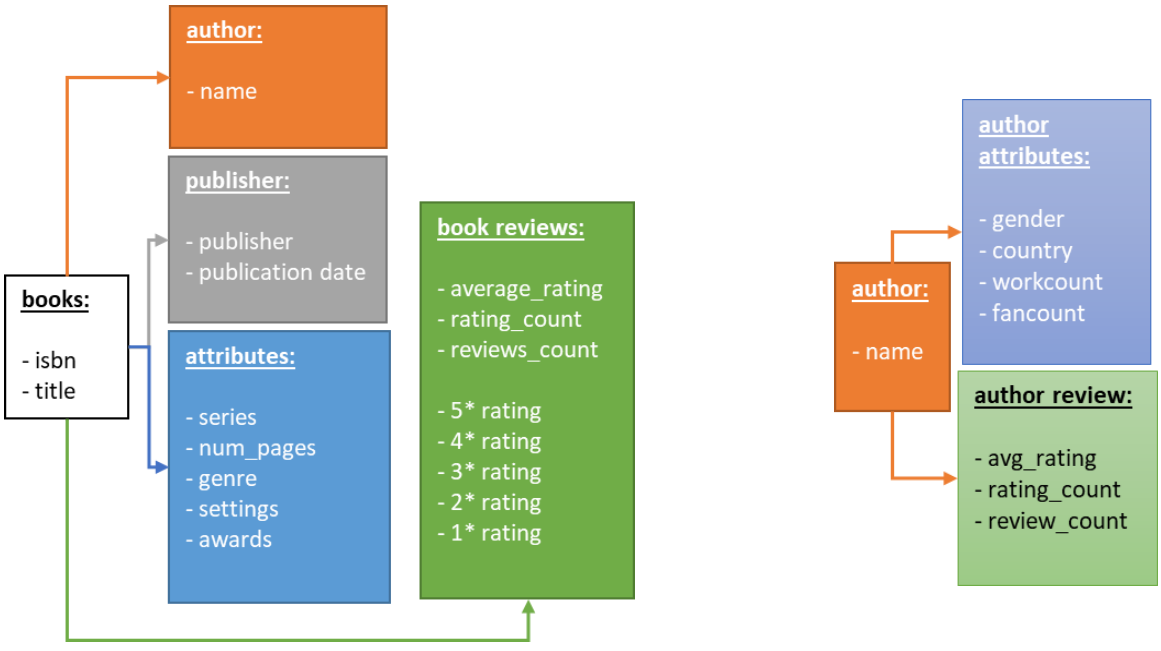
Audience: Book Publishing Companies

Description:

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 is not available, the next viable option is data on published books and authors. The purpose of this project is to provide the publishers an understanding of factors that influence the profits by considering the rating counts for published books.

Data Sources:

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



Use-Case:

- ❖ 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.

Hypothesis:

1. Is the average rating count 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 rating counts. 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 the number of pages into 'n' bins, is the average rating 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 a greater rating count than standalone books. Is this true?
5. Are the most frequently published genres reflective of rating counts? No. of books published for a genre vs mean rating count for the genre

Hypothesis Testing Detailed:

1. 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

   Test:
   - Independent 2-tailed Z-test to compare mean reader count among the 2 groups.

   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.

2. The publisher is interested to know if new-comer authors (published < 5 books) have a similar rating count as an average author (published average #books). This is to help avoid the new-comer trap! (when a publisher discards the book solely reasoning that the author has no experience)
   - Work_count provides the #works (books, articles, revisions etc) of an author. It can be classified into 3 bins:
     - New-comers: <5 (or <2 std dev from mean)
     - Average: within 1 std. Dev from mean (both sides)
     - Legendary: > 1 std. Dev from mean

   Null Hypothesis: Average rating count of a new-comer author is equal to that of an average author.

   Test:
   - Independent 2-tailed Z-test to compare mean rating count b/w New-comers and Average authors.

   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.

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

3. The publisher receives books with pages ranging from 1-10K+! Bulky books require a lot of time to be reviewed. And if they do not sell more, it's just a waste of effort. The publisher is interested to see if the bulkier books have less rating count on an average as compared to average-sized books.
   Books can be classified into 3 bins:
   - Light: <80 pages (or <2 std. Dev from mean)
   - Average: within 1 std. Dev from mean page count (both sides)
   - Bulky: > 1 std. Dev from mean page count

   Null Hypothesis: Average rating 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 rating count b/w Bulky and Average sized books.

   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.

4. Genre of the book is an obvious factor that can have a grossing affect on the rating count. The most popular ones being: Fiction and Non-Fiction (we'll club the rest in 'others'). Intuitively, most people prefer reading fiction over non-fiction and others. The publisher is interested to know if this is indeed true. 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 a publisher has toward fiction books is ill-formed.

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

   Test:
   - 2 Pairwise comparison of means b/w group 1 v/s 2, and group 1 v/s 3

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

Grouping Hypothesis 3 & 4: Bulkier books may have a different selling pattern across fiction & non-fiction genres. We also want to club Hypothesis 3 & 4 and check the effect of book genre & size on average rating count.

Optional Tasks:

5. 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	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

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 for mean reader count b/w Class I and Class II books

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.

6. 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? We plan to implement a linear regression model for the same:

➔ **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 b/w 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.