MS Data Science 2021 Cohort
BIOST 557 A Winter 2022: Applied Statistics & Experimental Design
Dated: February 9, 2022

UNIVERSITY of WASHINGTON

PROJECT PROPOSAL

Project Title: **Book Ratings Analysis**

Team Members:

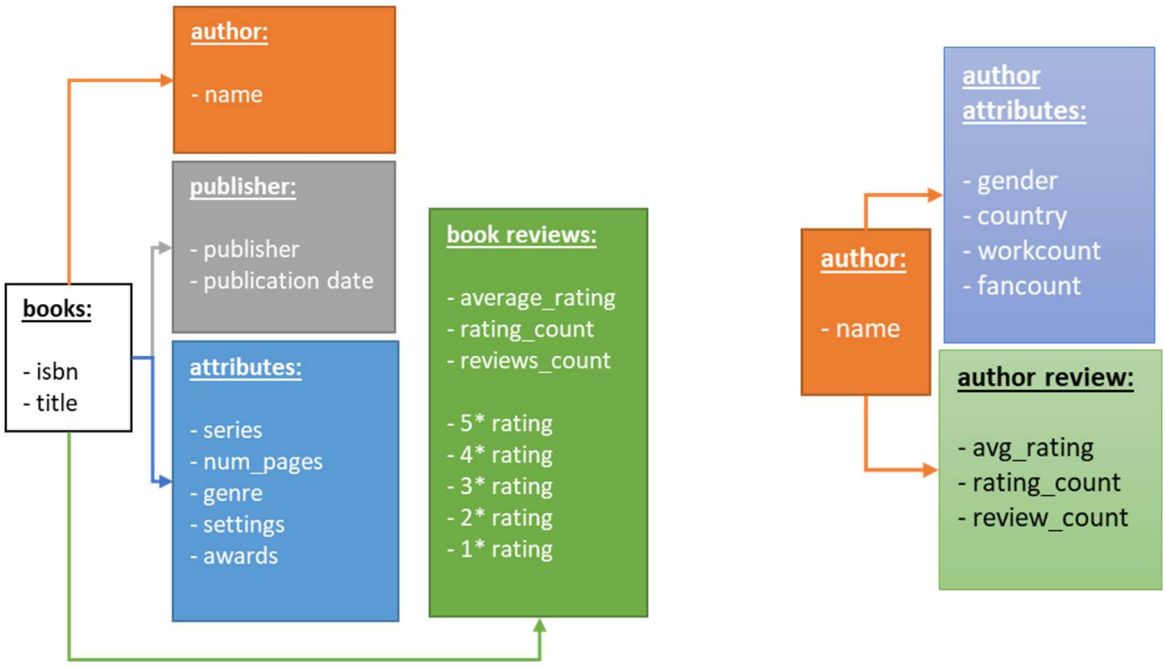
- Anuhya Bhagavatula *(anuhyabs@uw.edu)*
- Juhi Choubey *(choubju1@uw.edu)*
- Eli Corpron *(ecorpron@uw.edu)*
- Aishwarya Singh *(aish25@uw.edu)*
- Hunter Yobei Thompson *(hunteryt@uw.edu)*

Project Description:

The purpose of this project is to provide an understanding of factors that affect the success rating of books and their authors. Considering the entity "book" or "author" as a product, we can relate multiple attributes to it, such as: Ratings, Reviews, Genres, Publisher, Price, #Units sold, #Pages etc. Our aim is to develop statistically significant inferences around the role these metrics play in influencing the ratings of the book/author.

Data Sources:

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



Research Question #1: Are male authors more popular than female authors?

- Dimensions: gender, country of origin, time-period, genre
- Measures: work count, average rating, rating/review count

Hypothesis Tests:

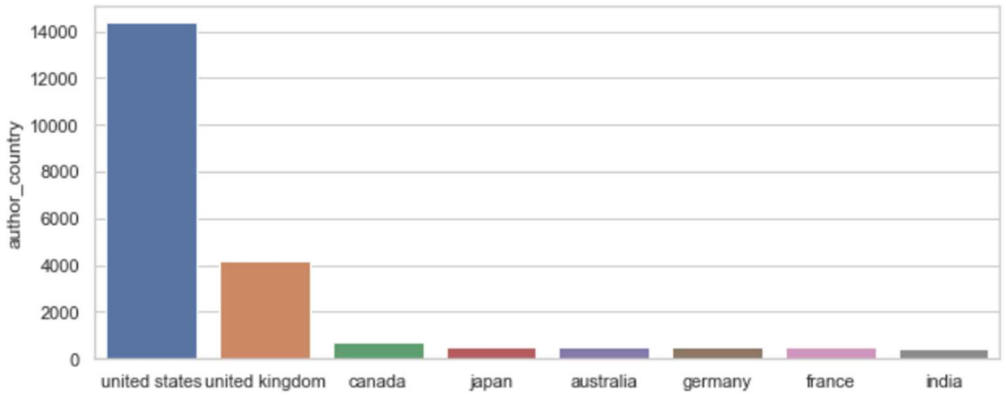
1. Are there more #male authors than #female authors? (EDA)
2. Have male authors published more #books than female authors in the last 20 years? (Proportion test)
3. Are books published by male authors on an average read more than those by female authors? (Measuring review/rating count: Z-test)
4. Is the mean average ratings of men and women authors equal? (Z-test)

Special case:

5. Women authors write more romance-based work than male authors (chi-square test)

Open questions:

- *2 measures of popularity: rating count & review count (working to integrate both or select 1)*
- *Author 'fan count' is available. But the logic behind this metric is unknown. We have decided to not utilize this metric. Any thoughts?*
- *Majority of data is for US only: should the tests be done for only US?*



Research Question #2: What types of books are most popular among today's readers?

- Dimensions: genre, book type (series or solo), time-period
- Measures: average rating, rating/review count, #ratings (5 star/4 star/…)

Hypothesis Tests:

Genre-based:

1. Are fiction books published more than non-fiction books? (EDA)
2. Is fiction read more than non-fiction? (Measuring review/rating count: Z-test)
3. Is the mean average rating of fiction & non-fiction books equal? (Z-test)
4. Do men & women authors contribute equally to the most popular genre?
5. Do non-fiction books have more extreme ratings (5/1) than fiction books? (ANOVA)

Special case:

6. Does the Harry Potter series have a higher rating than the average book? (2 sample z-test?)

Type of Books (Serial/Solo books):

1. Are solo books published more than volume-based books? (1 sided proportion test)
2. Mean average rating of solo and volume-based books are equal. (Z-test)
3. Mean review/rating counts is equal for solo and volume-based books. (Z-test)

Open questions:

- *Page count: Original hypothesis checked whether the #pages affect the book popularity. However, we believe that #pages are not good metric to determine this factor. We have instead decided to proceed with the 'series' metric which tells us if the book is part of the series or not.*
- *We have 465 unique genres in the dataset which we have combined into two main categories: fiction and non-fiction.*