

MS Data Science 2021 Cohort

BIOST 557 A Winter 2022: Applied Statistics & Experimental Design

Dated: January 26, 2022

**PROJECT PROPOSAL**

Project Title: **Book Ratings Analysis**

Team 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)*)*

Project Description:

The purpose of this project is to provide an understanding of factors that affect the success rating of a book. Considering the entity “book” as a product, we can relate multiple attributes to it, such as: Ratings, Reviews, Authors, 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.

Questions/ Base Hypothesis:

Many of us are ardent book readers. We thought it would be engaging to analyze

the best-selling books of the last century based on their explicit features, in a similar manner to how people analyze music, movies, and songs. People have unique tastes, but are there definable characteristics that correlate with a higher rating? To answer that, we came up with several different hypothesis to test:

1. Have books of a particular genre received a higher average rating than the books of another genre?

2. Do books with more than 500 pages have a higher average rating?

3. Are the books of a particular author more popular than others?

4. Are books from one publisher preferred over another?

5. Does the marketplace impact the rating of a book?

Data Sources:

Goodreads, Amazon Books, and Google Books are versatile platforms where almost every reader across the globe shares their views and ratings on books read. As part of this project, we will use the data gathered from these platforms which includes information on the book: \*ISBN, publishers, authors, reviews and average ratings.

| **Source** | **Location** | **Features** | **Identified Issues** |
| --- | --- | --- | --- |
| Goodreads | <https://www.kaggle.com/jealousleopard/goodreadsbooks> | Book title, authors, average rating, #pages, publication date, publisher |  |
| Amazon Users Ratings | <https://www.kaggle.com/saurabhbagchi/books-dataset> | User level rating with age, geographical location |  |
| Google Books | <https://www.kaggle.com/bilalyussef/google-books-dataset?select=google_books_1299.csv> | Unique columns include book prices and genres | Small data size (only 246 books) |
| NYT Times | <https://www.kaggle.com/dhruvildave/new-york-times-best-sellers> | Week-level data for top 5 books every week | Many missing attributes: genre, publisher details, ratings |
| 7k Books | <https://www.kaggle.com/dylanjcastillo/7k-books-with-metadata> | Book & author details, Avg rating, Categories (genres), Publication year |  |

\* ISBN: International Standard Book Number (A unique numeric commercial book identifier). We intend to use ISBN to map the above data-sets, if required, to procure additional information.