**Dataset Description**: We have created two data sets, “books\_by\_user.csv” and “books.csv.” One focuses on individual user reviews, and the other focuses on data for individual books from the lists "<https://www.goodreads.com/choiceawards/best-books-XXXX>" for the years XXXX from 2011 to 2024.

In the first dataset, “books\_by\_user.csv,” we have collected the following features: Title of the book, full name of the author, rating the user gave, user id, book ISBN, total number of ratings, average rating, number of pages, review, date the user started reading, date the user added the book, book format, publication date.

In the second dataset, “books.csv,” we have collected the following features: Category, genres, votes, title, description, author name, author about, author avatar url, number of pages, number of ratings, number of review, publication date, cover image url, goodreads url for book, ratings histogram, number of users with book on want to read list, number of people currently reading the book.

We scraped data from book pages from 2011 to 2024 (6475 books). We could add more, but this seems like enough books for now.

**Problem statement**:

We seek to use data scraped from GoodReads to create a book recommendation algorithm for users by determining genre clusters and analyzing trends in which books are popular over time. We will compare our genre clusters against existing genre tags, and analyze trends both within our clusters and within existing genre tags. This algorithm should both help readers to find books they want to read as well as help drive sales for publishers and booksellers.

**Stakeholders**:

· Readers

· Publishers

· Booksellers

· Libraries and other groups providing books to a community

**KPIs**:

* Hits: When trained on partial data of a test user, how many recommendations appear in the withheld data set
* User preference prediction match: When train on partial data of a test user, how closely can we predict their likes/dislikes
* Match between our clusters and genre tags: Genre tags have been collected in the “books.csv” dataset.