Book recommendation system

Diego Fuentes

Overview of the subject

Over 1m Books ratings containing general Book Data, and Users Demographics.

Problem statement

With countless books available, finding one that matches individual tastes can be tedious. Existing systems often miss the mark on capturing a reader's interests.

Opportunity

A Book recommendation system that have the potential to deliver personalized suggestions that will resonate deeply with readers' preferences. By using factors such as author, publication year, genre, and other user ratings.

▶ k-nearest neighbors & Collaborative filtering

Book Title

Book Author

Book Category

Publisher

Language

Year of Publication

User Id

Age

Country - State

Ratings

▶ Model Result

Readers

Better matches between reader preferences and recommendations, increasing user engagement and retention.
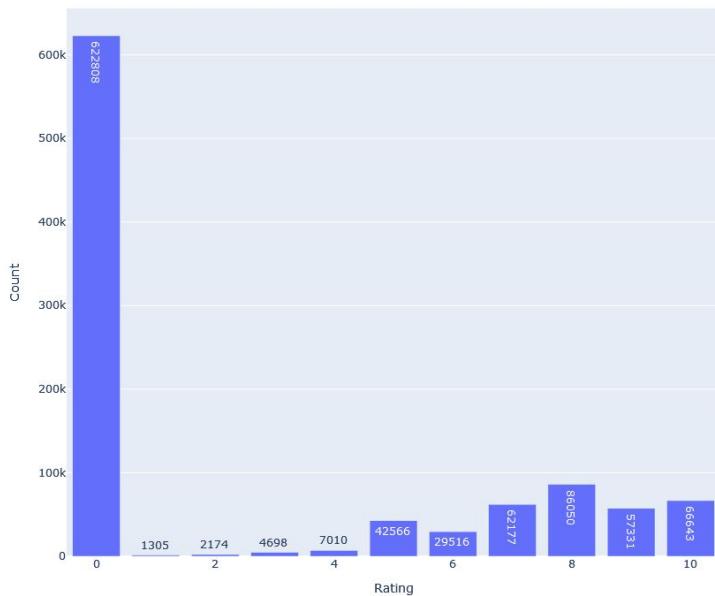
Business

Enhanced recommendations could boost sales by matching users with books they're more likely to purchase.
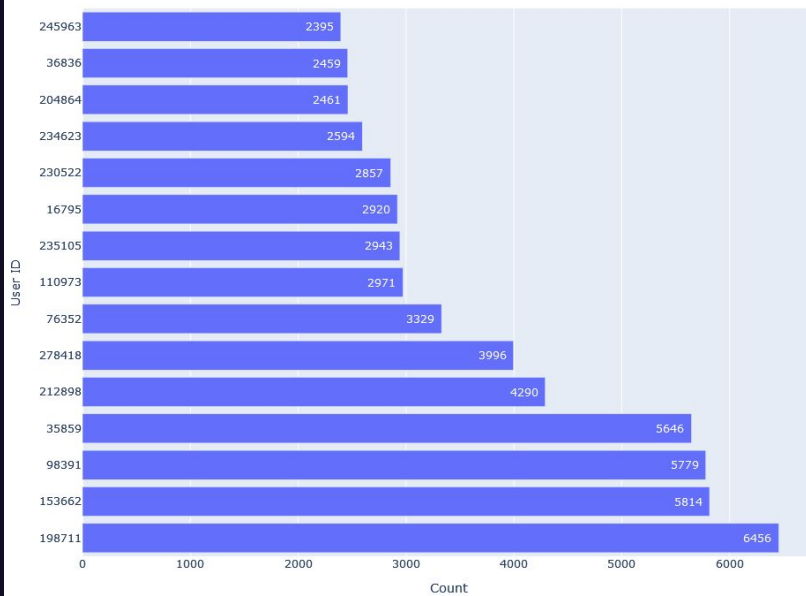
Exploratory data analysis Numericals

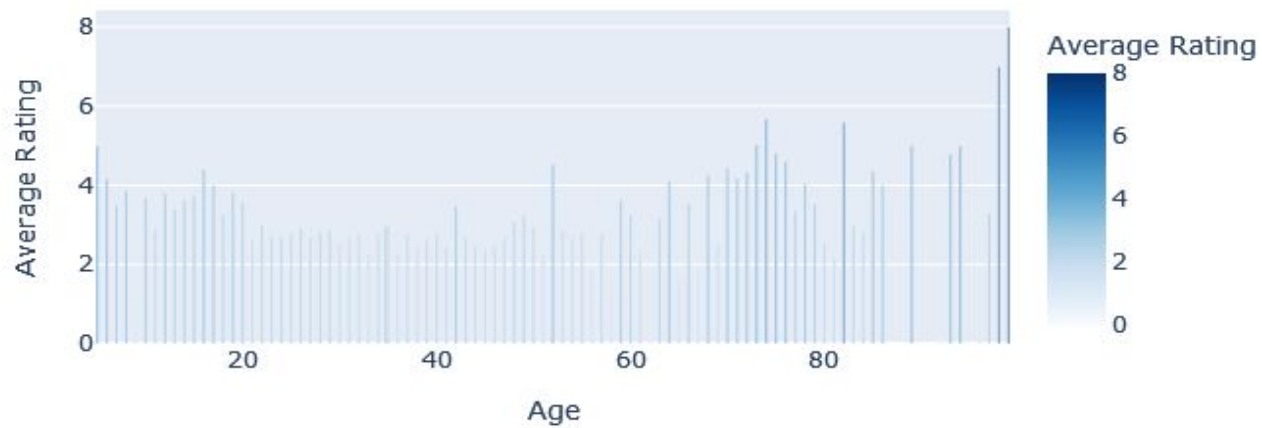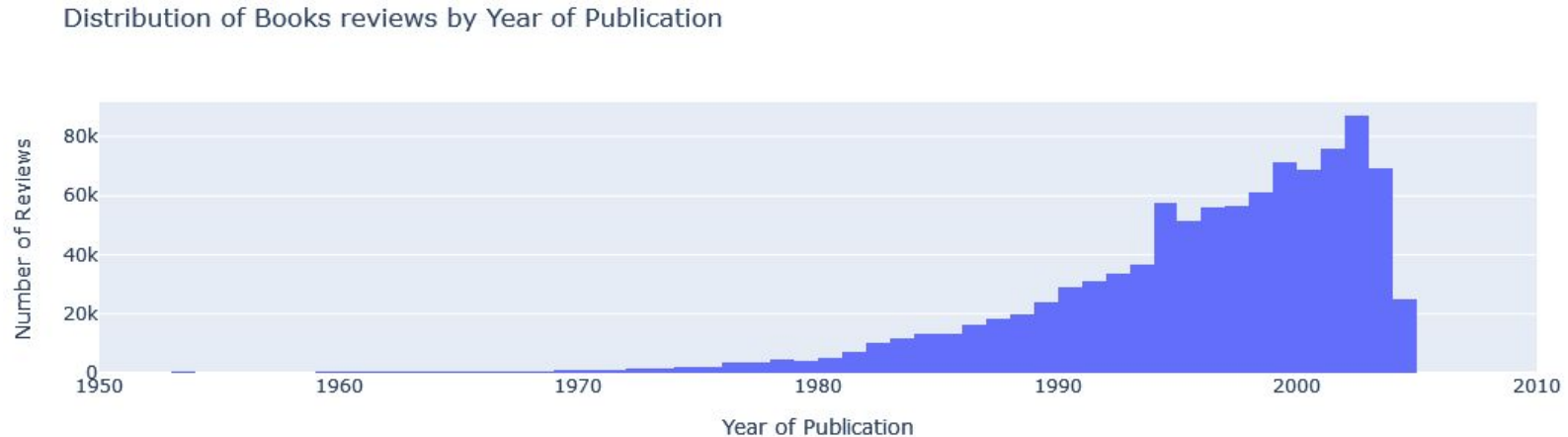Rating Distribution

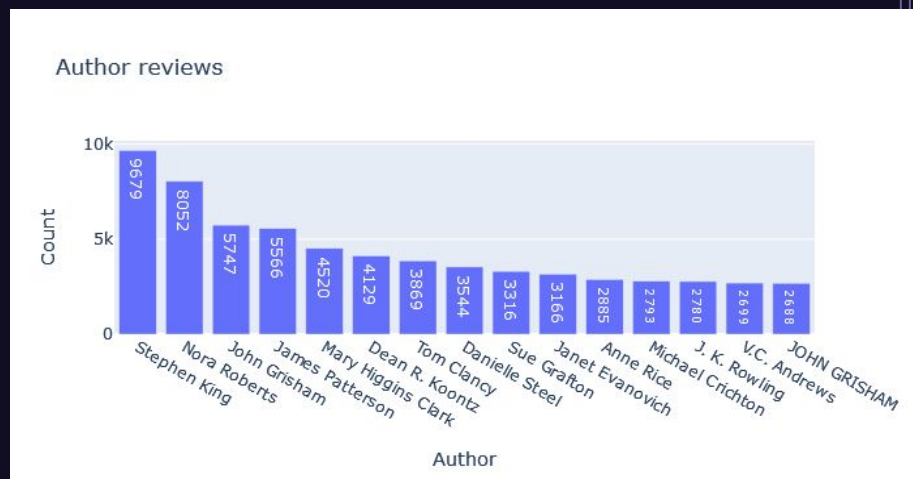Users with Most Reviews

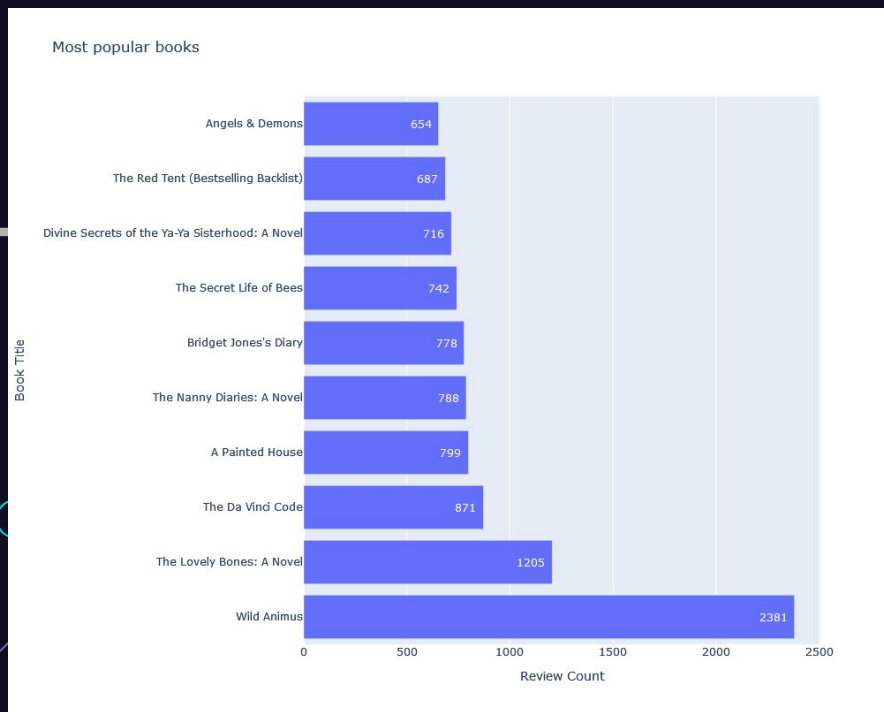# Age

Average Book Ratings by Age Group

▶ Years of publication

This bar graph trend suggests that users tend to prefer reading newer books.

Distribution of Books reviews by Year of Publication
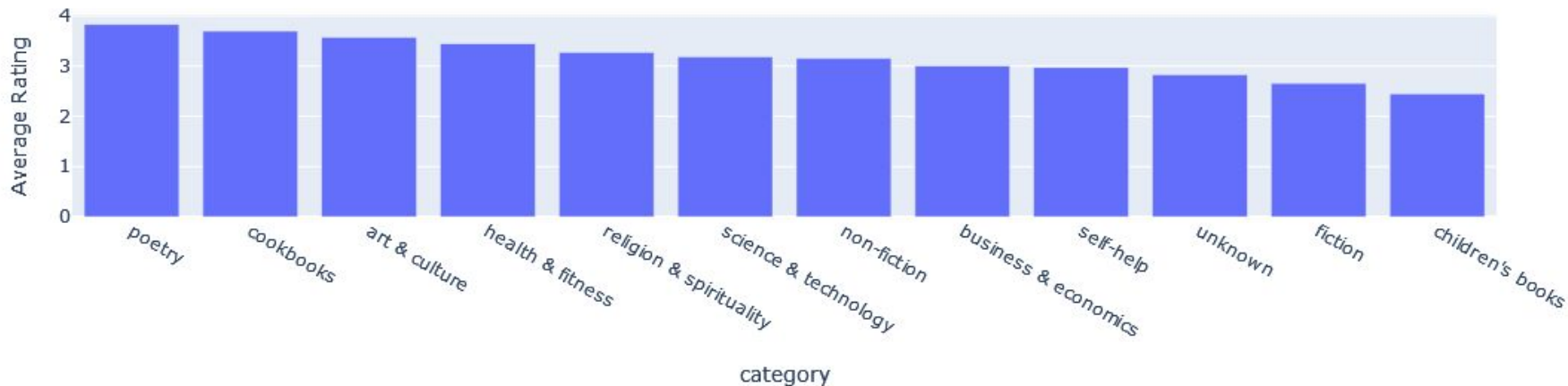
▶ Categoricals, Book Title and Authors

Both book titles and authors have a low mean number of reviews, but they exhibit high variance in the data, along with some very high maximum values.

Most popular books

Book Title	Review Count
Angels & Demons	654
The Red Tent (Bestselling Backlist)	687
Divine Secrets of the Ya-Ya Sisterhood: A Novel	716
The Secret Life of Bees	742
Bridget Jones's Diary	778
The Nanny Diaries: A Novel	788
A Painted House	799
The Da Vinci Code	871
The Lovely Bones: A Novel	1205
Wild Animus	2381

Author reviews

Author	Count
Stephen King	9679
Nora Roberts	8052
John Grisham	5747
James Patterson	5566
Mary Higgins Clark	4520
Dean R. Koontz	4129
Tom Clancy	3869
Danielle Steel	3544
Sue Grafton	3316
Janet Evanovich	3166
Anne Rice	2885
Michael Crichton	2793
J. K. Rowling	2780
V.C. Andrews	2699
JOHN GRISHAM	2688

▶ Categories

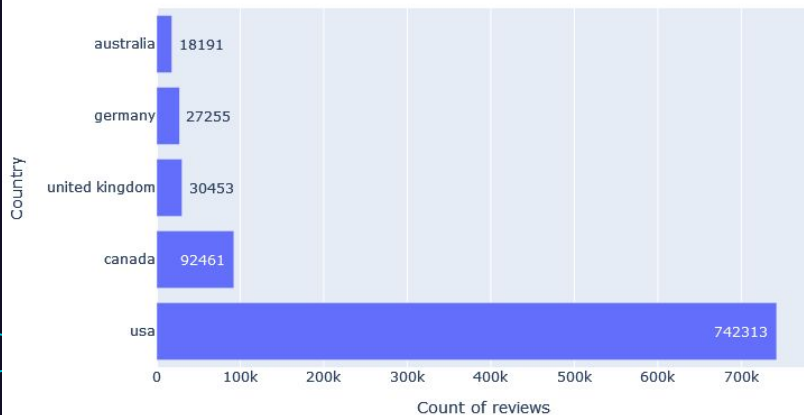6,347 unique categories consolidated into broader categories.

Top Categories by Average Rating

Geographical Data

Categories with more frequency

Country	Count of reviews
australia	18191
germany	27255
united kingdom	30453
canada	92461
usa	742313

States with More Reviews

State	Number of Reviews
minnesota	26036
georgia	26404
virginia	27169
new york	29338
illinois	30007
washington	31833
missouri	32878
florida	34188
texas	43989
california	106775
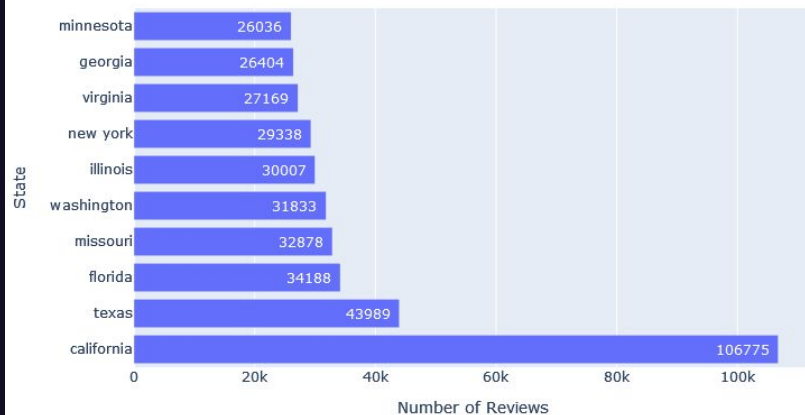
▶ Next steps in terms of data processing

Titles

Feature Hashing, mapping each book title to a fixed number of numerical columns.

Or dropping the column

Author, Publishers

Frequency or Mean Encoding for these variables as these columns have many unique values.

Countries

Group the data by broader categories to encapsulate countries into continents and use one-hot encoding.

If the locations to use is USA, we could use Binary encoding for states or one-hot encoding.