What should I Read Next?

A book recommendation engine based on GoodReads Reviews and Ratings

Presentation by: Aditi Jaiswal,Ankit Kumar

Contents

Introduction and Objective 01

About the Dataset 02

Data Cleaning & Preprocessing 03

04 Exploratory Data Analysis

05 Model Development

06 Output, Conclusion & Scope

THE BOOK LOVER'S DILEMMA

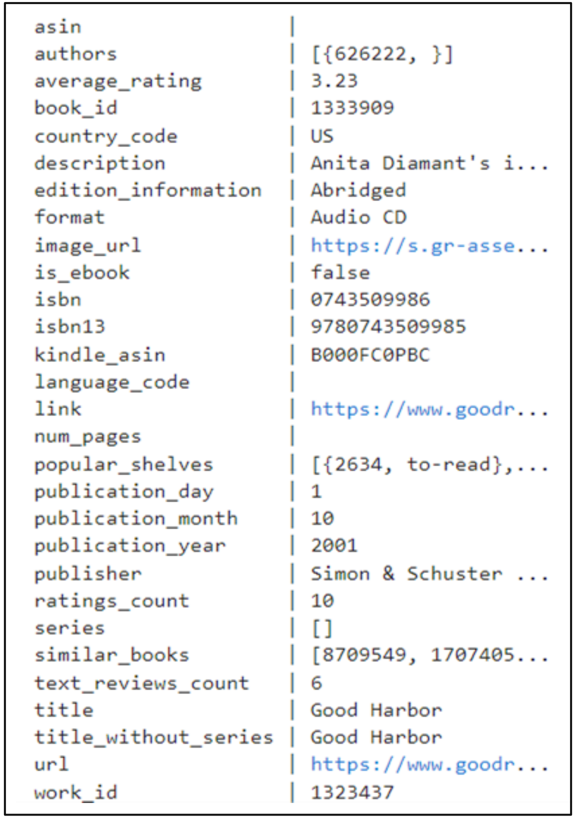
Problem Statement

Identify genres of books and implement a hybrid recommendation engine for better user experience

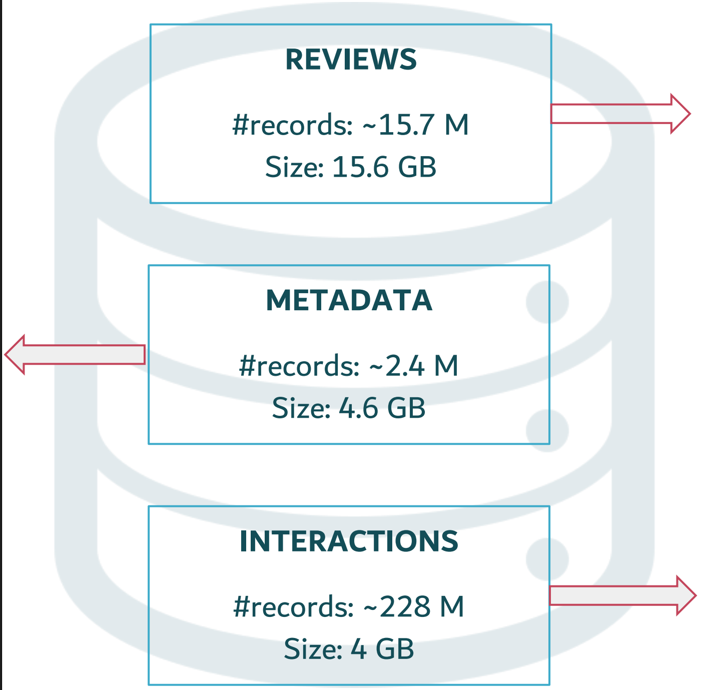
Objective

Develop a hybrid book recommendation system using collaborative filtering and content-based methods, considering user preferences, historical interactions, and genres for accurate and diverse recommendations

Data - Outlook



```
asin                 |
authors              | [{626222, }]
average_rating       | 3.23
book_id              | 1333909
country_code         | US
description          | Anita Diamant's i...
edition_information  | Abridged
format               | Audio CD
image_url            | https://s.gr-asse...
is_ebook             | false
isbn                 | 0743509986
isbn13               | 9780743509985
kindle_asin          | B000FC0PBC
language_code        |
link                 | https://www.goodr...
num_pages            |
popular_shelves      | [{2634, to-read},...
publication_day      | 1
publication_month    | 10
publication_year     | 2001
publisher            | Simon & Schuster ...
ratings_count        | 10
series               | []
similar_books        | [8709549, 1707405...
text_reviews_count   | 6
title                | Good Harbor
title_without_series | Good Harbor
url                  | https://www.goodr...
work_id              | 1323437
```



REVIEWS

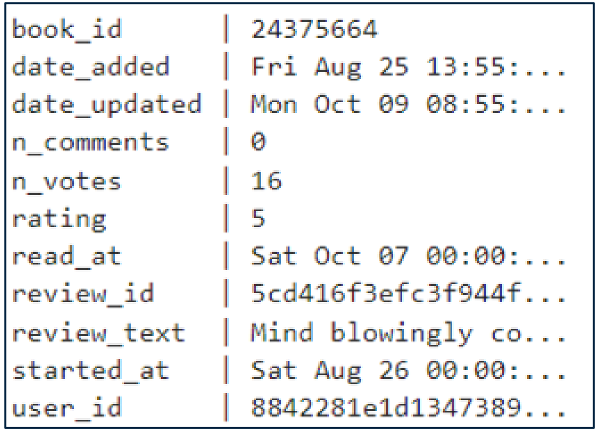
#records: ~15.7 M
Size: 15.6 GB

METADATA

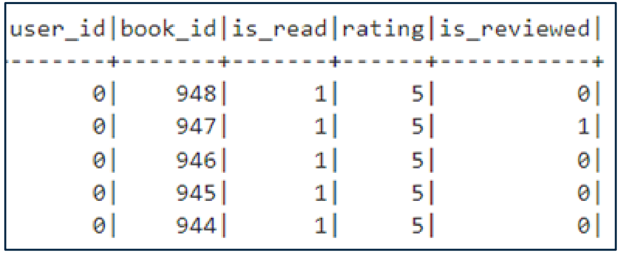
#records: ~2.4 M
Size: 4.6 GB

INTERACTIONS

#records: ~228 M
Size: 4 GB

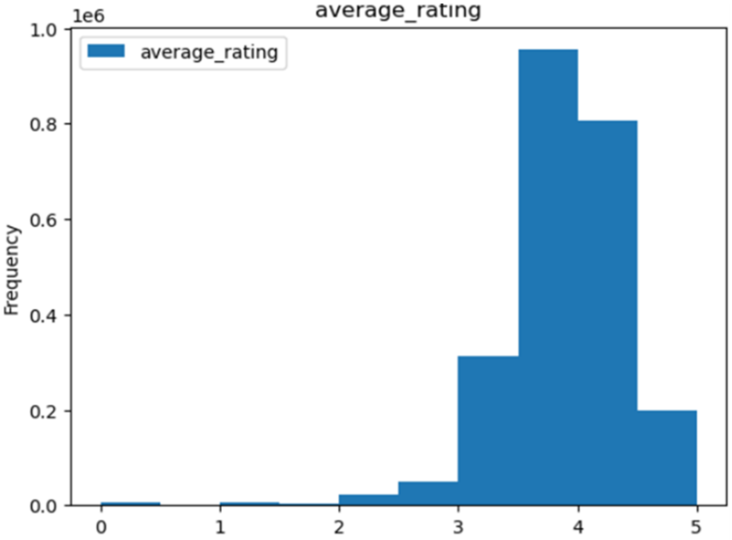


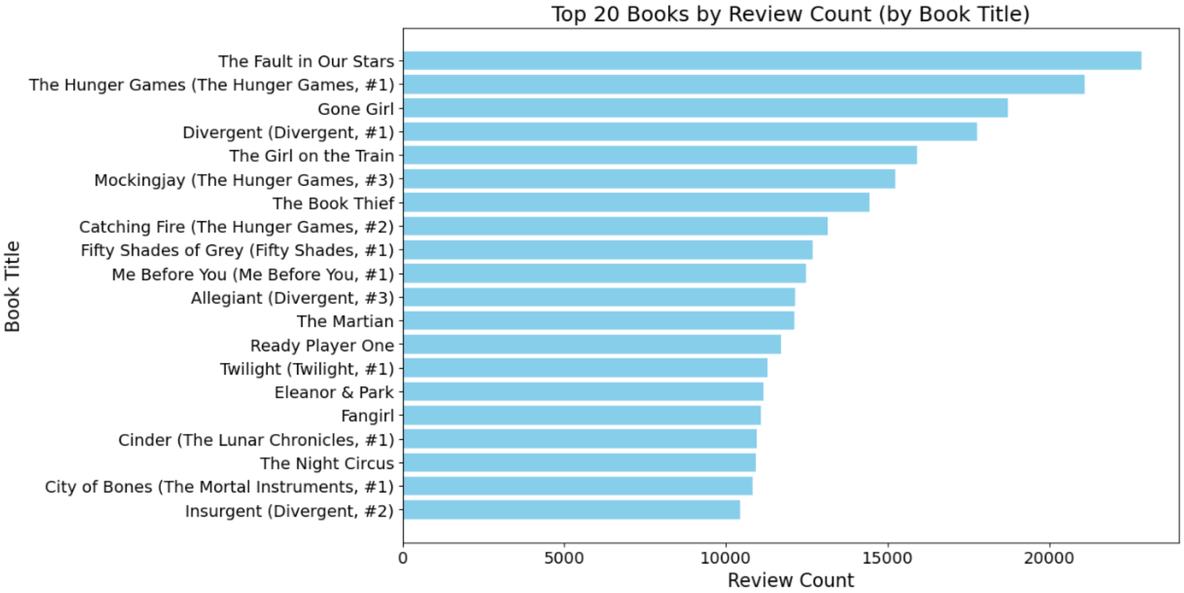
```
book_id      | 24375664
date_added   | Fri Aug 25 13:55:...
date_updated | Mon Oct 09 08:55:...
n_comments   | 0
n_votes      | 16
rating       | 5
read_at      | Sat Oct 07 00:00:...
review_id    | 5cd416f3efc3f944f...
review_text  | Mind blowingly co...
started_at   | Sat Aug 26 00:00:...
user_id      | 8842281e1d1347389...
```



```
user_id|book_id|is_read|rating|is_reviewed|
-------+-------+-------+------+-----------+
      0|    948|      1|     5|          0|
      0|    947|      1|     5|          1|
      0|    946|      1|     5|          0|
      0|    945|      1|     5|          0|
      0|    944|      1|     5|          0|
```

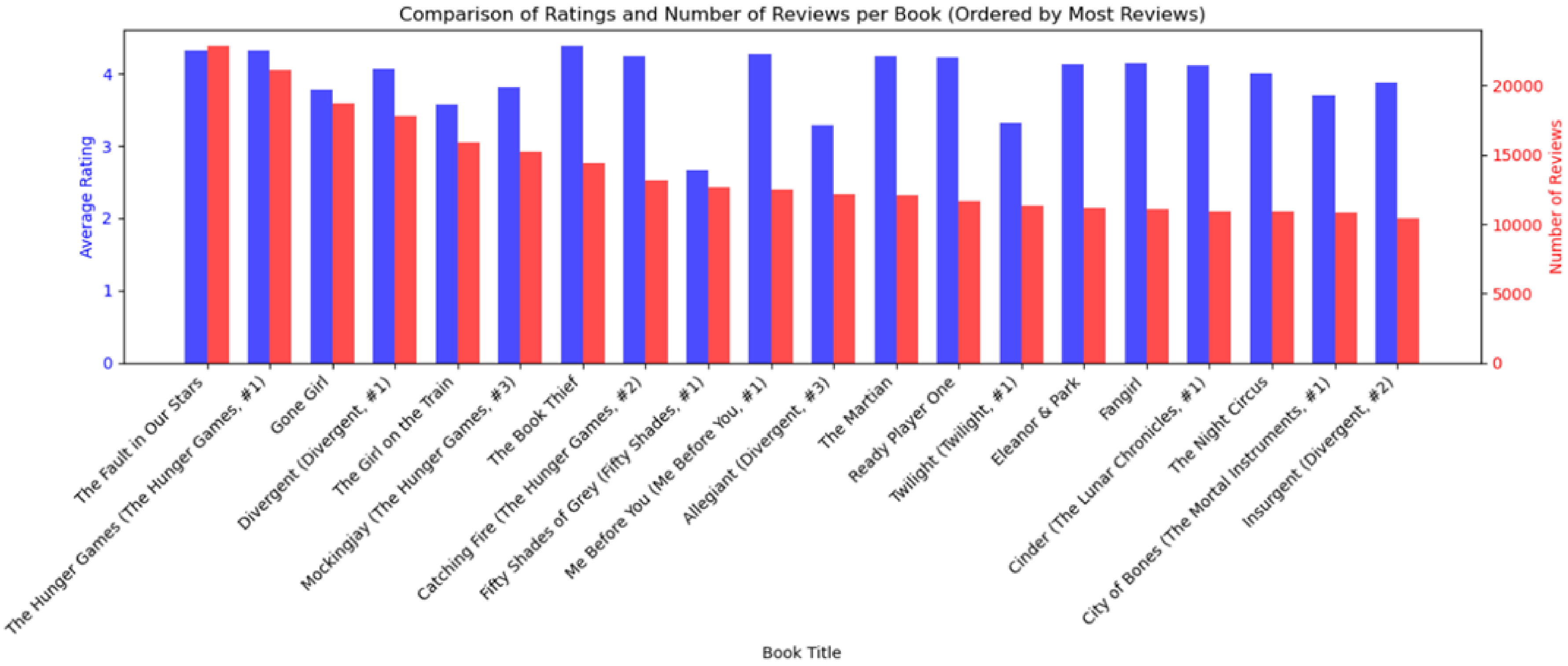
How are books being rated..?



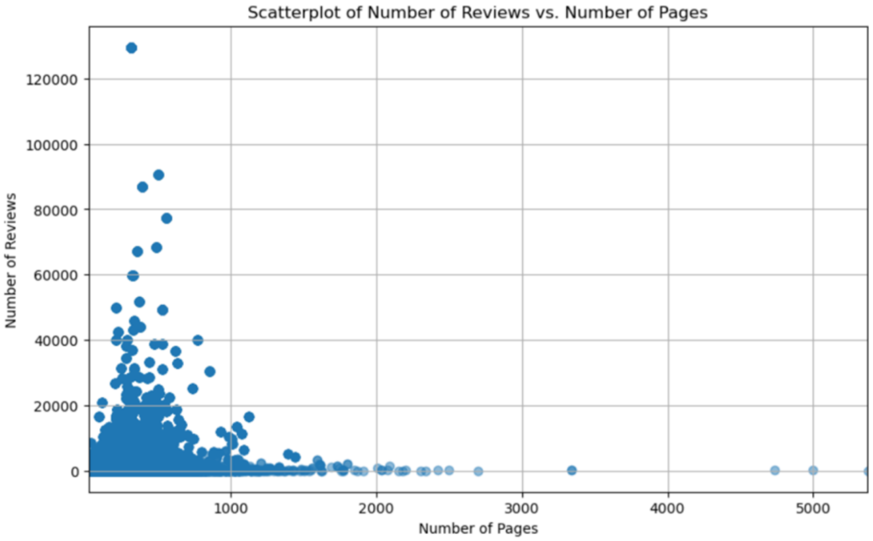


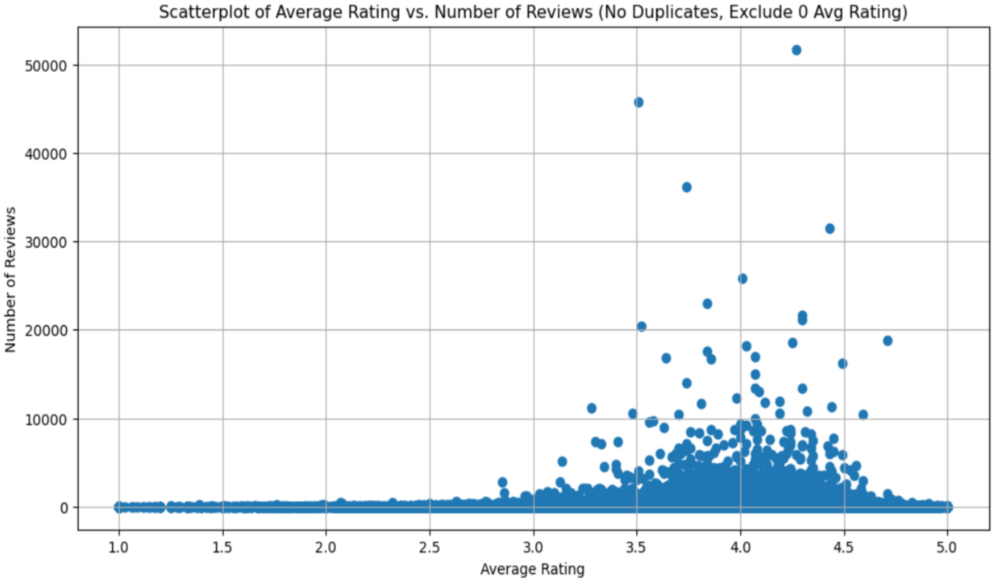
Most of the books on Goodreads have a rating between 3.5 and 4.5.

Ratings and Review for the Most Popular Books



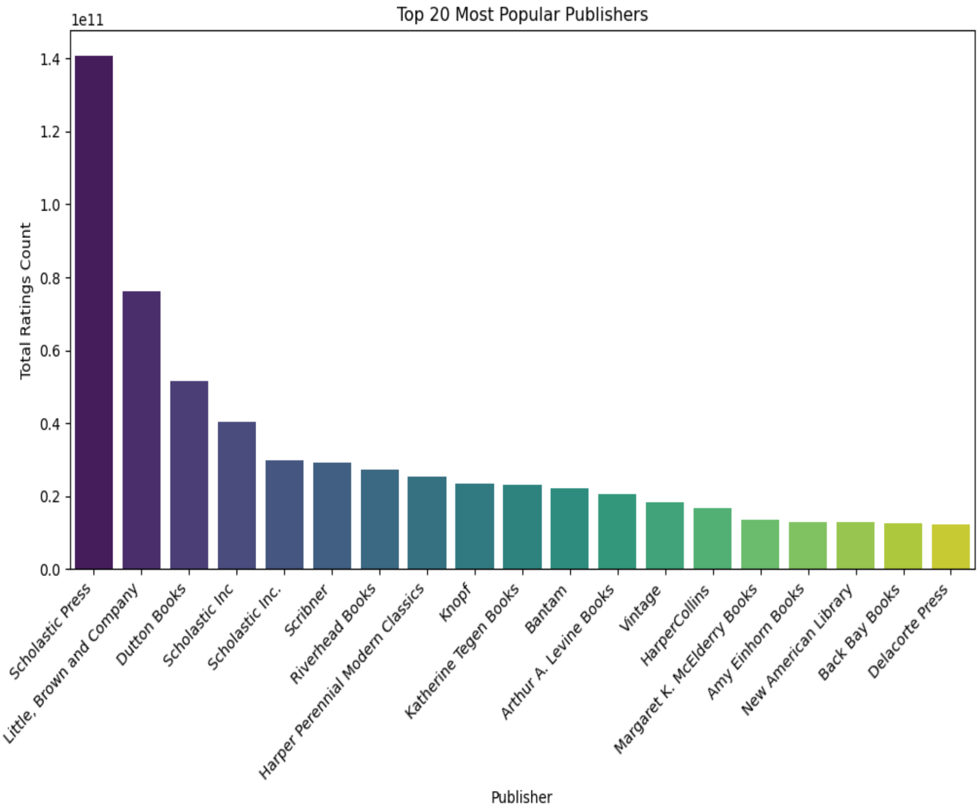
User Reading Patterns

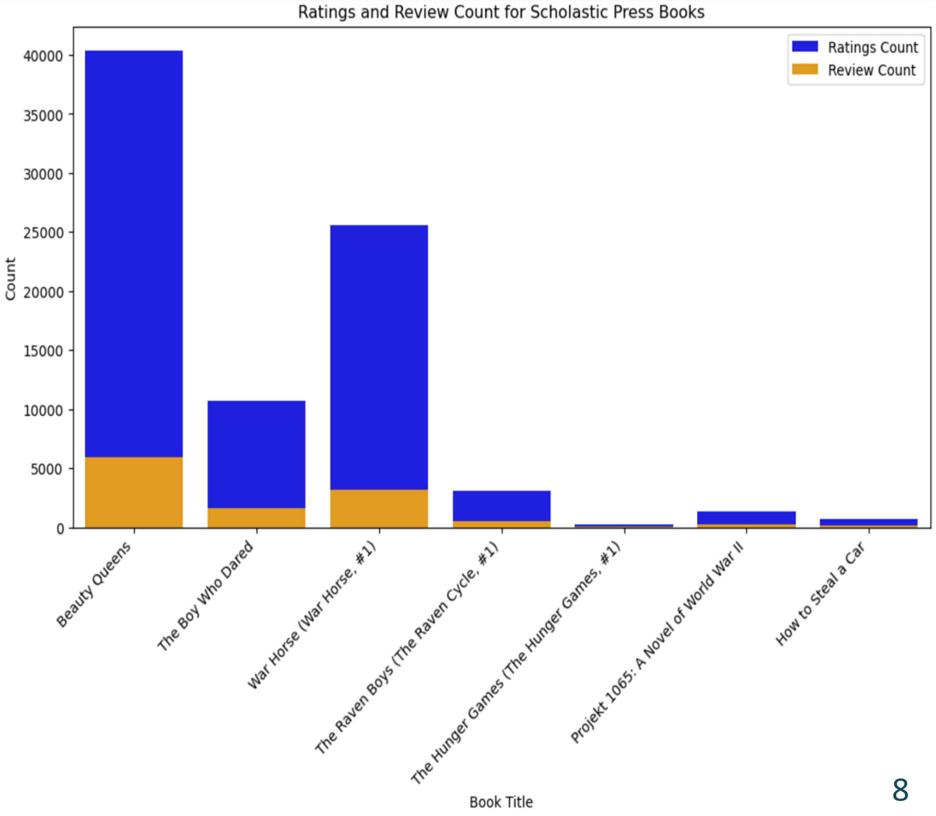




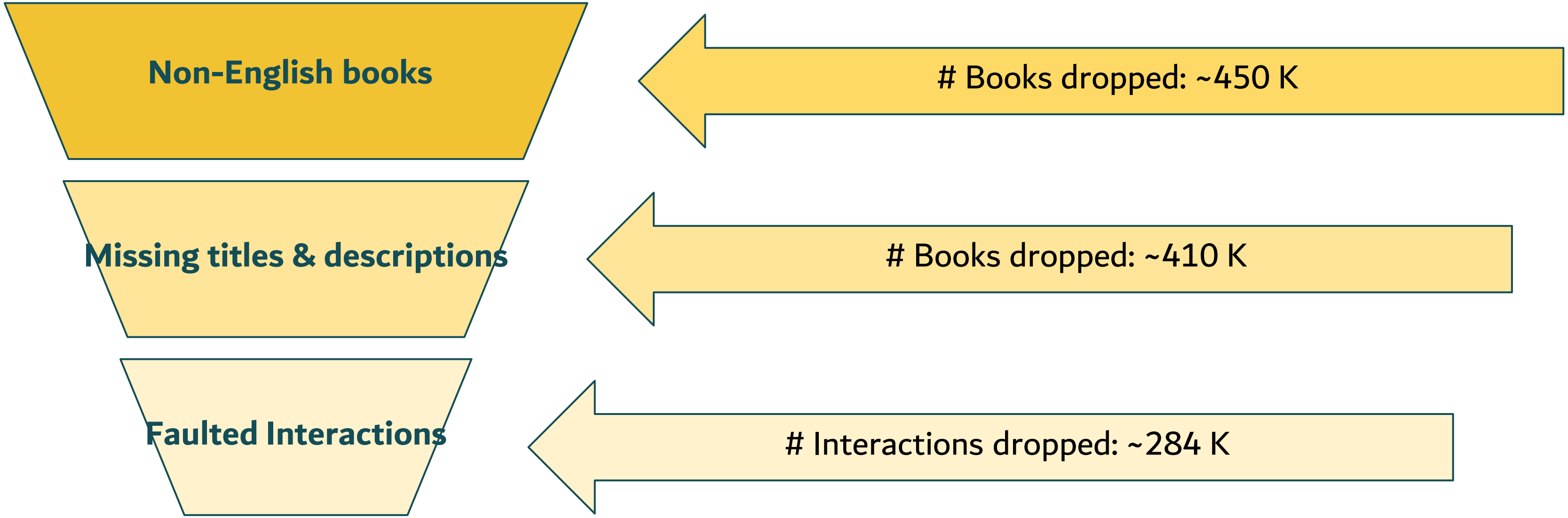
- Books with average number of pages around 250 receive higher number of reviews.
- Books with higher ratings also tend to receive more reviews from readers.

Stochastic Press is the Top Performing Publisher





Data Cleaning and Preprocessing



Non-English books

# Books dropped: ~450 K

Missing titles & descriptions

# Books dropped: ~410 K

Faulted Interactions

# Interactions dropped: ~284 K

Genre Identification - Handling Missing Language Code

Purpose of Clustering - Identify **salient genres** of books for building a hybrid recommendation system.

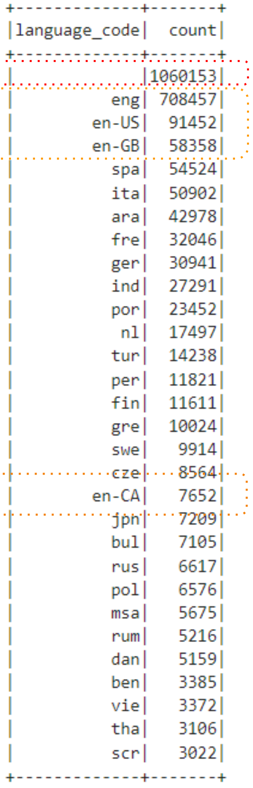
K-Means clustering is chosen for its ability to **group similar items** based on features, making it **ideal for genre identification** instead of Gaussian clustering.

Utilized a pre trained language decoder **(LanguageDetectorDL)** to identify and **extract the language** from book with **missing descriptions**.

Ensures **clustering** is performed on **English descriptions** only, maintaining consistency in the analysis.

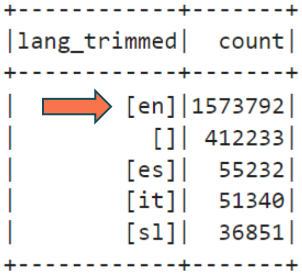
~1M Missing Values (43%)

~858k English Books



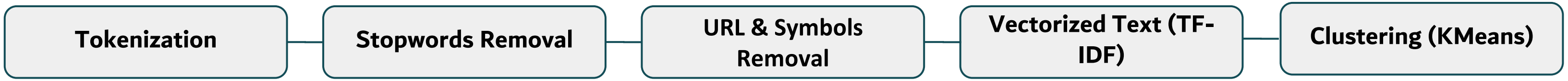
```
+-------------+-------+
|language_code|  count|
+-------------+-------+
|             |1060153|
|          eng| 708457|
|        en-US|  91452|
|        en-GB|  58358|
|          spa|  54524|
|          ita|  50902|
|          ara|  42978|
|          fre|  32046|
|          ger|  30941|
|          ind|  27291|
|          por|  23452|
|           nl|  17497|
|          tur|  14238|
|          per|  11821|
|          fin|  11611|
|          gre|  10024|
|          swe|   9914|
|          cze|   8564|
|        en-CA|   7652|
|          jpn|   7209|
|          bul|   7105|
|          rus|   6617|
|          pol|   6576|
|          msa|   5675|
|          rum|   5216|
|          dan|   5159|
|          ben|   3385|
|          vie|   3372|
|          tha|   3106|
|          scr|   3022|
+-------------+-------+
```

LanguageDetectorDL Results



```
+------------+-------+
|lang_trimmed|  count|
+------------+-------+
|        [en]|1573792|
|          []| 412233|
|        [es]|  55232|
|        [it]|  51340|
|        [sl]|  36851|
+------------+-------+
```

Genre Identification - Clustering



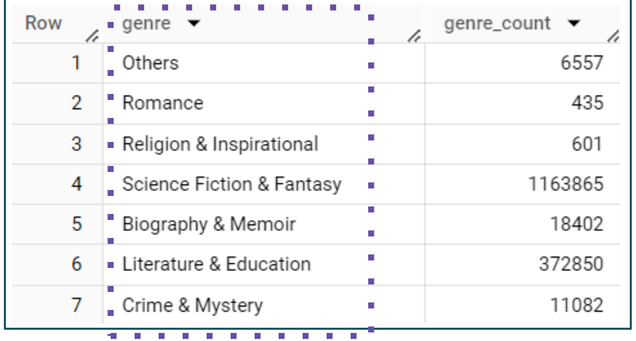
K-means **partitions data** into **k clusters** based on **similarity**.

Represent the text documents as **numerical vectors** using techniques like **TF-IDF.**

Computed the **silhouette score** for different values of k and choose the one with the highest score.

The final clusters acquired represent the **book genres** employed in **content-based filtering** for hybrid **recommendation systems.**

(Silhouette Score = 0.401)
Clusters = 7



Row	genre	genre_count
1	Others	6557
2	Romance	435
3	Religion & Inspirational	601
4	Science Fiction & Fantasy	1163865
5	Biography & Memoir	18402
6	Literature & Education	372850
7	Crime & Mystery	11082

Recommendation Model - Overview

Popularity Based

A recommendation technique that suggests items based on their popularity or, more precisely, their frequency of being chosen or interacted with by users.

Popularity Filtering is implemented for new users with no prior interaction

Collaborative Filtering

This approach uses ALS algorithm to recommend books based on patterns it has identified in the user-item interactions by analyzing the behavior of users and identifying latent factors that represent their preferences.

70% weightage is given to results from this model

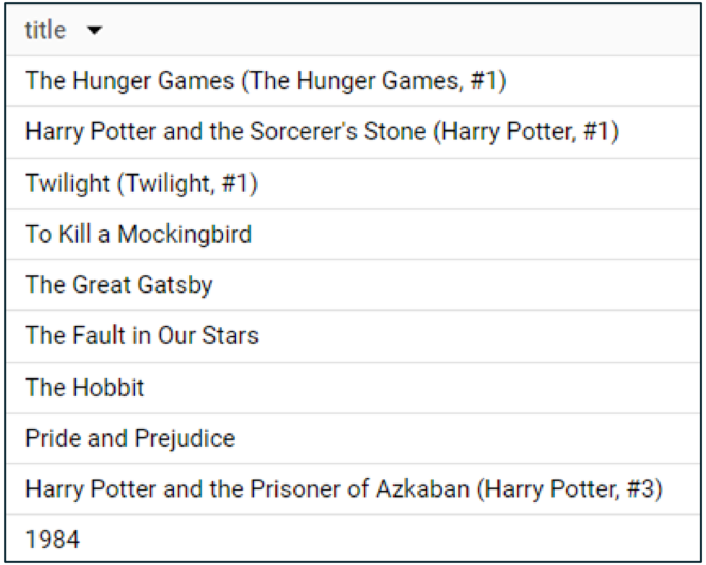
Content-Based

This approach recommends books based on the content of the items themselves. It analyzes the genre of the books, and matches them with the user's preferences.

30% weightage is given to results from this model

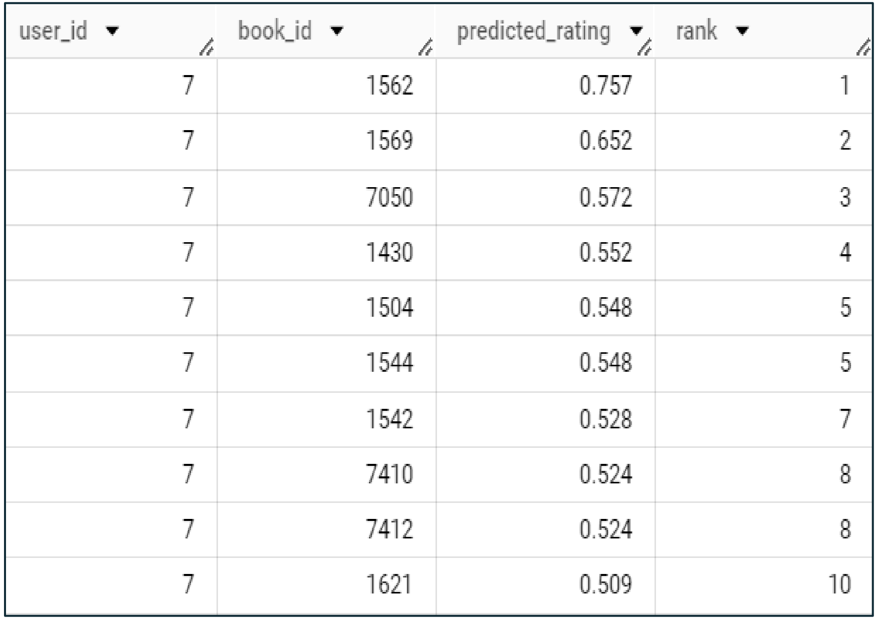
Hybrid Recommendation System - Output

Popularity Based Filtering



title
The Hunger Games (The Hunger Games, #1)
Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
Twilight (Twilight, #1)
To Kill a Mockingbird
The Great Gatsby
The Fault in Our Stars
The Hobbit
Pride and Prejudice
Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)
1984

Hybrid Recommendation Output



user_id	book_id	predicted_rating	rank
7	1562	0.757	1
7	1569	0.652	2
7	7050	0.572	3
7	1430	0.552	4
7	1504	0.548	5
7	1544	0.548	5
7	1542	0.528	7
7	7410	0.524	8
7	7412	0.524	8
7	1621	0.509	10

Future Work - Scope

FAKE REVIEWS DETECTION

Critical for a reliable and fair ecosystem for users, authors, and publishers. It contributes to the overall health of the platform by promoting authenticity.

COMMUNITY DETECTION

Understanding user communities can help in targeted marketing strategies and improve collaborative filtering

IDENTIFY TRENDING BOOKS AND AUTHORS

Identify trending books and authors based on recent user activity, ratings, and reviews

RARITY AND UNIQUENESS RECOMMENDATIONS

It seeks out hidden gems and literary treasures that might not be on the user's radar and help them explore niche genres

ADD NODES TO GRAPH FRAME

Add nodes for authors and publishers; to better understand the relation between users, books, authors and publishers

SENTIMENT ANALYSIS ON REVIEWS

Understand the sentiment and opinions expressed by users towards specific books

THANK YOU