# Book Recommendation System

Satyam Shrivastava, Pritish Arora

Northeastern University

# Overview

**Which book to read next?**

- In the quest for a new book, we often seek recommendations from friends or tirelessly search the internet or library shelves.
- Despite our efforts, finding a book that aligns with our preferences remains a challenge due to diverse interests.

Source: Image by Tenor

# Goal – Why this is relevant?

**Need for Recommendations:**

- Recognizing the uniqueness of individual interests, we encounter a need for a tailored solution.
- Introducing a system that goes beyond conventional searches, considering our choices for personalized book suggestions.

**Essence of Recommender Systems:**

- A good recommender system must consider how users interact with the recommendations, and leverage user data to deliver personalized suggestions, ensuring a more engaging and relevant reading experience.

**Project Objective:**

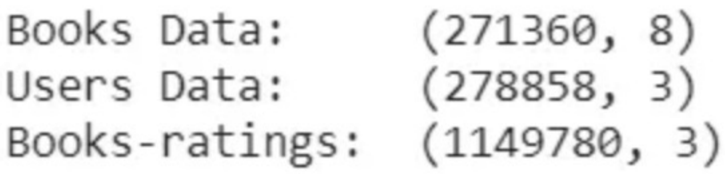
- Our project centers around creating an advanced book recommendation system.
- By harnessing the power of machine learning and the Book-Crossing dataset, we **aim to redefine how readers discover their next favorite book**.

# Dataset

Our project relies on the comprehensive Book-Crossing dataset, which comprises 3 files: **Books**, **Users**, and **Ratings**. It provides a rich source of information, including user demographics, book details, and user ratings.

**Dataset Components:**

- **Books**: Featuring book details and consists 8 columns: ISBN, Book title, Book author, Year of publication, Publisher, and 3 Image URL columns for various cover sizes.
- **Users**: Contains the user's information with 3 columns: UserID, Location, and Age.
- **Ratings**: Stores user ratings of the books on a scale from 1 to 10, with 3 columns: UserID, ISBN, and Book Rating.



```
Books Data:      (271360, 8)
Users Data:      (278858, 3)
Books-ratings:   (1149780, 3)
```

Dataset Shape

# Project Workflow

This systematic approach guides this project from inception to model implementation:

**Data Collection**: Obtain the dataset, and create Users, Books, and Ratings tables.

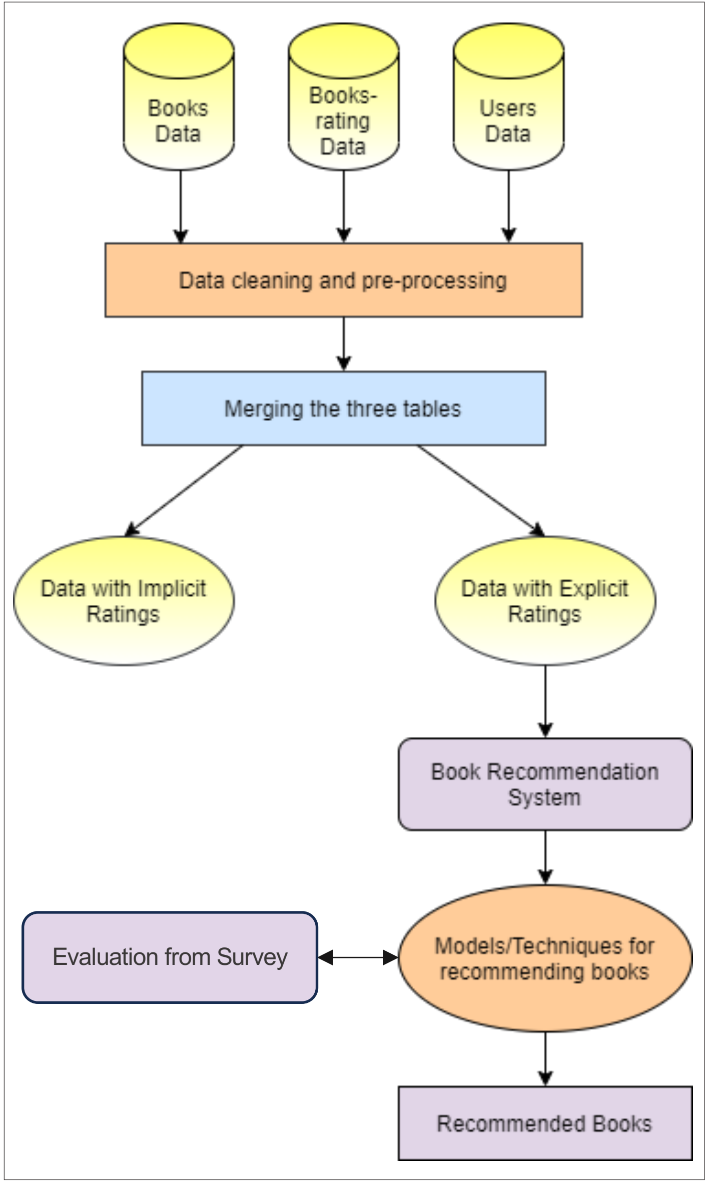
**Data Cleaning and Preprocessing**: Cleaning and refinement of each table, addressing null values, duplicates, and inconsistencies.

**Exploratory Data Analysis (EDA)**: To unveil hidden patterns, such as implicit and explicit ratings, and other distribution nuances.

**Model Development**: Implement models, including both low-risk options and high-risk, high-reward strategies.

**Evaluation**: Introducing surveys for subjective feedback, aligning with the absence of traditional labeled data.

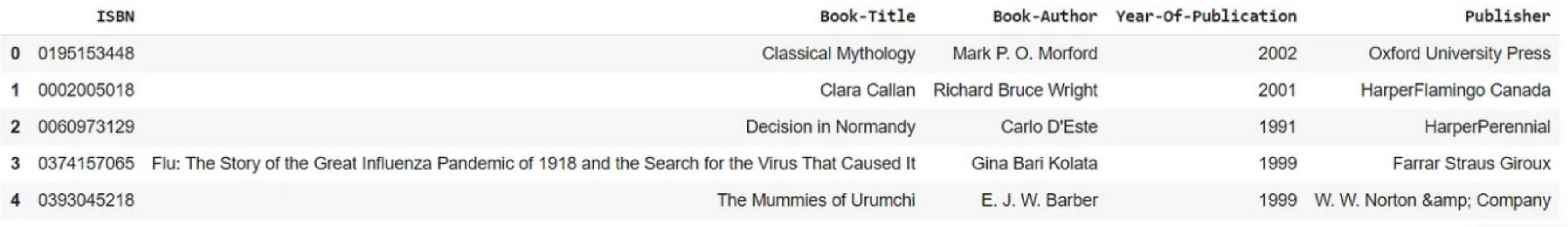
**Recommendations**: Recommended books based on distinct algorithms.


Books Data
Books-rating Data
Users Data
Data cleaning and pre-processing
Merging the three tables
Data with Implicit Ratings
Data with Explicit Ratings
Book Recommendation System
Evaluation from Survey
Models/Techniques for recommending books
Recommended Books

# Pre-processing & Cleaning

**Books Table:**

- **Image URL Features**: Removal of all three Image URL features.
- **Null Values Handling**: Addressed 3 null values, replaced with 'Other'.
- **Publication Year Cleanup**: Resolved inconsistencies in the Year of Publication column, manually correcting publisher entries and merging author names.
- **Year Type Conversion**: Conversion of the publication years to integers for consistency.
- **Invalid Year Replacement**: Replaced invalid years (< 2022 and not 0) with the mode (2002) for accuracy.
- **ISBN Standardization**: Uppercased ISBN and removed duplicates.



|  | ISBN | Book-Title | Book-Author | Year-Of-Publication | Publisher |
|---|---|---|---|---|---|
| 0 | 0195153448 | Classical Mythology | Mark P. O. Morford | 2002 | Oxford University Press |
| 1 | 0002005018 | Clara Callan | Richard Bruce Wright | 2001 | HarperFlamingo Canada |
| 2 | 0060973129 | Decision in Normandy | Carlo D'Este | 1991 | HarperPerennial |
| 3 | 0374157065 | Flu: The Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It | Gina Bari Kolata | 1999 | Farrar Straus Giroux |
| 4 | 0393045218 | The Mummies of Urumchi | E. J. W. Barber | 1999 | W. W. Norton &amp; Company |

Books Table

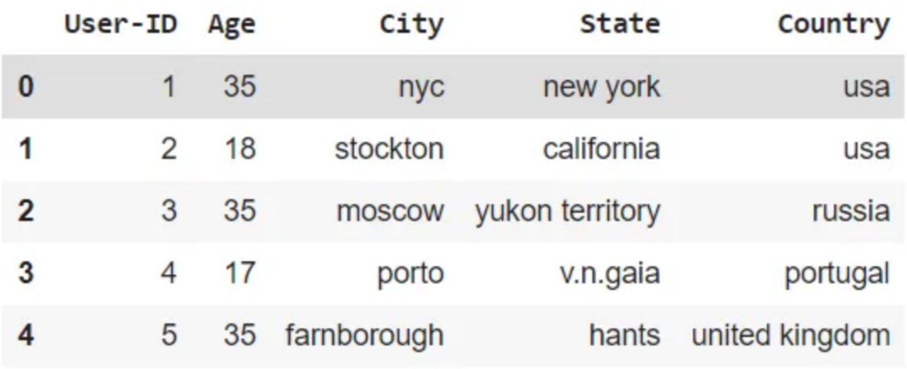
# Pre-processing & Cleaning

## Users Table:

- **Null Values and Age Cleanup**: Addressed over 100,000 null values in the Age column, replacing invalid ages (0 or 244) with the mean within a valid range (10 to 80).
- **Location Information**: Split location values (City, State, Country) and assigned 'Other' for null values. Removed duplicate entries.

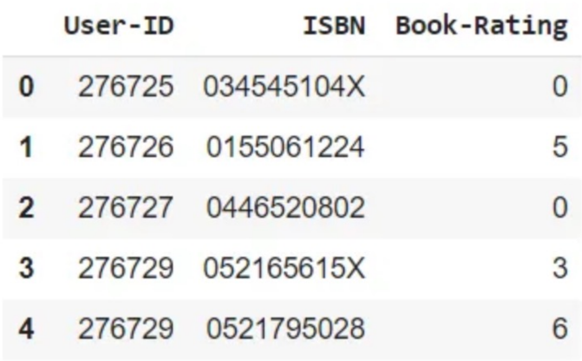
## Ratings Table:

- **Data Validation**: Ensured Rating and User-ID columns are of integer type.
- **ISBN Punctuation and Duplication**: Removed ISBN punctuation, considered entities available in the Books dataset. Eliminated duplicates.



| | User-ID | Age | City | State | Country |
|---|---|---|---|---|---|
| 0 | 1 | 35 | nyc | new york | usa |
| 1 | 2 | 18 | stockton | california | usa |
| 2 | 3 | 35 | moscow | yukon territory | russia |
| 3 | 4 | 17 | porto | v.n.gaia | portugal |
| 4 | 5 | 35 | farnborough | hants | united kingdom |

Users Table



| | User-ID | ISBN | Book-Rating |
|---|---|---|---|
| 0 | 276725 | 034545104X | 0 |
| 1 | 276726 | 0155061224 | 5 |
| 2 | 276727 | 0446520802 | 0 |
| 3 | 276729 | 052165615X | 3 |
| 4 | 276729 | 0521795028 | 6 |

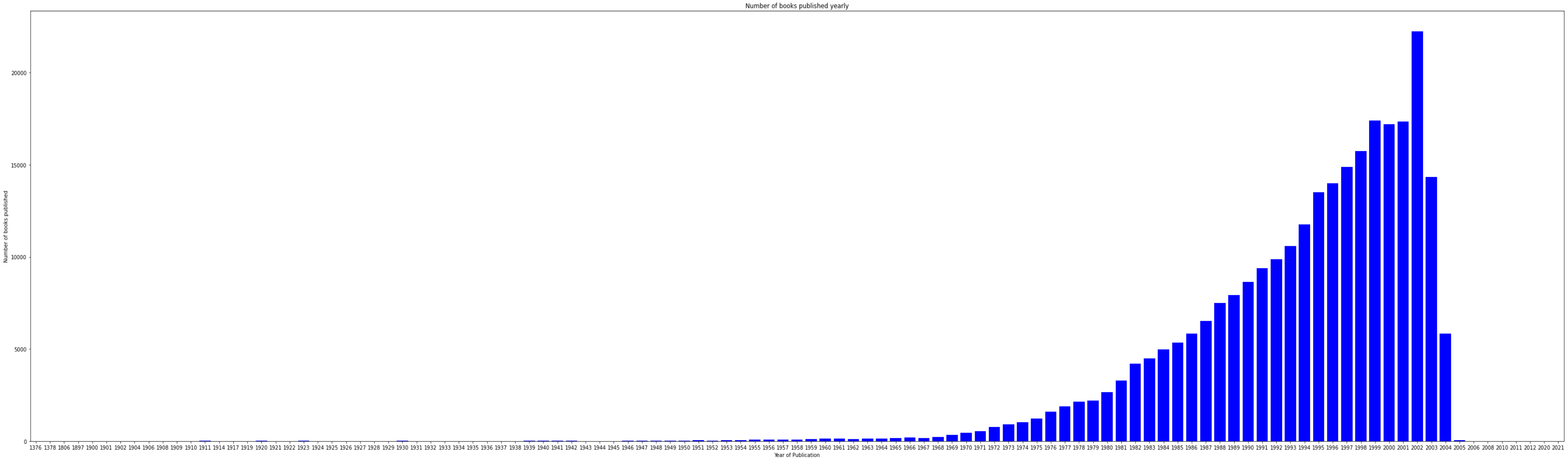
Ratings table

# Exploratory Data Analysis & Findings

**Books Published Over the Years**:

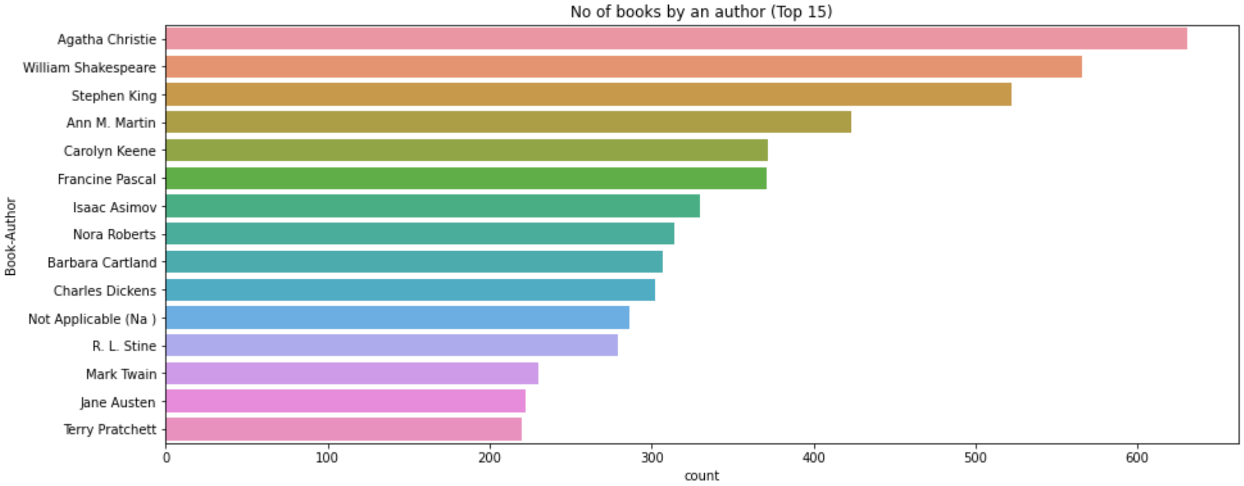
Here we explored the annual distribution of published books through a bar chart below.

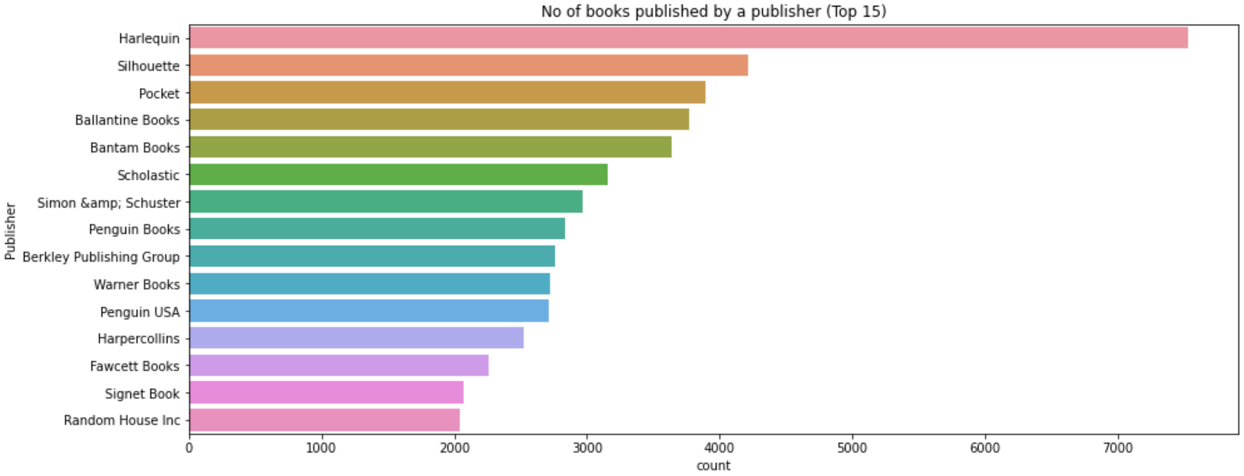
We can see that in the dataset most books are published in 2002, and there is a significant increase in each year's published books after 1968.



# Exploratory Data Analysis & Findings

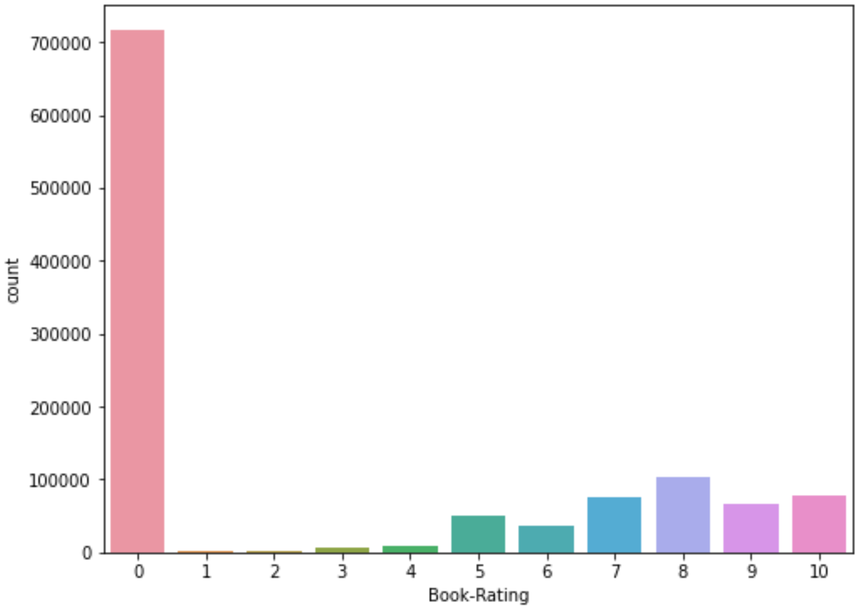
**Top Authors and Publishers**: Count plots showcasing the top 15 authors and publishers based on the number of books.

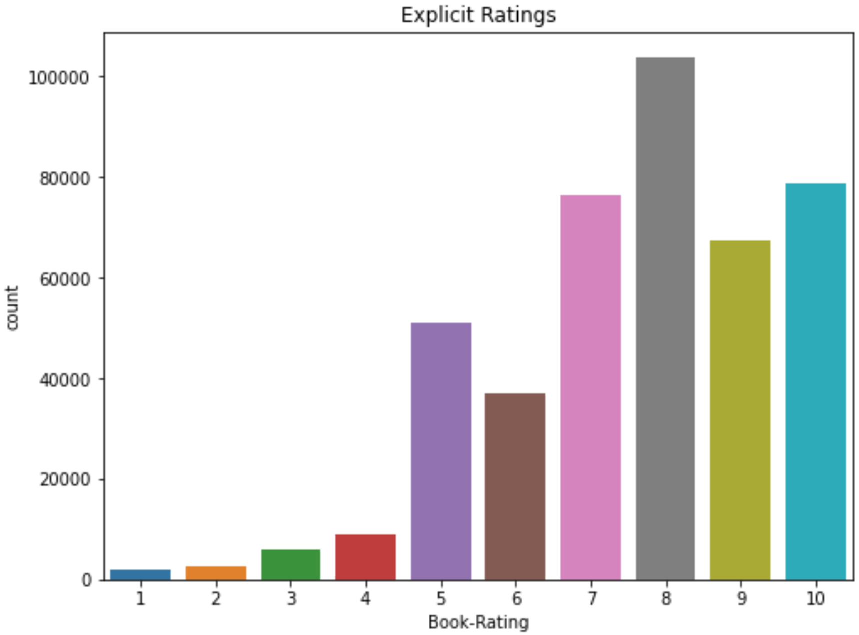




# Exploratory Data Analysis & Findings

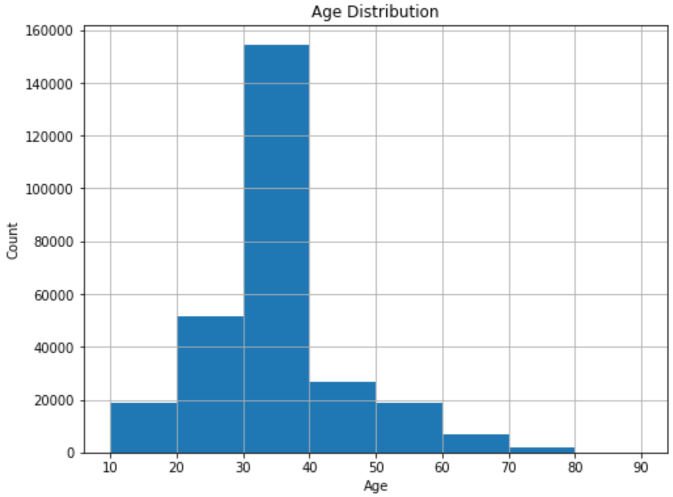
- **Implicit Ratings** (*Left Fig*): Entries where users have not provided numerical ratings (rated as 0). In this case, the system derives insights from user interactions, considering zero ratings as implicit feedback.
- **Explicit Ratings** (*Right Fig*): Entries where users have provided clear and direct ratings to books. These ratings are explicitly stated by the users, representing their preferences.
- Distinguished Implicit and Explicit Ratings in the dataset below:



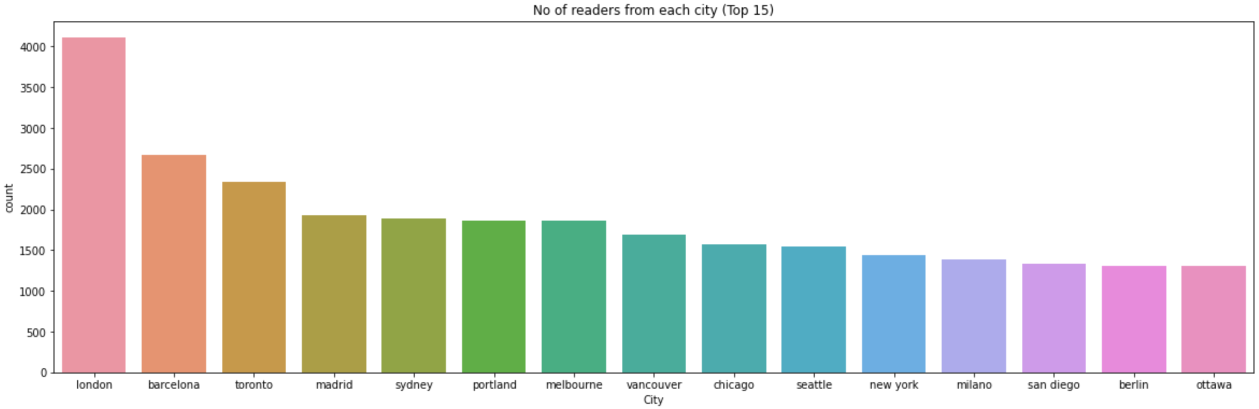


# Exploratory Data Analysis & Findings

- **User Demographics**: Visual representation of age distribution among readers.

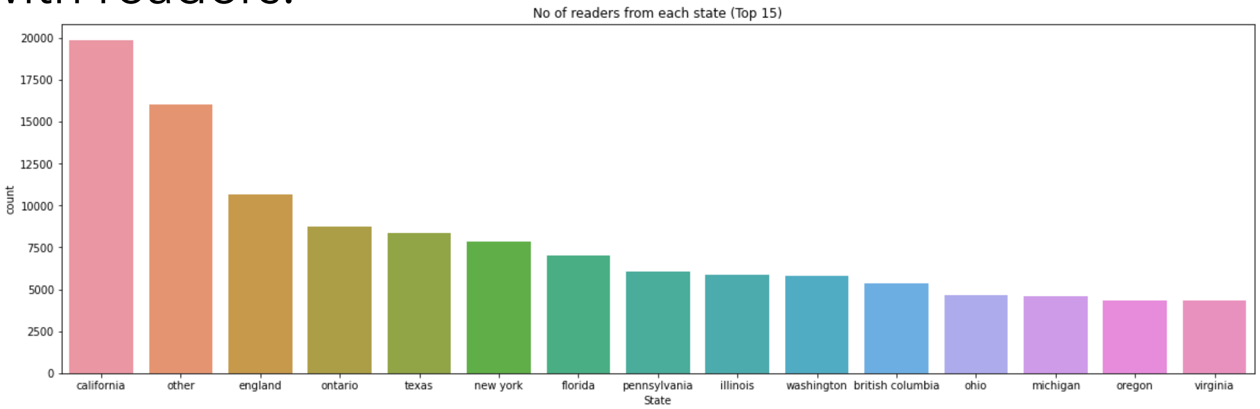


- **User Demographics by City**: Count plots displaying the top 15 cities with readers.

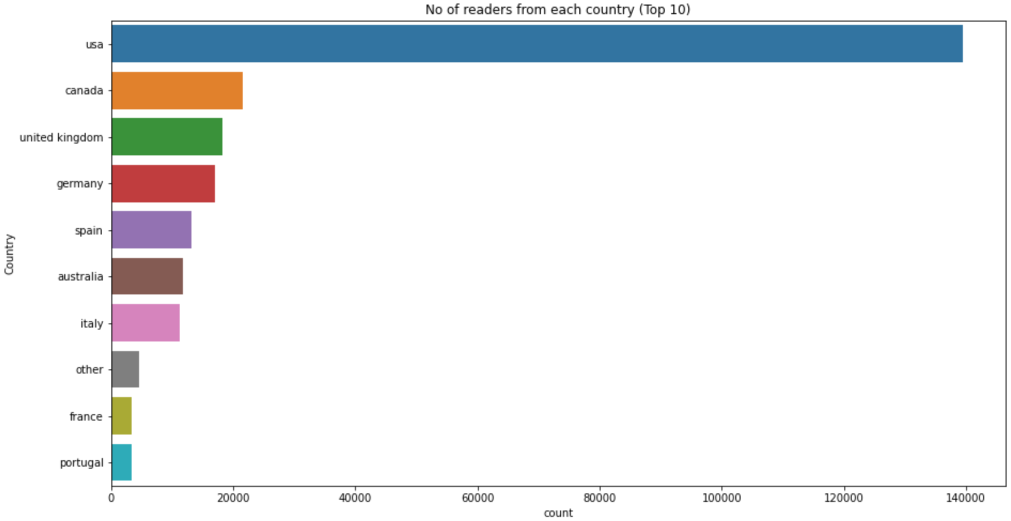


# Exploratory Data Analysis & Findings

- **User Demographics by State**: Count plots displaying the top 15 states with readers.

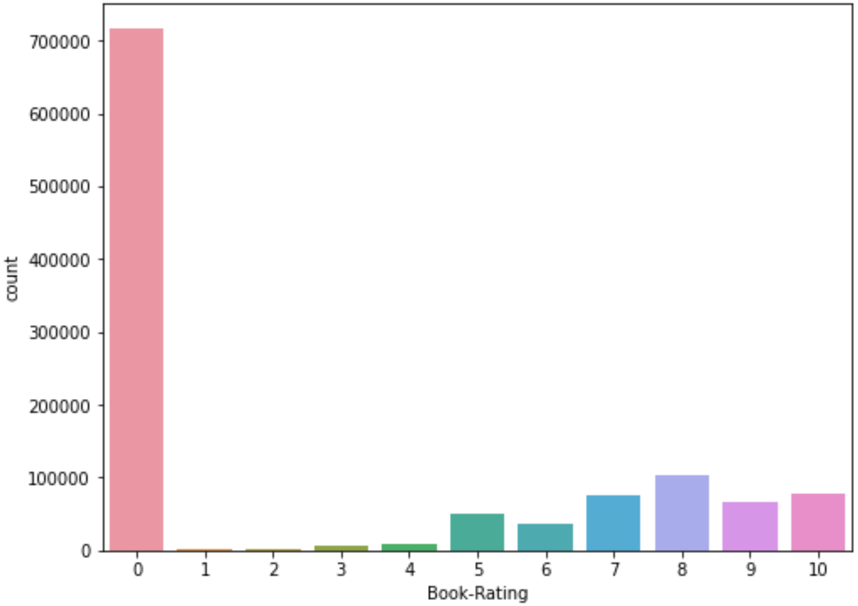


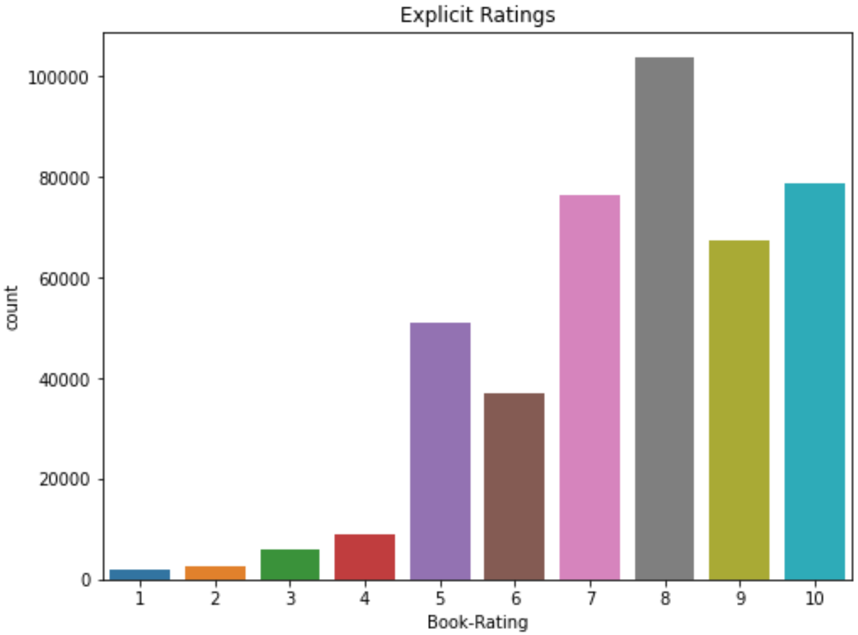
- **User Demographics by Country**: Count plots displaying the top 15 countries with readers.



# Implicit/Explicit Ratings

- After our exploration, we classified user interactions into two significant categories: Implicit and Explicit Ratings.
- Implicit Ratings dataset comprises 647,535 entries while Explicit Ratings dataset comprise 384,074 entries.
- **Exclusive Focus on Explicit Ratings**: By prioritizing explicit ratings (non-zero entries), we ensure precision in understanding user preferences. This intentional choice sharpens our recommendation algorithms, enhancing user experience.

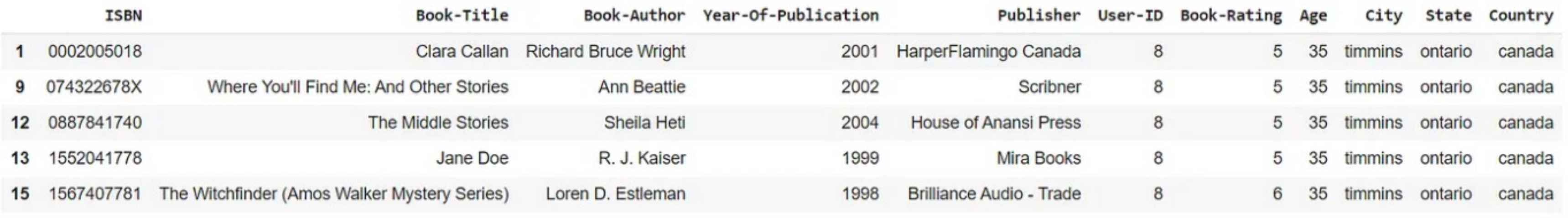




# Final Dataset

## Complete (Merged) Dataset:

- Integration of Books, Users, and Ratings tables into a cohesive dataset.



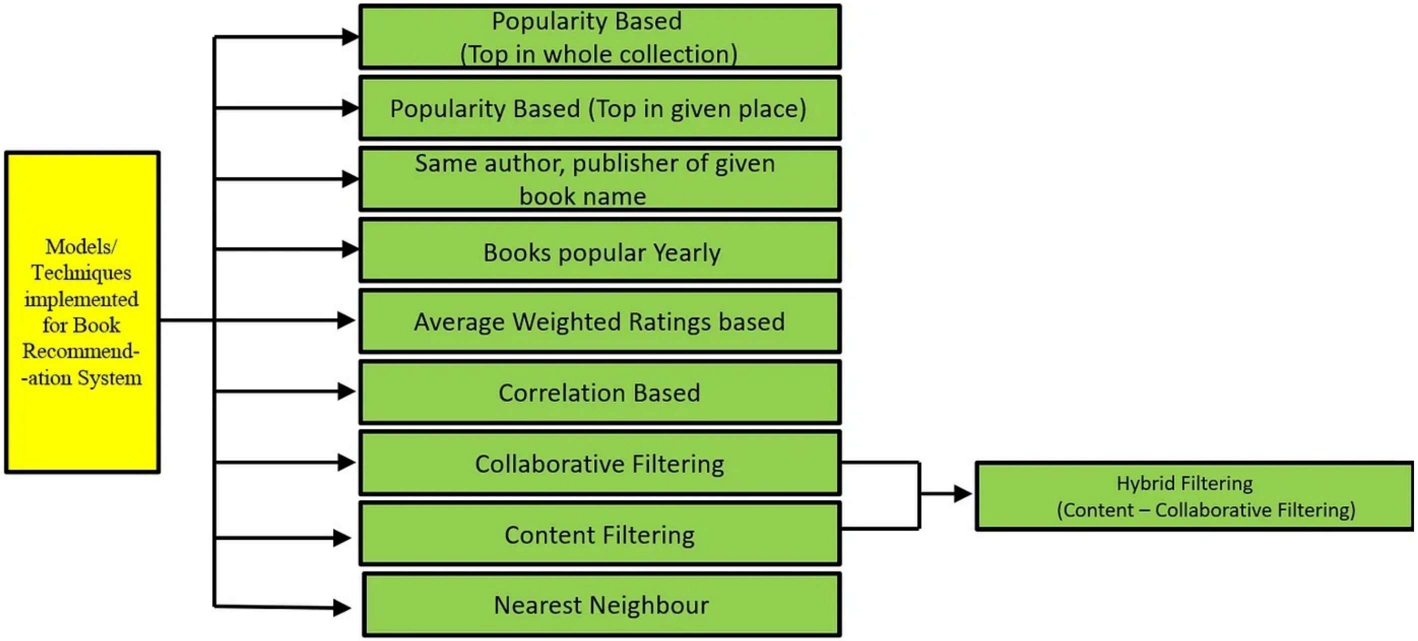
| | ISBN | Book-Title | Book-Author | Year-Of-Publication | Publisher | User-ID | Book-Rating | Age | City | State | Country |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0002005018 | Clara Callan | Richard Bruce Wright | 2001 | HarperFlamingo Canada | 8 | 5 | 35 | timmins | ontario | canada |
| 9 | 074322678X | Where You'll Find Me: And Other Stories | Ann Beattie | 2002 | Scribner | 8 | 5 | 35 | timmins | ontario | canada |
| 12 | 0887841740 | The Middle Stories | Sheila Heti | 2004 | House of Anansi Press | 8 | 5 | 35 | timmins | ontario | canada |
| 13 | 1552041778 | Jane Doe | R. J. Kaiser | 1999 | Mira Books | 8 | 5 | 35 | timmins | ontario | canada |
| 15 | 1567407781 | The Witchfinder (Amos Walker Mystery Series) | Loren D. Estleman | 1998 | Brilliance Audio - Trade | 8 | 6 | 35 | timmins | ontario | canada |

Complete Dataset

- In merging all three tables, our final dataset strategically discards tuples with zero ratings (focus on explicit ratings). This strategic decision centers on precision and clarity, emphasizing user-provided numerical ratings.
- Implicit Ratings Insight: Understanding the significance of implicit ratings, we acknowledge their role in capturing nuanced user interactions. While vital, our system opts for explicit ratings to sharpen recommendations, striking a balance between complexity and precision.

# Recommendation Models

- Our development unfolds in two ways: Low Risk/Low Reward for feasible suggestions and High Risk/High Reward for uncovering nuanced patterns that may go unnoticed with simpler models.



Models/ Techniques implemented for Book Recommend- -ation System

Popularity Based (Top in whole collection)

Popularity Based (Top in given place)

Same author, publisher of given book name

Books popular Yearly

Average Weighted Ratings based

Correlation Based

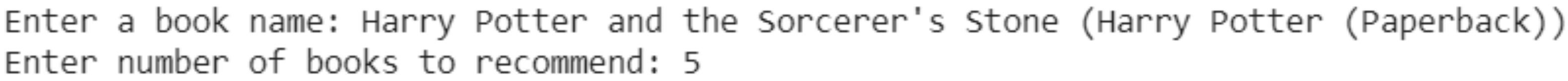
Collaborative Filtering

Content Filtering

Nearest Neighbour

Hybrid Filtering (Content – Collaborative Filtering)

- Input given for required models is:



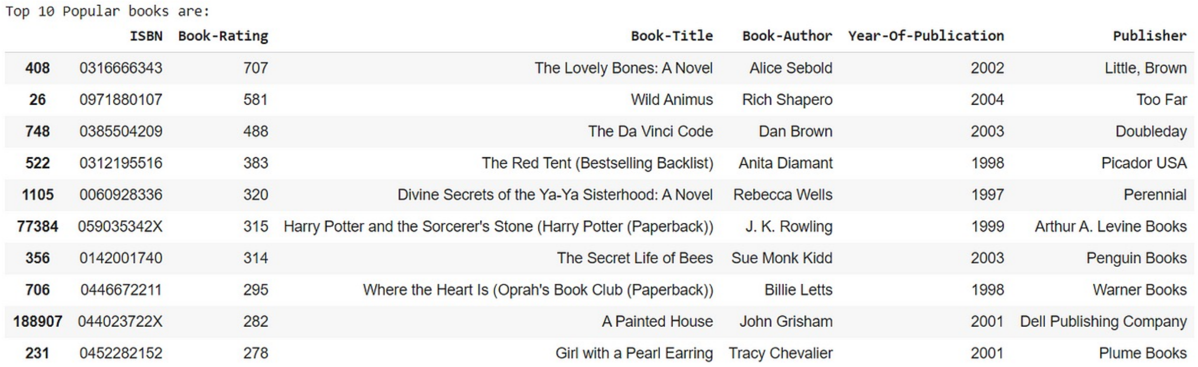
```
Enter a book name: Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
Enter number of books to recommend: 5
```

# Recommendation Models – Low Risk

Our journey into recommendation systems begins with low-risk models, ensuring a feasible foundation for user-centric suggestions.

1. Popularity-Based Approach:

- Global Popularity: Identifies top-rated books in whole collection.

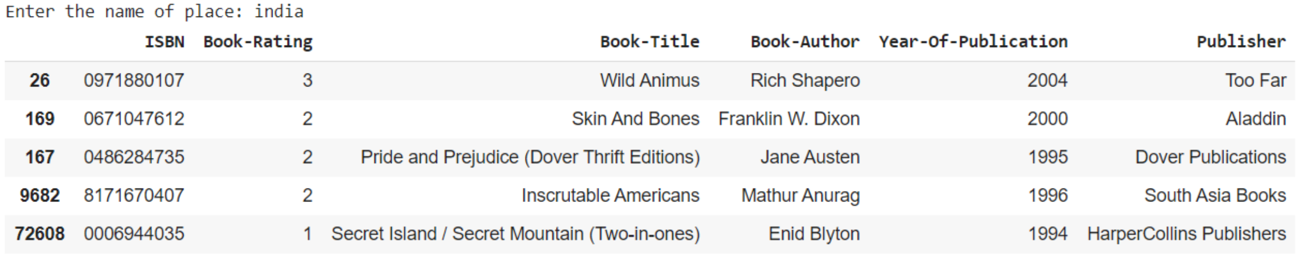


```
Top 10 Popular books are:
```

| | ISBN | Book-Rating | Book-Title | Book-Author | Year-Of-Publication | Publisher |
|---|---|---|---|---|---|---|
| 408 | 0316666343 | 707 | The Lovely Bones: A Novel | Alice Sebold | 2002 | Little, Brown |
| 26 | 0971880107 | 581 | Wild Animus | Rich Shapero | 2004 | Too Far |
| 748 | 0385504209 | 488 | The Da Vinci Code | Dan Brown | 2003 | Doubleday |
| 522 | 0312195516 | 383 | The Red Tent (Bestselling Backlist) | Anita Diamant | 1998 | Picador USA |
| 1105 | 0060928336 | 320 | Divine Secrets of the Ya-Ya Sisterhood: A Novel | Rebecca Wells | 1997 | Perennial |
| 77384 | 059035342X | 315 | Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback)) | J. K. Rowling | 1999 | Arthur A. Levine Books |
| 356 | 0142001740 | 314 | The Secret Life of Bees | Sue Monk Kidd | 2003 | Penguin Books |
| 706 | 0446672211 | 295 | Where the Heart Is (Oprah's Book Club (Paperback)) | Billie Letts | 1998 | Warner Books |
| 188907 | 044023722X | 282 | A Painted House | John Grisham | 2001 | Dell Publishing Company |
| 231 | 0452282152 | 278 | Girl with a Pearl Earring | Tracy Chevalier | 2001 | Plume Books |

We have sorted the dataset according to the total ratings each of the books have received in non-increasing order and then recommended top n books.

- Localized Popularity: Refines recommendations based on user-specified location (City/State/Country).



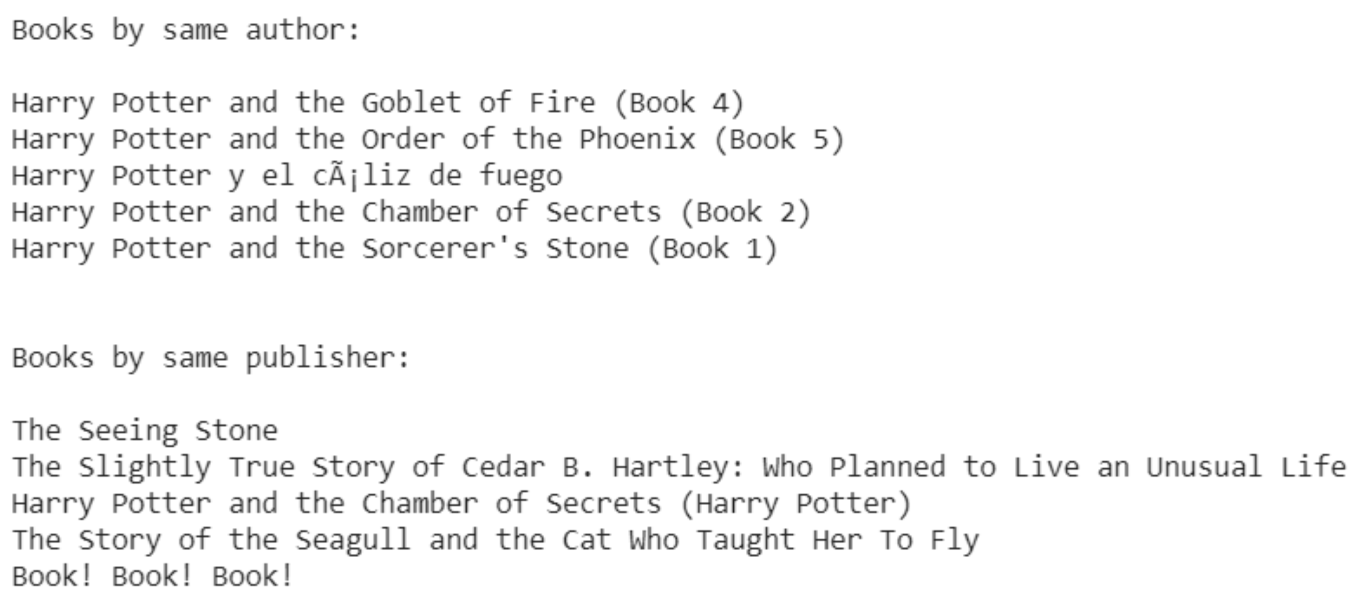
```
Enter the name of place: india
```

| | ISBN | Book-Rating | Book-Title | Book-Author | Year-Of-Publication | Publisher |
|---|---|---|---|---|---|---|
| 26 | 0971880107 | 3 | Wild Animus | Rich Shapero | 2004 | Too Far |
| 169 | 0671047612 | 2 | Skin And Bones | Franklin W. Dixon | 2000 | Aladdin |
| 167 | 0486284735 | 2 | Pride and Prejudice (Dover Thrift Editions) | Jane Austen | 1995 | Dover Publications |
| 9682 | 8171670407 | 2 | Inscrutable Americans | Mathur Anurag | 1996 | South Asia Books |
| 72608 | 0006944035 | 1 | Secret Island / Secret Mountain (Two-in-ones) | Enid Blyton | 1994 | HarperCollins Publishers |

We have filtered the dataset according to a given place (city, state, or country) and then sorted it according to total ratings they have received by the users in decreasing order of that place and recommended top n books.

# Recommendation Models – Low Risk

2. Author and Publisher Recommendations:
   - Recommends books from the same author or publisher as user input.
   - Enhances user exploration of titles related to preferred authors or publishers.

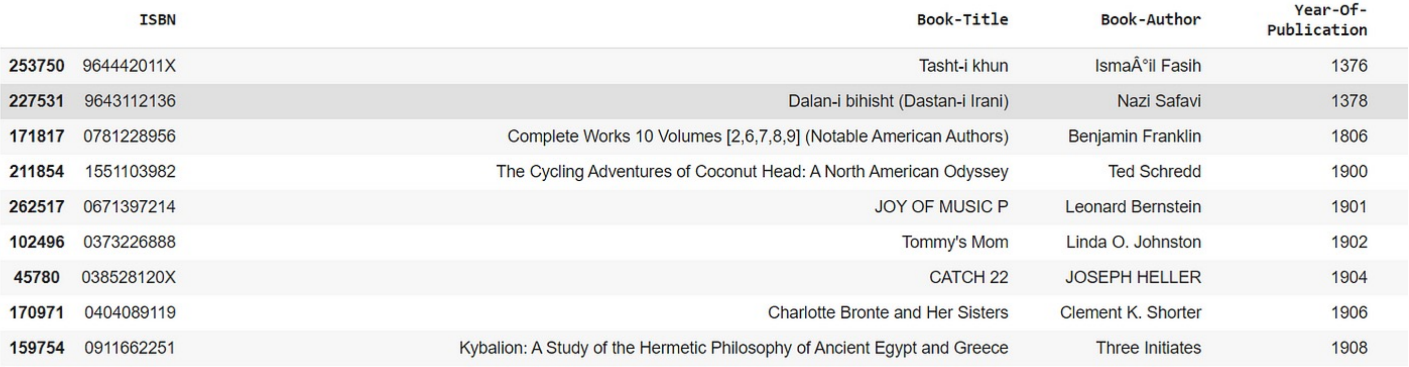


```
Books by same author:

Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Order of the Phoenix (Book 5)
Harry Potter y el cÃ¡liz de fuego
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Sorcerer's Stone (Book 1)


Books by same publisher:

The Seeing Stone
The Slightly True Story of Cedar B. Hartley: Who Planned to Live an Unusual Life
Harry Potter and the Chamber of Secrets (Harry Potter)
The Story of the Seagull and the Cat Who Taught Her To Fly
Book! Book! Book!
```

For this model, we have sorted the books by rating for the same author and same publisher of the given book and recommended top n books.

# Recommendation Models – Low Risk

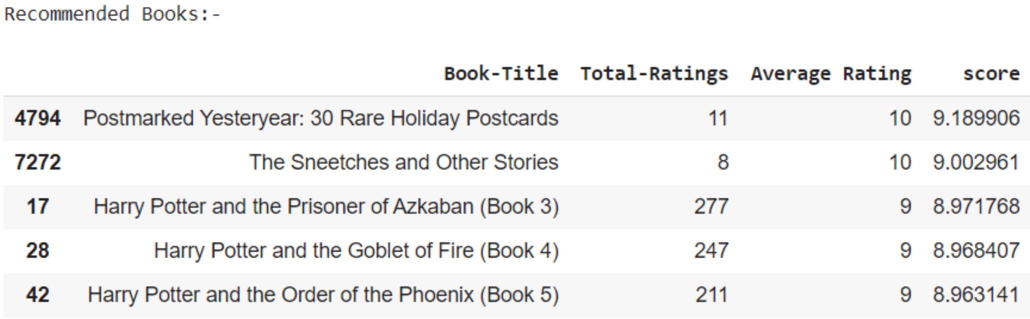
## 3. Yearly Popularity Insights: Most popular book for each publication year.



| | ISBN | Book-Title | Book-Author | Year-Of-Publication |
|---|---|---|---|---|
| 253750 | 964442011X | Tasht-i khun | IsmaÂ°il Fasih | 1376 |
| 227531 | 9643112136 | Dalan-i bihisht (Dastan-i Irani) | Nazi Safavi | 1378 |
| 171817 | 0781228956 | Complete Works 10 Volumes [2,6,7,8,9] (Notable American Authors) | Benjamin Franklin | 1806 |
| 211854 | 1551103982 | The Cycling Adventures of Coconut Head: A North American Odyssey | Ted Schredd | 1900 |
| 262517 | 0671397214 | JOY OF MUSIC P | Leonard Bernstein | 1901 |
| 102496 | 0373226888 | Tommy's Mom | Linda O. Johnston | 1902 |
| 45780 | 038528120X | CATCH 22 | JOSEPH HELLER | 1904 |
| 170971 | 0404089119 | Charlotte Bronte and Her Sisters | Clement K. Shorter | 1906 |
| 159754 | 0911662251 | Kybalion: A Study of the Hermetic Philosophy of Ancient Egypt and Greece | Three Initiates | 1908 |

For this model, we have grouped all the books published in the same year and recommended the top-rated book yearly.

## 4. Average Weighted Ratings: Utilizes a weighted formula considering average and total ratings.



```
Recommended Books:-
```

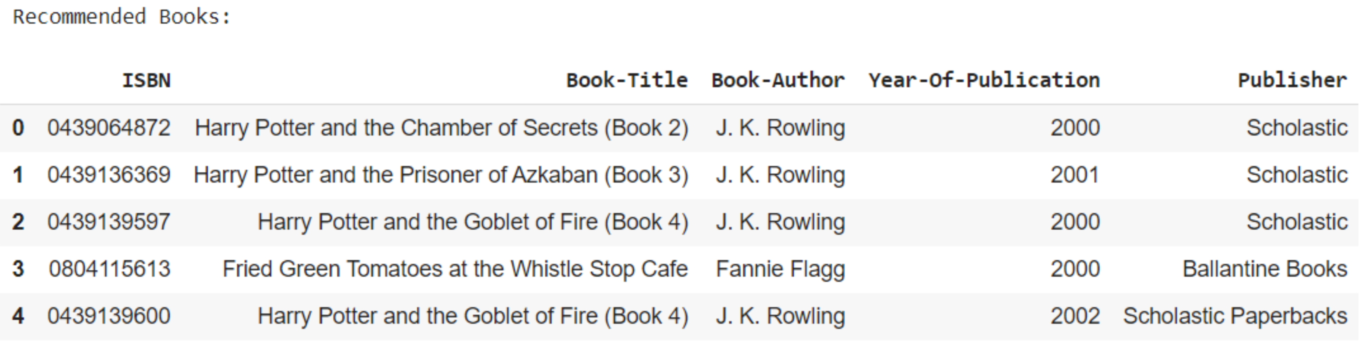
| | Book-Title | Total-Ratings | Average Rating | score |
|---|---|---|---|---|
| 4794 | Postmarked Yesteryear: 30 Rare Holiday Postcards | 11 | 10 | 9.189906 |
| 7272 | The Sneetches and Other Stories | 8 | 10 | 9.002961 |
| 17 | Harry Potter and the Prisoner of Azkaban (Book 3) | 277 | 9 | 8.971768 |
| 28 | Harry Potter and the Goblet of Fire (Book 4) | 247 | 9 | 8.968407 |
| 42 | Harry Potter and the Order of the Phoenix (Book 5) | 211 | 9 | 8.963141 |

We have calculated the weighted score using the below formula for all the books and recommended the books with the highest score.

**score = t/(t+m) ∗ a + m/(m+t) ∗ c**
**t** represents the total number of ratings received by the book
**m** represents the minimum number of total ratings considered to be included
**a** represents the average rating of the book and,
**c** represents the mean rating of all the books.

# Recommendation Models – Low Risk

## 5. Correlation Based:

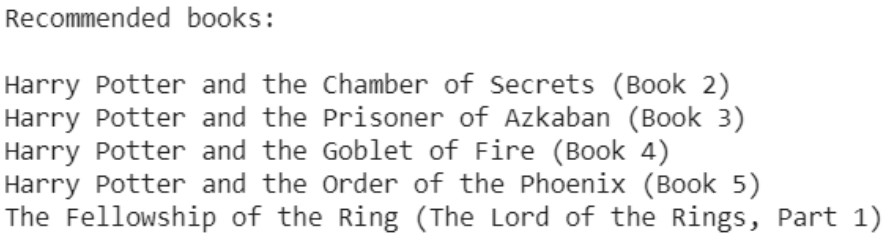


Recommended Books:

| | ISBN | Book-Title | Book-Author | Year-Of-Publication | Publisher |
|---|---|---|---|---|---|
| 0 | 0439064872 | Harry Potter and the Chamber of Secrets (Book 2) | J. K. Rowling | 2000 | Scholastic |
| 1 | 0439136369 | Harry Potter and the Prisoner of Azkaban (Book 3) | J. K. Rowling | 2001 | Scholastic |
| 2 | 0439139597 | Harry Potter and the Goblet of Fire (Book 4) | J. K. Rowling | 2000 | Scholastic |
| 3 | 0804115613 | Fried Green Tomatoes at the Whistle Stop Cafe | Fannie Flagg | 2000 | Ballantine Books |
| 4 | 0439139600 | Harry Potter and the Goblet of Fire (Book 4) | J. K. Rowling | 2002 | Scholastic Paperbacks |

For this model, we have created the correlation matrix for which we needed to reduce the dataset (because of limited resources). So, we have considered only those books which have total ratings of more than 50. Then from this data, we have created a user-book rating matrix. For the input book using the correlation matrix, top books are recommended.

## 6. Nearest Neighbors Based:

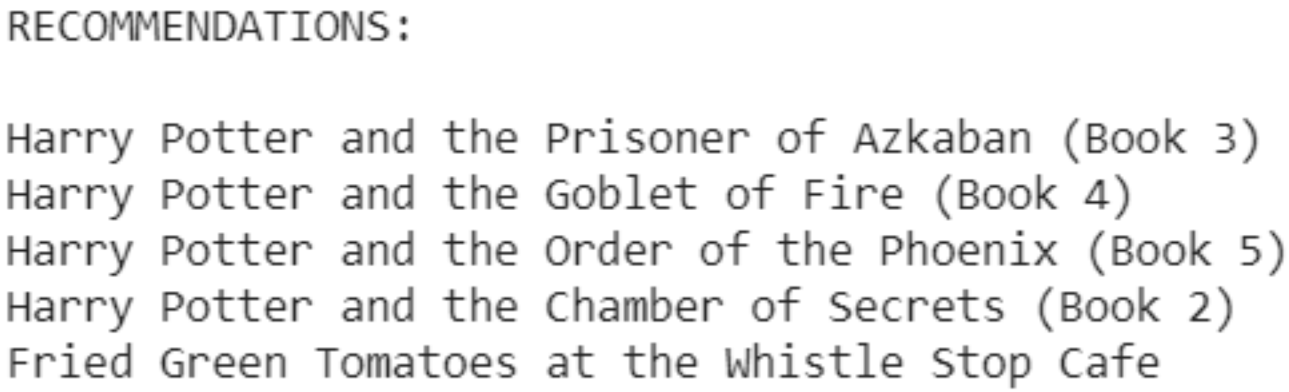


```
Recommended books:

Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Order of the Phoenix (Book 5)
The Fellowship of the Ring (The Lord of the Rings, Part 1)
```

To train the Nearest Neighbors model, we have created a compressed sparse row matrix taking ratings of each Book by each User individually. This matrix is used to train the Nearest Neighbors model and then to find n nearest neighbors using the cosine similarity metric.

# Recommendation Models – Low Risk

7. Collaborative Filtering (User-Item Filtering):
    - Built on user-item interactions using cosine similarity.
    - Identifies books similar to user’s input, emphasizing high similarity scores.



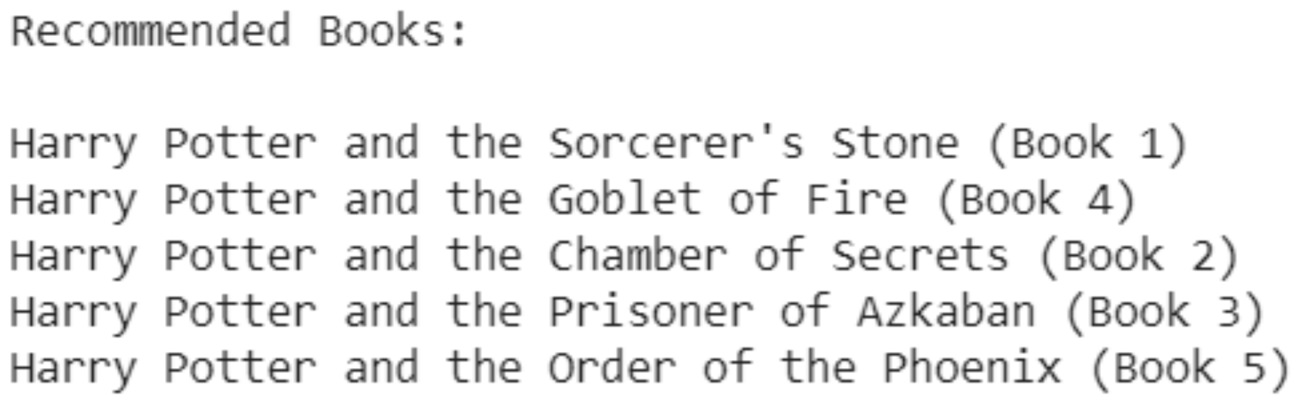
```
RECOMMENDATIONS:

Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Order of the Phoenix (Book 5)
Harry Potter and the Chamber of Secrets (Book 2)
Fried Green Tomatoes at the Whistle Stop Cafe
```

Collaborative Filtering Recommendation System works by considering user ratings and finds cosine similarities in ratings by several users to recommend books. To implement this, we took only those books' data that have at least 50 ratings in all (because of limited resources).

# Recommendation Models – High Risk

We explored High Risk/High Reward approaches for uncovering nuanced patterns that may go unnoticed with simpler models.

1. Content-Based Filtering:
   - Analyzing book content to refine recommendations based on user preferences.
   - Enhances personalization by considering book attributes and user traits.

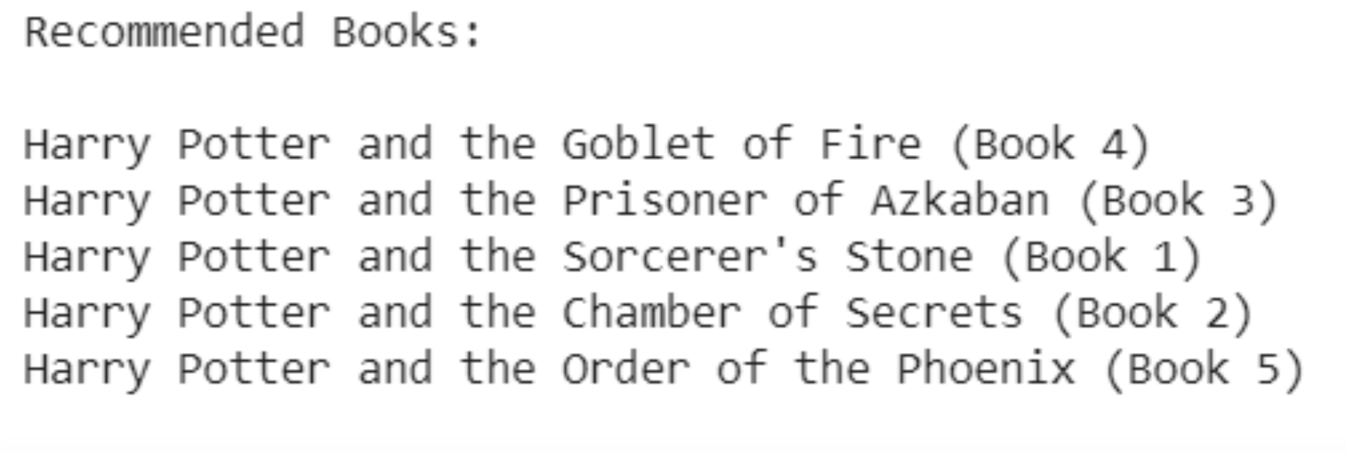


```
Recommended Books:

Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Order of the Phoenix (Book 5)
```

We have implemented a content-based recommendation system that recommends books by calculating similarities in Book Titles. For this, TF-IDF feature vectors are created for unigrams and bigrams of Book-Titles where only those books' data has been considered which are having at least 80 ratings (because of limited resources).

# Recommendation Models – High Risk

2. Hybrid Recommendation System:
   - Merging collaborative and content-based filtering for a comprehensive recommendation strategy.



```
Recommended Books:

Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Order of the Phoenix (Book 5)
```

We have built a hybrid recommendation system using both content-based filtering and collaborative filtering systems. A percentile score is given to the results obtained from both content and collaborative filtering models and is combined to recommend top n books.

# Evaluation

- Evaluating recommendation systems required a nuanced approach.

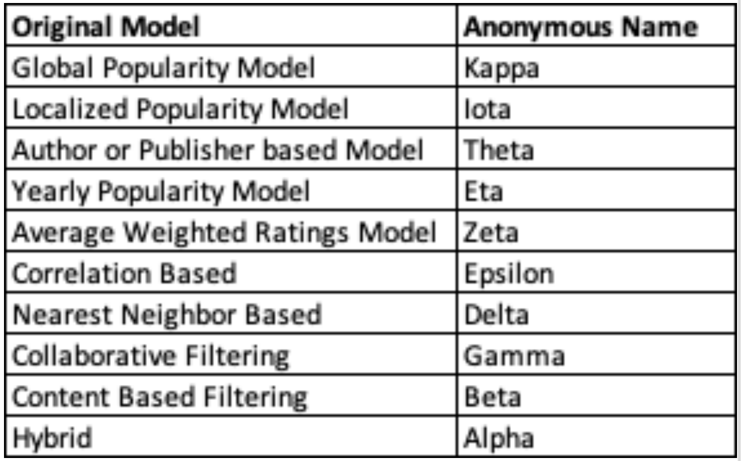
- Given the absence of labeled data for traditional methods, we employed a strategy–leveraging subjective insights through manual survey.

- We randomly reached out to 25 peer students available at Snell Library and requested their feedback, rating the effectiveness of book recommendations from our different models on a scale of 1 to 10 based on their given input book.

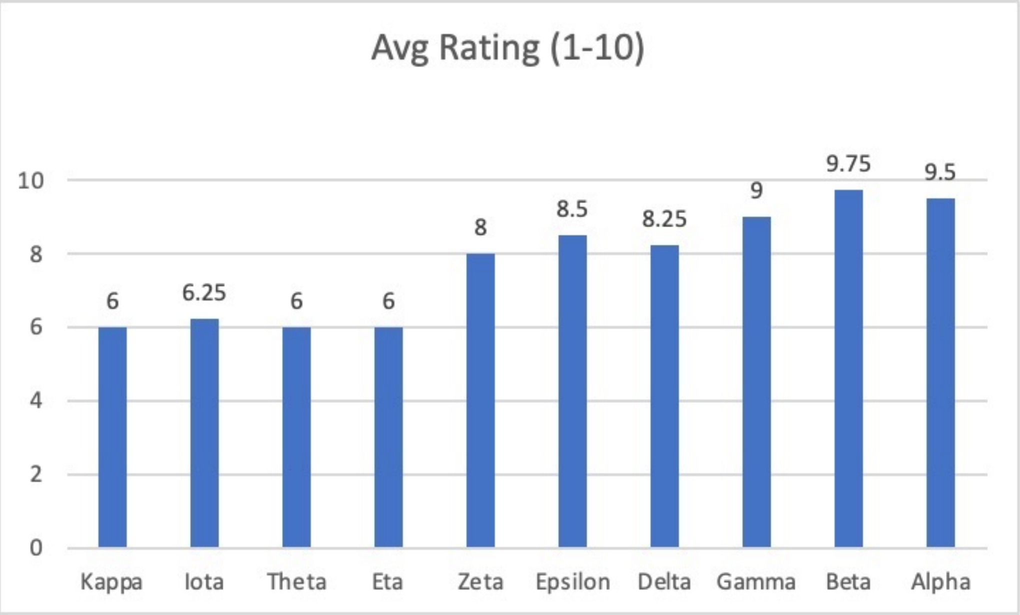
- This approach enhanced our understanding beyond numerical metrics, ensuring a comprehensive evaluation.

# Evaluation Results

- A visual depiction of our models' performance unfolds. Bar graph below shows average rating for each recommendation model in survey results, revealing the perceived effectiveness of each recommendation model.



| Original Model | Anonymous Name |
|---|---|
| Global Popularity Model | Kappa |
| Localized Popularity Model | Iota |
| Author or Publisher based Model | Theta |
| Yearly Popularity Model | Eta |
| Average Weighted Ratings Model | Zeta |
| Correlation Based | Epsilon |
| Nearest Neighbor Based | Delta |
| Collaborative Filtering | Gamma |
| Content Based Filtering | Beta |
| Hybrid | Alpha |



# Project Repository

GitHub Repository:

https://github.com/shrivastavasatyam/book_recommendation_system

Survey Link:

https://forms.gle/zWEopm4JUTfvk2WK6

# References

- Book Recommendation Dataset: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset
- Book Recommendation System – Ms. Sushama Rajpurkar, Ms. Darshana Bhatt, and Ms. Pooja Malhotra (2015) http://www.ijirst.org/articles/IJIRSTV1I11135.pdf
- How to Build a Book Recommendation System: https://www.analyticsvidhya.com/blog/2021/06/build-book-recommendation-system-unsupervised-learning-project/
- Recommendation System Wikipedia: https://en.wikipedia.org/wiki/Recommender_system
- Scikit-learn – Machine Learning library in Python: https://scikit-learn.org/stable/

# Thank You!