Capstone Project 4
Book Recommendation System

RADHIKA MENON

Presentation Outline

1. Problem Statement
2. Data Overview
3. Data preprocessing
4. Exploratory Data Analysis
5. Recommender Systems
6. Evaluation Metrics
7. Conclusion

Problem Statement

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries).

Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. The main objective is to create a book recommendation system for users.

This Book-crossing dataset contains three files: Users, Books, Ratings

Data Overview

Users

User-ID: Unique ID of each user
Location: Location of the user
Age: Age of the user

Books

ISBN: The International Standard Book Number is a unique numeric Identifier
Book-Title: Title of Book corresponding to an ISBN
Book-Author: Author of the book
Year-Of-Publication: Year of Publication of the book
Publisher: Publisher of the book

Data Overview contd.

Image-URL-S: Small cover image url to a book
Image-URL-M: Medium cover image url to a book
Image-URL-L: Large cover image url to a book

Ratings

User-ID: Unique ID of each user
ISBN: The International Standard Book Number is a unique numeric Identifier
Book-Rating: Book-Rating are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

Preprocessing

Books (books_df)

Replacing null and incorrect values with correct values.

```
#Filling the null value
books_df.loc[187689, 'Book-Author'] = 'Larissa Anne Downes'
```

```
#Replacing NaNs with correct  values
books_df.loc[128890, 'Publisher'] = 'Mundania Press LLC'
books_df.loc[129037, 'Publisher'] = 'Bantam'
```

```
# on searching for these  books we came to know about its authors
#ISBN '078946697X'
books_df.loc[books_df.ISBN == '078946697X','Year-Of-Publication'] = 2000
books_df.loc[books_df.ISBN == '078946697X','Book-Author'] = "Michael Teitelbaum"
books_df.loc[books_df.ISBN == '078946697X','Publisher'] = "DK Publishing Inc"
books_df.loc[books_df.ISBN == '078946697X','Book-Title'] = "DK Readers: Creating the X-Men, How It All Began (Level 4: Proficient Readers)"

#ISBN '0789466953'
books_df.loc[books_df.ISBN == '0789466953','Year-Of-Publication'] = 2000
books_df.loc[books_df.ISBN == '0789466953','Book-Author'] = "James Buckley"
books_df.loc[books_df.ISBN == '0789466953','Publisher'] = "DK Publishing Inc"
books_df.loc[books_df.ISBN == '0789466953','Book-Title'] = "DK Readers: Creating the X-Men, How Comic Books Come to Life (Level 4: Proficient Readers)"
#replacing with correct  values
books_df.loc[books_df.ISBN==' 9643112136','Year-Of-Publication'] = 2010
books_df.loc[books_df.ISBN=='964442011X', 'Year-Of-Publication'] = 1991
```

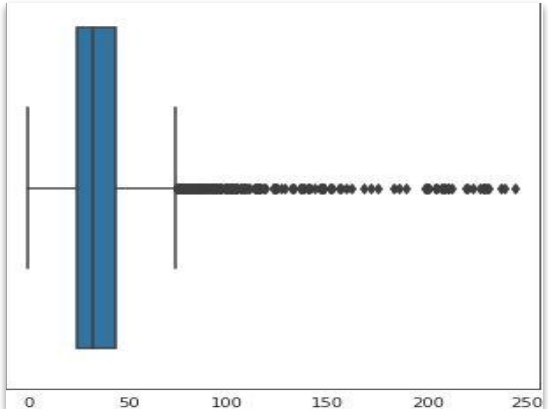
```
#Sustituting np.Nan in rows with year=0 or  greater than the current year,2022.
books_df.loc[(books_df['Year-Of-Publication'] > 2022) | (books_df['Year-Of-Publication'] == 0),'Year-Of-Publication'] = np.NAN

# replacing NaN values with median value of Year-Of-Publication
books_df['Year-Of-Publication'].fillna(int(books_df['Year-Of-Publication'].median()), inplace=True)
```

Preprocessing contd.

Users (users_df)

Replacing null and incorrect values.

Box plot of ‘Age’ column




```
# create a normal distribution pd.Series to fill Nan values with
normal_age_series = pd.Series(np.random.normal(loc=users_df.Age.mean(), scale=users_df.Age.std(), size=users_df[users_df.Age.isna()]['User-ID'].count()))

# take the absolute value of temp_age_series
abs_age_series=np.abs(normal_age_series)

# sort users df so as NaN values in age to be first and reset index to match with index of abs_age_series. Then using fillna()
users_df = users_df.sort_values('Age',na_position='first').reset_index(drop=True)
users_df.Age.fillna(round(abs_age_series), inplace = True)
```

Preprocessing contd.

Users

Correcting misspelled country names that were extracted from ‘Location’ column

```
#correcting the mispelled country names
users_df.loc[users_df['Country'].isin(['australii','autralia','western australia']), 'Country'] = 'australia'
users_df.loc[users_df['Country'].isin(['unite states','01776','02458','19104','23232','30064','85021','87510','united sates','united staes','united state','united statea','united stated','america'
                                       'united stated of america','united states','united states of america','us','us of a','us virgin islands',
                                       'usa  canada','usa currently living in england','uusa','usaa','wonderful usa','california','orange co']), 'Country'] = 'usa'
users_df.loc[users_df['Country'].isin(['united kindgdom', 'united kindgonm','united kingdom','u k']), 'Country'] = 'uk'
users_df.loc[users_df['Country'].isin(['the philippines', 'philippines', 'philippinies', 'phillipines', 'phils', 'phippines']), 'Country'] = 'philippines'
users_df.loc[users_df['Country'].isin(['','xxxxxx','universe','nowhere','x','y','a','öð¹ú','the','unknown',np.nan,'n/a','aaa','z','somewherein space']), 'Country'] = 'others'
users_df.loc[users_df['Country'].isin([ 'italia','italien','itlay']), 'Country'] = 'italy'
users_df.loc[users_df['Country'].isin([ ' china öð¹ú','chinaöð¹ú','chian']), 'Country'] = 'china'
users_df['Country'].replace([ 'the gambia','the netherlands','geermany','srilanka','saudia arabia','brasil','_ brasil','indiai','malaysian','hongkong','russian federation'],
                            ['gambia','netherlands','germany','sri lanka','saudi arabia','brazil','brazil','india','malaysia','hong kong','russia'],inplace=True)
```

Preprocessing contd.

Ratings (ratings_df)

Separating Explicit and implicit ratings

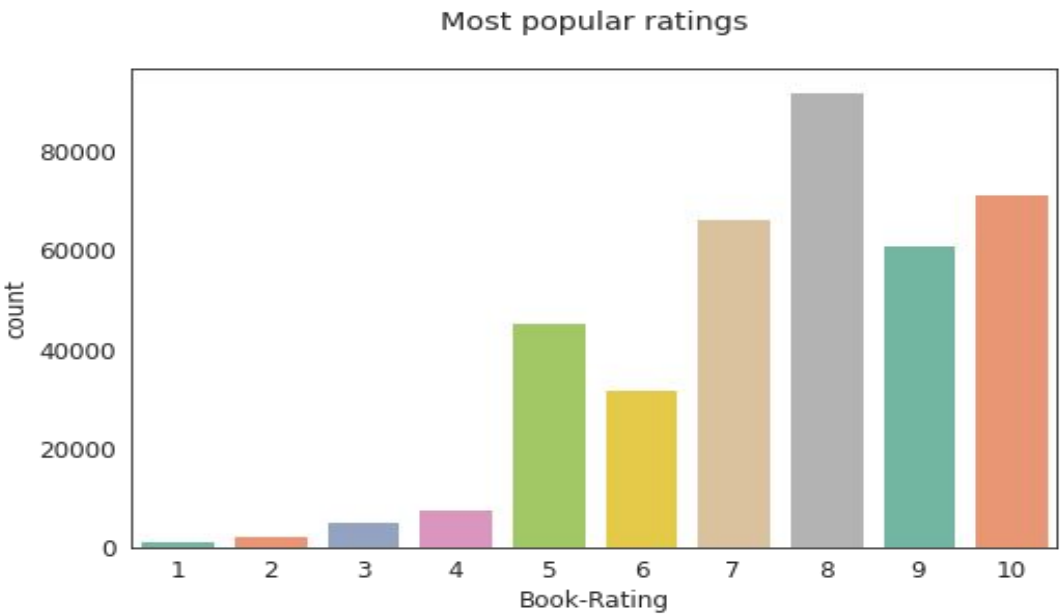
```
[ ] # lets see if all the books in rating_df are also in books_df
    rating_df_new = rating_df[rating_df['ISBN'].isin(books_df['ISBN'])]
```

```
explicit_rating = rating_df_new[rating_df_new['Book-Rating'] != 0]
implicit_rating = rating_df_new[rating_df_new['Book-Rating'] == 0]
print('Shape of explicit rating: {} and implicit rating: {}'.format(explicit_rating.shape, implicit_rating.shape))
```

EDA

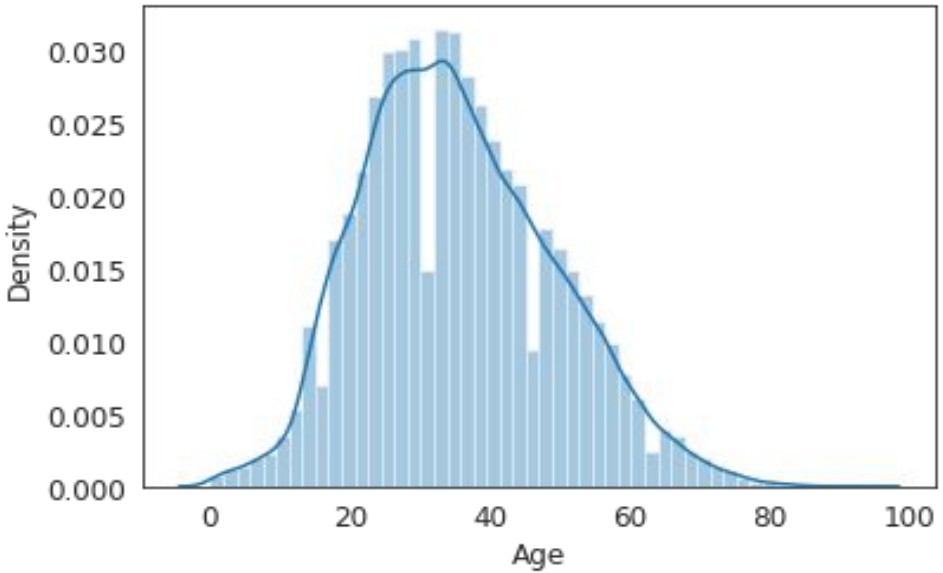
After merging the datasets

Rating distribution



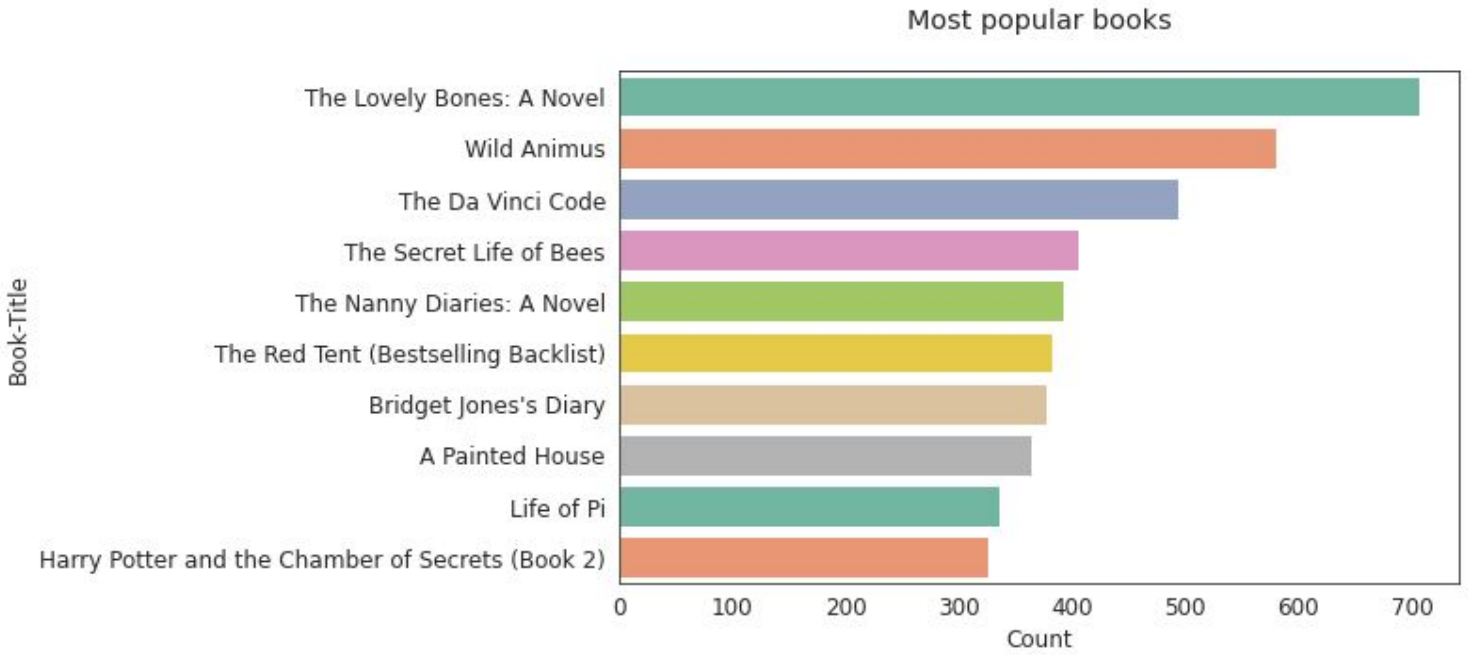
EDA

Age distribution of users



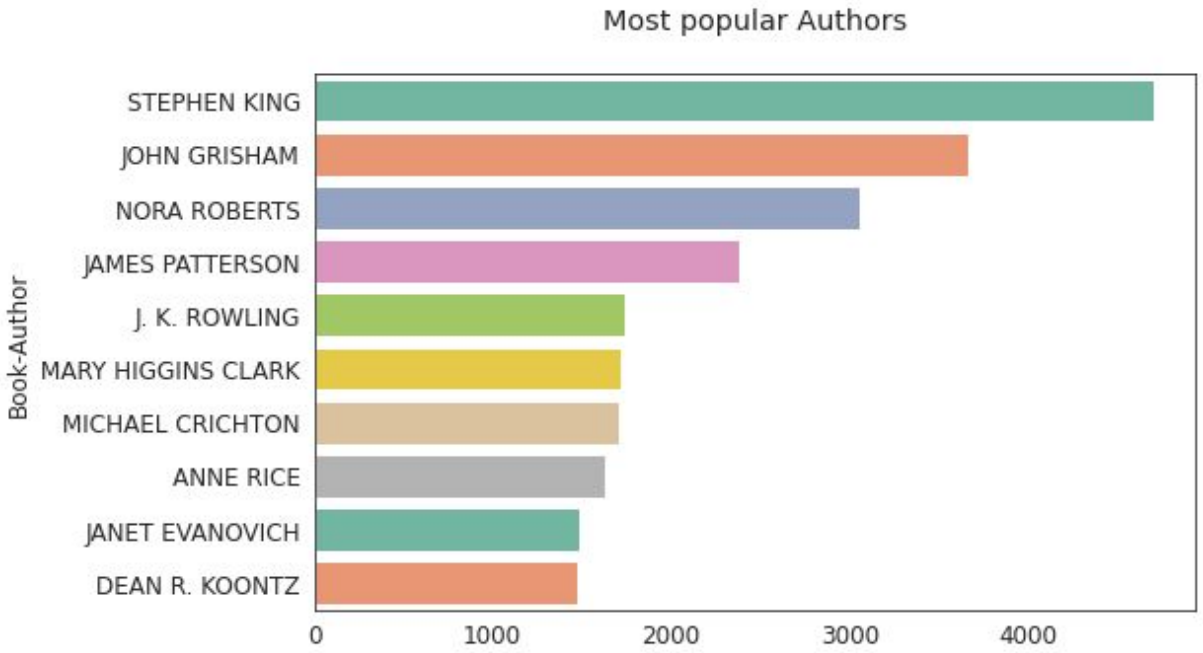
EDA

Most popular Books



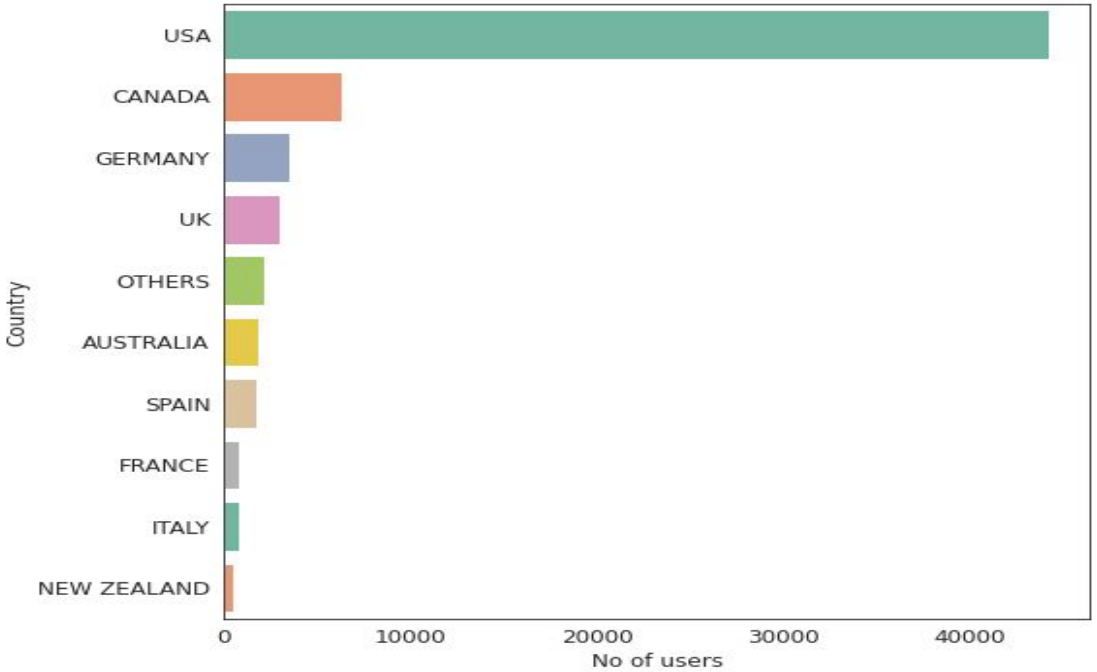
EDA

Most popular Authors



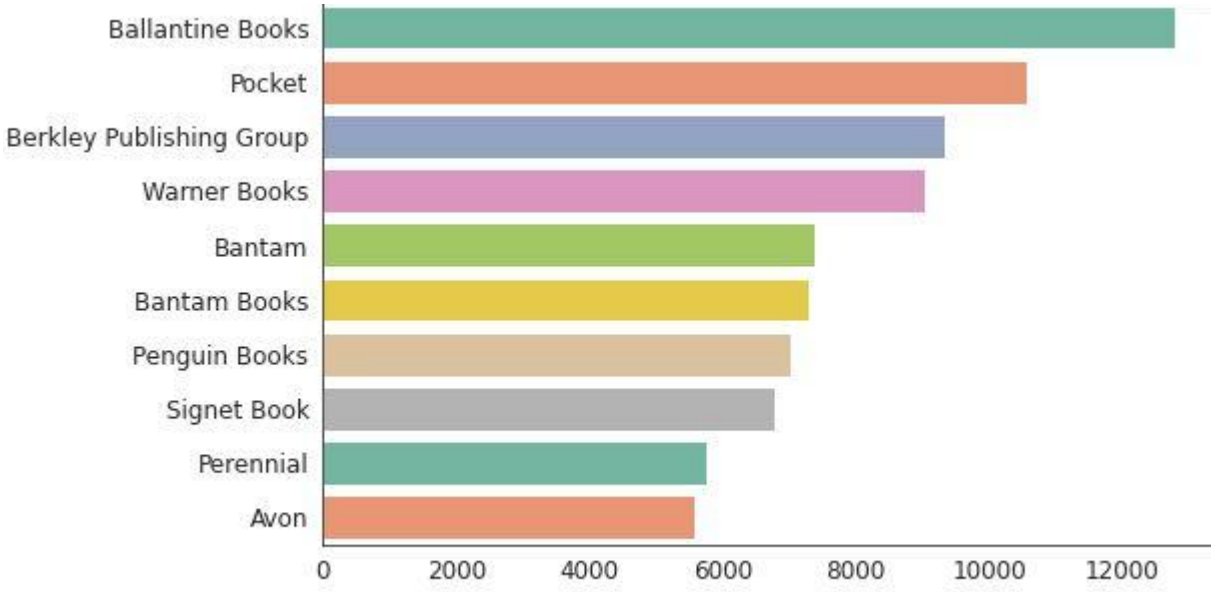
EDA

Countries with highest number of users



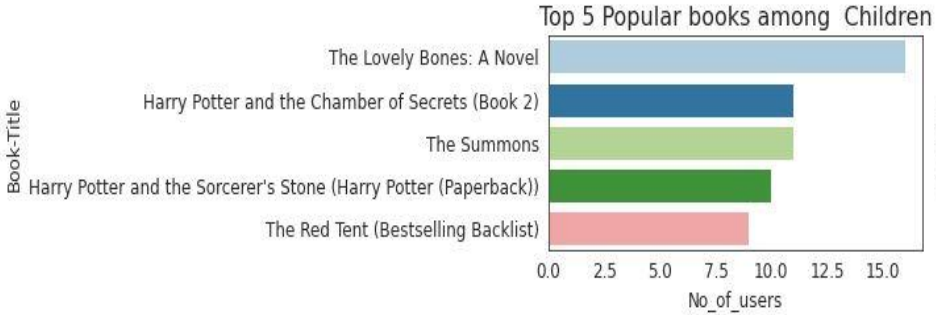
EDA

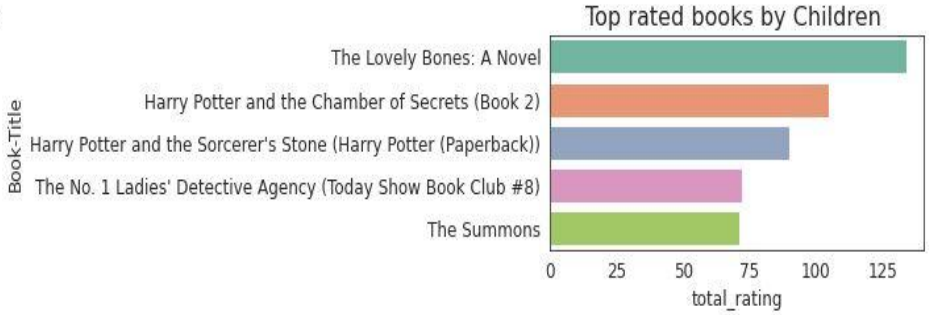
Publishers with most number of books

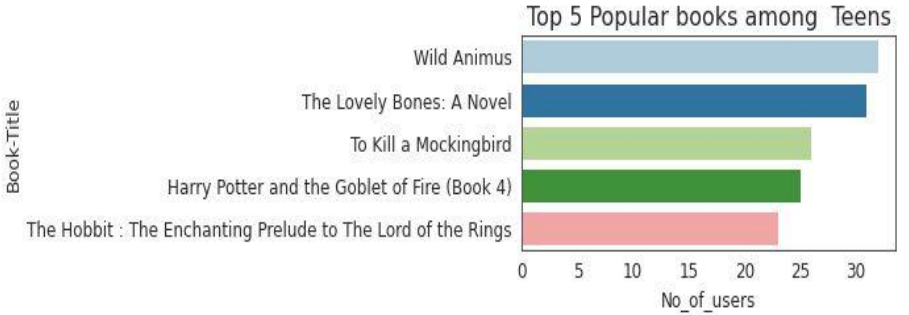


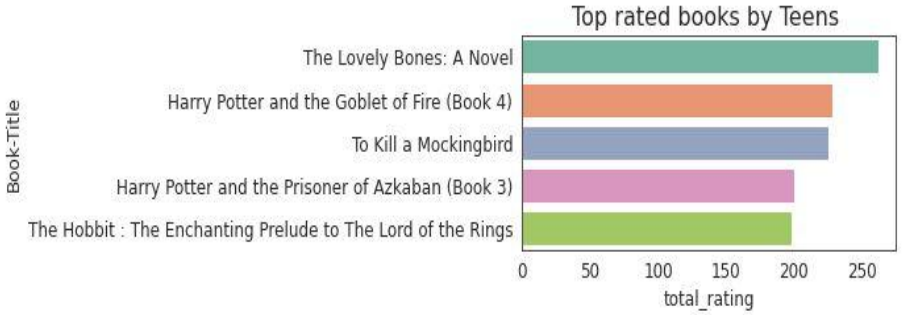
EDA

Most popular Books and Top rated books among Youth and Middle aged adults



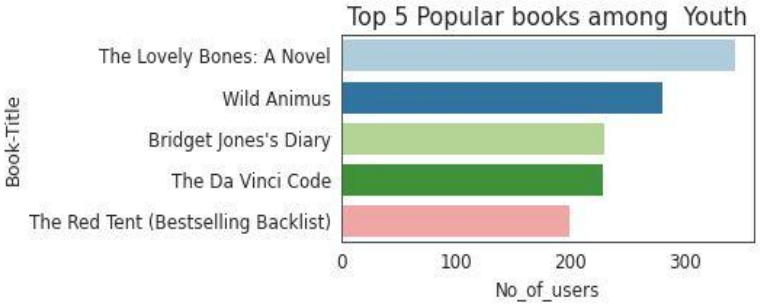


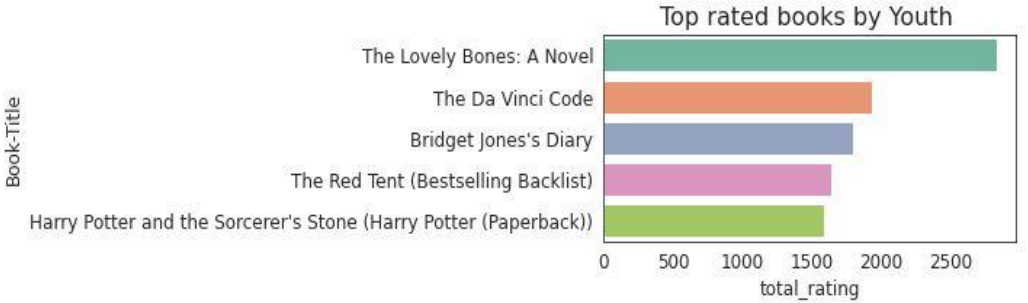


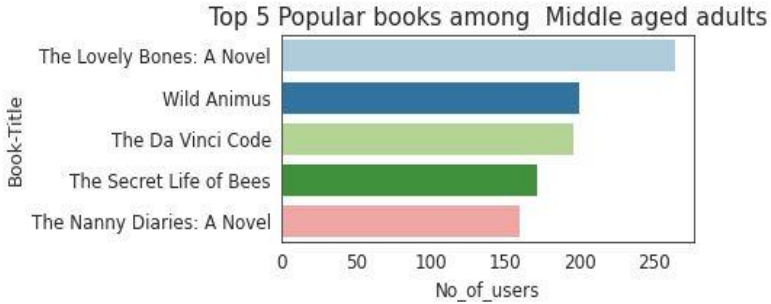


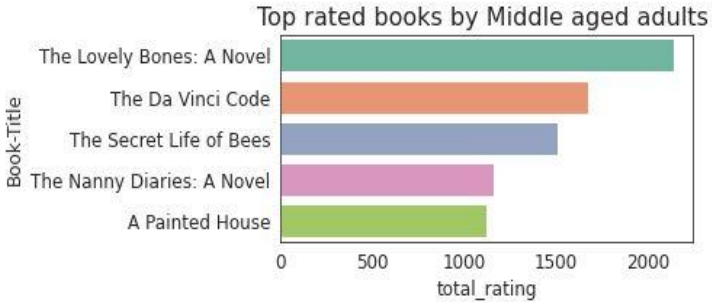
EDA

Most popular Books and Top rated books among Youth and Middle aged adults



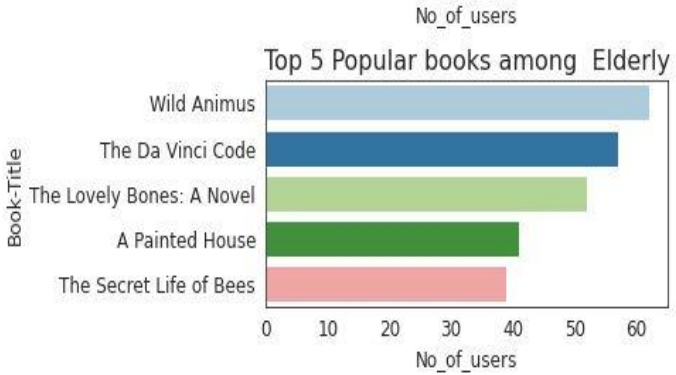


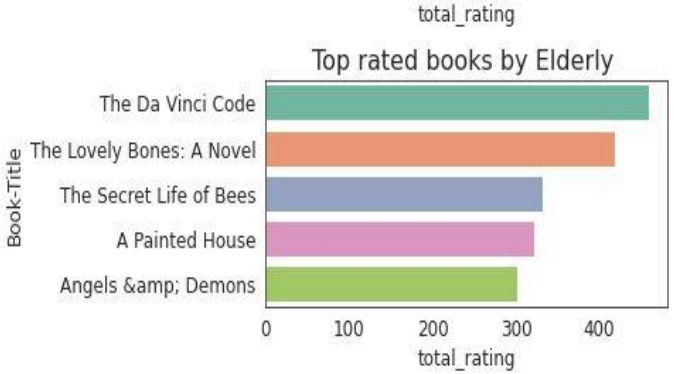




EDA

Most popular Books and Top rated books among the elderly





Conclusions based on EDA

- Most of the users have given a rating of 5 or above to the books.
- The majority of readers are between the ages of 25 and 40.
- Stephen King is the most popular author.
- The majority of readers who have given the books ratings are from United States (US).
- Ballantine Books has published most number of books.
- The Lovely Bones: A Novel is the most popular book.
- Lovely Bones :A Novel is also highly rated by users of all age groups.

Recommender Systems

Following recommender systems were chosen:

1. Popularity Based Recommender Systems
   - Country wise
   - Author wise
   - Weighted average rating

2. Collaborative Filtering based Recommender Systems
   - Memory Based (Item- Item): KNN based recommender system
   - Model based : SVD based recommender system

Recommender Systems

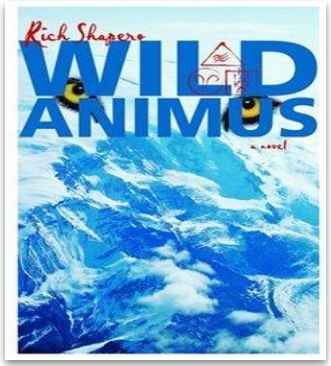
Popularity Based

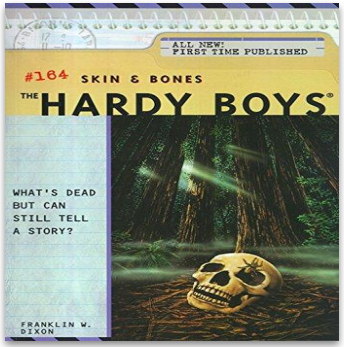
Country wise

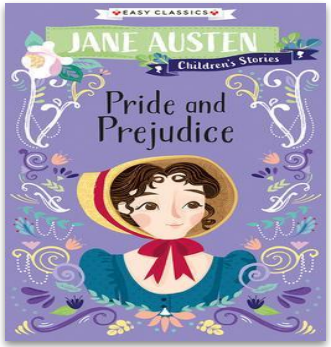
Input : INDIA

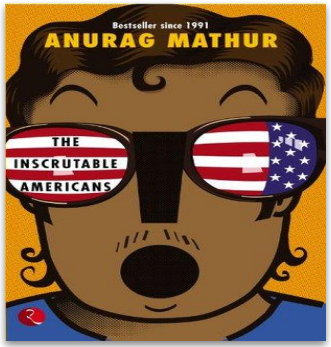
```
country_popular(df, 'INDIA')
```

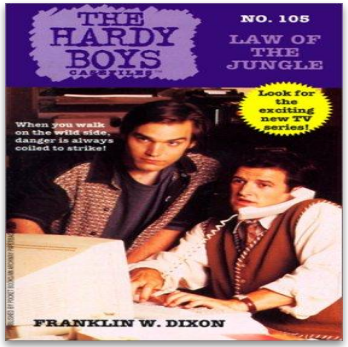
	ISBN	Book-Rating	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	0971880107	3	Wild Animus	RICH SHAPERO	2004.0	Too Far
1	0671047612	2	Skin And Bones	FRANKLIN W. DIXON	2000.0	Aladdin
2	0486284735	2	Pride and Prejudice (Dover Thrift Editions)	JANE AUSTEN	1995.0	Dover Publications
3	8171670407	2	Inscrutable Americans	MATHUR ANURAG	1996.0	South Asia Books
4	0006944035	1	Secret Island / Secret Mountain (Two-in-ones)	ENID BLYTON	1994.0	HarperCollins Publishers











Recommender Systems

Popularity Based

Weighted average rating approach

Weighted average rating method

Using Weighted average for each Book's Average Rating

W = (Rv + Cm)/(v + m)

where

W= Weighted Rating

R = Average of the Books rating

v = No of people who have rated the books(number of votes)

m = minimum no of votes to be listed

C = the mean rating across all the books

```
df_relevant_data.sort_values(by='weighted_average',ascending=False).head(10)
```

	Book-Title	Book-Author	avg_rating	ratings_count	weighted_average
46516	Harry Potter and the Chamber of Secrets Postcard Book	J. K. ROWLING	9.869565	23	9.52
122145	The Two Towers (The Lord of the Rings, Part 2)	J. R. R. TOLKIEN	9.653846	52	9.50
30142	Dilbert: A Book of Postcards	SCOTT ADAMS	9.923077	13	9.36
81784	Postmarked Yesteryear: 30 Rare Holiday Postcards	PAMELA E. APKARIAN-RUSSELL	10.000000	11	9.34
118127	The Return of the King (The Lord of the Rings, Part 3)	J.R.R. TOLKIEN	9.397436	78	9.31
17713	Calvin and Hobbes	BILL WATTERSON	9.583333	24	9.29
100902	The Authoritative Calvin and Hobbes (Calvin and Hobbes)	BILL WATTERSON	9.600000	20	9.25
72637	My Sister's Keeper : A Novel (Picoult, Jodi)	JODI PICOULT	9.545455	22	9.23
118123	The Return of the King (The Lord of The Rings, Part 3)	J. R. R. TOLKIEN	9.625000	16	9.20
120090	The Sneetches and Other Stories	DR. SEUSS	10.000000	8	9.17

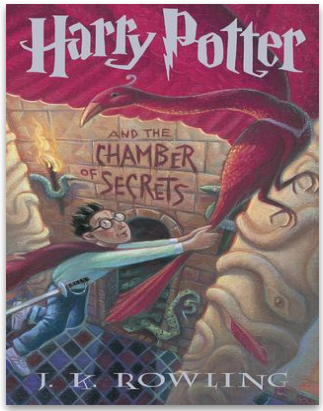
Recommender Systems

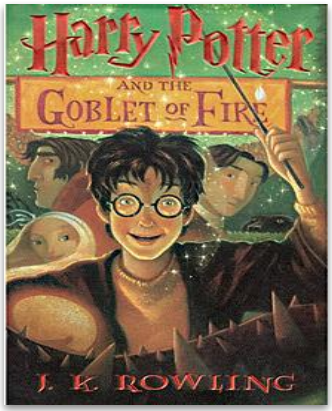
Popularity Based

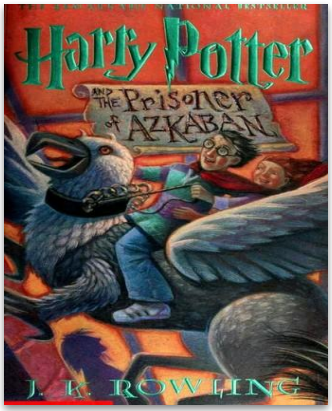
Author wise

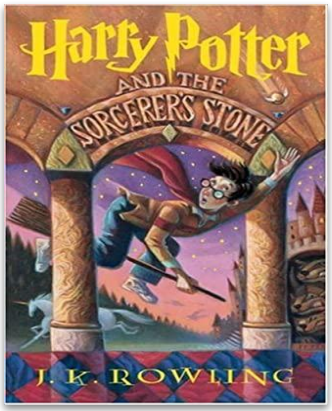
```
The author of the book Harry Potter and the Chamber of Secrets (Book 2) is J. K. ROWLING

Here are the top 5 books from the same author
```

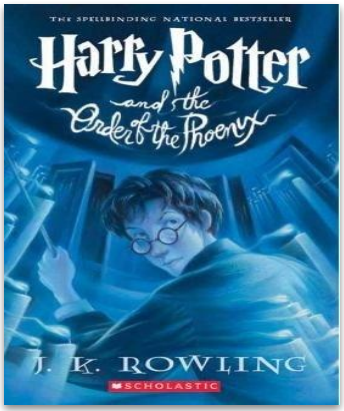
	Book-Title	weighted_average
46516	Harry Potter and the Chamber of Secrets Postcard Book	9.52
46520	Harry Potter and the Goblet of Fire (Book 4)	9.10
46532	Harry Potter and the Prisoner of Azkaban (Book 3)	9.02
46539	Harry Potter and the Sorcerer's Stone (Book 1)	9.02
46524	Harry Potter and the Order of the Phoenix (Book 5)	9.01











Recommender Systems

Memory Based CF - KNN (Euclidean distance based)

```
get_recommendations('Harry Potter and the Chamber of Secrets (Book 2)', 10)
```

```
The top 10 Recommended books for Harry Potter and the Chamber of Secrets (Book 2) are:

Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Dragons of a Lost Star (The War of Souls, Volume II)
Dr. Seuss's A B C (I Can Read It All by Myself Beginner Books)
J. K. Rowling: The Wizard Behind Harry Potter
The Second Generation
Lover Beware
Dragonquest Achille Cover
Betsy and Joe (Betsy &amp;amp; Tacy)
```

Recommender Systems

Memory Based CF - KNN (with cosine metric)

```
get_cosine_recommendations('Harry Potter and the Chamber of Secrets (Book 2)', 10)
```

```
Cosine Similarity based recommendations.

The top 10 Recommended books for Harry Potter and the Chamber of Secrets (Book 2) are:

Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
Harry Potter and the Order of the Phoenix (Book 5)
The Fellowship of the Ring (The Lord of the Rings, Part 1)
The Hobbit: or There and Back Again
Dragons of a Lost Star (The War of Souls, Volume II)
Dr. Seuss's A B C (I Can Read It All by Myself Beginner Books)
The Second Generation
```

Recommender Systems

Model based CF - Matrix Factorization (SVD)

Testing for User-ID :254

books that the user ID 254 has already rated

The Golden Compass (His Dark Materials, Book 1)
Making Minty Malone
Animal Farm
The Secret Life of Bees
She's Come Undone (Oprah's Book Club)
American Gods
The Hobbit: or There and Back Again
Harry Potter and the Sorcerer's Stone (Book 1)
The Bonesetter's Daughter
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
American Gods: A Novel
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Chamber of Secrets (Book 2)
The Dark Half
Harry Potter and the Prisoner of Azkaban (Book 3)
The Golden Compass (His Dark Materials, Book 1)
Familiar Lullaby (Fear Familiar) (Harlequin Intrigue, No 614)
The Fellowship of the Ring (The Lord of the Rings, Part 1)
The Duke
Complete Chronicles of Narnia
Stardust
Amazing Grace : Lives of Children and the Conscience of a Nation, The
Something Wicked This Way Comes

Recommending books for User ID: 254

	ISBN	Book-Title	Book-Author	Publisher
0	043935806X	Harry Potter and the Order of the Phoenix (Book 5)	J. K. ROWLING	Scholastic
1	059035342X	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	J. K. ROWLING	Arthur A. Levine Books
2	0439139600	Harry Potter and the Goblet of Fire (Book 4)	J. K. ROWLING	Scholastic Paperbacks
3	0446310786	To Kill a Mockingbird	HARPER LEE	Little Brown & Company
4	0385504209	The Da Vinci Code	DAN BROWN	Doubleday
5	0345339681	The Hobbit : The Enchanting Prelude to The Lord of the Rings	J.R.R. TOLKIEN	Del Rey
6	0385484518	Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	MITCH ALBOM	Doubleday
7	0316769487	The Catcher in the Rye	J.D. SALINGER	Little, Brown
8	0345339703	The Fellowship of the Ring (The Lord of the Rings, Part 1)	J.R.R. TOLKIEN	Del Rey
9	0345339711	The Two Towers (The Lord of the Rings, Part 2)	J.R.R. TOLKIEN	Del Rey

Recommender Systems

Evaluation metrics for SVD based recommender

Recall@k

Recall at k is the proportion of relevant items found in the set of top-k recommendations.

R = (# of top k recommendations that are relevant)/(# of all relevant items)

Global metrics:
{'modelName': 'Collaborative Filtering', 'recall@5': 0.30367936925098554, 'recall@10': 0.4110819097678493}

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	_person_id
36	30	79	545	0.055046	0.144954	11676
202	52	73	139	0.374101	0.525180	98391
271	27	34	93	0.290323	0.365591	153662
60	23	27	88	0.261364	0.306818	16795
474	20	26	73	0.273973	0.356164	95359
485	52	60	72	0.722222	0.833333	114368
390	32	33	61	0.524590	0.540984	104636
456	14	22	54	0.259259	0.407407	158295
660	40	44	54	0.740741	0.814815	123883
659	7	13	53	0.132075	0.245283	35859

Global metrics

Recall@5 of 30%
Recall@10 of 41%

Conclusion

- The initial step, of our project was Data preprocessing of the three datasets-books_df, users_df and ratings_df, wherein we removed duplicates and imputed the missing values & invalid entries with appropriate values and corrected spellings .
- Then, we used Popularity-based approach, Collaborative filtering approach to built different types of recommendation models.
- In the case of Memory-based approach, the Cosine similarity-based KNN performs better at recommending books that are similar than the Euclidean distance-based KNN.
- We evaluated the performance of Singular Value Decomposition based recommender and obtained a Global Recall@5 of 30% and Recall@10 of 41%.

Thank you