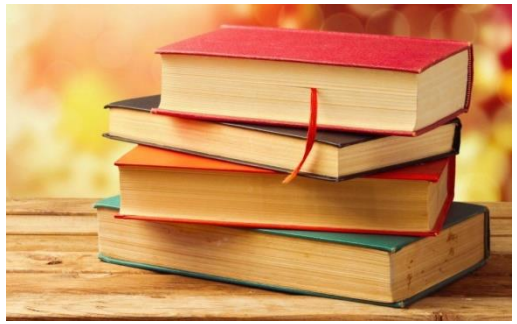
Capstone Project - 4

Book Recomendation system



Sujit Musale

(Self Project)

Index …….

- Problem Statement
- Data Overview
- Data Preprocessing
- Exploratory Data analysis
- Recommender system
- Evaluation Matrix
- Conclusion

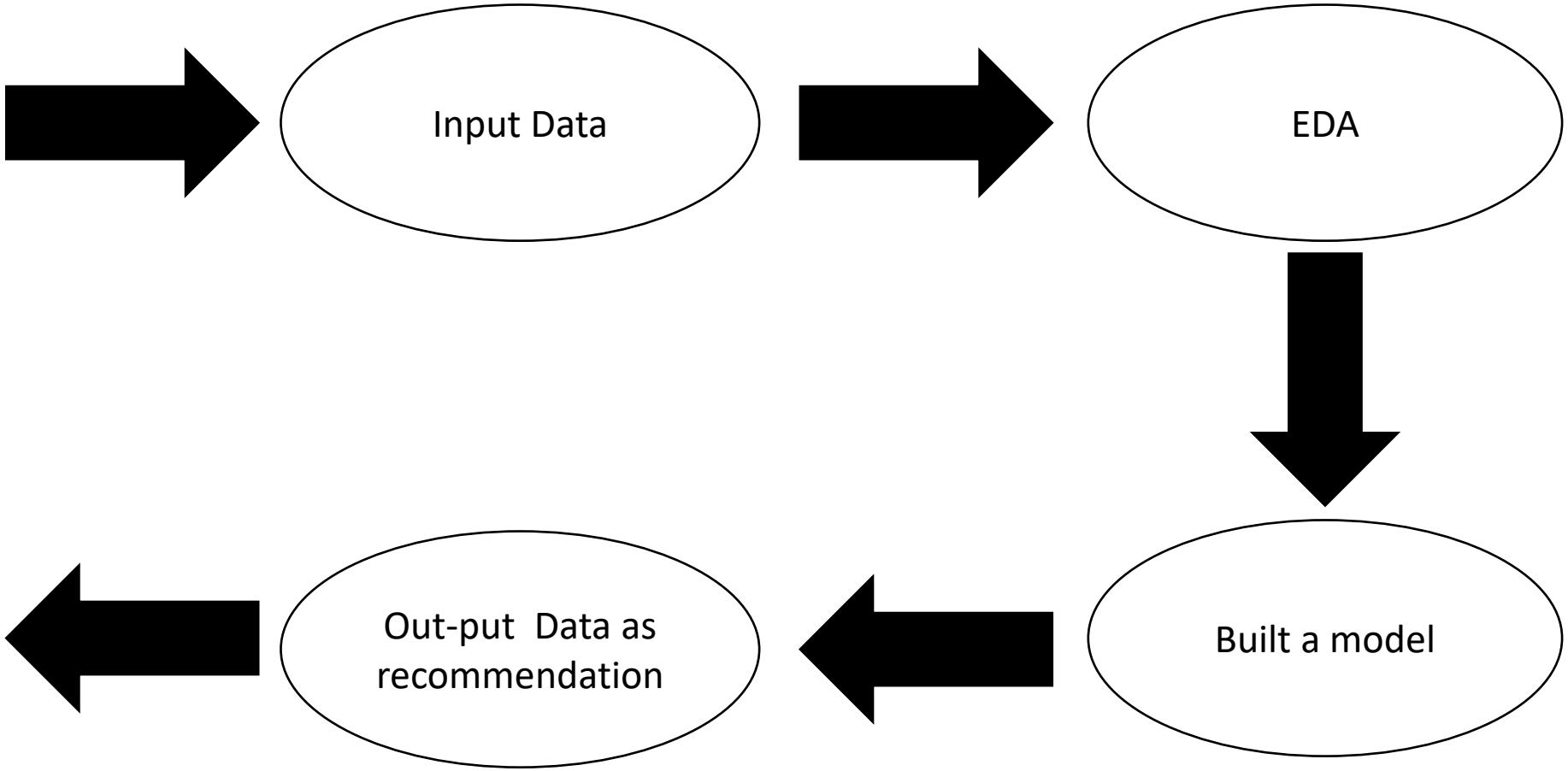
About Recommender System and project …….

- A book recommendation system is a type of recommendation system where we have to recommend similar books to the reader based on his interest.
- Recommendation system is used by online websites which provide ebooks like google play books, open library, good Read's, etc.
- We are using various machine learning model to built a recommender system.
- And also used various evaluation matrix to evaluate the result.

Problem statement.....

During the last few decades, with the rise of Youtube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys. 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.

Project (Recommender) process flow .....



Different Dataset Details .....

- **Users** - Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.
- **Books** - Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. Note that in the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website.
- **Ratings** -Contains the book rating information. Ratings (Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

Variables in Users …..

- **User-ID –** Unique Id of each book. (Same in Ratting Data set)
- **Location –** Location of users.
- **Age-** Age of users.

Variables in Books …..

- **ISBN –** The international Standard Book Number (Same in Rating Dataset)
- **Book-Title –** Title of the corresponding to ISBN
- **Book-Author-** Author Of the Book
- **Year-Of-Publication -** Year of publication of book
- **Publisher –** Name of the book publisher
- **Image-URL-S, Image-URL-M, Image-URL-L -** Small, Medium, Large cover url to book respectively

Variables in Ratings….

- **Book-Rating** – Book Rating are either explicit, expressed on scale from 1-10.(higher value Denote higher appreciation), Or implicit demoted by 0

Loading Data & treatment …..

User Data set –

- Columns – 3 and Rows – 278858
- Null value- 110762 in Age column
- Outlier – Outlier are present in Age column like age from 100 to 250
- we removed that age and fill with Mean value of age.
- Duplicate Value – No Duplicate value Present.
- We create country Name column by using location name and drop location name.

Loading Data & treatment …..

Books Data set –

- Columns – 8 and Rows – 271360
- Null value- 2 in publisher column & 3 in Image-URL-L
- Drop null value column as dataset is too big
- Outlier – outlier filled with median.
- we Drop all Image url columns as we don’t have any use in analysis.
- Duplicate Value – 313 rows Duplicate but its because of the revised book so will kept 1st revision ISBN.

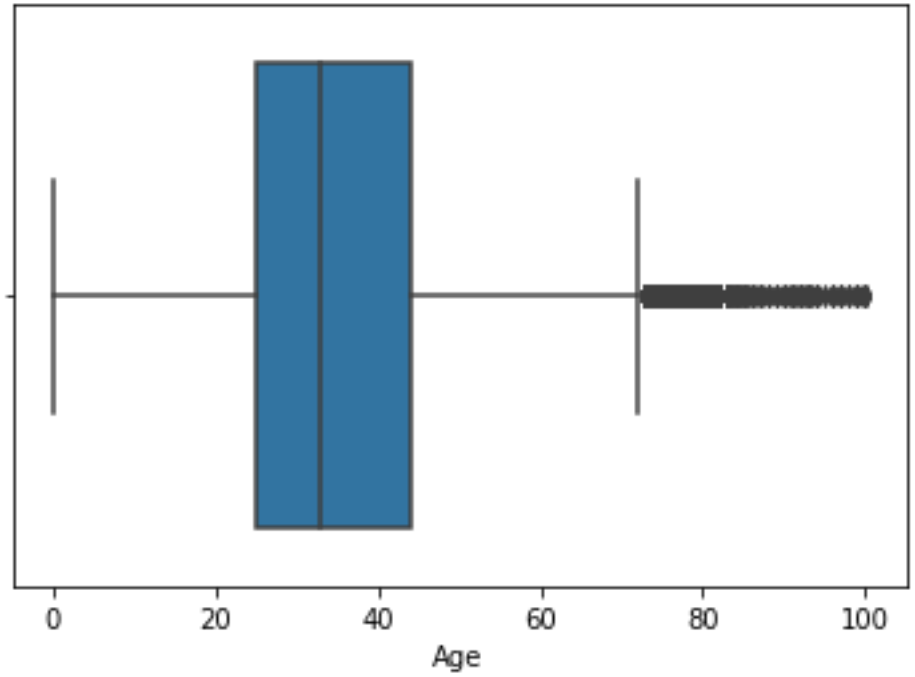
Loading Data & treatment …..

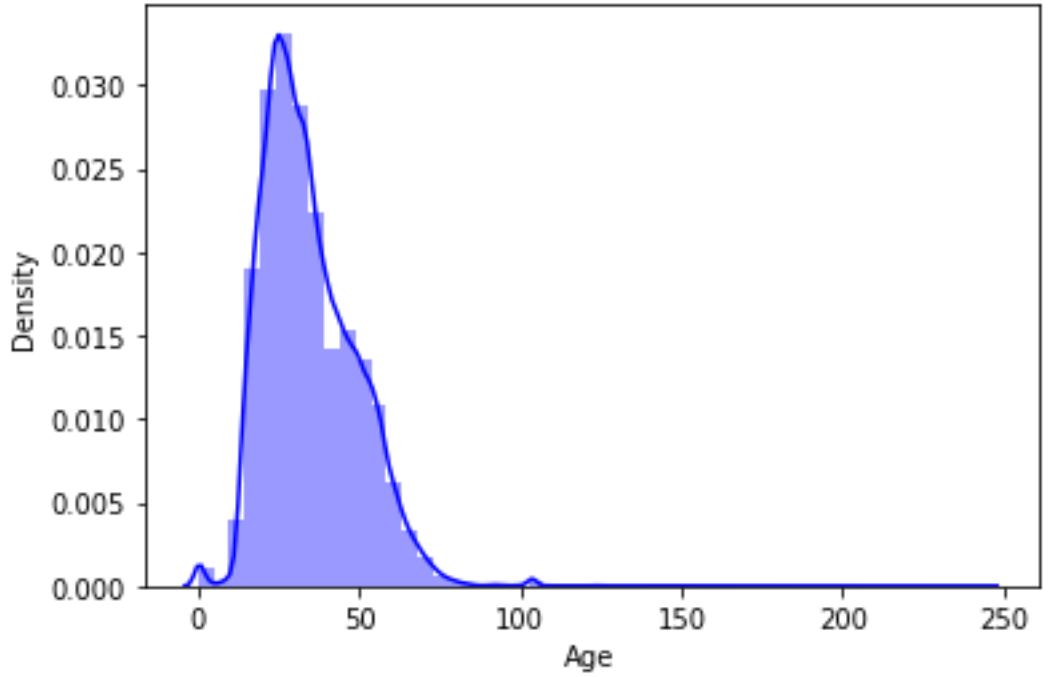
Rating Data set –

- Columns – 3 and Rows – 1149780
- Null value- No null values.
- Drop null value column as dataset is too big
- Outlier – we checked the Rating with Book data set we have 119170 number of ratings are out of our interest.
- We kept only rating related to book data set.
- Duplicate Value – No duplicate value present.

EDA .....

Box plot and Normal Distribution Plot for Age -

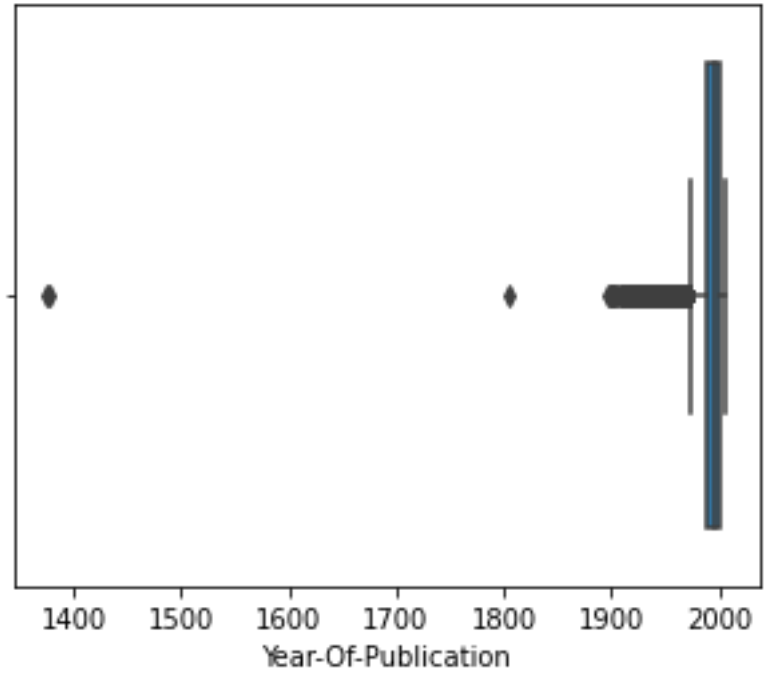


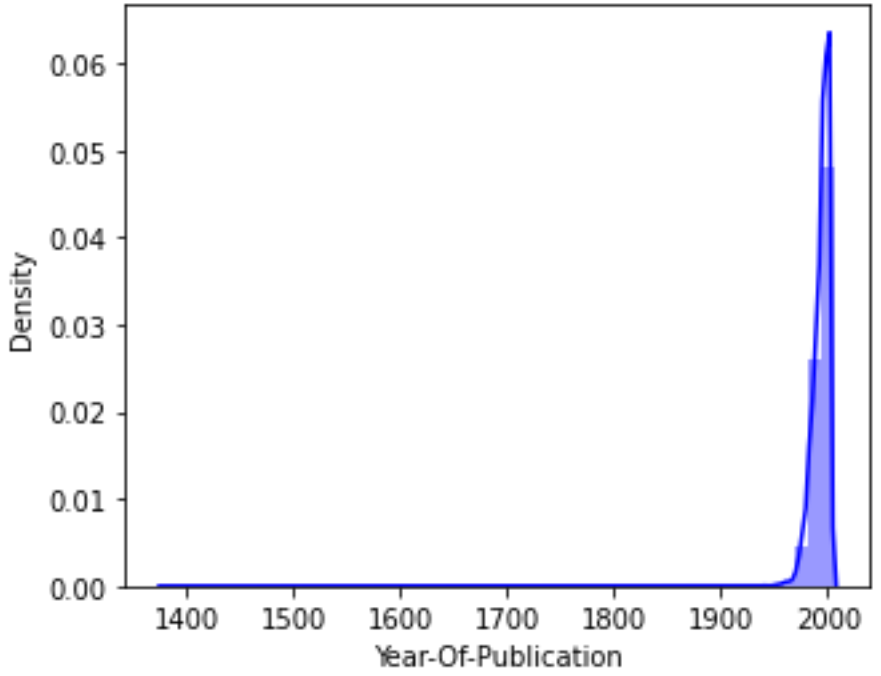


“ we know that people leaves up to max 100 year and also know 2nd thing child cant read up to age of 5 year so ,we will treat this as shown in both box plot and In Normal distribution plot.”

EDA .....

Box plot and Normal Distribution Plot for Year-Of-Publication -

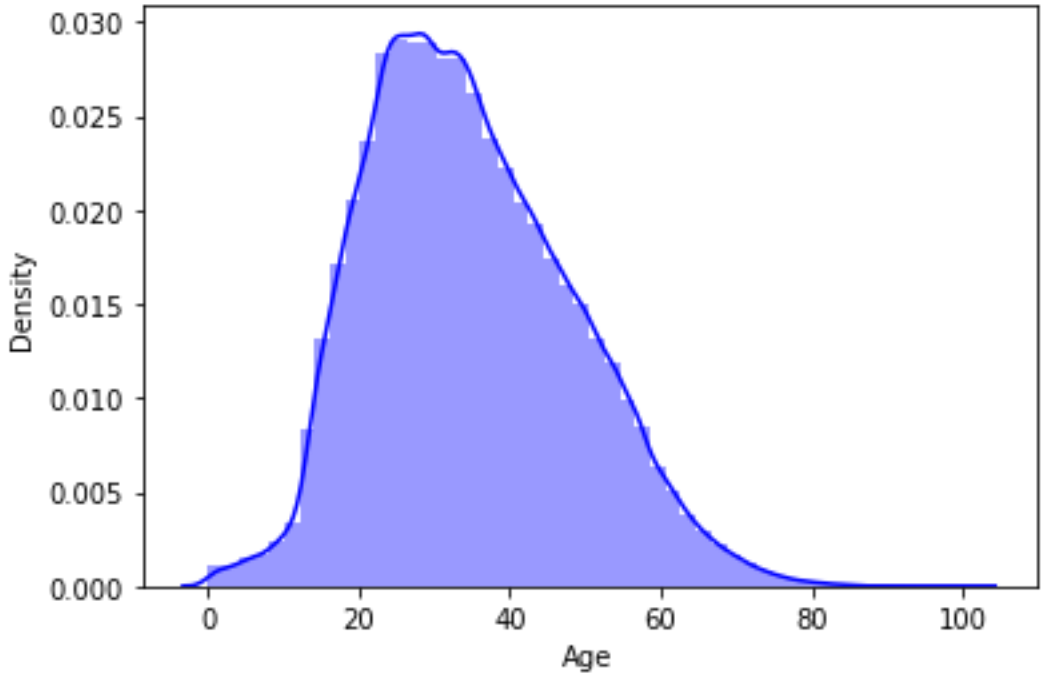




“here we can see that date is left skewed and some of the book published in year 1400 it is possible that the book are published at that time but some of
the book published after 2006 and some of having published year 0 which we can say as a outlier.”

EDA .....

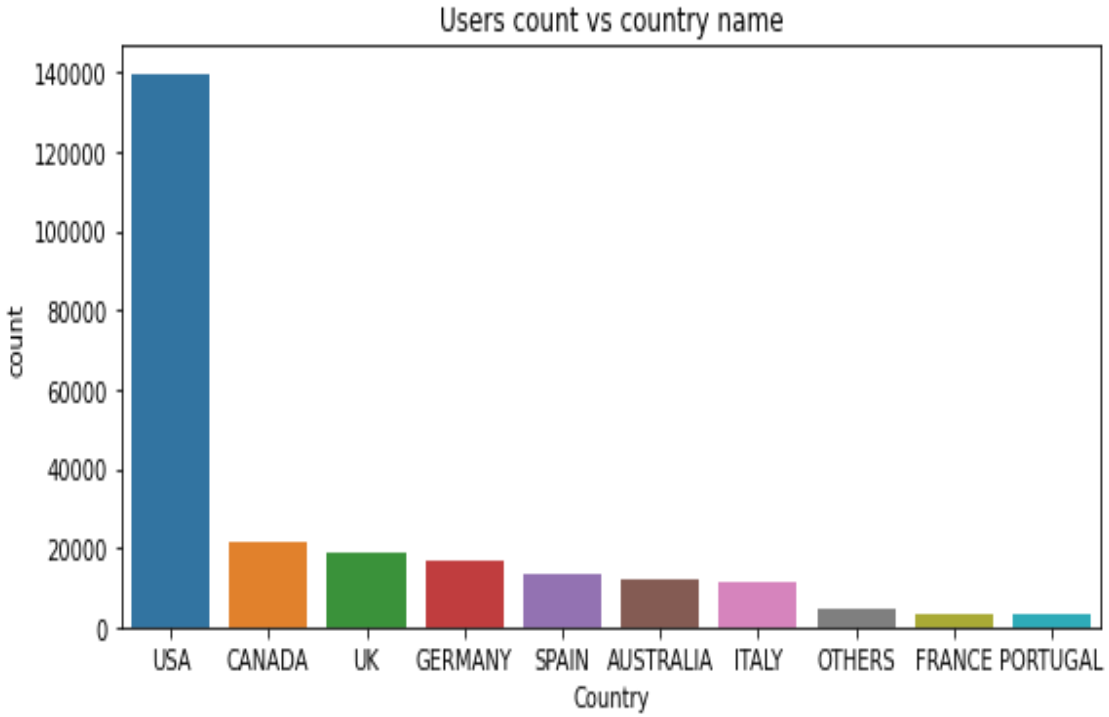
Box plot and Normal Distribution Plot for Age after outlier treatment -

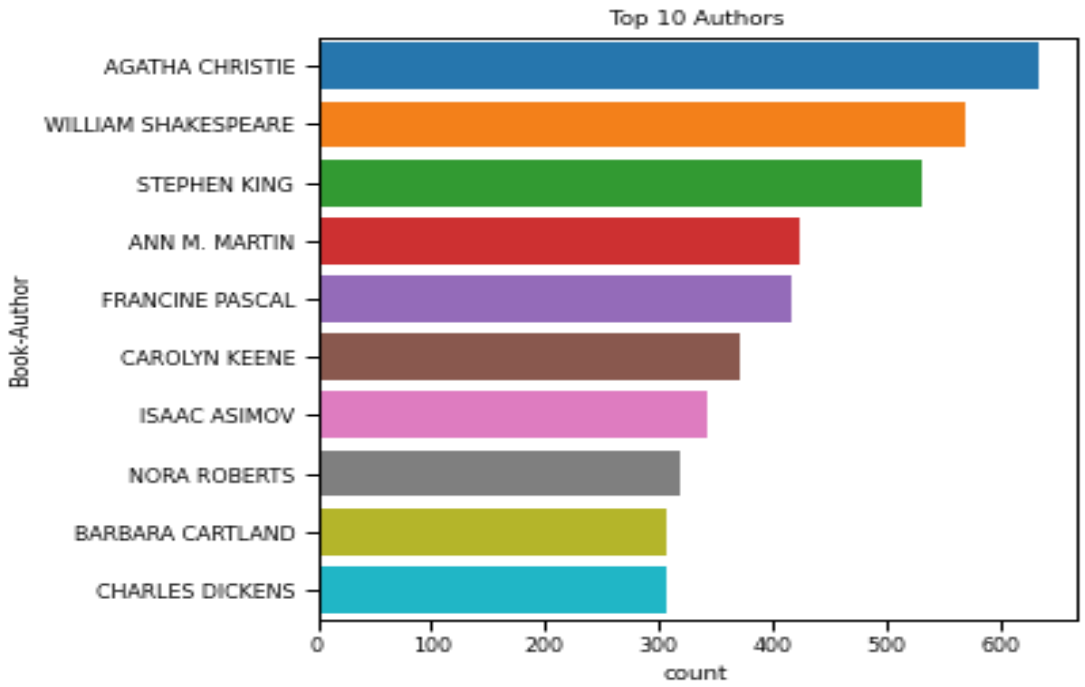


“Here from above plot we see data is normally distributed and most of the users age is in the range of 20 to 40 years .”

EDA .....

User Count vs Country:





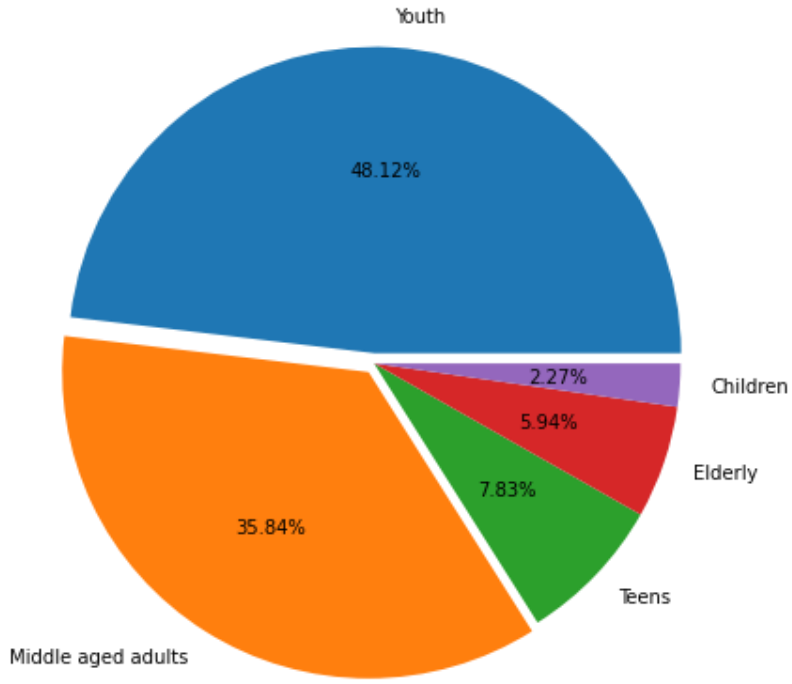
<u>Top Country Vs Users Count</u> – USA(High count)

<u>Top Books Count Vs Author</u>
Top Author is "AGATHA CHRISTHE"

EDA .....

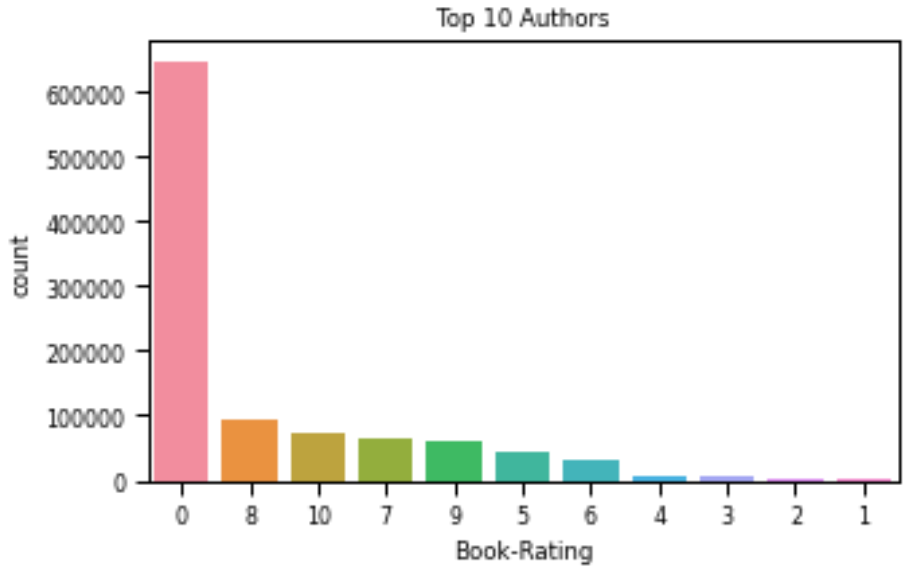
Users Distribution in terms of age group:

“ here we can see that youth and middle aged adults
Are having near about 80% contribution to
Total users.
“ where as children's are contributing very less
In the list.



Users Distribution in terms of age group:

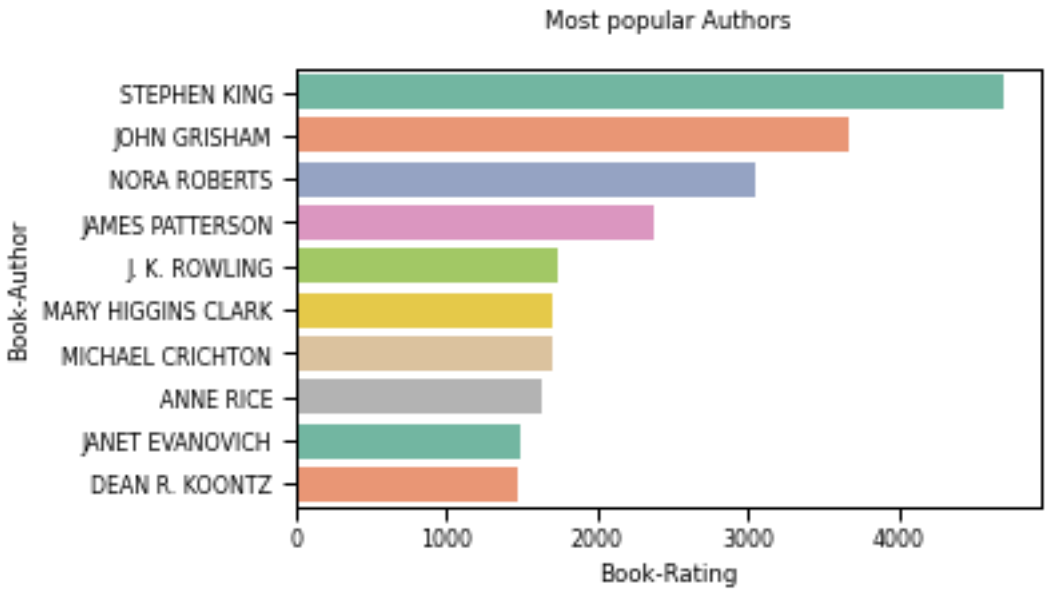
Here the count of zero is high so we can say that
most of the people don't use to give rating
also most of the people has given a rating of 8
That’s nice we can observe that people use to give
rating in between 6 to 10
“We called Zero ratings as a Implicit and
rest of rating as Explicit ratings.”



EDA …..

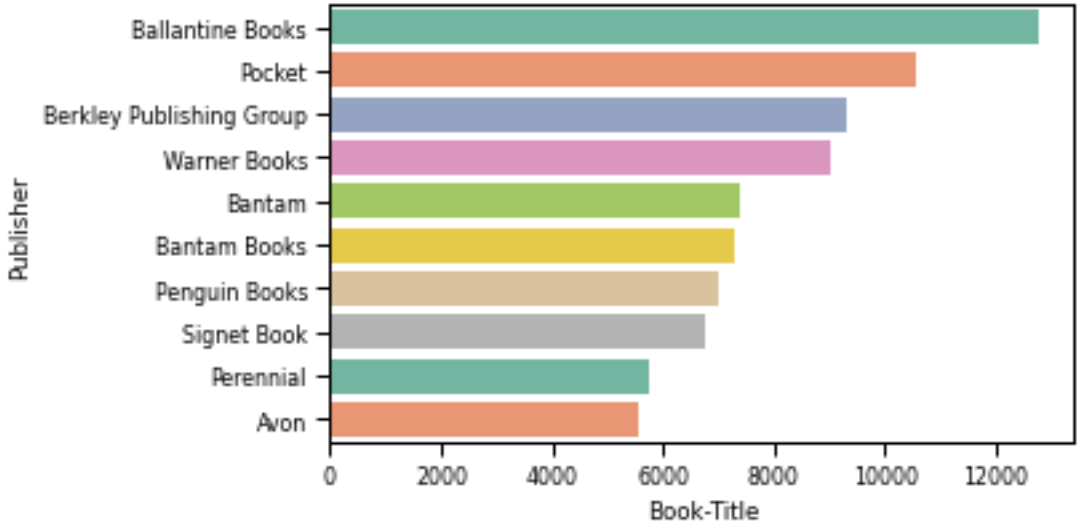
Book rating Count vs Book Author:

Here top Book Author in terms of Book rating count is “Stephen King”.
“ the graph showing top 10 Book Author”



Book Title Count vs Publisher:

Here top Publisher is “Ballantine Books” is the top publisher in terms of book title count.



.

Model

Collaborative Filtering (Model Based):

Matrix Factorization

List of book that user ID 643 already rated

811 George W. Bushisms : The Slate Book of The Acc...

819 Les Fourmis

823 Scottish Folk and Fairy Tales (Penguin Popular...

7289 Pride & Prejudice (Wordsworth Classics)

7856 FORREST GUMP (Movie Tie in)

15641 Vipere Au Poing 15648 Bilbo, Le Hobbit

15655 Sacred Clowns (Joe Leaphorn/Jim Chee Novels)

Recommending books for User ID: 643

	ISBN	Book-Title	Book-Author	Publisher
0	043935806X	Harry Potter and the Order of the Phoenix (Boo...	J. K. ROWLING	Scholastic
1	0439064864	Harry Potter and the Chamber of Secrets (Book 2)	J. K. ROWLING	Scholastic
2	0439136350	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. ROWLING	Scholastic
3	0439139597	Harry Potter and the Goblet of Fire (Book 4)	J. K. ROWLING	Scholastic
4	0590353403	Harry Potter and the Sorcerer's Stone (Book 1)	J. K. ROWLING	Scholastic
5	059035342X	Harry Potter and the Sorcerer's Stone (Harry P...	J. K. ROWLING	Arthur A. Levine Books
6	0439139600	Harry Potter and the Goblet of Fire (Book 4)	J. K. ROWLING	Scholastic Paperbacks
7	0439064872	Harry Potter and the Chamber of Secrets (Book 2)	J. K. ROWLING	Scholastic
8	0439136369	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. ROWLING	Scholastic
9	0446310786	To Kill a Mockingbird	HARPER LEE	Little Brown & Company

'recall@5': 0.3053438458169076, 'recall@10': 0.4153745072273324 3

Model

Memory Based Filtering KNN (Euclidean Distance Based):

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

1. Harry Potter and the Prisoner of Azkaban (Book 3)
2. Harry Potter and the Goblet of Fire (Book 4)
3. Harry Potter and the Sorcerer's Stone (Book 1)
4. Dr. Seuss's A B C (I Can Read It All by Myself Beginner Books)
5. The Second Generation
6. Lover Beware
7. Finders Keepers
8. J. K. Rowling: The Wizard Behind Harry Potter
9. So Much to Tell You
10. Dragonquest Achille Cover

Model

Popularity Based Filtering:

Weighted Average

Most favored books based on the weighted rating scores,

	Book-Title	Book-Author	avg_rating	ratings_count	weighted_average
46514	Harry Potter and the Chamber of Secrets Postca...	J. K. ROWLING	9.870	23	9.520
122142	The Two Towers (The Lord of the Rings, Part 2)	J. R. R. TOLKIEN	9.654	52	9.500
30141	Dilbert: A Book of Postcards	SCOTT ADAMS	9.923	13	9.360
81782	Postmarked Yesteryear: 30 Rare Holiday Postcards	PAMELA E. APKARIAN-RUSSELL	10.000	11	9.340
118124	The Return of the King (The Lord of the Rings,...	J.R.R. TOLKIEN	9.397	78	9.310
17713	Calvin and Hobbes	BILL WATTERSON	9.583	24	9.290
100900	The Authoritative Calvin and Hobbes (Calvin an...	BILL WATTERSON	9.600	20	9.250
72635	My Sister's Keeper : A Novel (Picoult, Jodi)	JODI PICOULT	9.545	22	9.230
118120	The Return of the King (The Lord of The Rings,...	J. R. R. TOLKIEN	9.625	16	9.200
120087	The Sneetches and Other Stories	DR. SEUSS	10.000	8	9.170

This is the list of most favored books based on the weighted rating scores. The book 'Harry Potter and the Chamber of Secrets Postcard Book' seems to have top this chart.

Conclusion

The initial step,of our project was Data preprocessing of the three datasets-books_df,users_df and ratings_df, where in we removed duplicates and imputed the missing values & invalid entries with appropriate values, corrected spellings. Then we performed Exploratory Data Analysis to find out the countries with maximum users ,popular books ,popular authors and popular publishing companies.We also analysed the rating distribution, age distribution of users and the popular books amongst various age groups . Then, we used Popularity-based approach , Collaborative filtering approach to built different types of recommendation models. *We evaluated the performance of Singular Value Decomposition based recommender and obtained a Global Recall@5 of 30 % and Recall@10 of 41%

Thanks