Capstone Project
Book Recommendation System

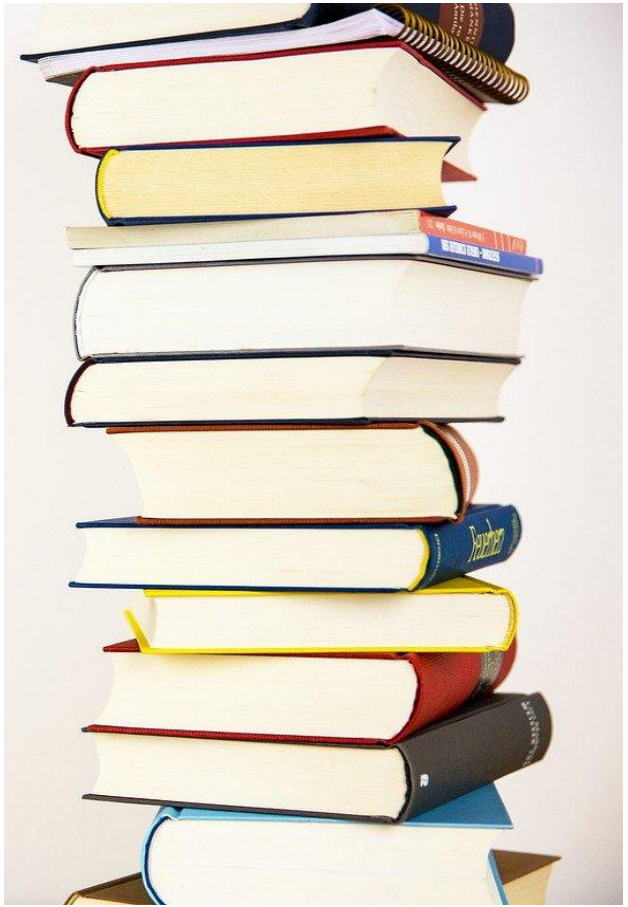
Team

Kanishka Raj, Sridhar Nagar, Raushan Kumar

AImaBetter

Content

- ❑ Problem Statement
- ❑ Data Pipeline
- ❑ Data Description
- ❑ Exploratory Data Analysis
- ❑ Rating Dataset
- ❑ Data Cleaning
- ❑ Models Performed
- ❑ Model Evaluation & Result
- ❑ Evaluation Matrix of All the models
- ❑ Challenges
- ❑ Conclusion
- ❑ Future Scope



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). Recommendation 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.

Data Pipeline

- Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to see the trend.

- Data Processing: In this part we went through each attributes and encoded the categorical features.

- Model Creation: Finally in this part we created the various models. These various models are being analysed and we tried to study various models so as to get the best performing model for our project.

Data Description

The Book-Crossing dataset comprises 3 files:

Users_dataset: Contains the users information

- User-ID (unique for each user)
- Location (contains city, state and country separated by commas)
- Age Shape of Dataset - (278858, 3)

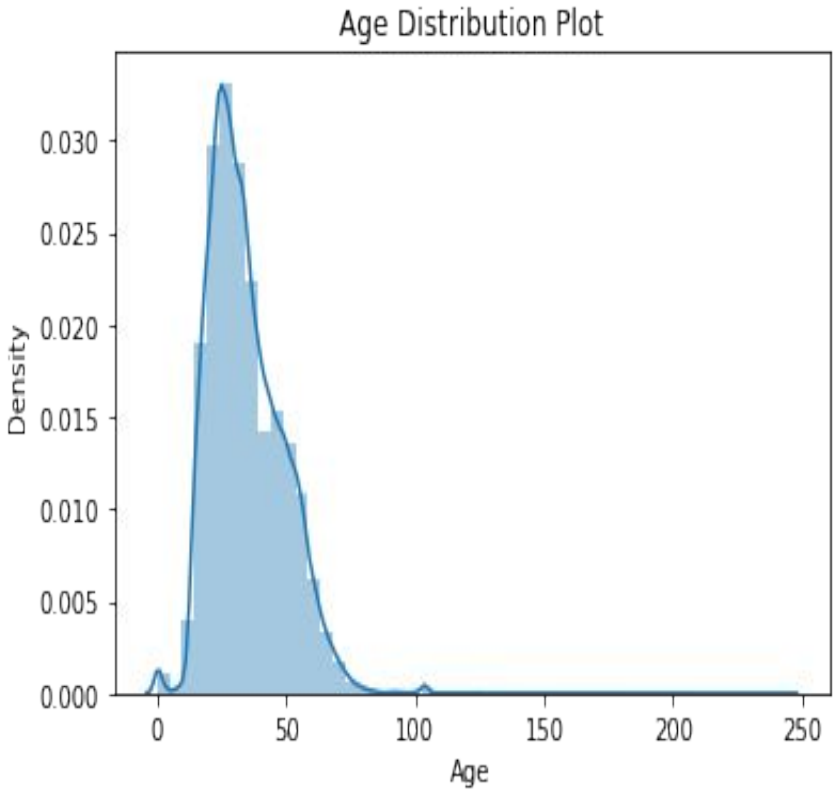
Books_dataset:

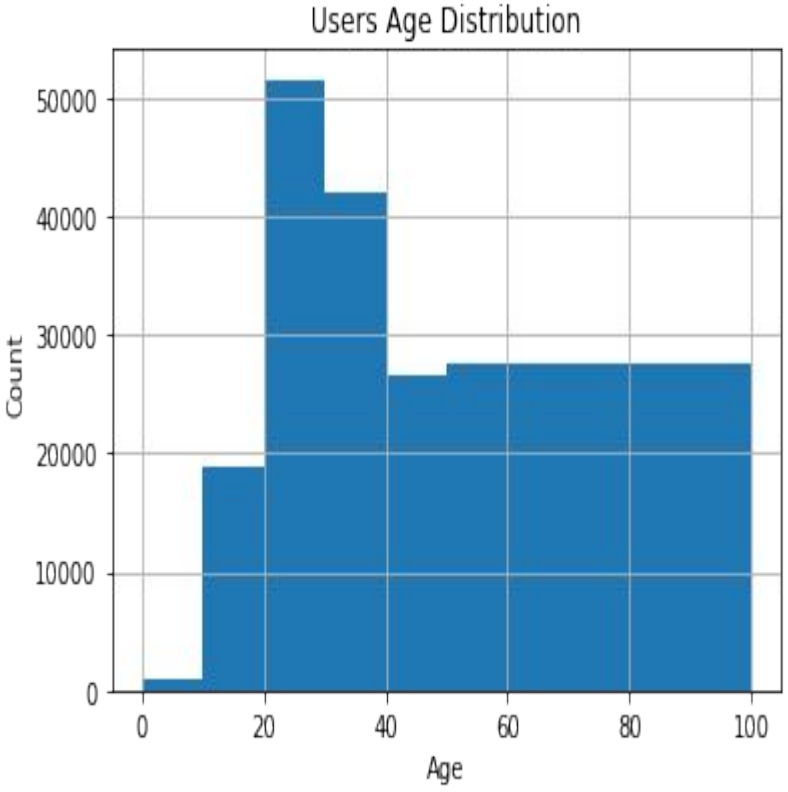
- ISBN (unique for each book)
- Book-Author
- Publisher
- Image-URL-S
- Shape of Dataset - (271360, 8)
- Book-Title
- Year-Of-Publication
- Image-URL-M
- Image-URL-L

Ratings_dataset: Contains the book rating information.

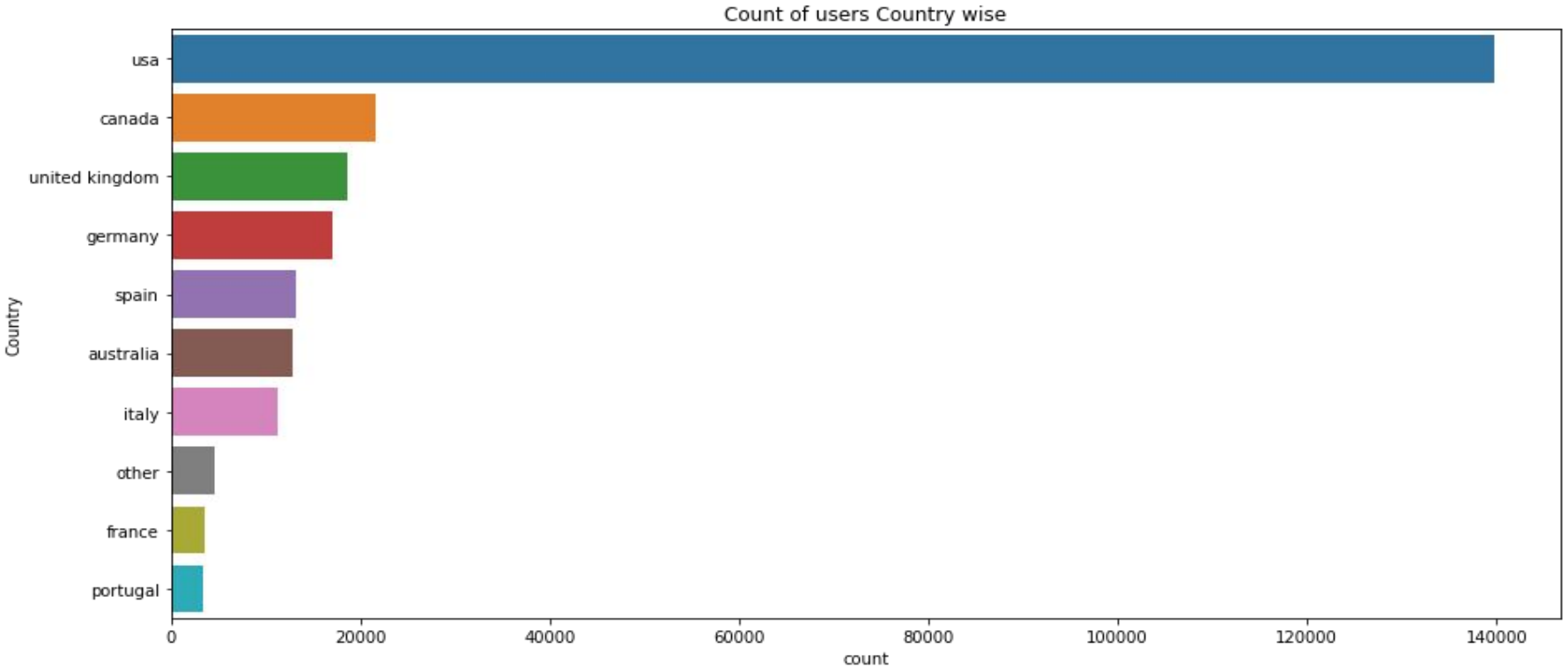
- User-ID ISBN
- Book-Rating
- Shape of Dataset - (1149780, 3)

EDA- Users_df (Age) & Users_df (Age)

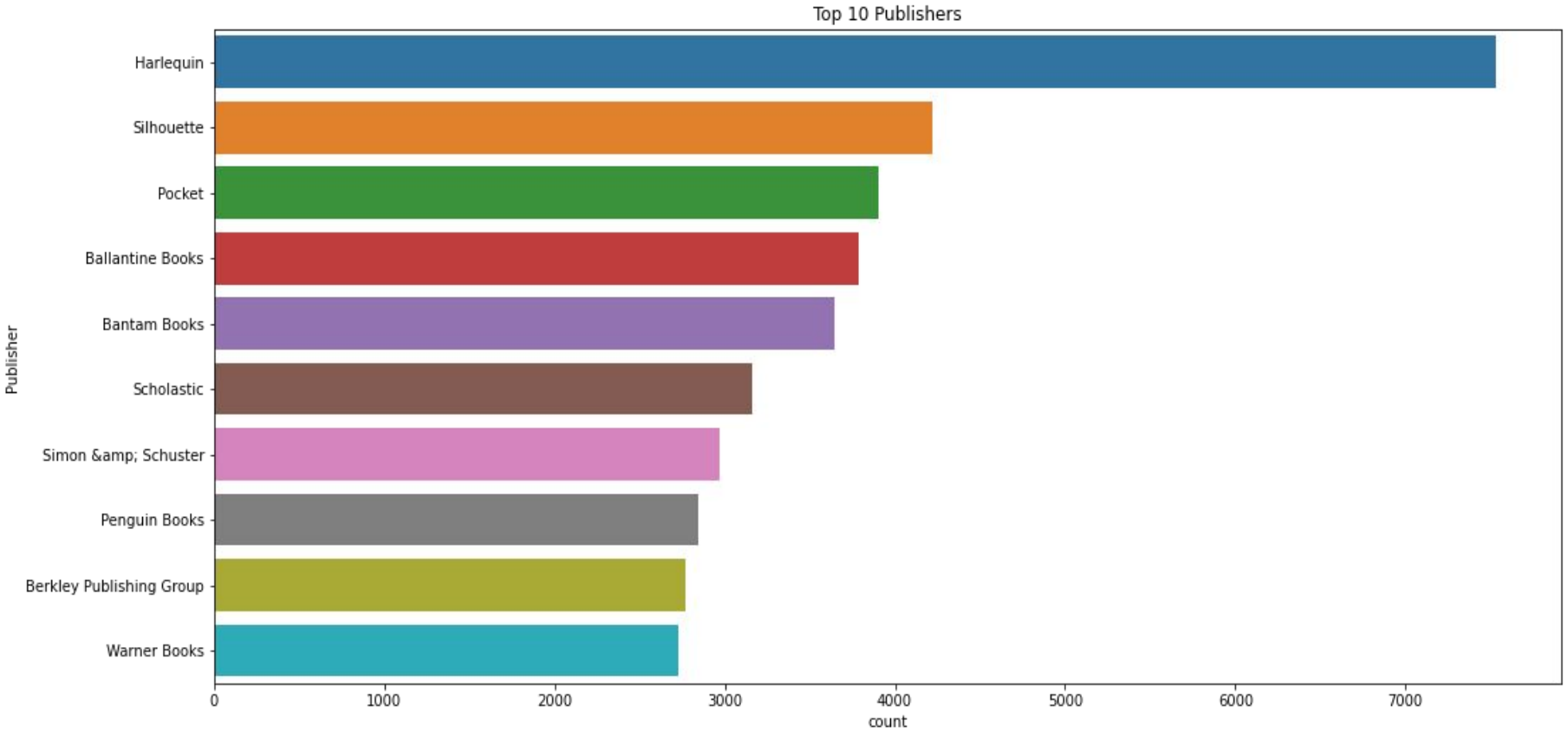




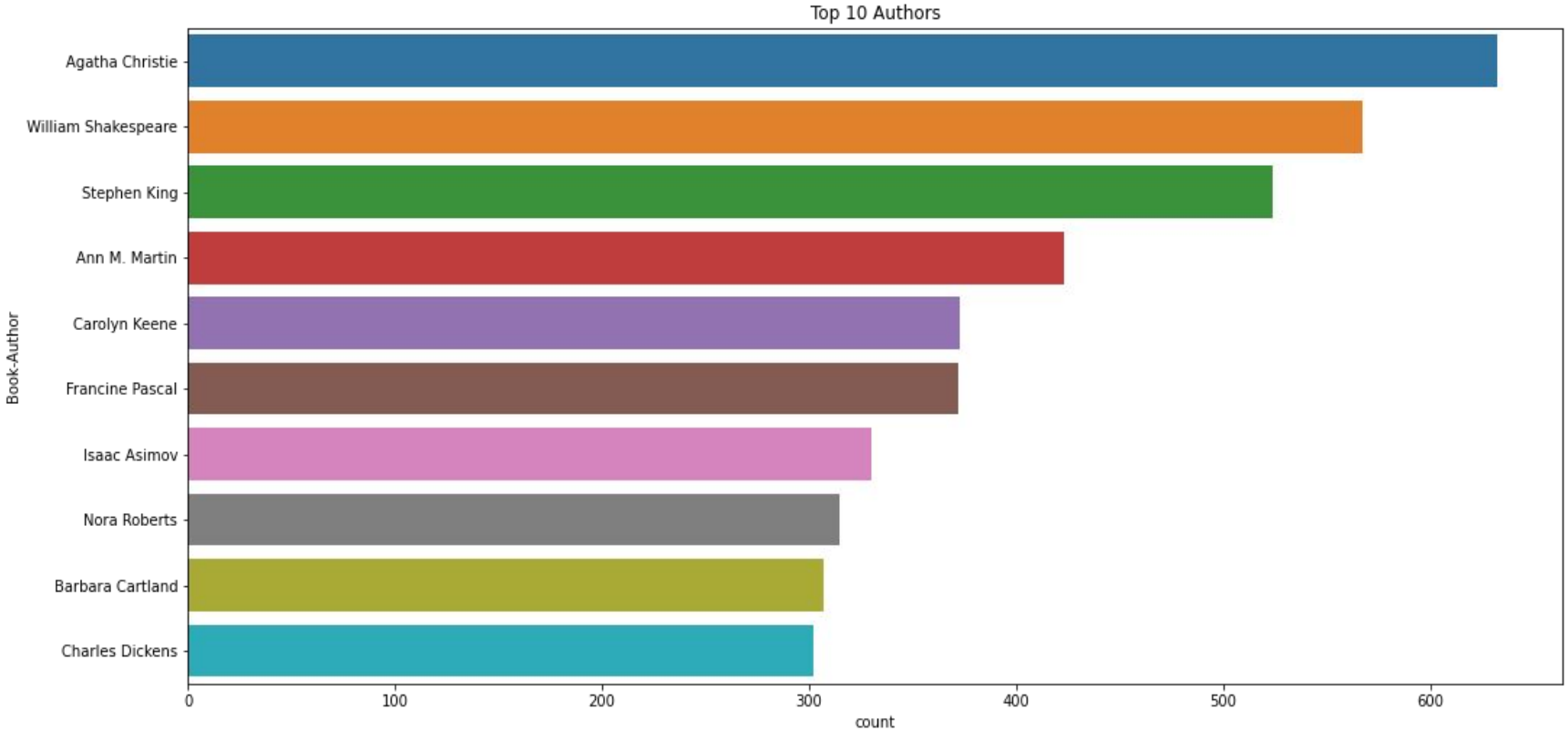
EDA- Users_df (Location)



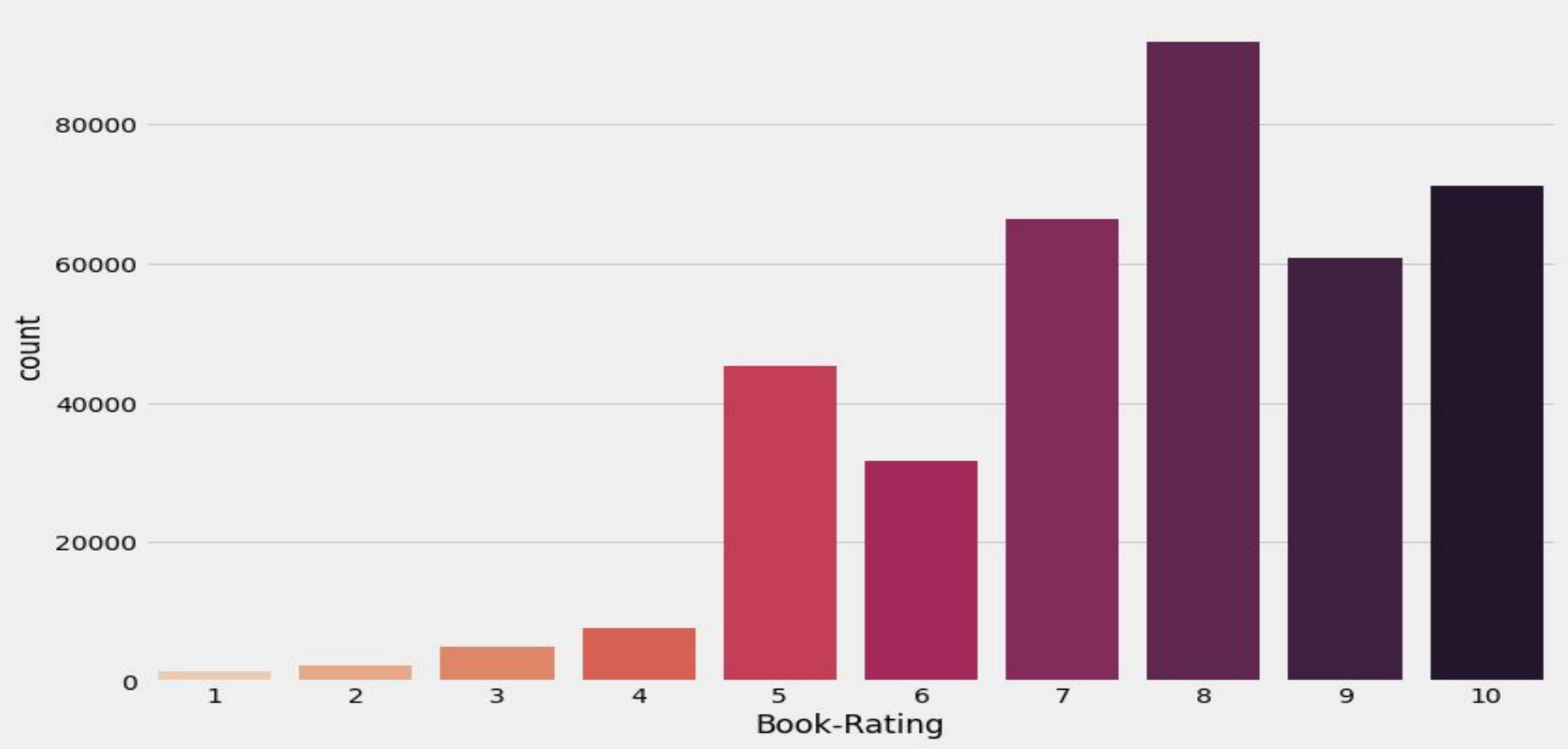
EDA- Book_df (Publishers)



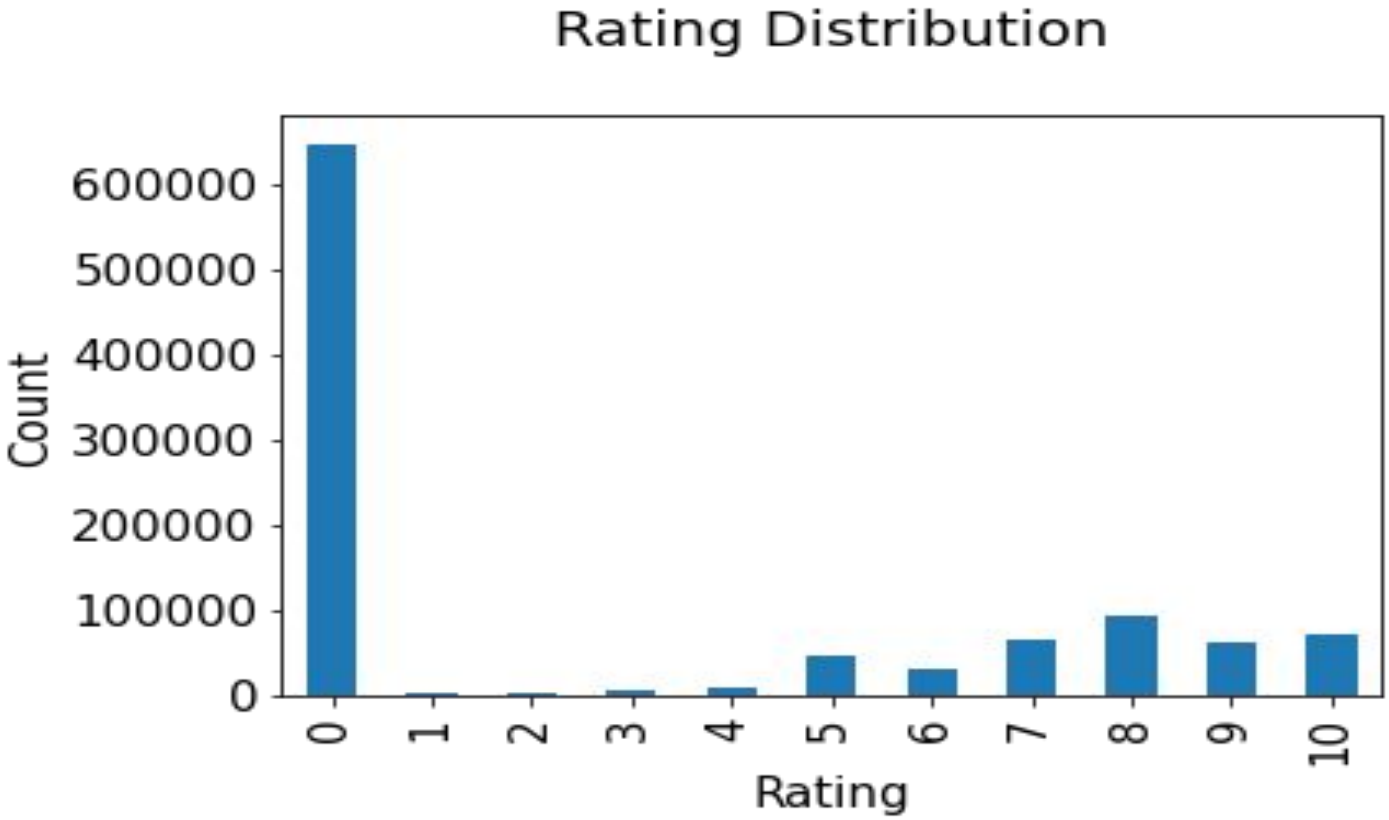
Book_df (Authors)



Ratings_df (Book_Rating)



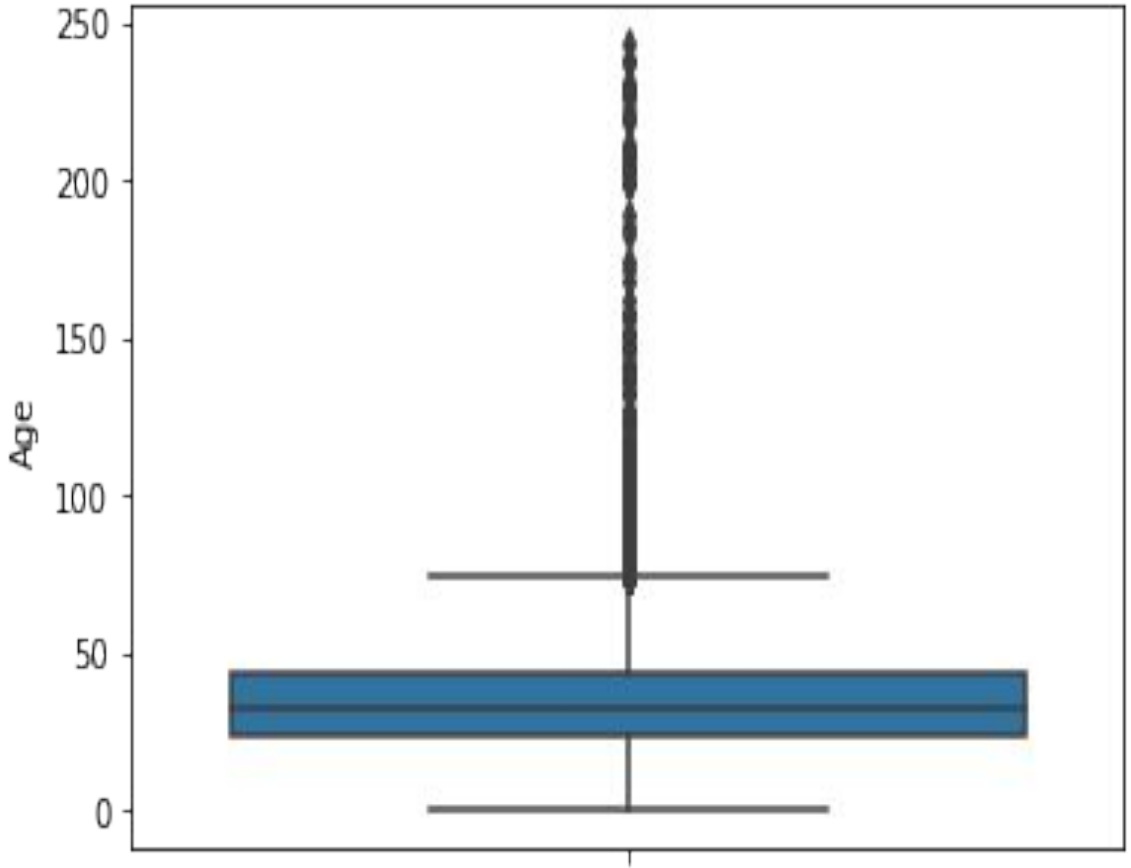
Rating Dataset



Data Cleaning - Null Value Imputation

index	Missing Values	% of Total Values	Data_type	
0	User-ID	0	0.0	int64
1	Age	0	0.0	float64
2	Country	0	0.0	object
3	ISBN	0	0.0	object
4	Book-Rating	0	0.0	int64
5	Avg_Rating	0	0.0	float64
6	Total_No_Of_Users_Rated	0	0.0	int64
7	Book-Title	0	0.0	object
8	Book-Author	0	0.0	object
9	Year-Of-Publication	0	0.0	float64
10	Publisher	0	0.0	object

Data Cleaning Checking Outliers (Missing values)



Replacing strings by int values

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Image-URL-M	Image-URL-L
209538	078946697X	DK Readers: Creating the X-Men, How It All Beg...	2000	DK Publishing Inc	http://images.amazon.com/images/P/078946697X.0...	http://images.amazon.com/images/P/078946697X.0...	http://images.amazon.com/images/P/078946697X.0...	NaN
221678	0789466953	DK Readers: Creating the X-Men, How Comic	2000	DK Publishing Inc	http://images.amazon.com/images/P/0789466953.0...	http://images.amazon.com/images/P/0789466953.0...	http://images.amazon.com/images/P/0789466953.0...	NaN

Model’s Performed

- Popularity Based Recommendation
- Model based collaborative filtering
- Collaborative Filtering-(Item-Item based)
- Collaborative Filtering-(User-Item based)

Popularity Based Recommendation

The popularity index used for our books dataset was **weighted rating**. The formula for weighted rating is:

$$WR = [(v * R)/(v + m)] + [(m * c)/(v + m)]$$

Where,

WR is weighted rating;

v is the number of votes for the books;

m is the minimum votes required to be listed in the chart;

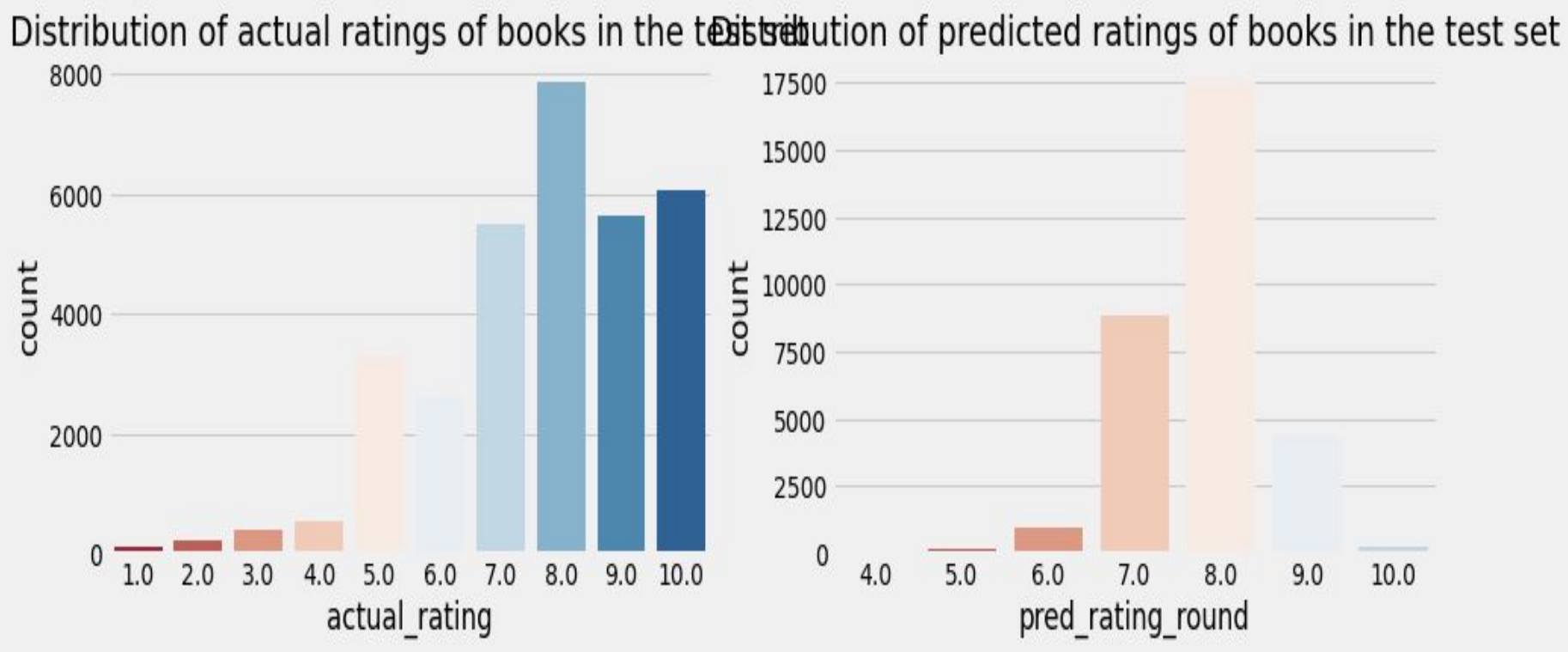
R is the average rating of the book; and

C is the mean vote across the whole report.

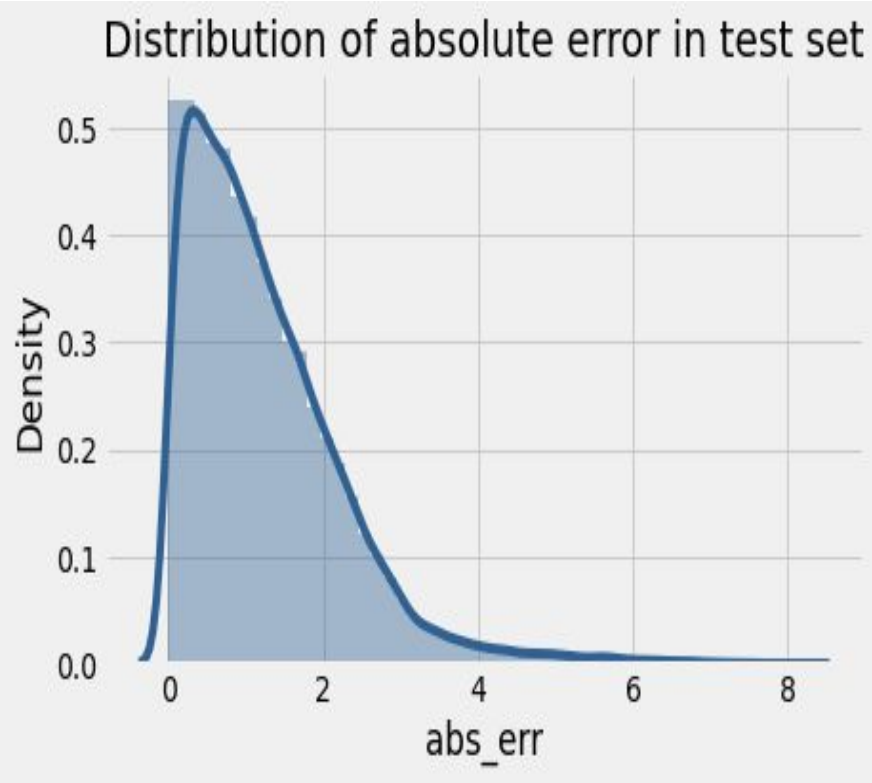
Model based collaborative filtering

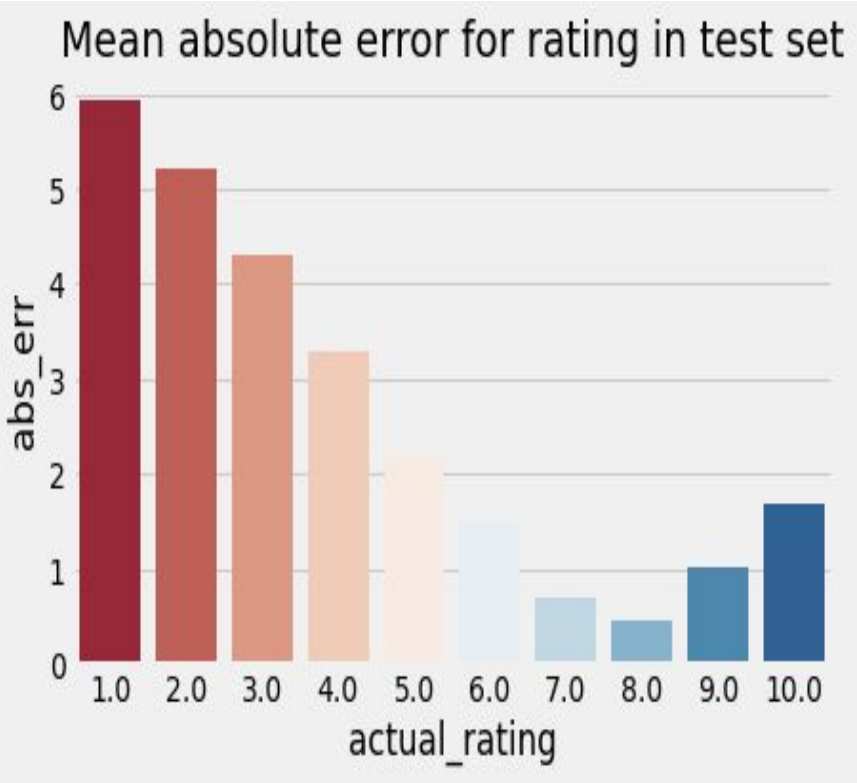
	user_id	isbn	actual_rating	pred_rating	impossible	pred_rating_round	abs_err
810	232107	0345453743	10.0	8.659923	False	9.0	1.340077
15173	86728	0061099643	10.0	7.713388	False	8.0	2.286612
4186	261522	0786866586	7.0	7.013394	False	7.0	0.013394
13	28700	1551669293	7.0	7.943029	False	8.0	0.943029
9865	243312	0765342987	1.0	7.527790	False	8.0	6.527790

SVD Model Results



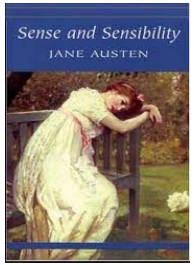
SVD Model Results

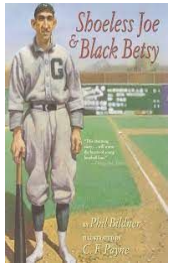


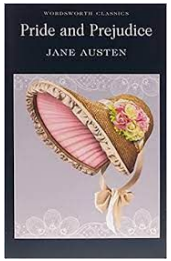


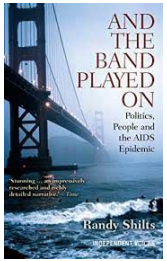
Train set: Top rated books

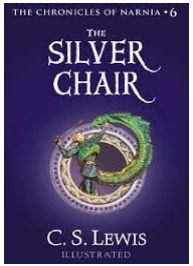
	user_id	isbn	book_rating	Avg_Rating	Total_No_Of_Users_Rated	book_title	pred_rating
113549	193458	0140298479	9	7.539823	113	Bridget Jones: The Edge of Reason	NaN
113594	193458	1853260169	10	8.153846	13	Sense and Sensibility (Wordsworth Classics)	NaN
113558	193458	0345342569	9	7.947368	19	Shoeless Joe	NaN
113592	193458	1853260002	10	8.217391	23	Pride & Prejudice (Wordsworth Classics)	NaN
113583	193458	0671880314	9	8.305556	36	Schindler's List	NaN
113548	193458	014011369X	9	9.125000	8	And the Band Played on: Politics, People, and ...	NaN
113546	193458	0064471098	9	8.733333	15	The Silver Chair	NaN
113541	193458	0064471047	9	8.714286	42	The Lion, the Witch, and the Wardrobe (The Chr...	NaN
113584	193458	0679429220	9	7.794393	107	Midnight in the Garden of Good and Evil: A Sav...	NaN
113545	193458	006447108X	9	8.833333	18	The Last Battle	NaN

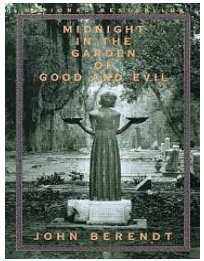






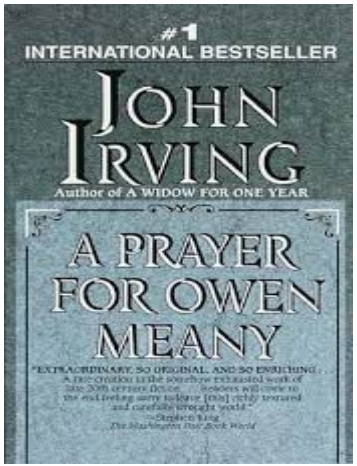


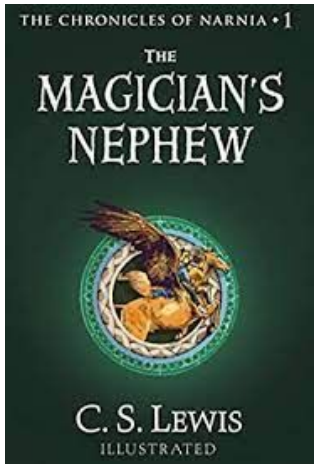


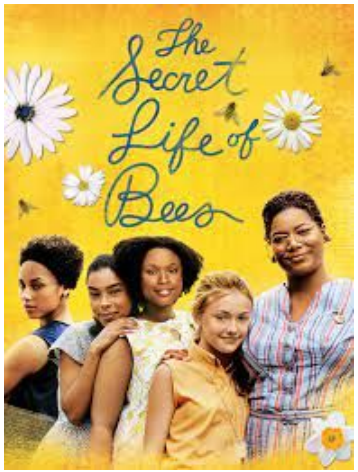


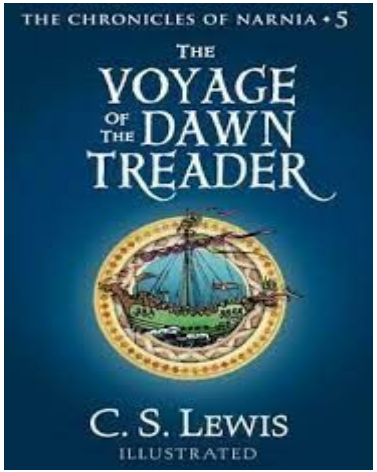
Test set: predicted top rated books

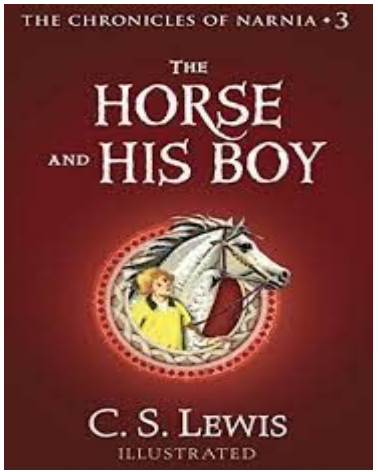
	user_id	isbn	book_rating	Avg_Rating	Total_No_Of_Users_Rated	book_title	pred_rating
113559	193458	0345361792	10	8.607735	181	A Prayer for Owen Meany	8.539484
113547	193458	0064471101	9	8.729730	37	The Magician's Nephew (rack) (Narnia)	8.294170
113552	193458	0142001740	9	8.452769	307	The Secret Life of Bees	8.214533
113544	193458	0064471071	9	8.733333	15	The Voyage of the Dawn Treader (rack) (Narnia)	8.086277
113543	193458	0064471063	9	8.518519	27	The Horse and His Boy	8.021627





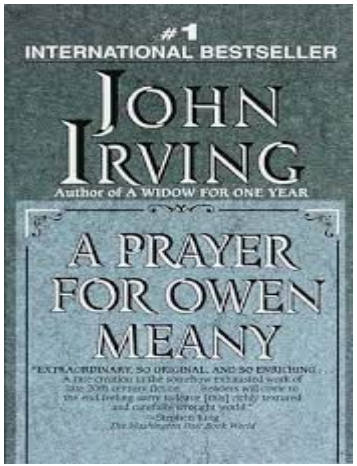


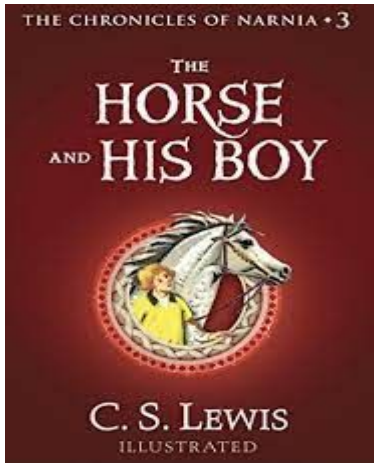


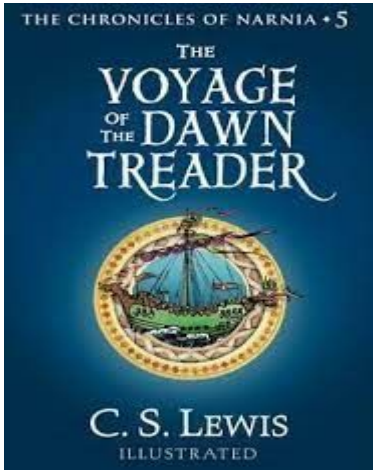


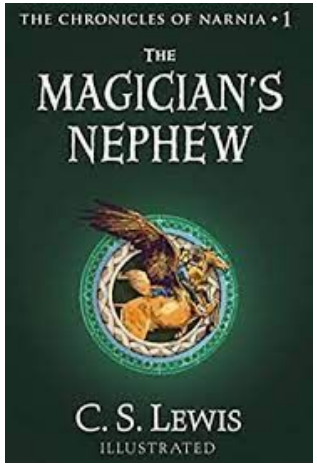
Test set: actual top rated books

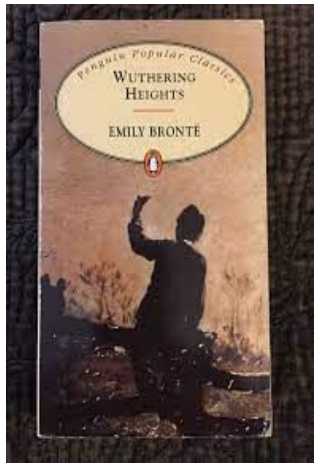
	user_id	isbn	book_rating	Avg_Rating	Total_No_Of_Users_Rated	book_title	pred_rating
113559	193458	0345361792	10	8.607735	181	A Prayer for Owen Meany	8.539484
113543	193458	0064471063	9	8.518519	27	The Horse and His Boy	8.021627
113544	193458	0064471071	9	8.733333	15	The Voyage of the Dawn Treader (rack) (Narnia)	8.086277
113547	193458	0064471101	9	8.729730	37	The Magician's Nephew (rack) (Narnia)	8.294170
113550	193458	0140620125	9	8.133333	15	Wuthering Heights (Penguin Popular Classics)	7.628546











Collaborative Filtering - (Item-Item based)

	user_id	isbn	book_rating	Avg_Rating	Total_No_Of_Users_Rated
16	276747	0060517794	9	8.000000	30
19	276747	0671537458	9	7.176471	17
20	276747	0679776818	8	7.476190	21
59	276772	0553572369	7	6.625000	8
61	276772	3499230933	10	7.166667	6

	userID	ISBN	bookRating	Avg_Rating	Total_No_Of_Users_Rated
16	276747	0060517794	9	8.000000	30
19	276747	0671537458	9	7.176471	17
20	276747	0679776818	8	7.476190	21
59	276772	0553572369	7	6.625000	8
61	276772	3499230933	10	7.166667	6

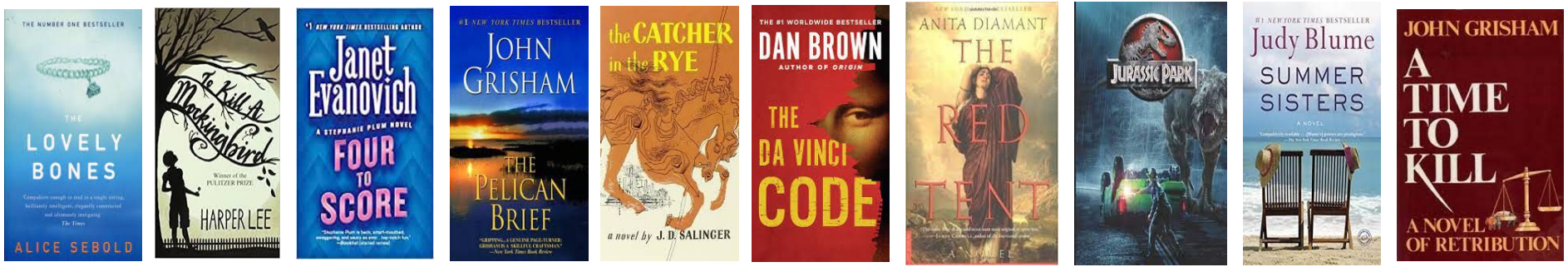
Implementing KNN

	bookTitle	TotalRatingCount
0	A Light in the Storm: The Civil War Diary of ...	4
1	Always Have Popsicles	1
2	Apple Magic (The Collector's series)	1
3	Ask Lily (Young Women of Faith: Lily Series, ...	1
4	Beyond IBM: Leadership Marketing and Finance ...	1

	userID	ISBN	bookRating	bookTitle	TotalRatingCount
0	276725	034545104X	0	Flesh Tones: A Novel	60
1	2313	034545104X	5	Flesh Tones: A Novel	60
2	6543	034545104X	0	Flesh Tones: A Novel	60
3	8680	034545104X	5	Flesh Tones: A Novel	60
4	10314	034545104X	9	Flesh Tones: A Novel	60

Collaborative Filtering-(User-Item based)

	ISBN	Book-Title	recStrength
0	0316666343	The Lovely Bones: A Novel	0.175
1	0446310786	To Kill a Mockingbird	0.105
2	0312966970	Four To Score (A Stephanie Plum Novel)	0.102
3	0440214041	The Pelican Brief	0.100
4	0316769487	The Catcher in the Rye	0.095
5	0385504209	The Da Vinci Code	0.092
6	0312195516	The Red Tent (Bestselling Backlist)	0.091
7	0345370775	Jurassic Park	0.090
8	0440226430	Summer Sisters	0.090
9	0440211727	A Time to Kill	0.088



Model Evaluation & Result

```
Global metrics:
{'modelName': 'Collaborative Filtering', 'recall@5': 0.23708545146453644, 'recall@10': 0.3059791817961753}
```

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	User-ID
10	264	339	1389	0.190	0.244	11676
31	192	244	1138	0.169	0.214	98391
45	22	28	380	0.058	0.074	189835
30	87	103	369	0.236	0.279	153662
70	27	35	236	0.114	0.148	23902
7	27	46	204	0.132	0.225	235105
47	25	31	203	0.123	0.153	76499
50	26	35	193	0.135	0.181	171118
42	59	73	192	0.307	0.380	16795
43	19	33	188	0.101	0.176	248718

Challenges

- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.

- Understanding the metric for evaluation was a challenge as well.

- Decision making on missing value imputations and outlier treatment was quite challenging as well.

- As dataset was quite big enough which led more computation time.

- In EDA, the Top-10 most rated books were essentially novels. Books like The Lovely Bone and The Secret Life of Bees were very well perceived.
- Majority of the readers were of the age bracket 20-35 and most of them came from North American and European countries namely USA, Canada, UK, Germany and Spain.
- If we look at the ratings distribution, most of the books have high ratings with maximum books being rated 8. Ratings below 5 are few in number.
- Author with the most books was Agatha Christie, William Shakespeare and Stephen King.
- We can conclude that item-item based collaborative filtering performed better than user-user based collaborative filtering because of lower computation among the memory based approach.
- For modelling, it was observed that for model based collaborative filtering SVD technique worked way better than NMF with lower Mean Absolute Error (MAE)

Future Scope

- The future of recommender systems lie in integrating self actualization to do justice to serendipity while recommending which will also support rather than replace human decision-making by understanding preferences.

- Recommender systems can be broadly divided into: Collaborative filtering, Content-based filtering, Hybrid recommender systems, Personality-based recommender systems. Each type of filtering algorithm is used according to the specific need of the application or the product.

- Here we are trying to achieve a recommendation which is not extremely personalised which may feel intrusive to the user and not very generic that it doesn't really account the user's distinct taste. Striking a balance between the two is what needs to be achieved.

- Furthermore there will be plenty of work needed to be done on the famous 'cold start problem' in order to somehow manage collecting just the right amount of implicit information and data to recommend users even if there is little or no direct information available on users.

THANK YOU