Capstone Project -4

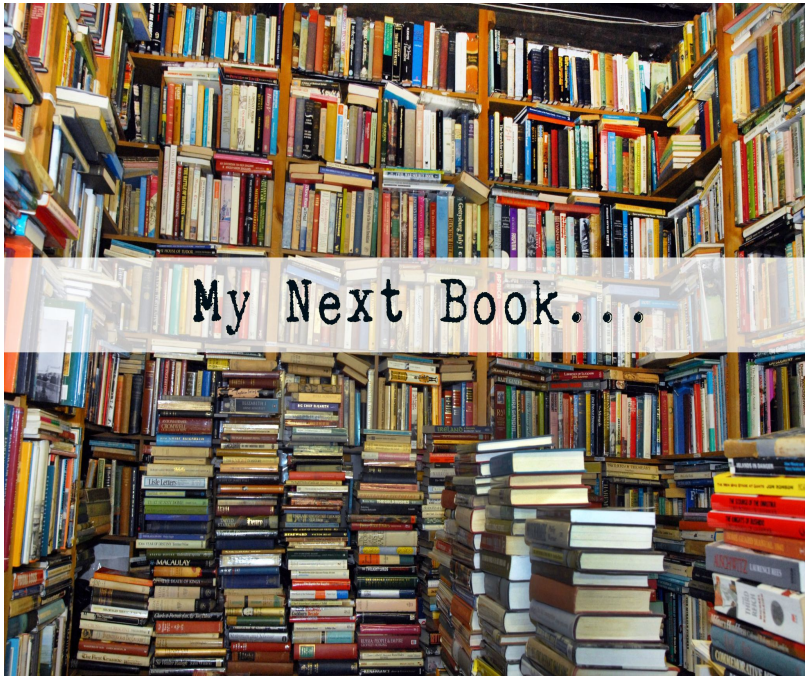
BOOK RECOMMENDATION SYSTEM

Anson sibi

Content

- Problem statement
- Data Summary
- Analysis of different datasets
- Data Cleaning
- Outlier treatment
- Imputing missing values
- Different Recommendation Model
- 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 become much more important in our lives in terms of providing highly personalized and relevant content.

The main objective is to create a recommendation system to recommend relevant books to users based on popularity and user interests.

Data Summary

The dataset is comprised of three csv files:: User_df, Books_df, Ratings_df

Users_dataset.

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

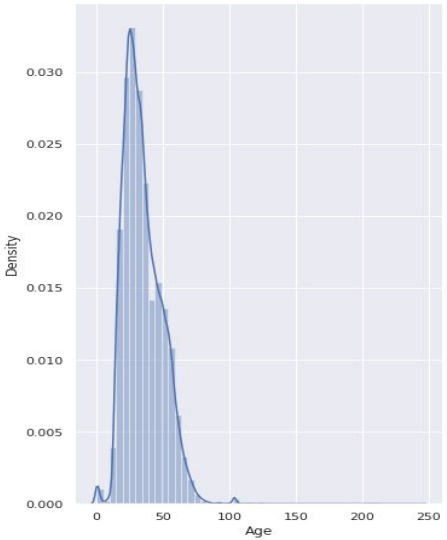
Ratings_dataset.

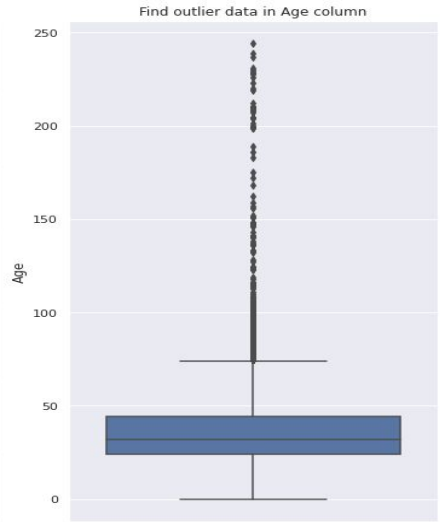
- User-ID
- ISBN
- Book-Rating
- Shape of Dataset - (1149780, 3)

Users (Age)

- Percentage of missing values in Age column is 39.71%.
- Age distribution is positively skewed
- Most active readers lie in age group 20- 40
- There are outliers in the Age column.
- Age distribution is positively skewed median is used to remove outliers and replace missing values.

	Features	Missing	Percentage of total values
0	User-ID	0	0.000000
1	Location	0	0.000000
2	Age	110762	39.719857





Data Cleaning

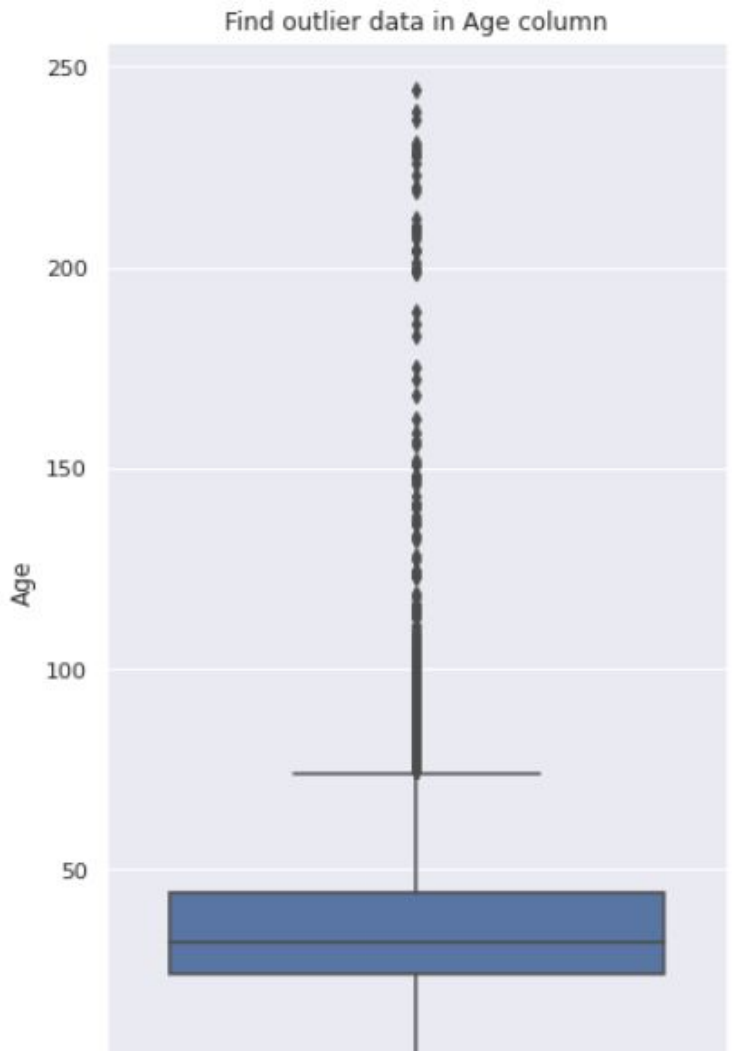
1. **Null Value Imputation:**

Age column has almost 40% missing values

	Features	Missing	Percentage of total values
0	User-ID	0	0.000000
1	Location	0	0.000000
2	Age	110762	39.719857

Imputing missing values

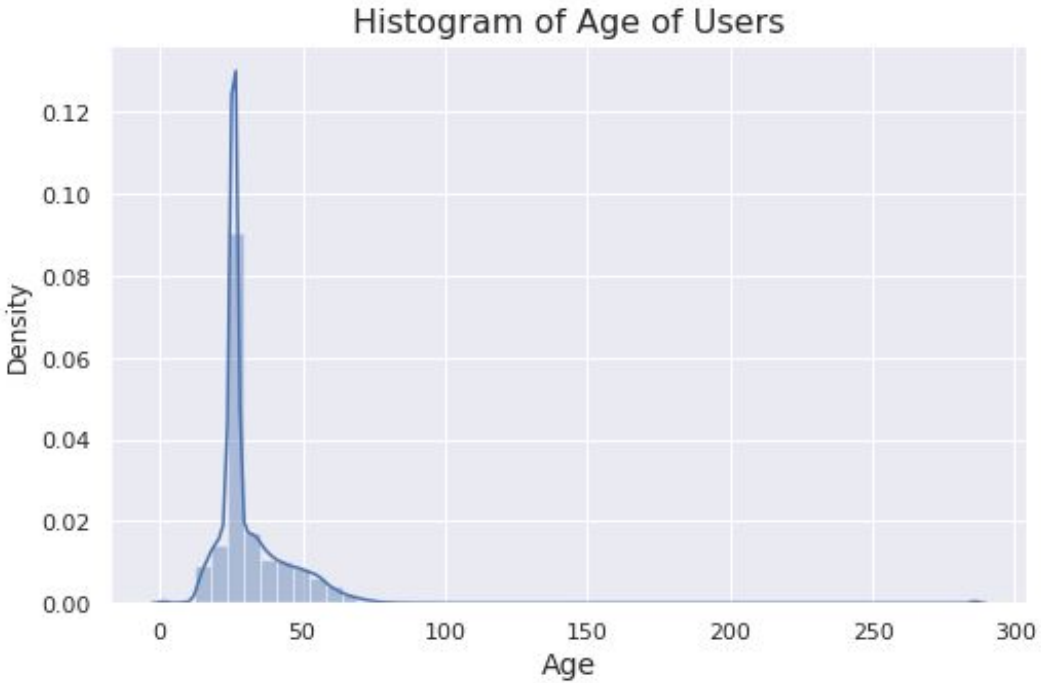
- Outliers in Age column
- Age has positive Skewness (right tailed)
- Age distribution is positively skewed median is used to remove outliers and replace missing values.



Users (Age)

This is the plot after dealing with missing values and outliers

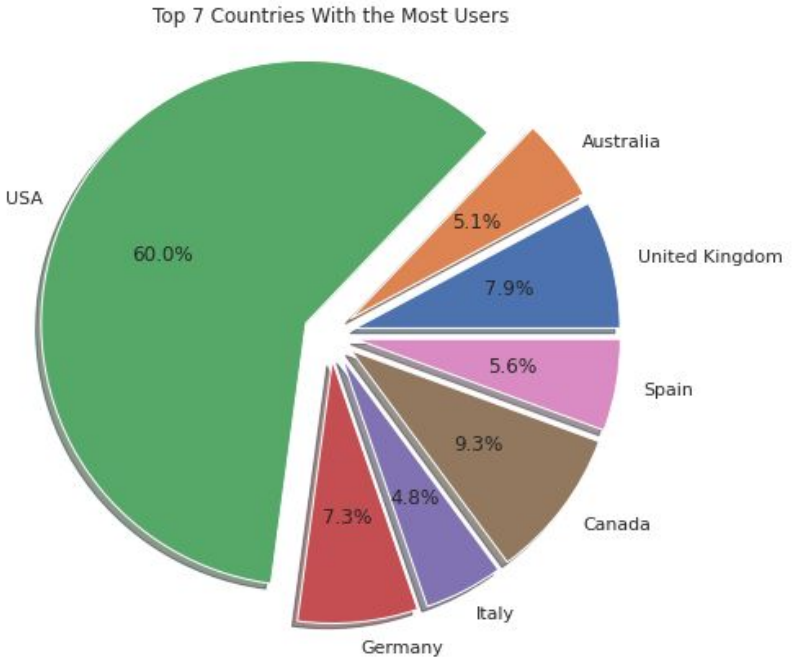
- The Age range distribution is right skewed.
- Most active readers lie in age group 20- 40



Users (Location)

- Splitting Location column and analysing country.

- Most of them came from North American and European countries namely
USA, Canada, UK, Germany, and Spain.



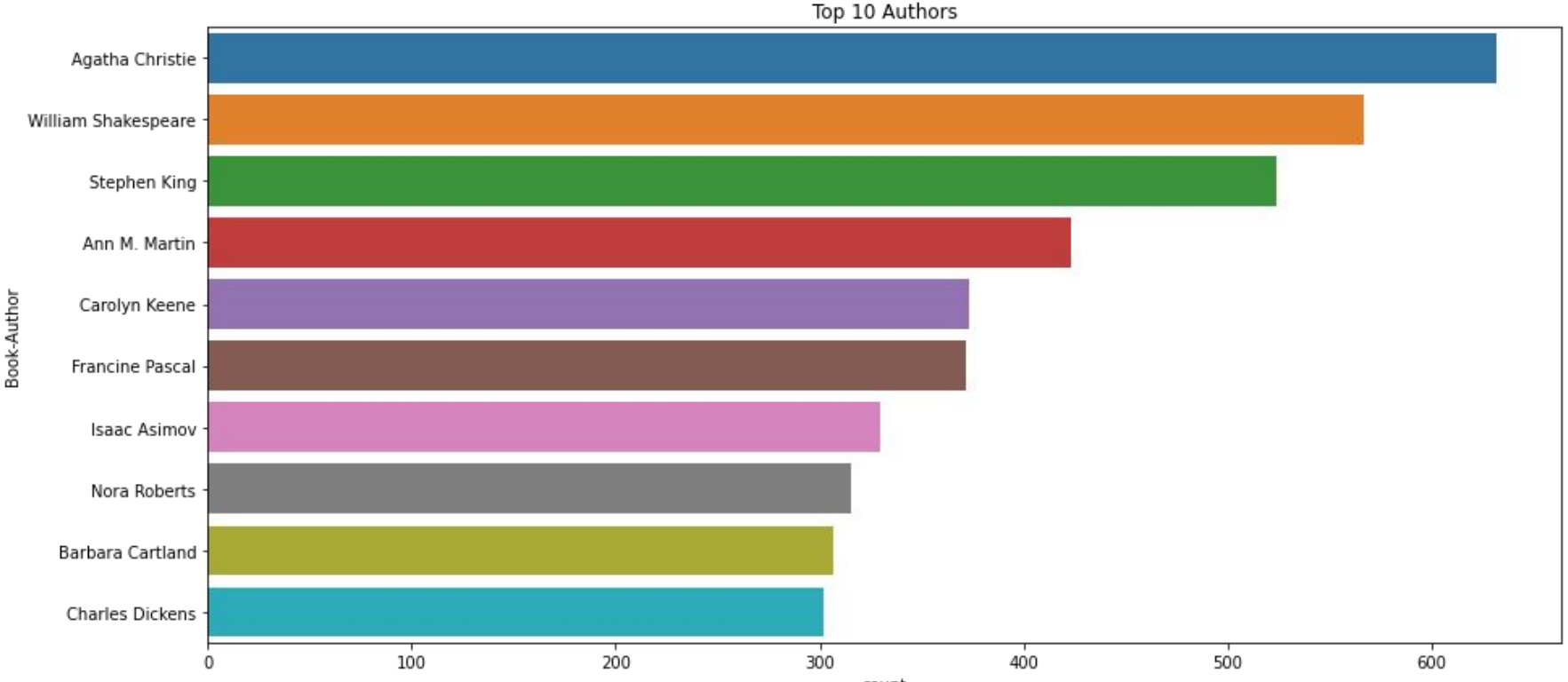
Creating new column from Location.

- Three new columns named City , state , and country Is created from the column Location.

	User-ID	Age	city	state	country
0	1	26	nyc	new york	usa
1	2	18	stockton	california	usa
2	3	26	moscow	yukon territory	russia
3	4	17	porto	v.n.gaia	portugal
4	5	26	farnborough	hants	united kingdom
...	...	...	...	...	...
278853	278854	26	portland	oregon	usa
278854	278855	50	tacoma	washington	united kingdom
278855	278856	26	brampton	ontario	canada
278856	278857	26	knoxville	tennessee	usa
278857	278858	26	dublin	n/a	ireland

Books(Authors)

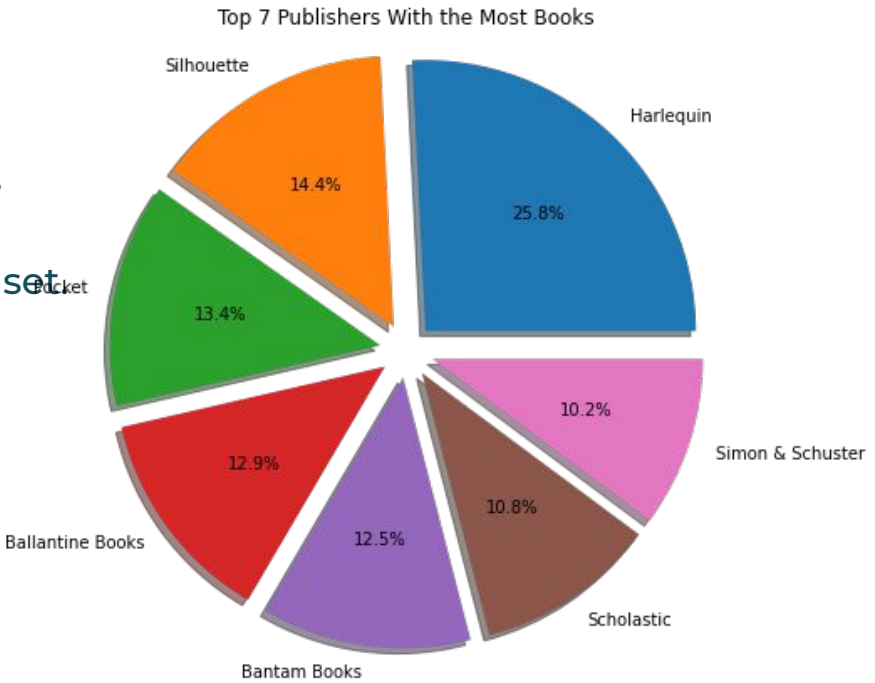
Agatha Christie wrote highest number of books in our given dataset



Books(Publishers)

- Harlequin published highest number of books in our given dataset(25.8%).

- Harlequin,Silhouette,Pocket,Ballantine Books, Bantam Books, Scholastic,Simon & Schuster are the top 7 publishers given in the book dataset.

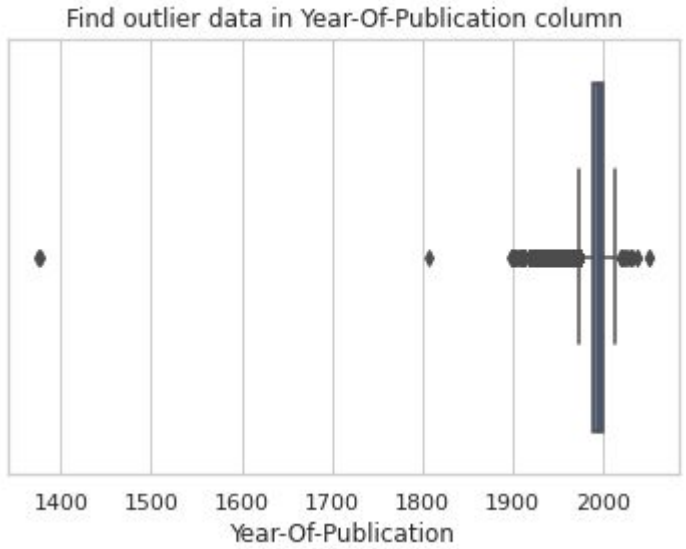


Data cleaning(Year of Publications)

- The datatype of Year of Publication column has Changed from string to integer.

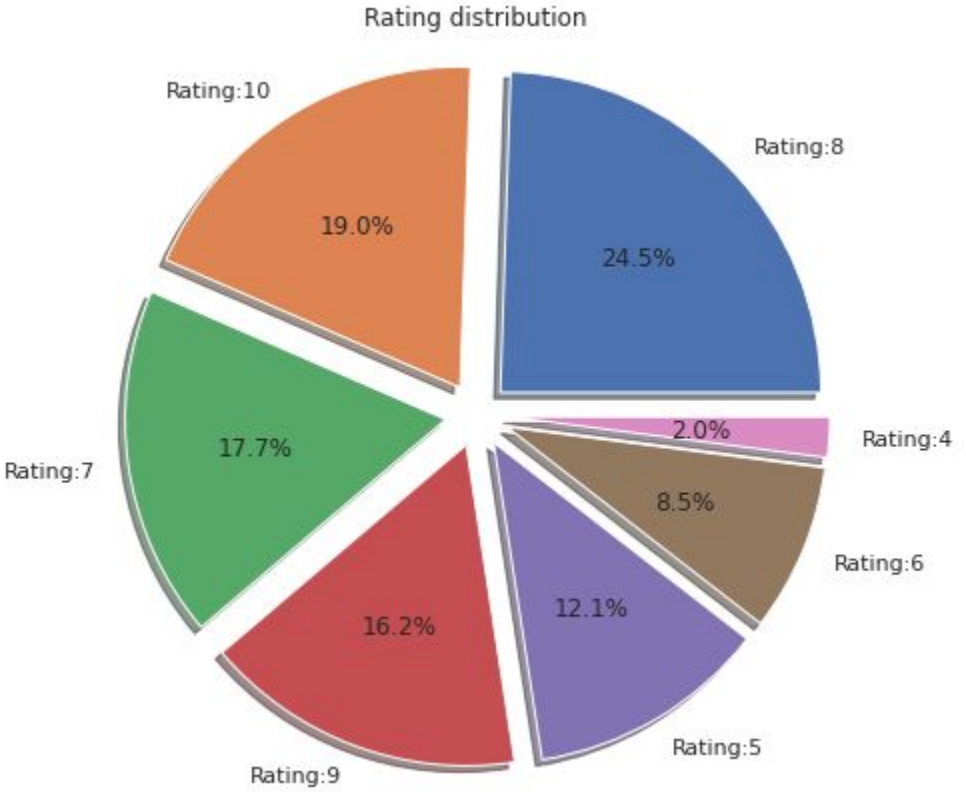
- Year of Publication has zeros as well as there is some error by including names in years.

- There are outliers in Year of Publication. The outliers are removed such that books published after 2020 has been removed.The lower bound Outliers can’t be removed as there are book publishing at that time



Ratings(Book_Rating)

- Higher ratings are more common amongst users
- If we look at the rating distribution, most of the books have high ratings with a maximum number of books being rated 8. Ratings below 5 are few in number.
- Rating 8 has been rated the highest number of times(24.5%)



Merging Datasets

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	User-ID	Book-Rating	Age	city	state	country
307568	0060977035	Still Missing	Beth Gutcheon	1996	Perennial	149084	9.0	32.0	germantown	maryland	usa
451273	0670863734	The Cook's Companion	Stephanie Alexander	1998	Viking Australia	236340	10.0	29.0	st kilda	victoria	australia
143019	0898157803	Menopaws: The Silent Meow	Martha Sacks	1995	Ten Speed Press	218552	6.0	48.0	san antonio	texas	usa
257564	076530080X	The Secret of Life	Paul J. McAuley	2001	Tor Books	83521	7.0	26.0	midvale	utah	usa
27523	1570362084	The Best of Larry King Live: The Greatest Inte...	Larry King	1995	Turner Pub	1021	10.0	26.0	rowland	north carolina	usa

- Final_df is formed by merging all datasets.
- All the Book-Ratings with 0 values are removed.
- This dataset will be used for different models.

1.)Popularity Based Recommendation

Book weighted average formula:

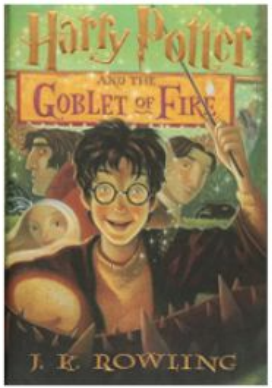
Weighted Rating(WR)=[vR/(v+m)]+[mC/(v+m)]

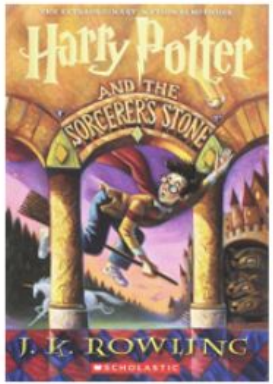
Where,

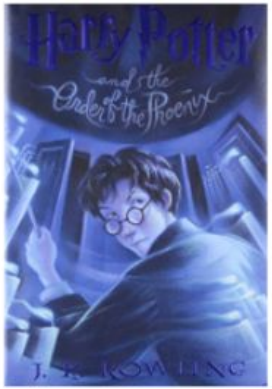
v is the number of votes for the books;
m is the minimum votes required to be listed in the chart;
R is the average rating of the book; and
C is the mean vote across the whole report.

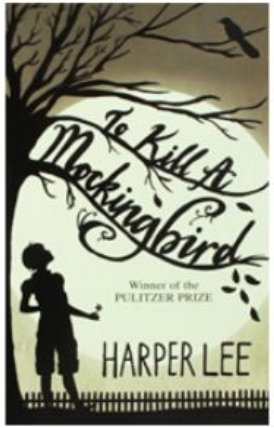
Top books(Popularity Based Recommendation)

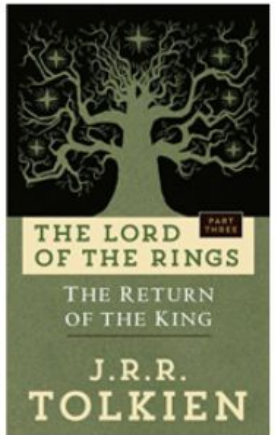
	Book-Title	Total_No_Of_Users_Rated	Avg_Rating	Score
0	Harry Potter and the Goblet of Fire (Book 4)	137	9.262774	8.741834
1	Harry Potter and the Sorcerer's Stone (Harry P...	313	8.939297	8.716469
2	Harry Potter and the Order of the Phoenix (Boo...	206	9.033981	8.700402
3	To Kill a Mockingbird	214	8.943925	8.640679
4	Harry Potter and the Prisoner of Azkaban (Book 3)	133	9.082707	8.609689
5	The Return of the King (The Lord of the Rings,...	77	9.402597	8.596515











2)Memory based collaborative Filtering

Collaborative Filtering-(item-Item based)

- The similarity between item pairs can be found in different ways. One of the most common methods is to use cosine similarity.

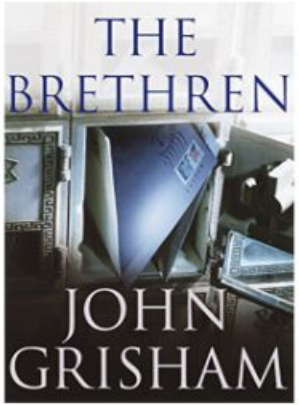
$$Similarity(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| * ||\vec{B}||}$$

- Here k nearest neighbors is used to find similar books.Which generates predictions based on the ratings of similar products.

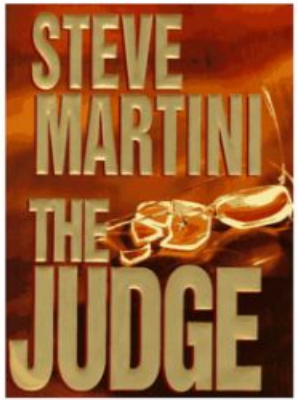
Item based Recommendation('A Painted House')

	Book Recommendations	Similarity Score
0	The Brethren	0.904865
1	The Judge	0.916810
2	The Firm	0.920558
3	The Summons	0.922756
4	The Pelican Brief	0.922822

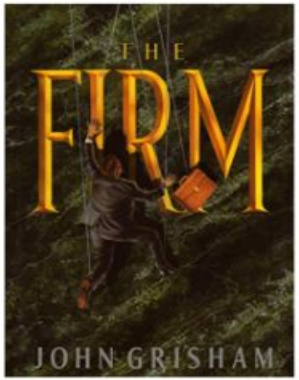
Book Recommendations



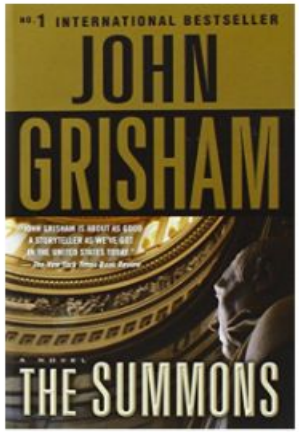
Rating: 7.4



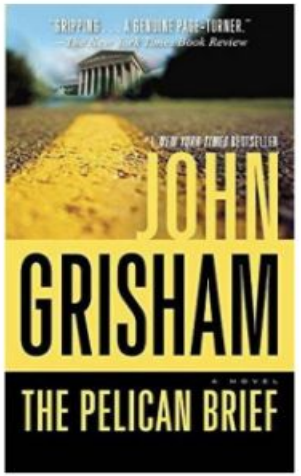
Rating: 7.1



Rating: 7.8



Rating: 7.3



Rating: 7.7

2)Memory based collaborative Filtering

Collaborative Filtering-(user-user based)

- User-Based Collaborative Filtering is a technique used to predict the items that a user might like on the basis of ratings given to that item by the other users who have similar taste with that of the target user.

- For a User-User CF algorithm, similarity, sim_{xy} between the users x and y who have both rated the same items is calculated first. To calculate this similarity different metrics are used. We will be using correlation-based similarity metrics to compute the similarity between user x and user y.

User based Recommendation(User_id=153662)

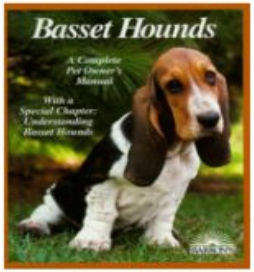
	Book Recommendations
0	Dear Dad: Letters from an Adult Child
1	Coraline
2	Wizard and Glass (The Dark Tower, Book 4)
3	Basset Hounds: Everything About Purchase, Care...
4	Been There Should've Done That: 505 Tips for M...
5	The Stand: Complete and Uncut

Book Recommendations

Average Rating: nan

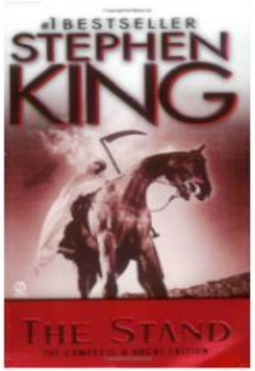
Average Rating: 8.2

Average Rating: 8.3



Average Rating: nan

Average Rating: nan



Average Rating: 8.5

3.)Model based collaborative filtering

- It's clear that for the given dataset much better results can be obtained with SVD approach - both in terms of accuracy and fit time.

SVD

```
test_rmse     1.619140
test_mae      1.253389
fit_time     13.059181
test_time     0.852580
dtype: float64
```

NMF

```
test_rmse     2.479644
test_mae      2.085980
fit_time     17.656860
test_time     0.660090
dtype: float64
```

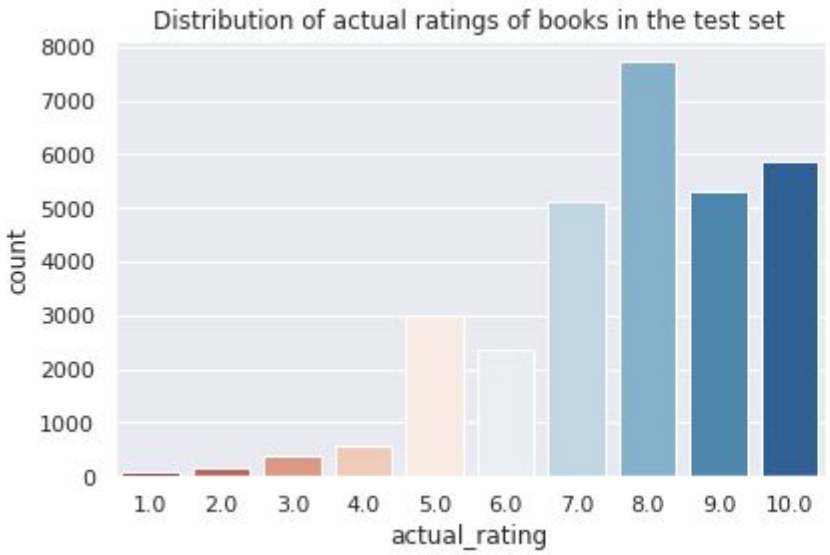
SVD Model Results

Svd model results

	user_id	book_title	actual_rating	pred_rating	impossible	pred_rating_round	abs_err
0	269087	My Dream of You	9.0	7.450325	False	7.0	1.549675
1	107244	B Is for Burglar (Kinsey Millhone Mysteries (P...	10.0	8.985337	False	9.0	1.014663
2	14051	Airframe	9.0	7.493452	False	7.0	1.506548
3	221114	Love and Marriage	7.0	7.698944	False	8.0	0.698944
4	88499	Tuesdays with Morrie: An Old Man, a Young Man,...	9.0	8.296064	False	8.0	0.703936
5	84897	Bel Canto: A Novel	8.0	7.085035	False	7.0	0.914965

SVD Model Results

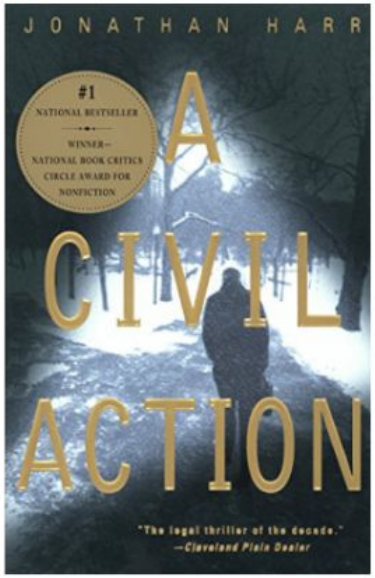
SVD Model Results



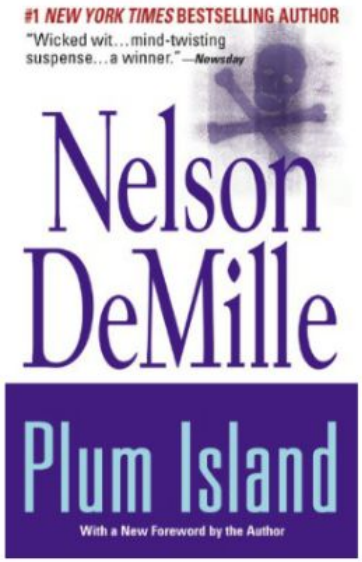
Recommending books to user-id(23511)

	Book Recommendations	Actual_Rating_by_user	Predicted_rating
0	A Civil Action	10.0	8.748597
1	Plum Island	9.0	8.608478
2	Skipping Christmas	10.0	8.447469

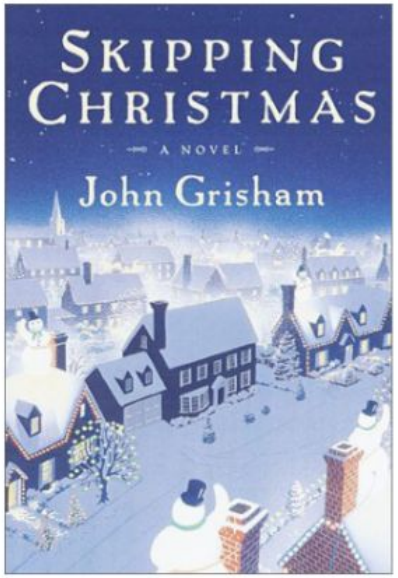
Book Recommendations



Average Rating: 7.7



Average Rating: 7.7

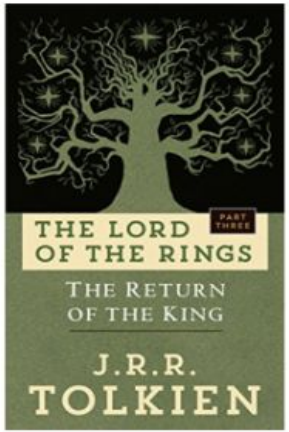


Average Rating: 7.5

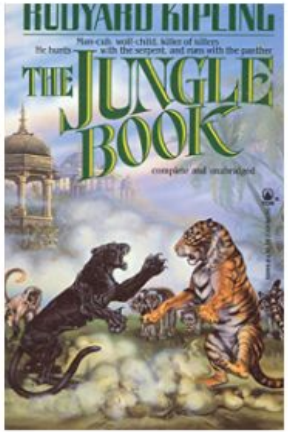
Recommending books to user-id(137118)

	Book Recommendations	Actual_Rating_by_user	Predicted_rating
0	The Return of the King (The Lord of the Rings,...	5.0	7.530680
1	Jungle Book	5.0	6.960180
2	I Know Why the Caged Bird Sings	7.0	6.917741
3	Acceptable Risk	7.0	6.201210
4	The Devil's Code	5.0	6.187194

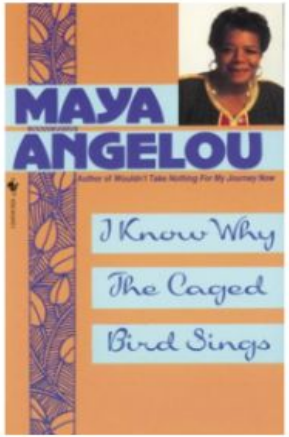
Book Recommendations



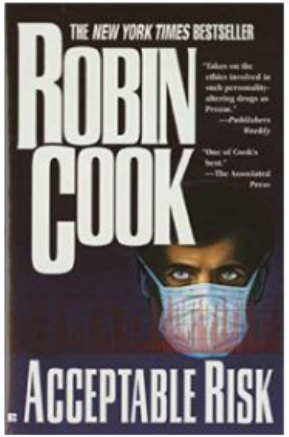
Average Rating: 9.2



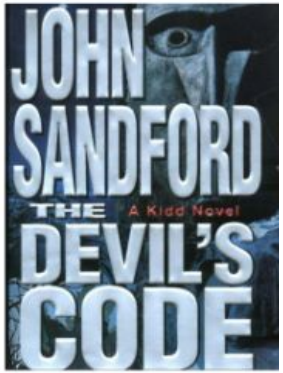
Average Rating: 8.3



Average Rating: 8.1



Average Rating: 7.2



Average Rating: 7.3

Evaluation Of model

Model Results

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	User-ID
14	25	34	149	0.167785	0.228188	98391
51	29	40	98	0.295918	0.408163	153662
118	13	20	84	0.154762	0.238095	16795
3	11	18	72	0.152778	0.250000	95359
15	22	32	72	0.305556	0.444444	114368
166	9	11	54	0.166667	0.203704	235105
114	25	30	52	0.480769	0.576923	123883
34	7	11	49	0.142857	0.224490	60244
61	8	11	48	0.166667	0.229167	204864
38	9	18	40	0.225000	0.450000	78973

Conclusion

- 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.

- For modelling, it was observed that for model based collaborative filtering SVD technique worked way better than NMF with lower Mean Absolute Error (MAE) .

Challenges

- Handling of sparsity was a major challenge as well since the user interactions were not present for the majority of the books.
- Understanding the metric for evaluation was a challenge as well.
- Since the data consisted of text data, data cleaning was a major challenge in features like Location etc..
- Decision making on missing value imputations and outlier treatment was quite challenging as well.

Future Scope

- Given more information regarding the books dataset, namely features like Genre, Description etc, we could implement a content-filtering based recommendation system and compare the results with the existing collaborative-filtering based system.

- We would like to explore various clustering approaches for clustering the users based on Age, Location etc., and then implement voting algorithms to recommend items to the user depending on the cluster into which it belongs.

Thank You
Q & A