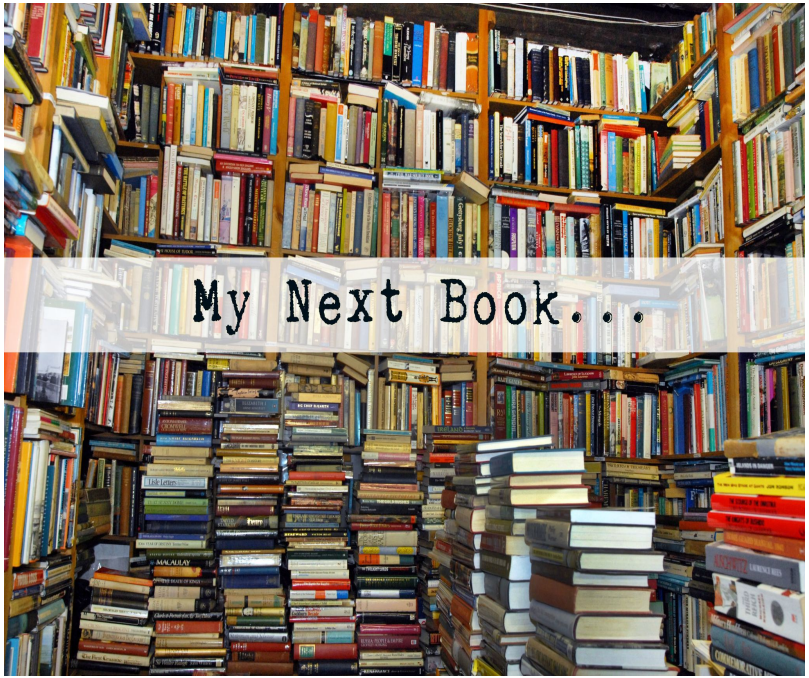
Capstone Project

BOOK RECOMMENDATION SYSTEM

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.

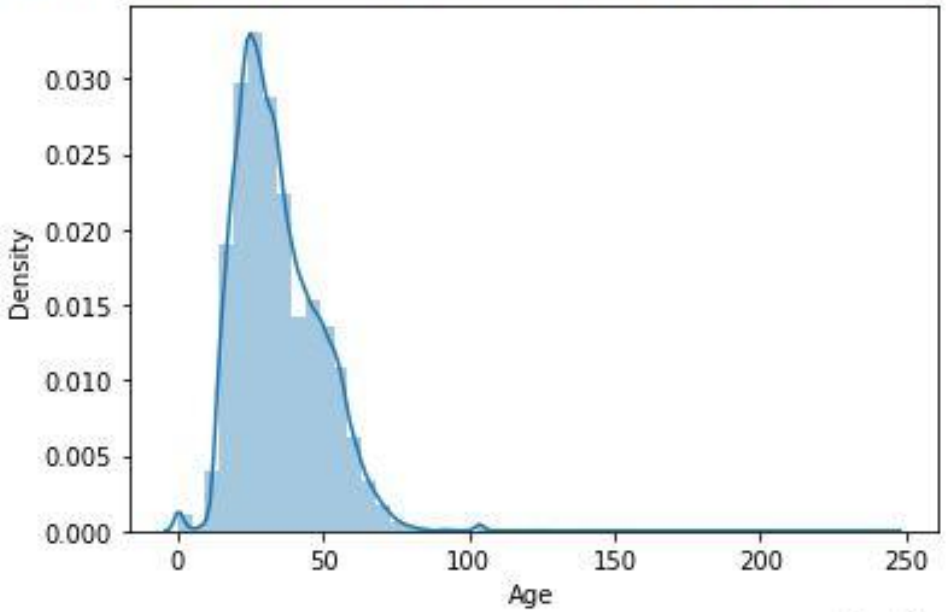
- User-ID
- ISBN
- Book-Rating
- Shape of Dataset - (1149780, 3)

Observations from Users_df (Age)

- The Age range given here is from 0 To 250.
- Outliers in the Age column.

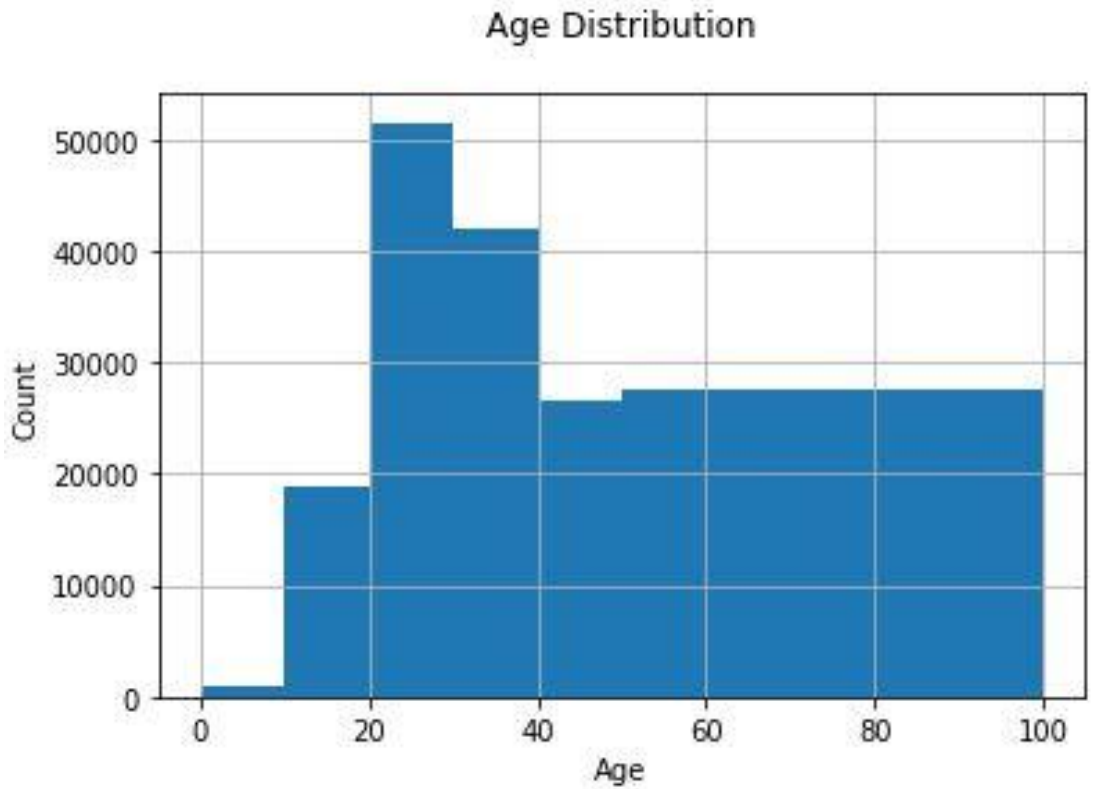
```
1 sns.distplot(users.Age)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5a11ac00d0>
```



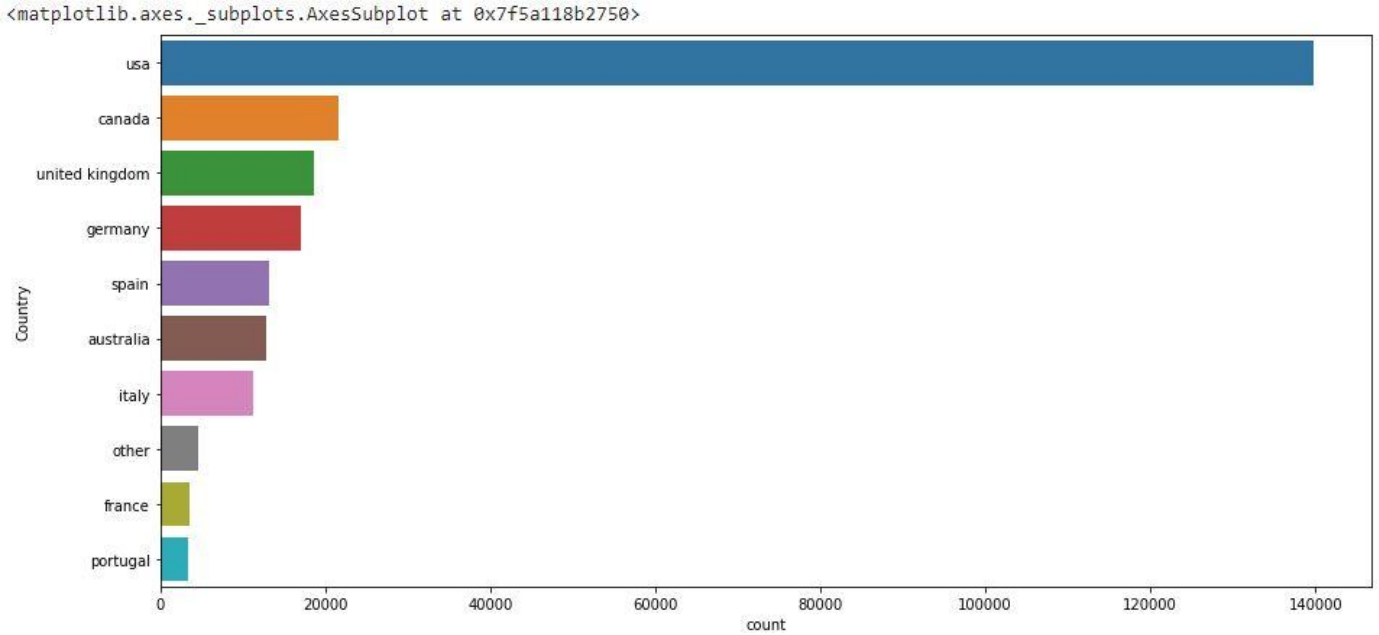
Observations from Users_df (Age)

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



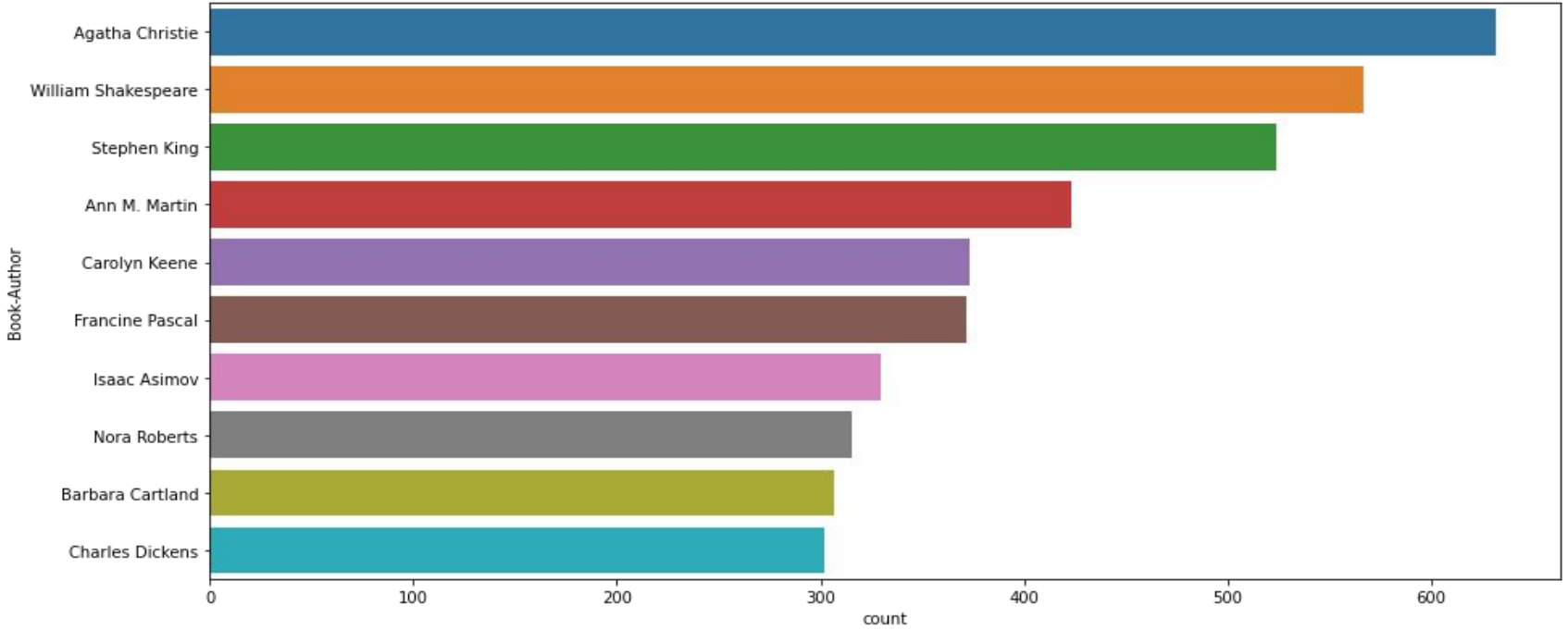
Observations from Users_df (Location)

- Splitting Location column and analysing country.
- Most active readers are from USA.



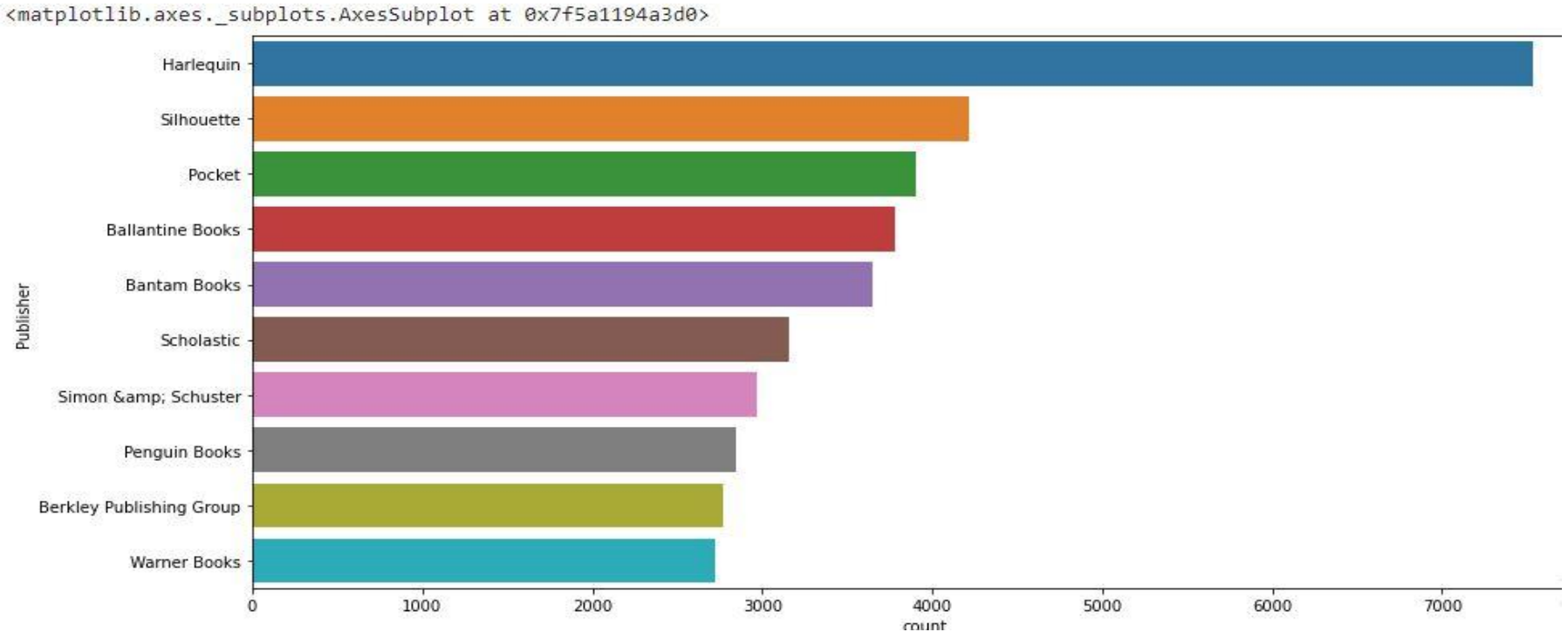
Observations from Book_df (Authors)

Agatha Christie wrote highest number of books in our given dataset



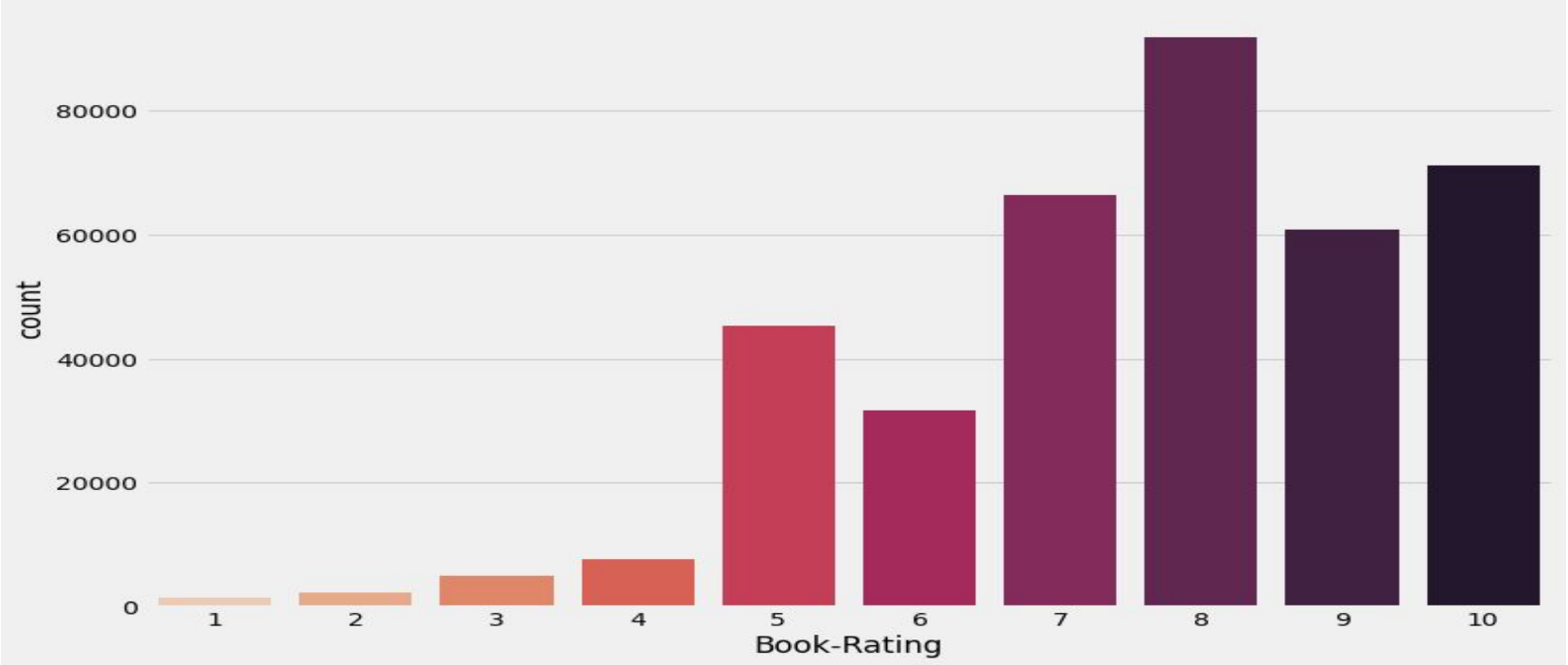
Observations from Book_df (Publishers)

Harlequin published highest number of books in our given dataset



Observations from Ratings_df (Book_Rating)

- Higher ratings are more common amongst users
- Rating 8 has been rated the highest number of times



Data Cleaning

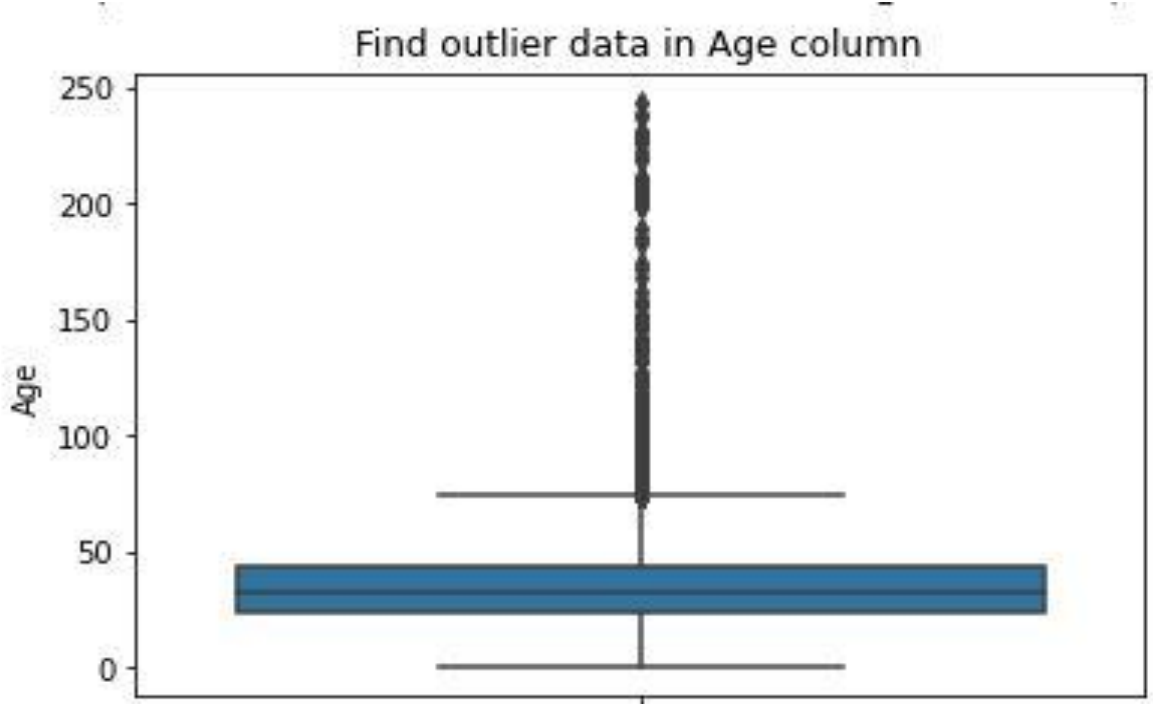
1. Null Value Imputation:

Age column has 40% missing values

	index	Missing Values	% of Total Values	Data_type
0	Age	110762	39.72	float64
1	User-ID	0	0.00	int64
2	Location	0	0.00	object

Imputing missing values

- Outliers in Age column
- Age has positive Skewness (right tail) so we can use median to fill Nan values,



Data Cleaning

1. Null Value Imputation:

```
books_df.isnull().sum()
```

```
ISBN                 0
Book-Title           0
Book-Author          1
Year-Of-Publication  0
Publisher            2
Image-URL-S          0
Image-URL-M          0
Image-URL-L          3
dtype: int64
```

Replacing strings by int values

	ISBN	Book-Title	Book-Author	Year-Of-Publication	
209538	078946697X	DK Readers: Creating the X-Men, How It All Beg...	2000	DK Publishing Inc	ht
221678	0789466953	DK Readers: Creating the X-Men, How Comic Book...	2000	DK Publishing Inc	h

Different Models

1.)Popularity Based Recommendation

Book weighted average formula:

$$\textbf{Weighted Rating(WR)}=[vR/(v+m)]+[mC/(v+m)]$$

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.

Different Models

	Book-Title	Total_No_Of_Users_Rated	Avg_Rating	Score
0	Harry Potter and the Goblet of Fire (Book 4)	137	9.262774	8.741835
1	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	313	8.939297	8.716469
2	Harry Potter and the Order of the Phoenix (Book 5)	206	9.033981	8.700403
3	To Kill a Mockingbird	214	8.943925	8.640679
4	Harry Potter and the Prisoner of Azkaban (Book 3)	133	9.082707	8.609690
5	The Return of the King (The Lord of the Rings, Part 3)	77	9.402597	8.596517
6	Harry Potter and the Prisoner of Azkaban (Book 3)	141	9.035461	8.595653
7	Harry Potter and the Sorcerer's Stone (Book 1)	119	8.983193	8.508791
8	Harry Potter and the Chamber of Secrets (Book 2)	189	8.783069	8.490549
9	Harry Potter and the Chamber of Secrets (Book 2)	126	8.920635	8.484783
10	The Two Towers (The Lord of the Rings, Part 2)	83	9.120482	8.470128
11	Harry Potter and the Goblet of Fire (Book 4)	110	8.954545	8.466143
12	The Fellowship of the Ring (The Lord of the Rings, Part 1)	131	8.839695	8.441584
13	The Hobbit : The Enchanting Prelude to The Lord of the Rings	161	8.739130	8.422706
14	Ender's Game (Ender Wiggins Saga (Paperback))	117	8.837607	8.409441
15	Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	200	8.615000	8.375412
16	Charlotte's Web (Trophy Newbery)	68	9.073529	8.372037
17	Dune (Remembering Tomorrow)	75	8.973333	8.353301
18	A Prayer for Owen Meany	181	8.607735	8.351465
19	Fahrenheit 451	164	8.628049	8.346969

Different Models

2.)Model based collaborative filtering

SVD

```
test_rmse    1.602152
test_mae     1.239638
fit_time     5.437686
test_time    0.472132
dtype: float64
```

NMF

```
test_rmse    2.626532
test_mae     2.242070
fit_time     8.057059
test_time    0.546524
dtype: float64
```

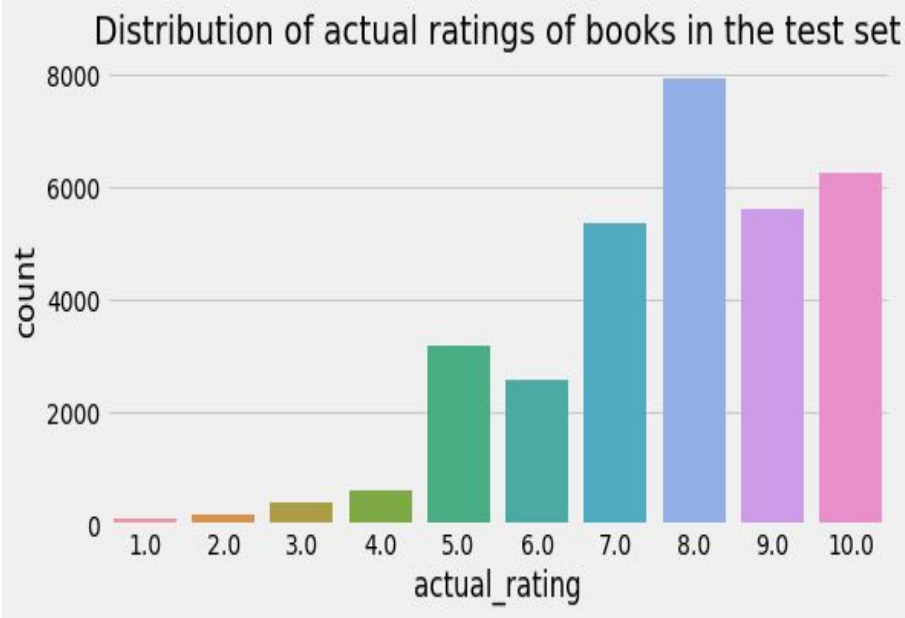
Different Models

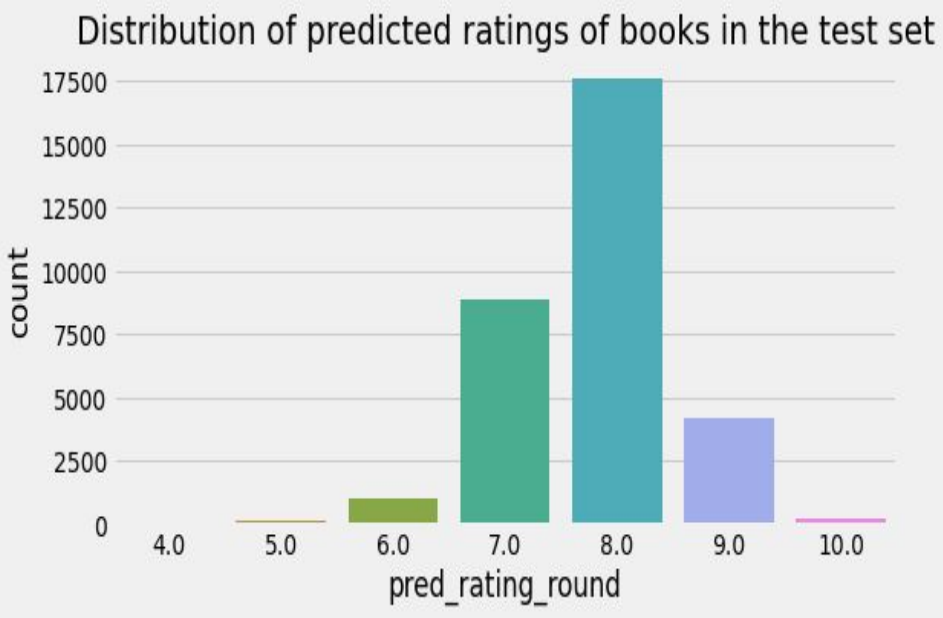
SVD Model Results

	user_id	isbn	actual_rating	pred_rating	impossible	pred_rating_round	abs_err
15594	62862	0385335482	8.0	7.978811	False	8.0	0.021189
30626	193938	0385497288	8.0	7.882566	False	8.0	0.117434
27451	234401	0812540026	8.0	7.316338	False	7.0	0.683662
14130	89602	0060987529	8.0	6.649098	False	7.0	1.350902
18074	86189	0312186886	10.0	7.303280	False	7.0	2.696720

Different Models

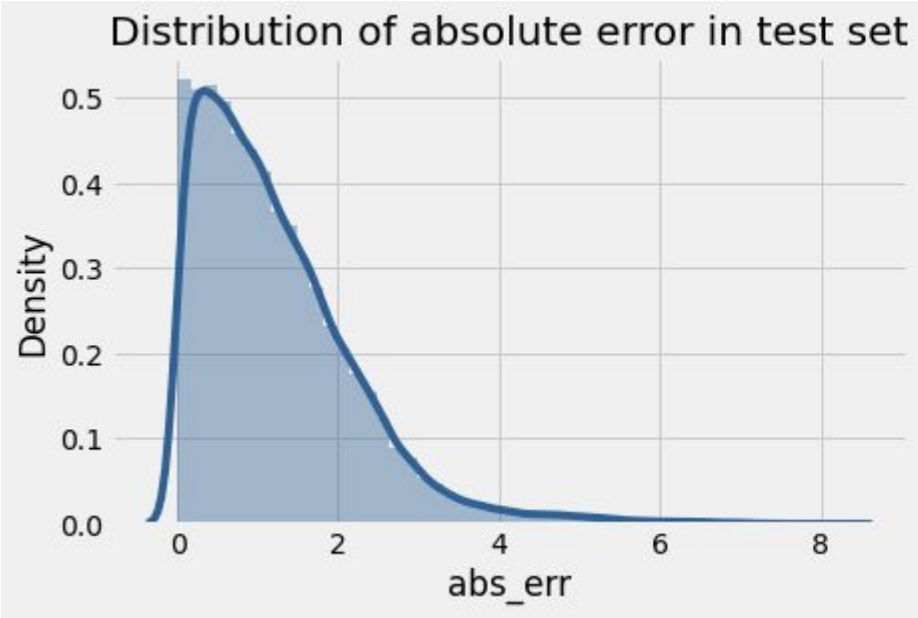
SVD Model Results

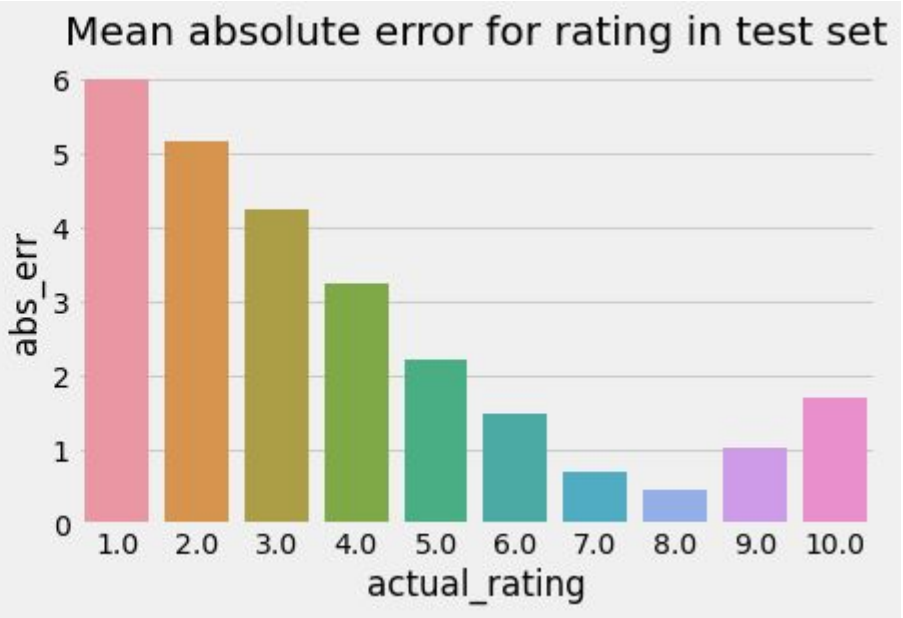




Different Models

SVD Model Results

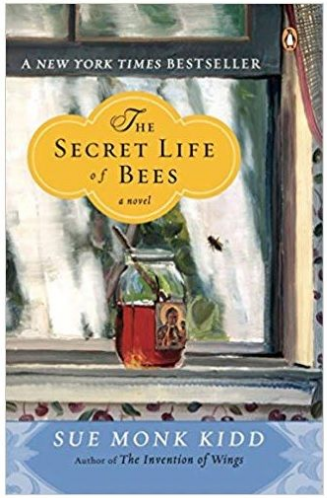


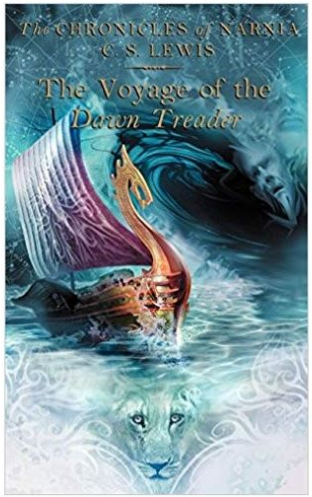


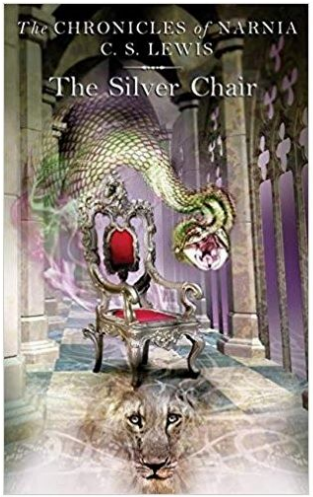
Different Models

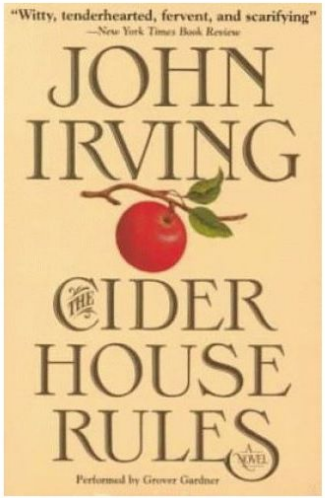
User-ID - 193458

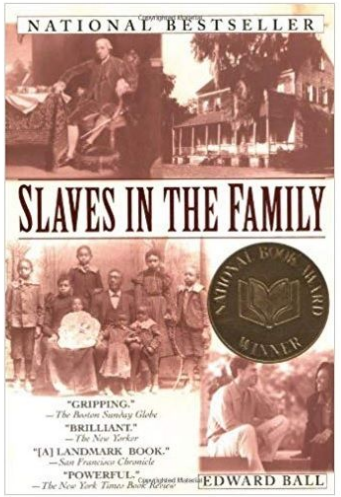
Test set: predicted top rated books





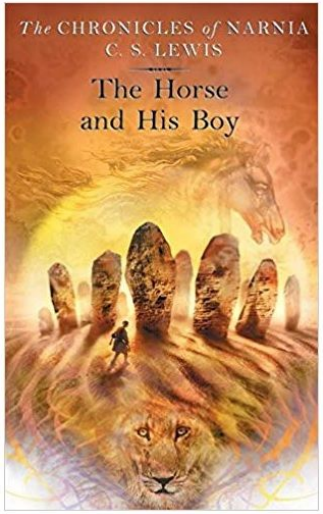


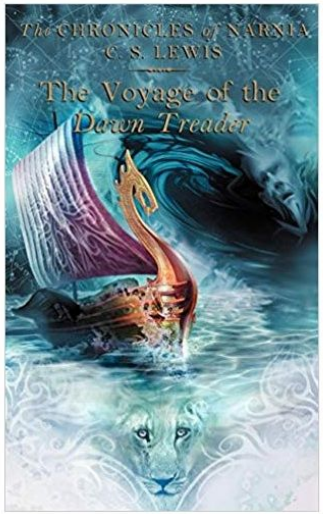


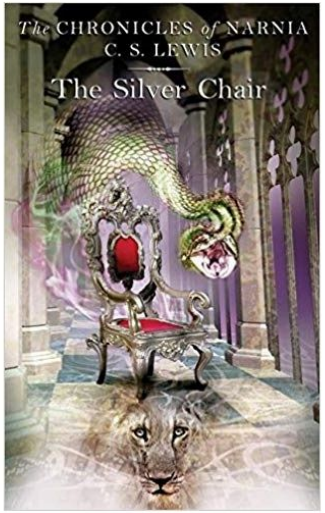


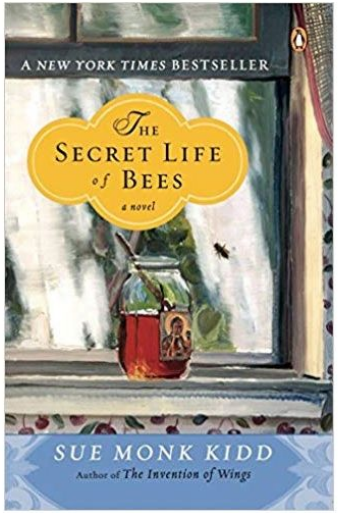
Different Models

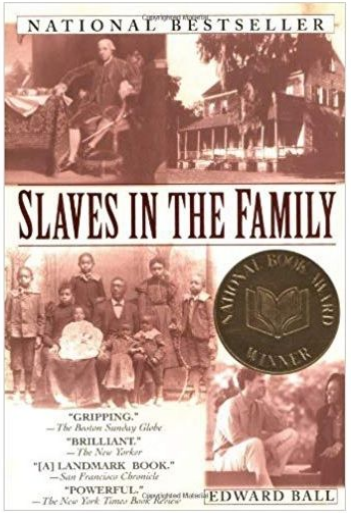
Test set: actual top rated books











Collaborative Filtering-(Item-Item based)

3.)Collaborative Filtering-(Item-Item based)

- Cosine Similarity
- Nearest Neighbour

```
Recommendations for Angels &amp;amp; Demons:

1: The Da Vinci Code, with distance of 0.8275555141289059:
2: Digital Fortress : A Thriller, with distance of 0.83781217691282:
3: Deception Point, with distance of 0.8422605379839627:
4: Prey: A Novel, with distance of 0.9216969275206289:
5: The Cat Who Knew a Cardinal, with distance of 0.9280814355076102:
```

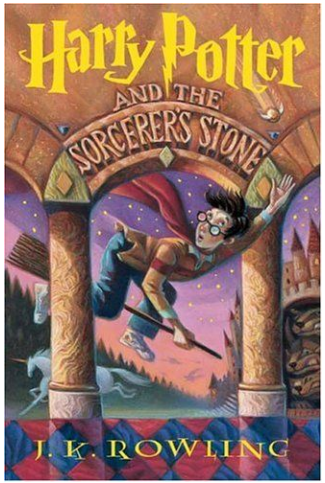
Different Models

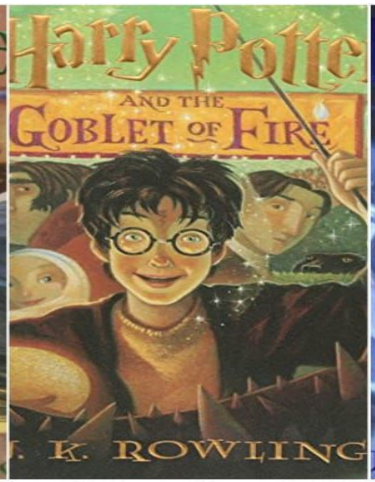
SVD and Correlation

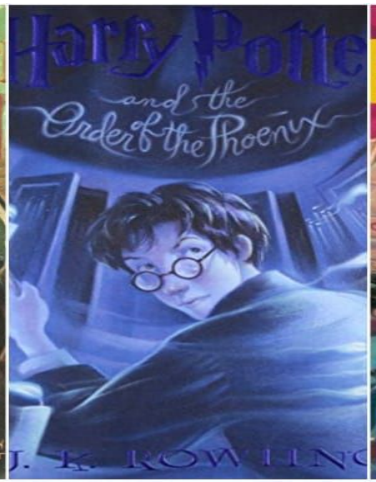
Recommendations for Harry Potter and the Sorcerer's Stone (Book 1)

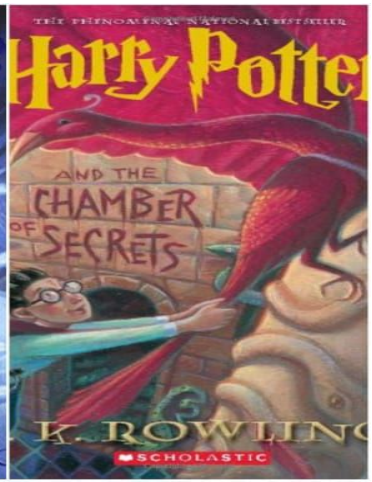
Input

Output









Different Models

4.)Collaborative Filtering-(User-Item based)

```
Enter User ID from above list for book recommendation  69078
Recommendation for User-ID =  69078
         ISBN                                          Book-Title  recStrength
0  0446310786                                To Kill a Mockingbird        0.842
1  0345370775                                        Jurassic Park        0.802
2  0312966970          Four To Score (A Stephanie Plum Novel)        0.675
3  0316769487                              The Catcher in the Rye        0.673
4  0345361792                             A Prayer for Owen Meany        0.646
5  0440214041                                   The Pelican Brief        0.621
6  044021145X                                           The Firm        0.617
7  0440211727                                      A Time to Kill        0.617
8  0060928336  Divine Secrets of the Ya-Ya Sisterhood: A Novel        0.606
9  0312924585                               Silence of the Lambs        0.600
```

Different Models

Model Results

Global metrics:
{'modelName': 'Collaborative Filtering', 'recall@5': 0.2357298474945534, 'recall@10': 0.3057371096586783}

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	User-ID
10	252	343	1389	0.181	0.247	11676
31	189	245	1138	0.166	0.215	98391
45	17	30	380	0.045	0.079	189835
30	83	104	369	0.225	0.282	153662
70	29	33	236	0.123	0.140	23902
7	30	49	204	0.147	0.240	235105
47	22	32	203	0.108	0.158	76499
50	23	35	193	0.119	0.181	171118
42	55	68	192	0.286	0.354	16795
43	23	31	188	0.122	0.165	248718

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