Capstone Project
Book recommendation system

Individual Contributor
Mohammed Javeed

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.

It is our main objective to develop an algorithm that helps users find relevant books based on their interests and popularity.

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.

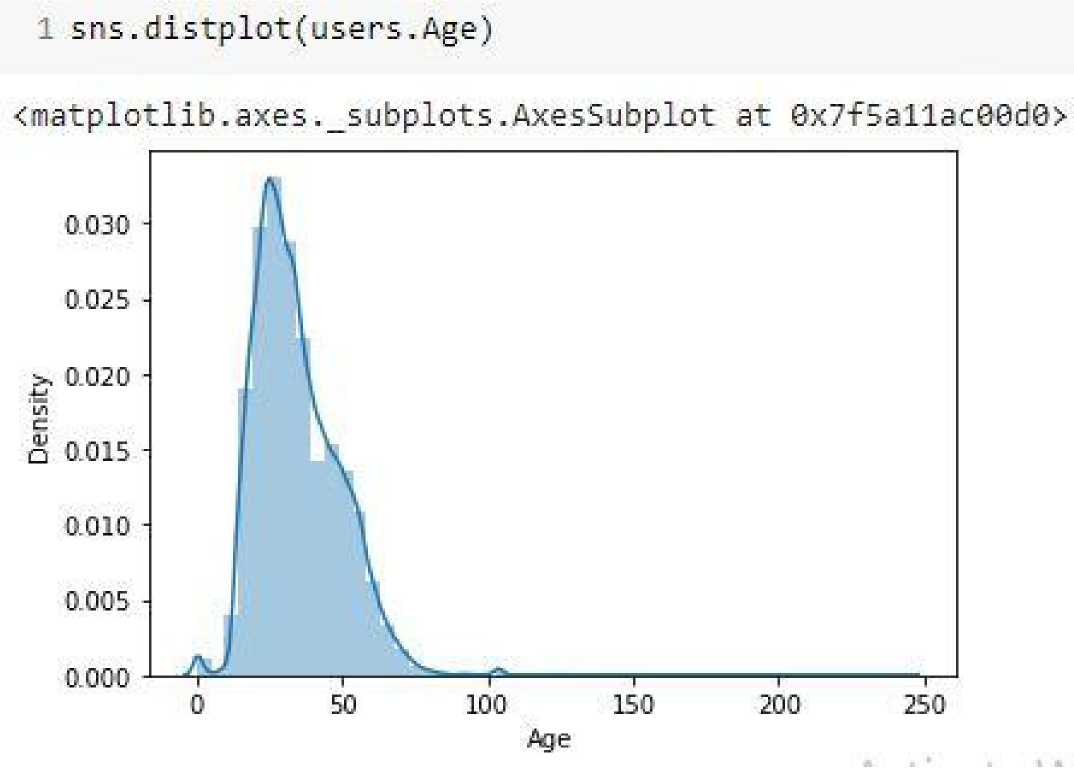
- ISBN (unique for each book)
- Image-URL-S
- Book-Title
- Image-URL-M
- Book-Author
- Image-URL-L
- Year-Of-Publication
- Publisher
- Shape of Dataset - (271360, 8)

Ratings_dataset.

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

Observations from Users_df (Age)

- The Age range given here is from 0 To 250.
- The Age column contains outliers.

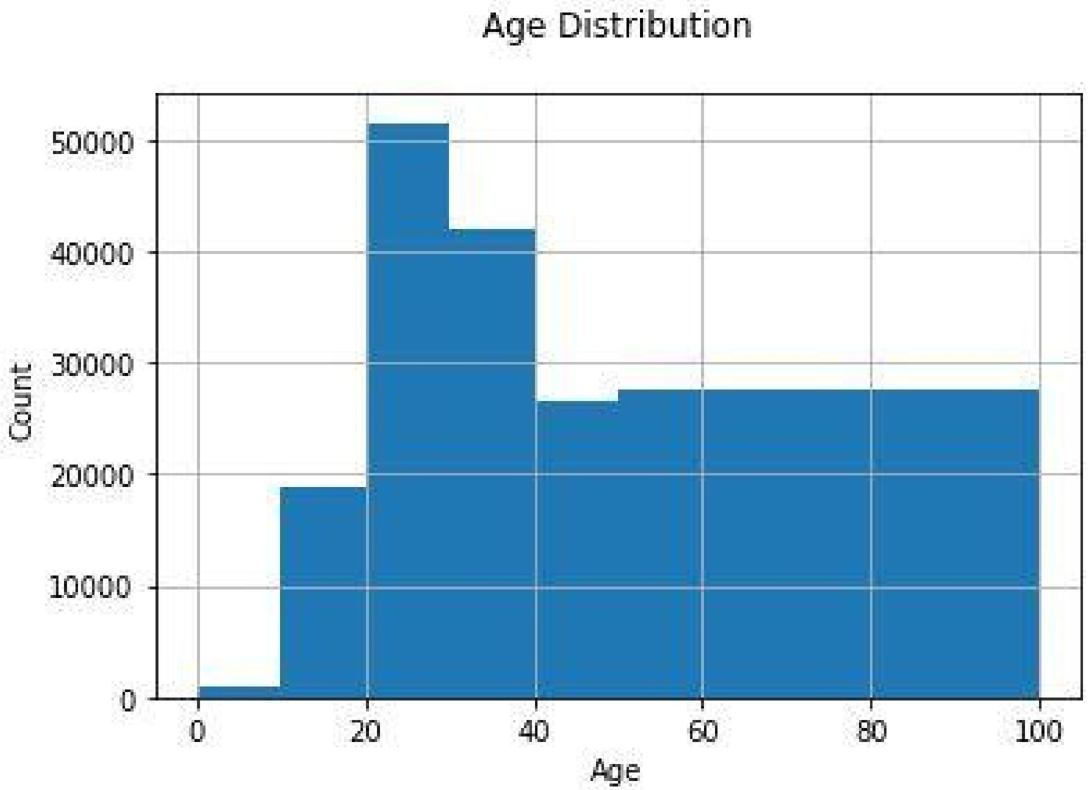


```
sns.distplot(users.Age)
```

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

Observations from Users_df (Age)

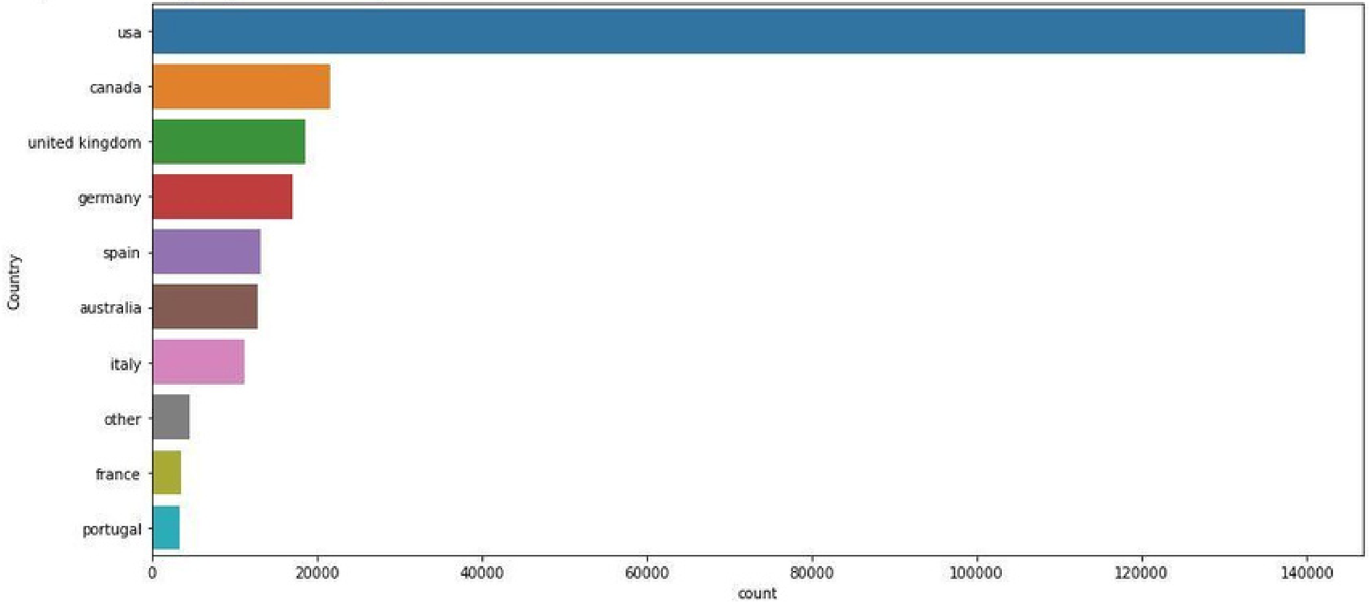
- The Age range distribution is right skewed
- The majority of active readers are in their twenties group of 20- 40



Observations from Users_df (Location)

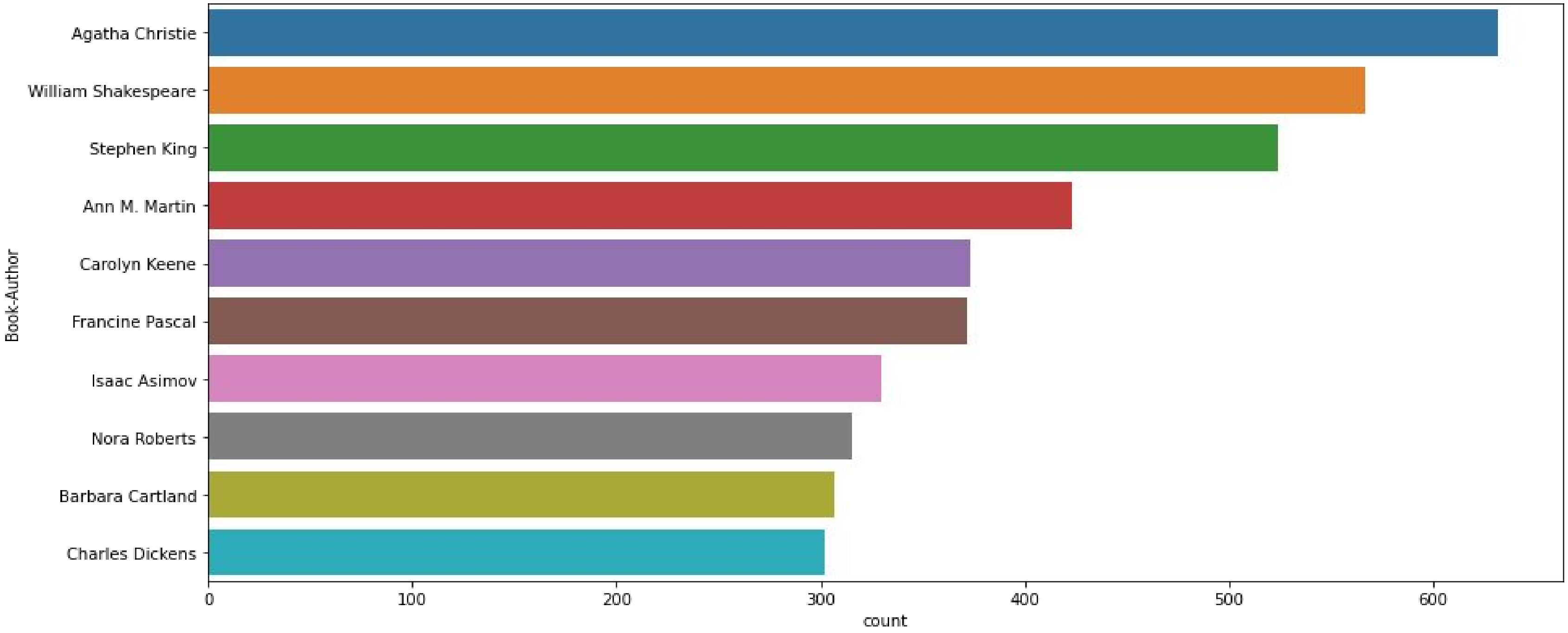
- Analyzing country by splitting Location column.
- The majority of active readers come from the United States.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5a118b2750>
```



Observations from Book_df (Authors)

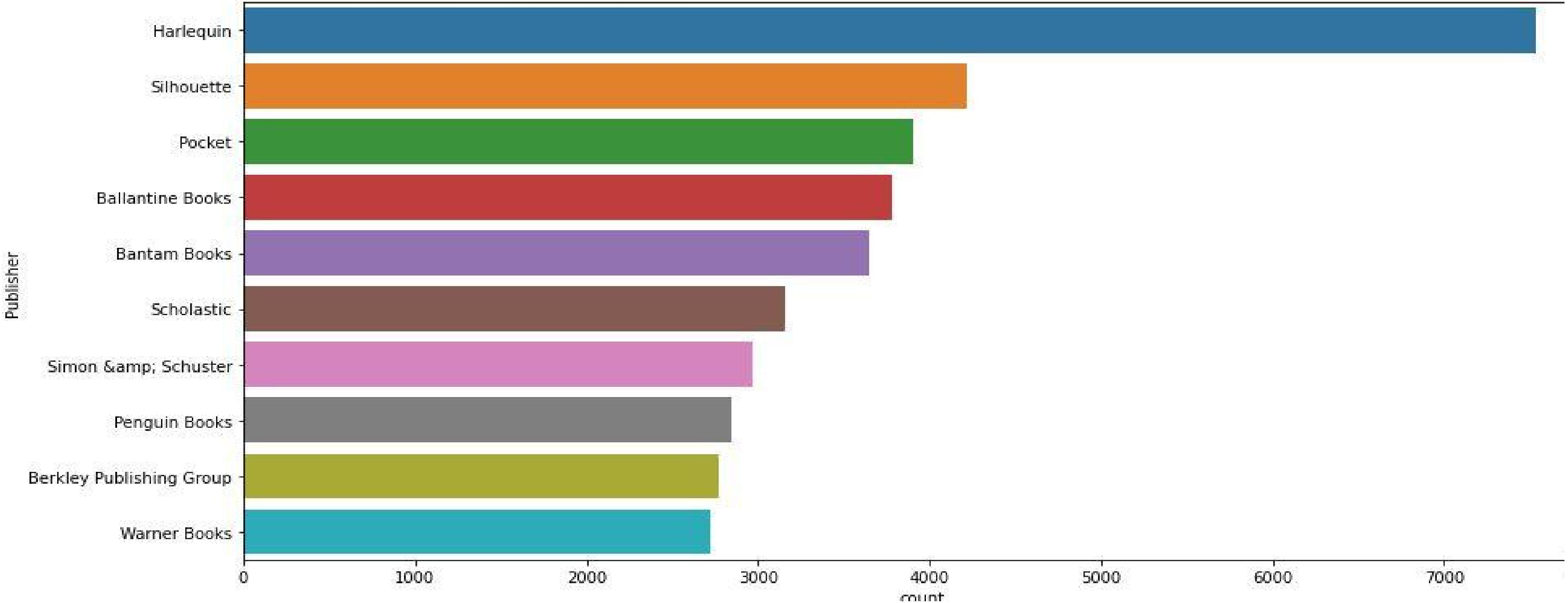
According to our dataset, Agatha Christie was the author of the most books



Observations from Book_df (Publishers)

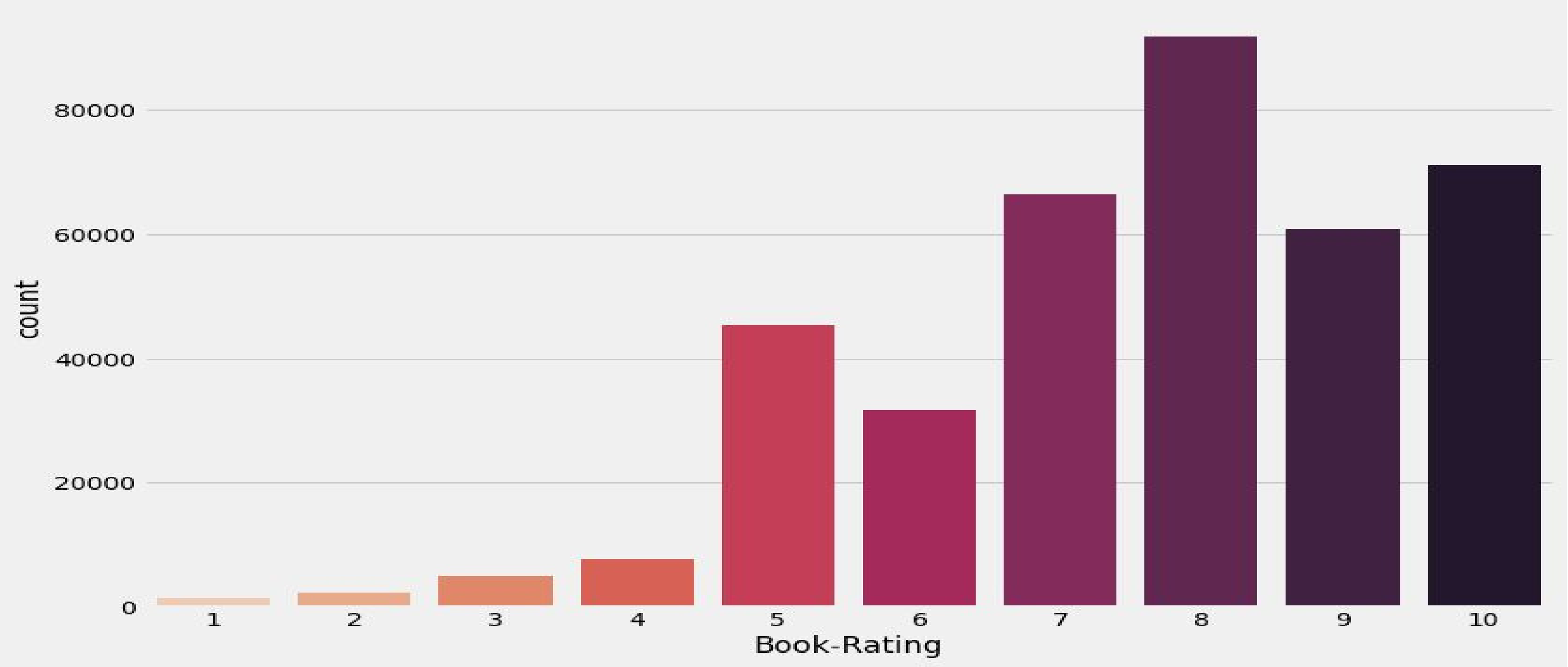
As far as the number of books published in our dataset goes, Harlequin is the top publisher

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5a1194a3d0>
```



Observations from Ratings_df (Book_Rating)

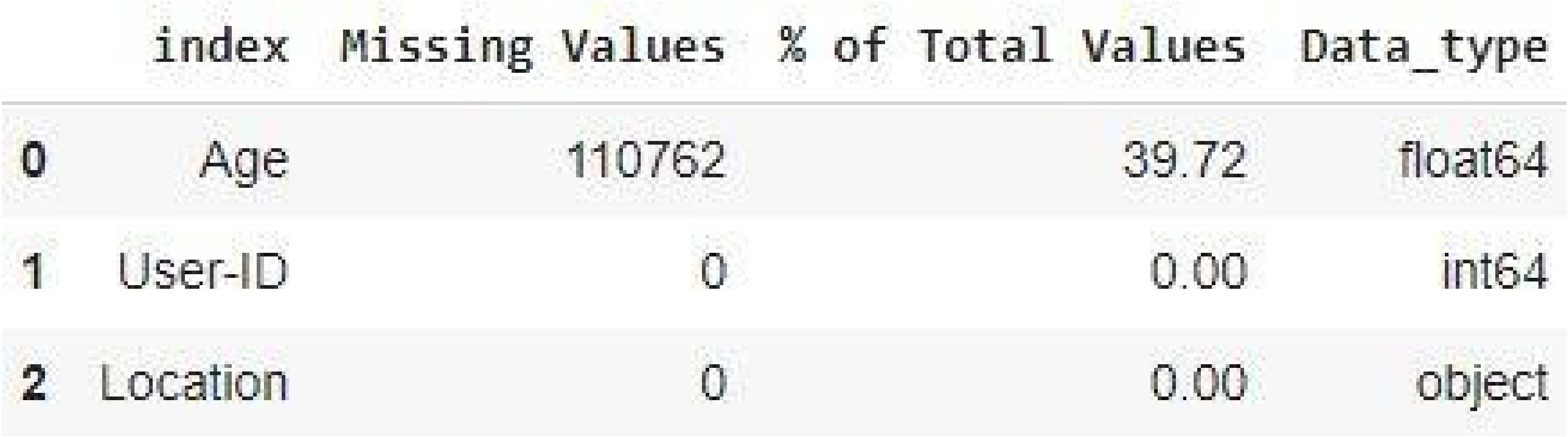
- Higher ratings are more common amongst users
- Rating 8 has been rated the highest number of times



Data Cleaning

1.Null Value Imputation:

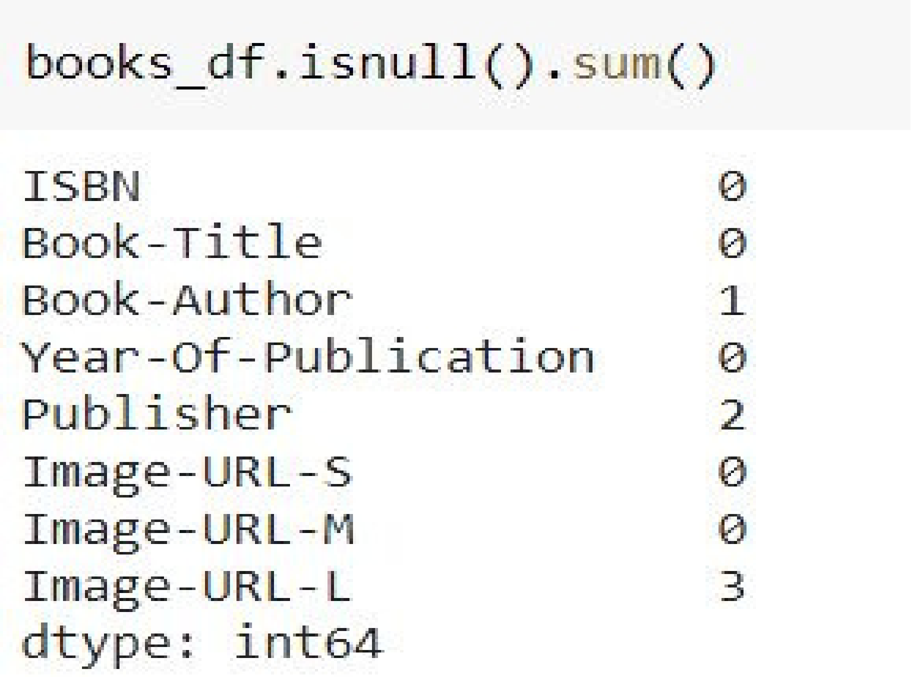
Age column has 40% missing values



	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

Data Cleaning

1. **Null Value Imputation:**

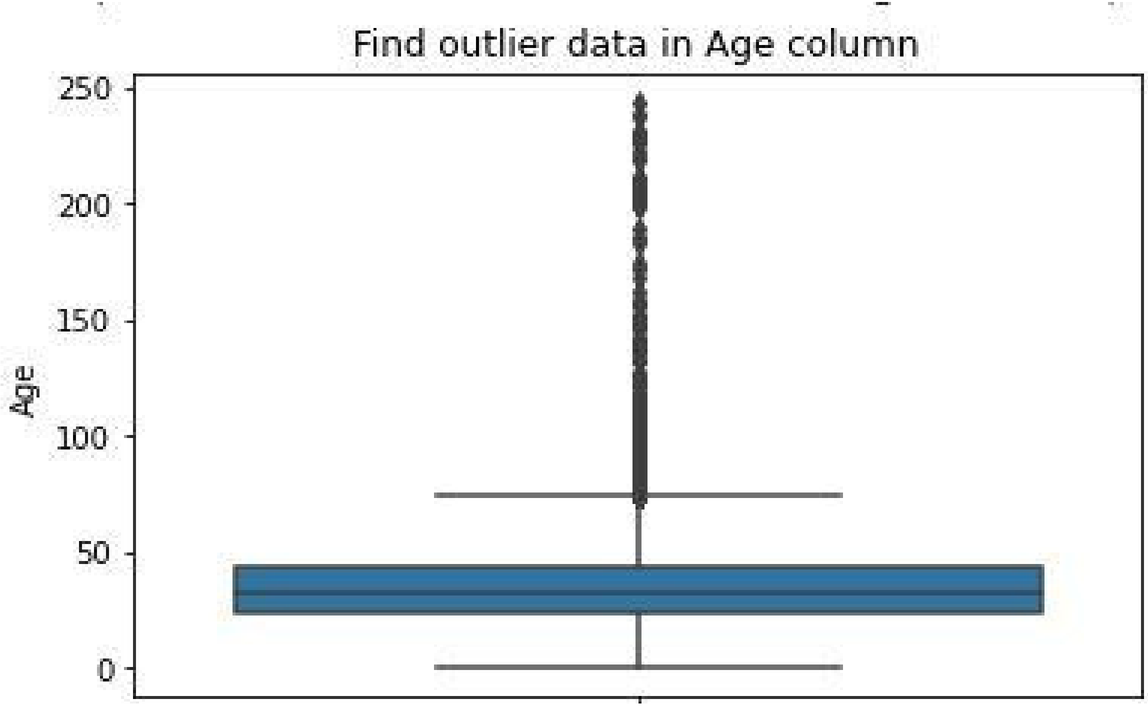


```
books_df.isnull().sum()
```

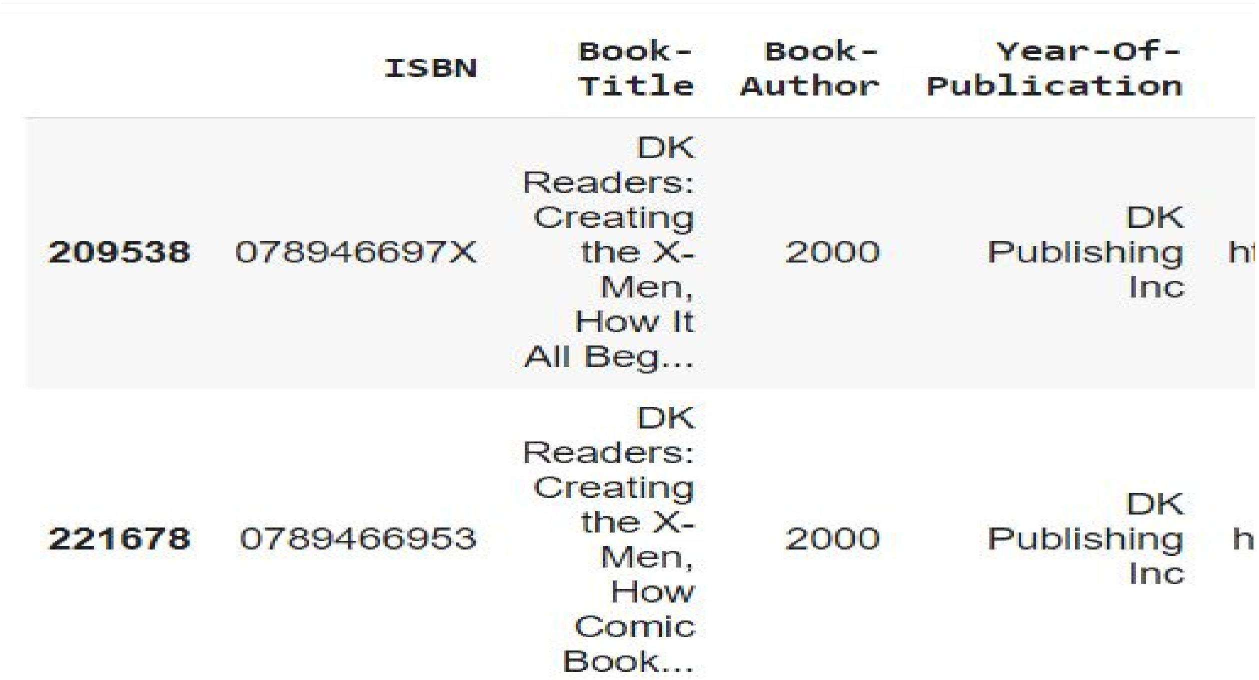
```
ISBN                   0
Book-Title             0
Book-Author            1
Year-Of-Publication    0
Publisher              2
Image-URL-S            0
Image-URL-M            0
Image-URL-L            3
dtype: int64
```

Imputing missing values

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



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

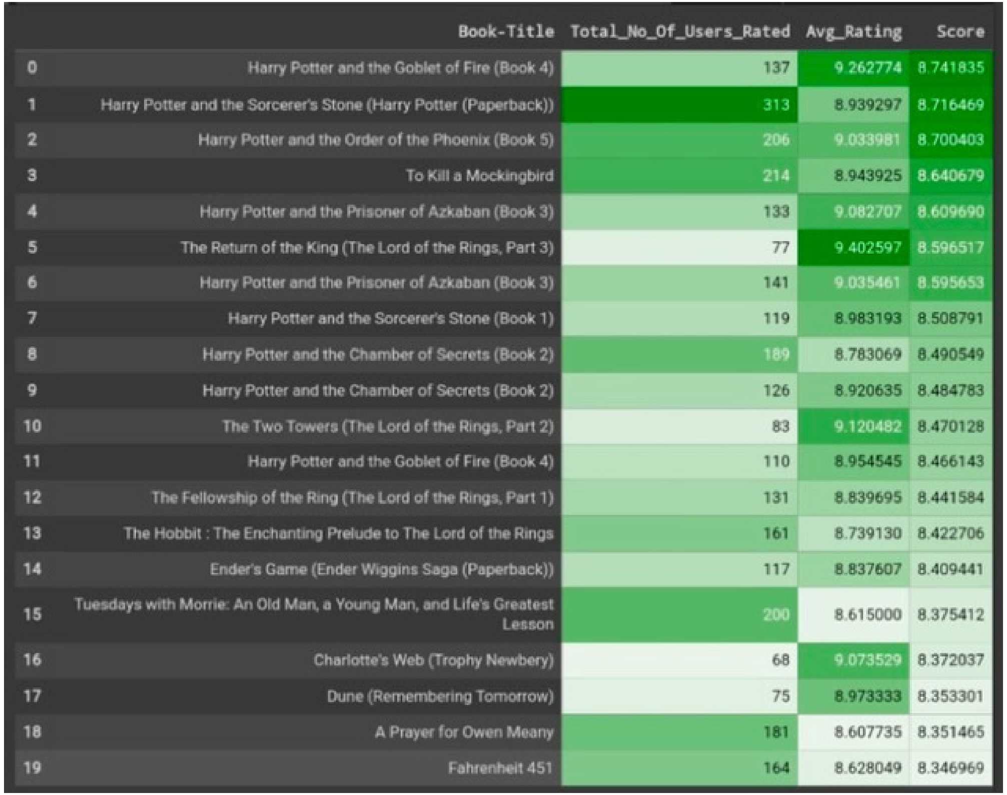
Book weighted average formula:

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

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

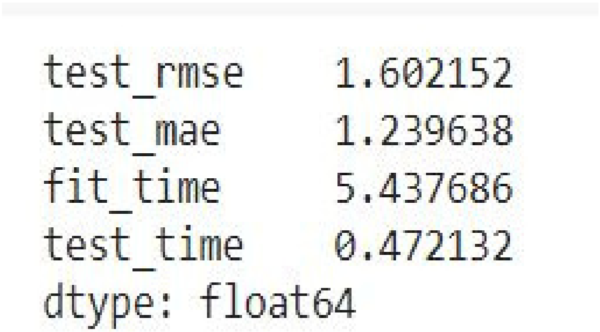


	Book-Title	Total_No_Of_Users_Rated	Avg_Rating	Score
0	Harry Potter and the Goblet of Fire (Book 4)	137	9.262774	8.741835
1	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	313	8.939297	8.716469
2	Harry Potter and the Order of the Phoenix (Book 5)	206	9.033981	8.700403
3	To Kill a Mockingbird	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

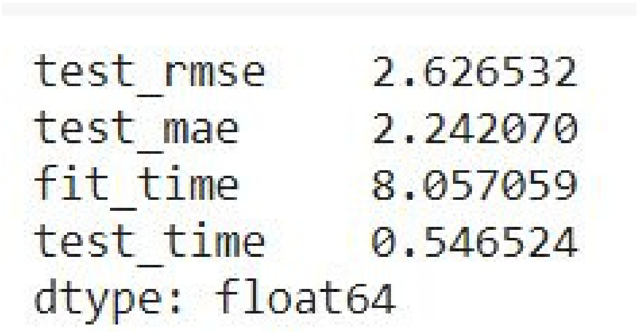
2.)Model based collaborative filtering

SVD



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

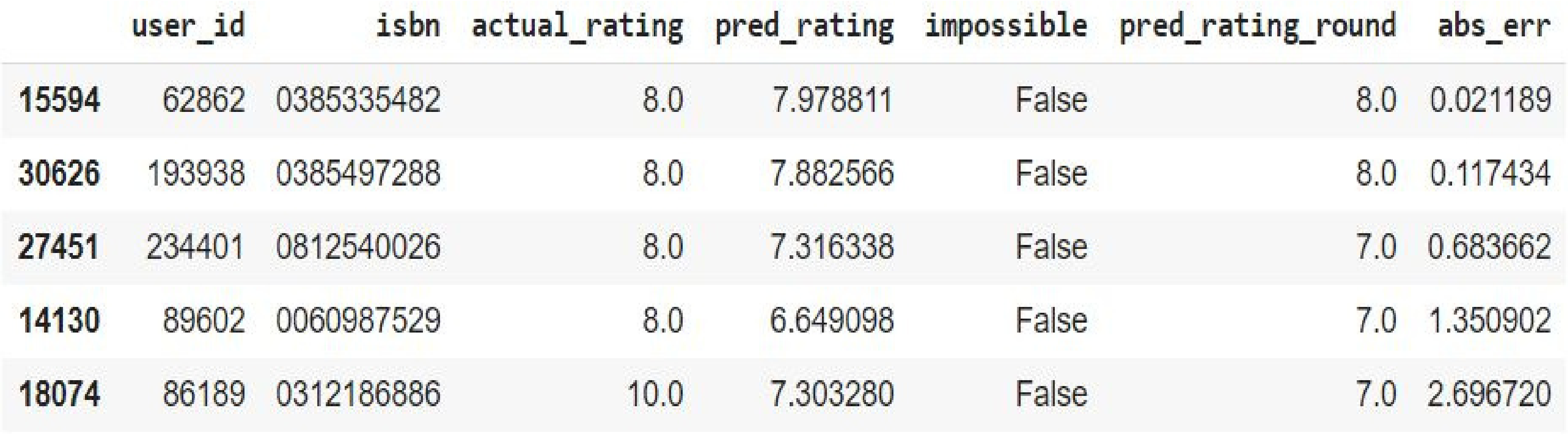
NMF



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

Different Models

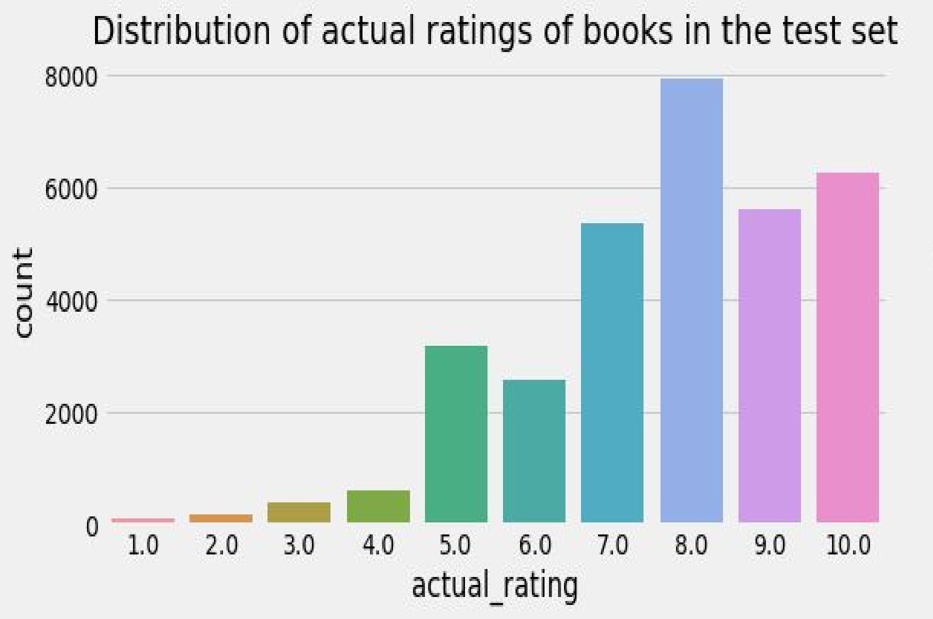
SVD Model Results



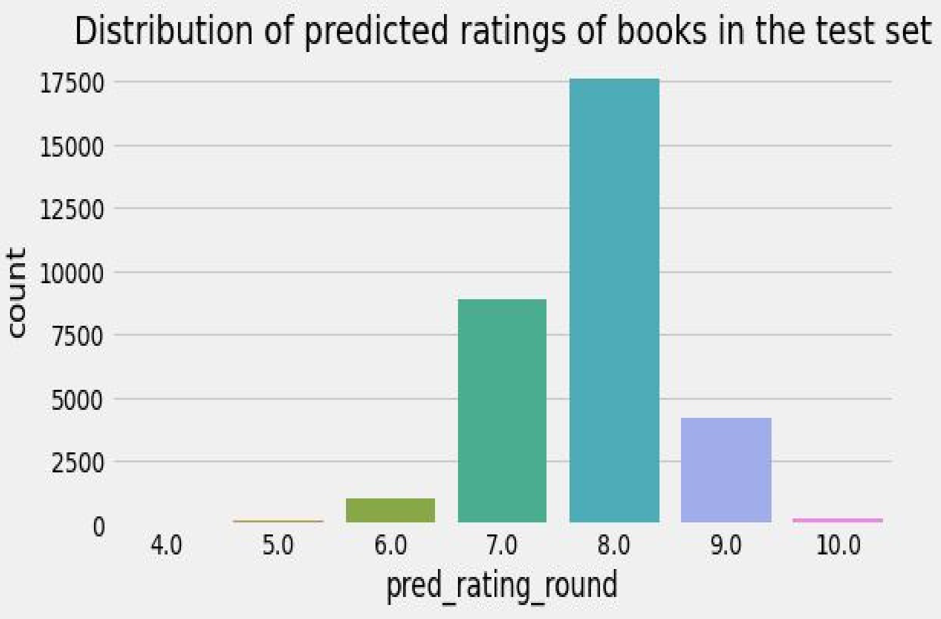
	user_id	isbn	actual_rating	pred_rating	impossible	pred_rating_round	abs_err
15594	62862	0385335482	8.0	7.978811	False	8.0	0.021189
30626	193938	0385497288	8.0	7.882566	False	8.0	0.117434
27451	234401	0812540026	8.0	7.316338	False	7.0	0.683662
14130	89602	0060987529	8.0	6.649098	False	7.0	1.350902
18074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

Different Models

SVD Model Results



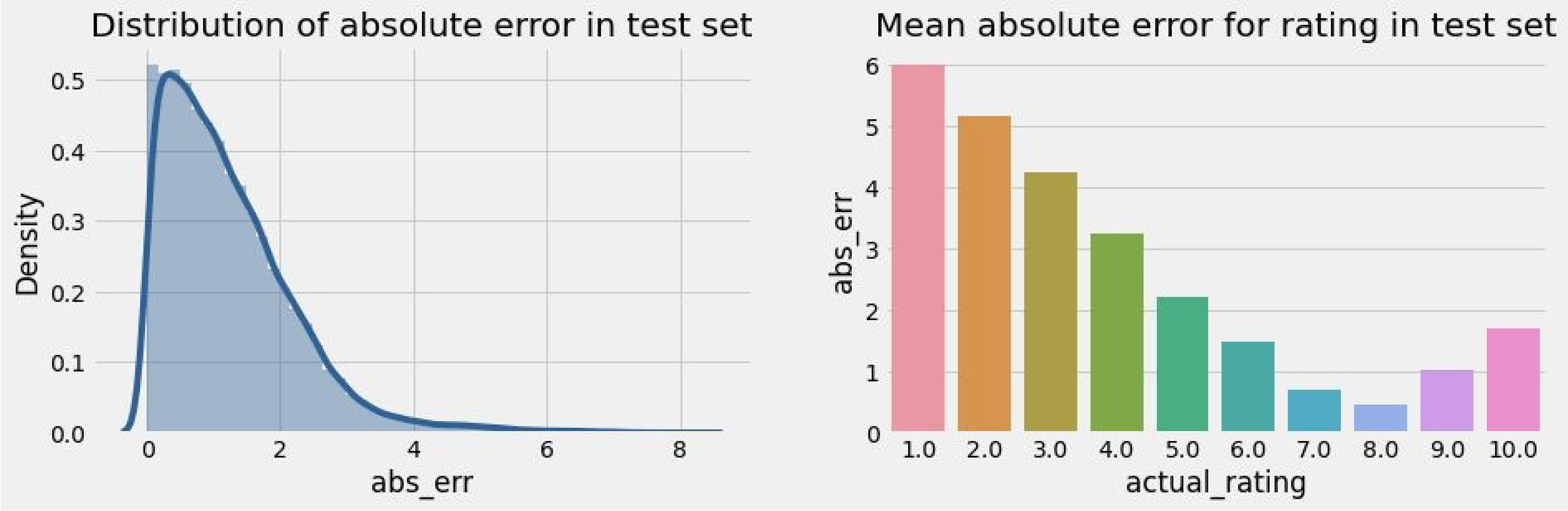
Distribution of actual ratings of books in the test set



Distribution of predicted ratings of books in the test set

Different Models

SVD Model Results



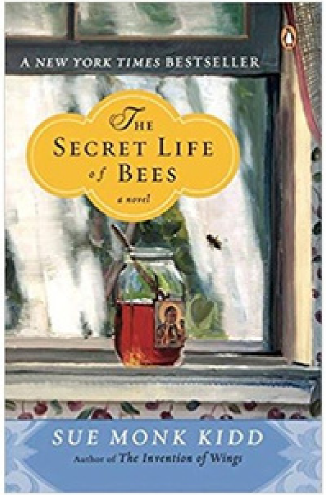
Distribution of absolute error in test set

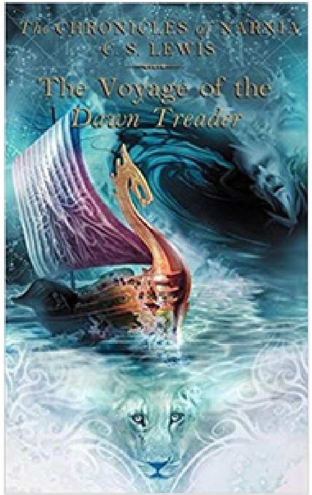
Mean absolute error for rating in test set

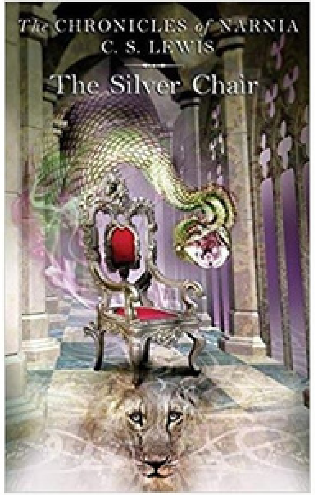
Different Models

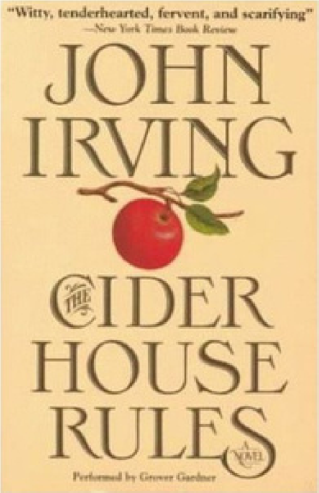
User-ID - 193458

Test set: predicted top rated books

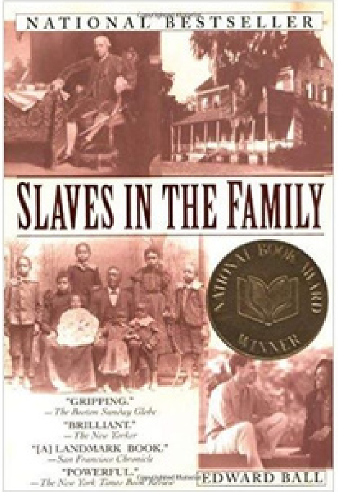








"Witty, tenderhearted, fervent, and scarifying"
—New York Times Book Review
JOHN IRVING
THE CIDER HOUSE RULES
A NOVEL
Performed by Grover Gardner



Different Models

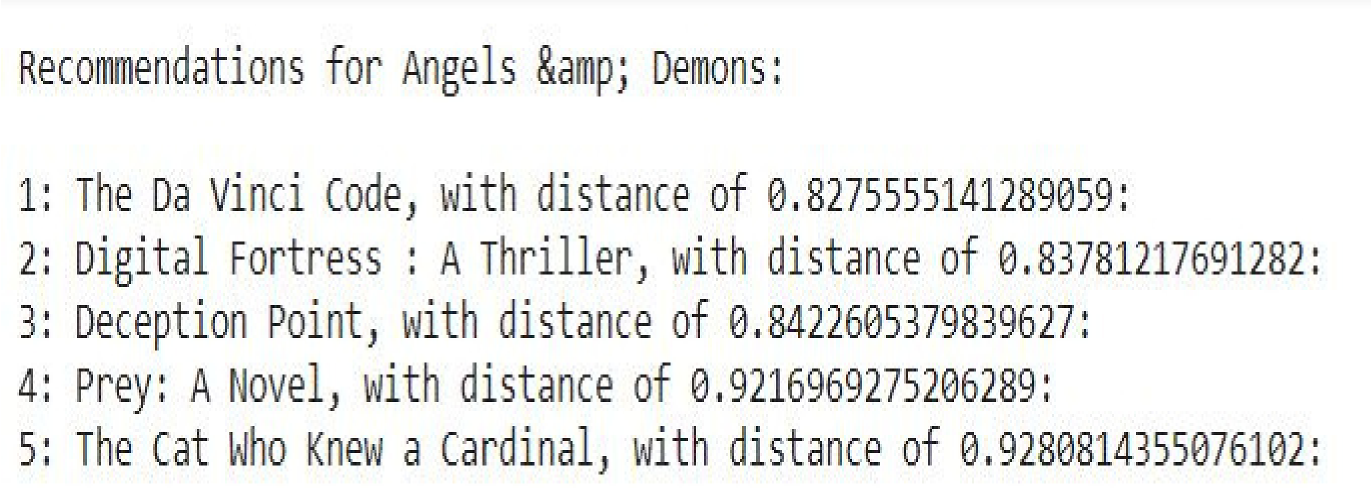
Test set: actual top rated books

The CHRONICLES of NARNIA
C. S. LEWIS
The Horse and His Boy
The CHRONICLES of NARNIA
C. S. LEWIS
The Voyage of the Dawn Treader
The CHRONICLES of NARNIA
C. S. LEWIS
The Silver Chair
A NEW YORK TIMES BESTSELLER
The Secret Life of Bees
a novel
SUE MONK KIDD
Author of The Invention of Wings
NATIONAL BESTSELLER
SLAVES IN THE FAMILY
NATIONAL BOOK AWARD WINNER
"GRIPPING." —The Boston Sunday Globe
"BRILLIANT." —The New Yorker
"[A] LANDMARK BOOK." —San Francisco Chronicle
"POWERFUL." —The New York Times Book Review
EDWARD BALL

Collaborative Filtering-(Item-Item based)

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

- Cosine Similarity Nearest
- Neighbour



```
Recommendations for Angels &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:
```

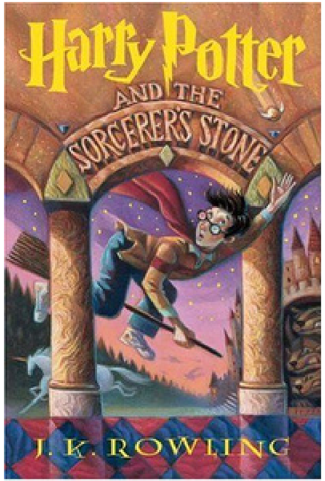
Different Models

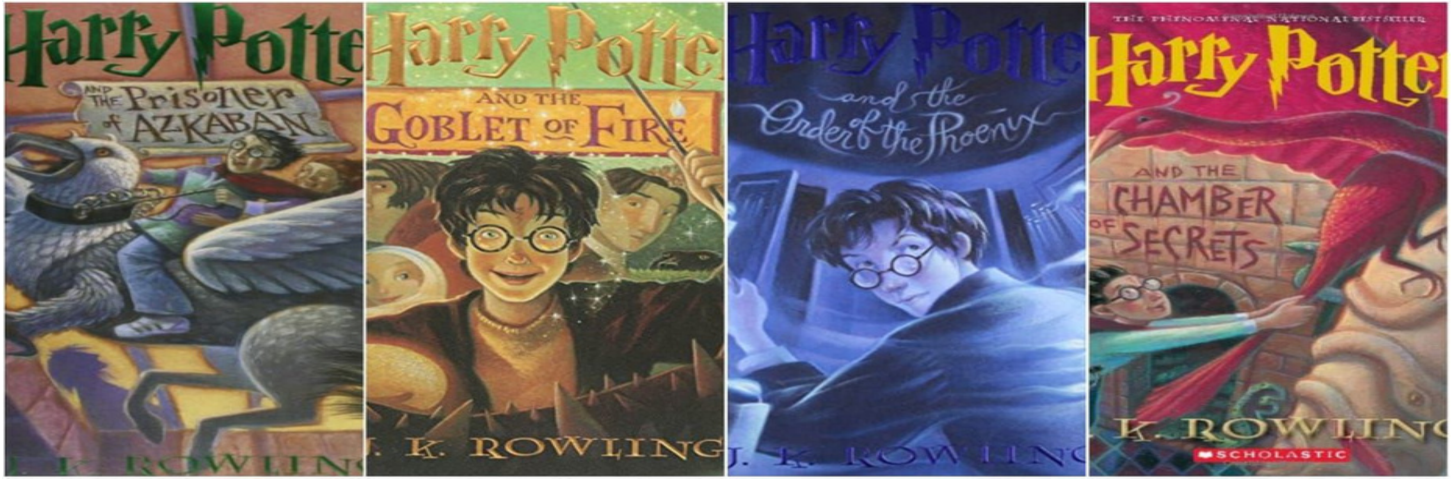
SVD and Correlation

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

Input

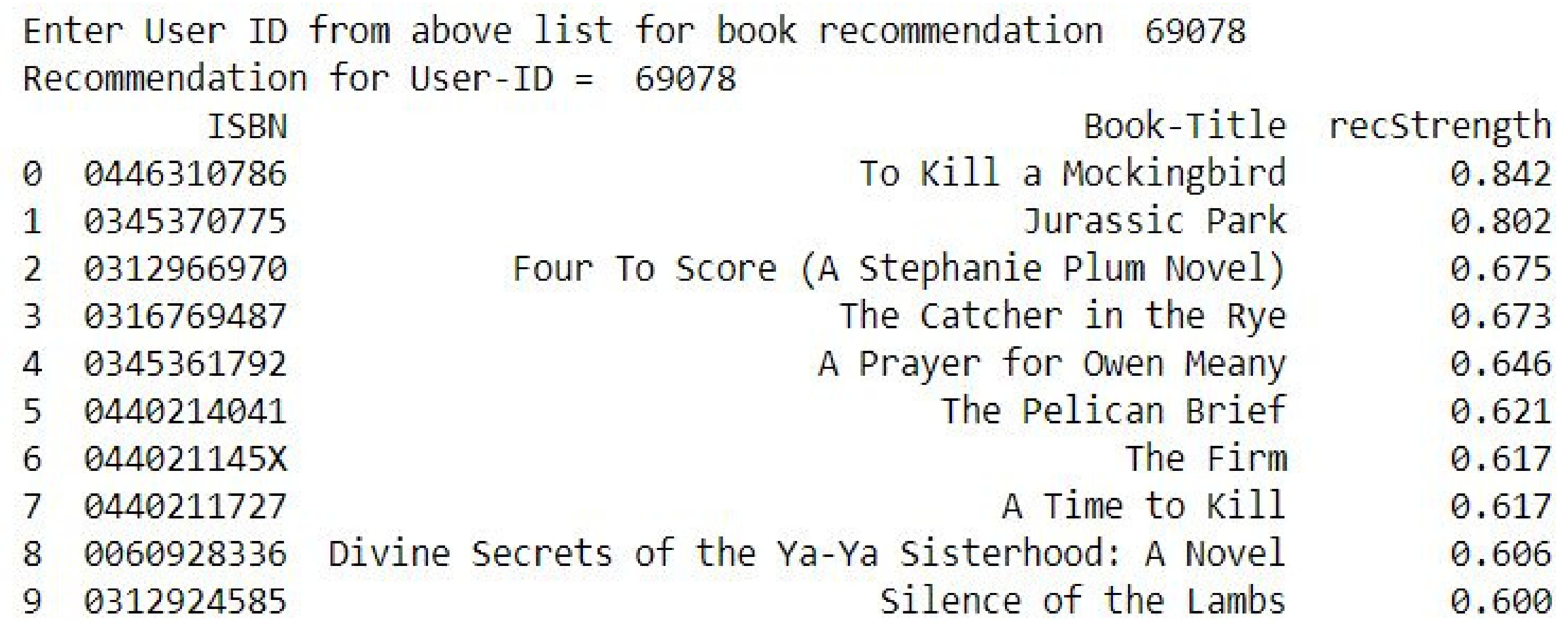
Output





Different Models

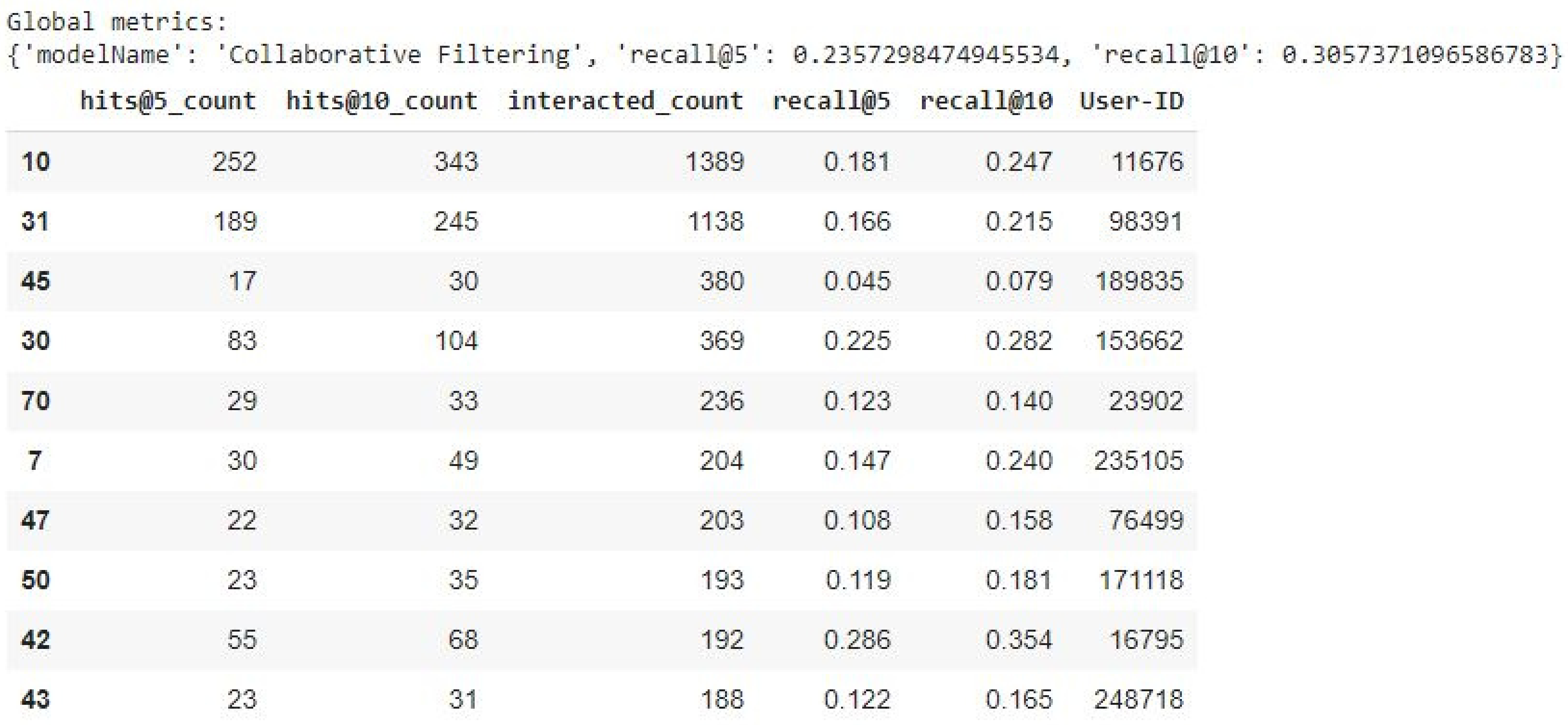
4.)Collaborative Filtering-(User-Item based)



```
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

- The Top-10 most rated books in EDA were primarily novels. The Secret Life of Bees and The Lovely Bone were highly regarded books.

- Most of the readers were between the ages of 20 and 35 and most of them had ties to North American and European countries, especially the United States, Canada, the United Kingdom, Germany, and Spain.

- According to the ratings distribution, most books received high ratings, with the maximum book receiving an 8. The number of ratings below 5 is relatively low.

- William Shakespeare, Stephen King, and Agatha Christie wrote the most books.

- A model-based collaborative filtering solution based on SVD technique performed significantly better than NMF with lower Mean Absolute Error (MAE).

Challenges

- As most of the books did not have user interactions, handling sparsity was another challenge.
- In addition, understanding the metric for evaluation was challenging.
- Because the data contained text information, features such as Location posed a major challenge for data cleaning.
- It was quite challenging to determine how to impute missing values and deal with outliers.

Future Scope

- A content-filtering based recommendation system could be implemented based on more information about the books dataset, such as Genre, Description etc., and compared with the existing collaborative-filtering system.

- Based on the age, location, etc., of the users, we intend to explore various clustering approaches and then implement voting algorithms that recommend items based on the cluster in which they are located.

Thank You