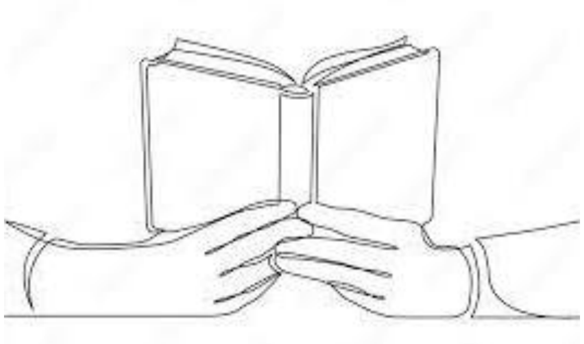
Capstone Project - 4

Book Recommendation System



Manas Ranjan Behera

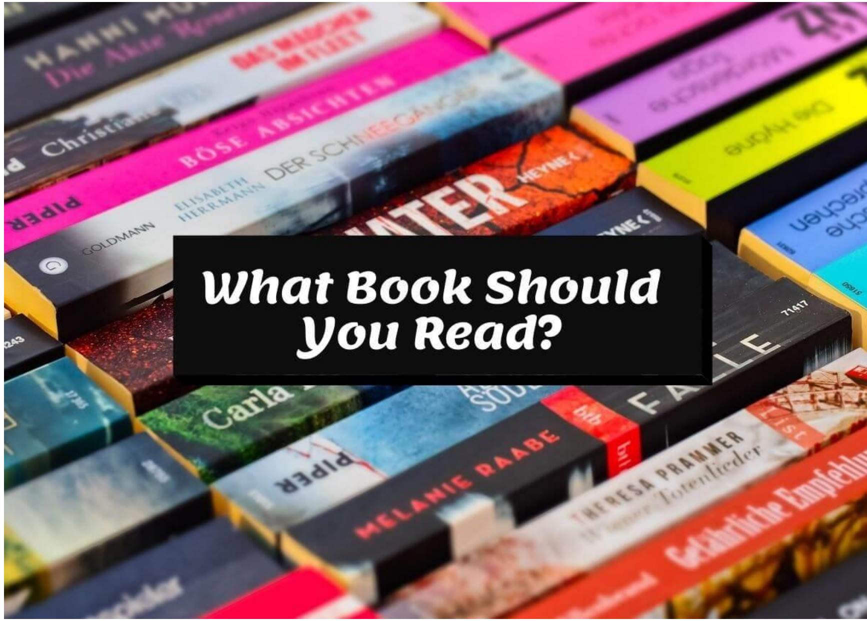
Contains

- Problem statement
- Data Summary
- Analysis of different datasets
- Data Cleaning & Outlier treatment
- Imputing missing values
- Different Recommendation Model
- Challenges
- Conclusion

1.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.



2.Dataset Summary :-

Our 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)

Dataset Summary (Continued)

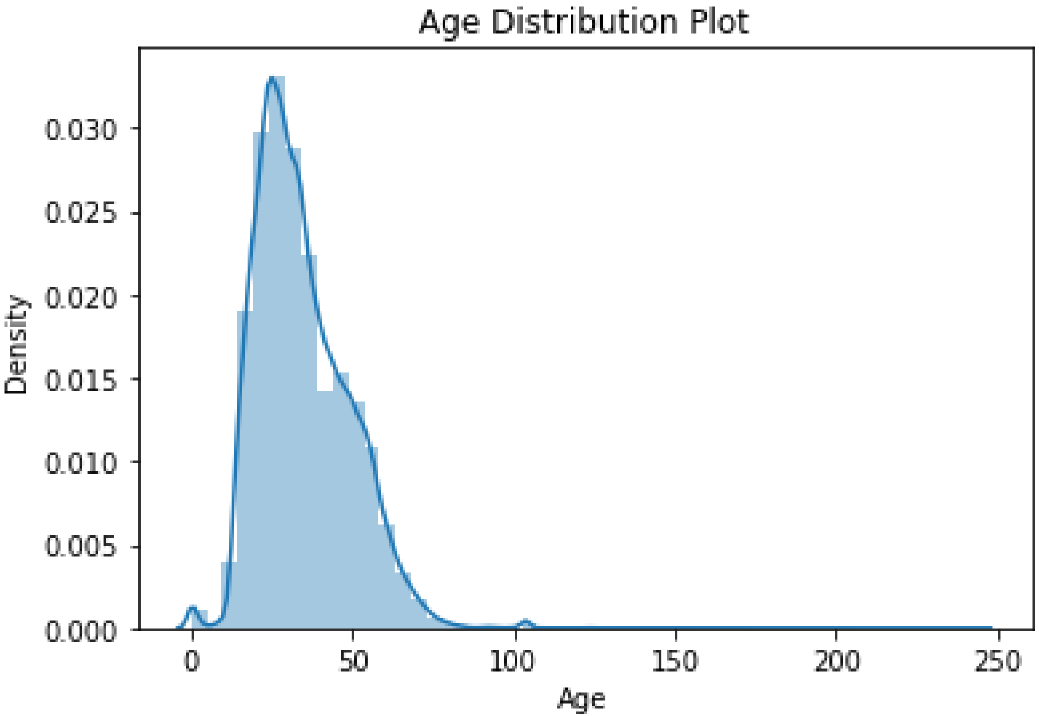
Ratings_Dataset

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

3.Analysis of different datasets

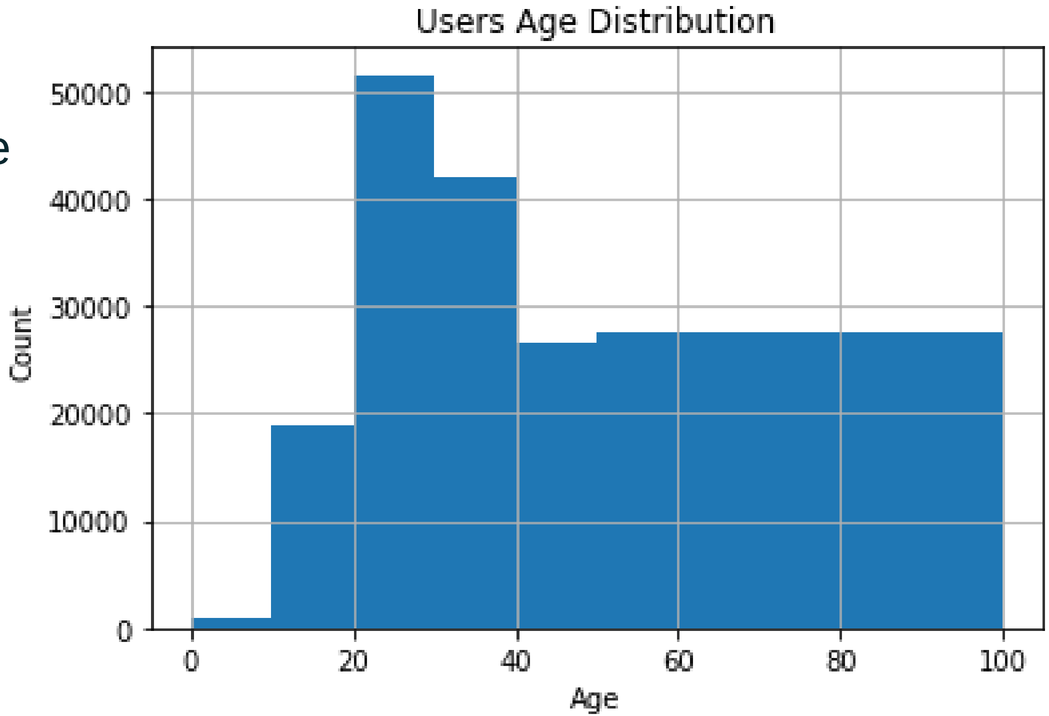
User-DF (Age)

- The Age range is given from 0 to 250.
- There are some data points at 0-5 and above 100 age. Consider it as outlier.(Generally The life span is between 1 to 100)



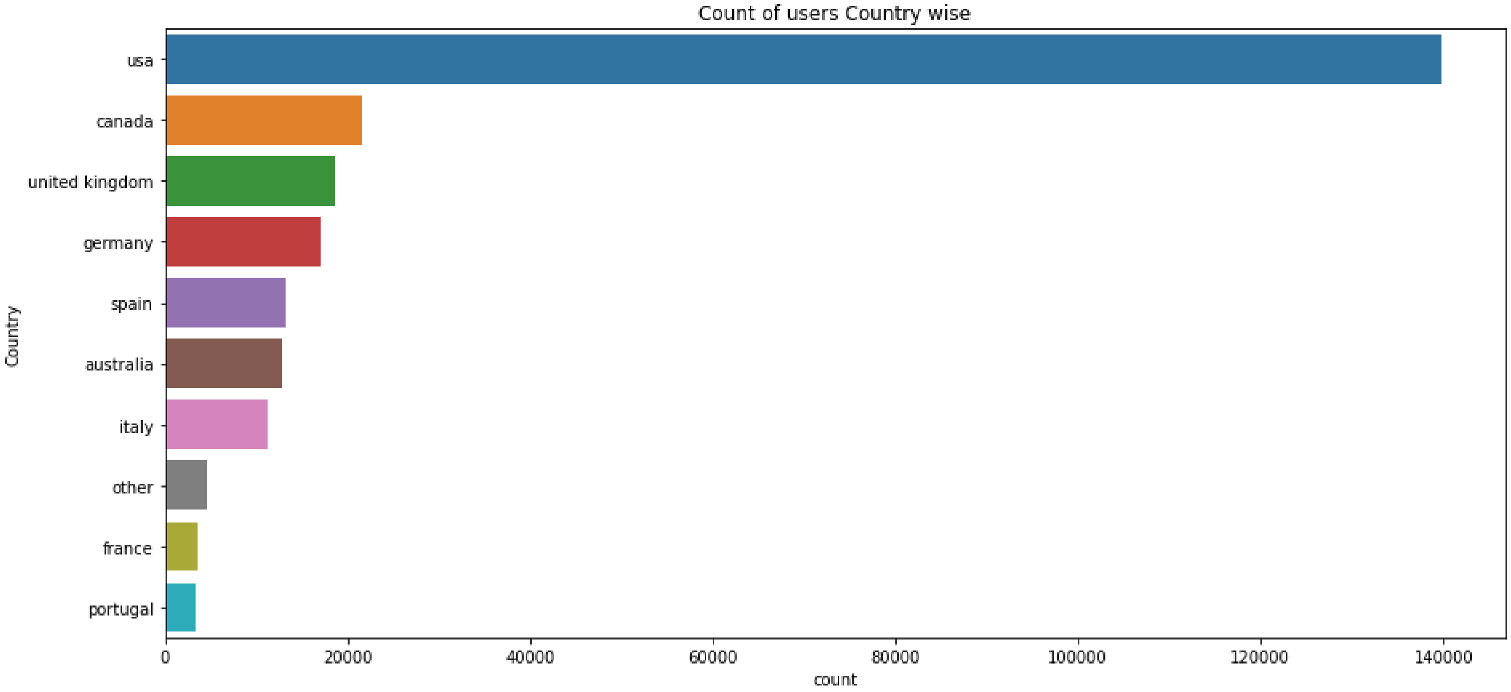
Observations from Users_df (Age)

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



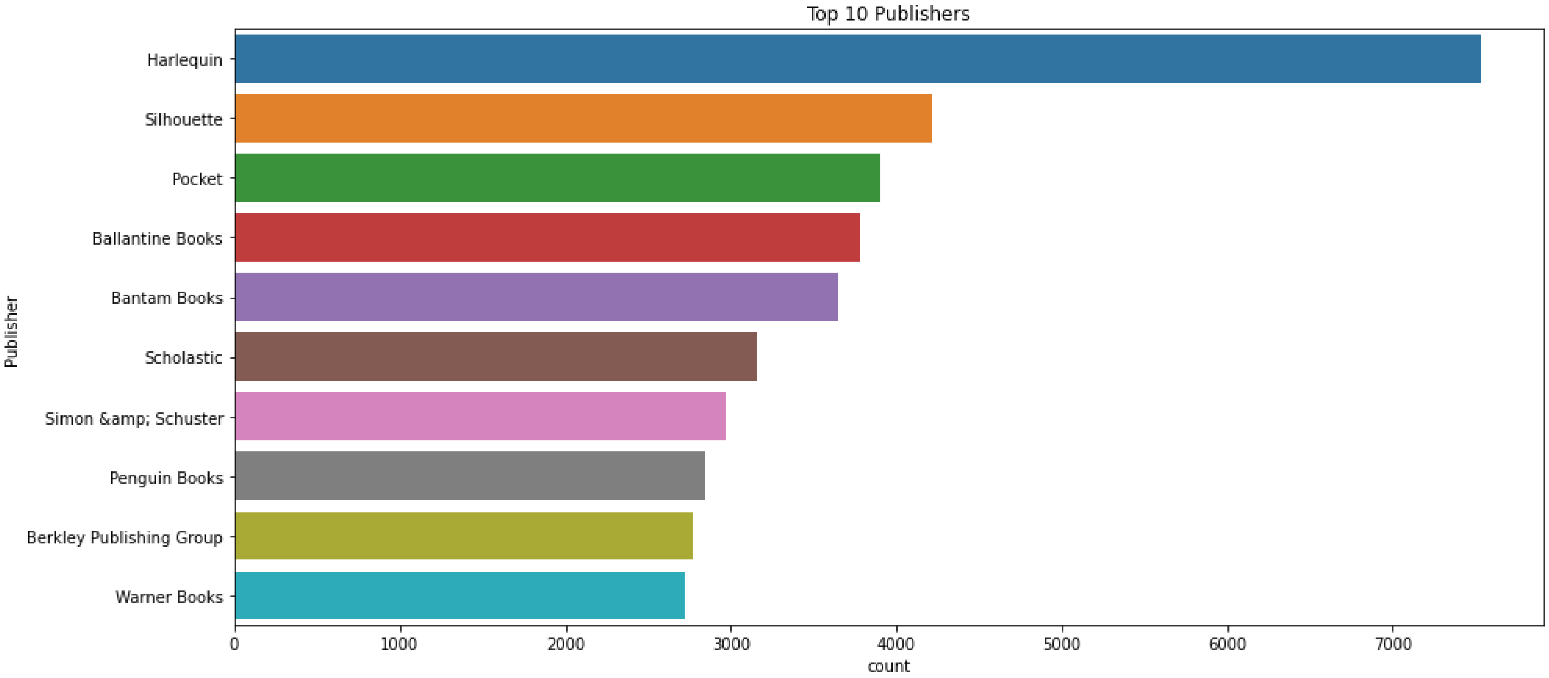
Observations from Users_df (Location)

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



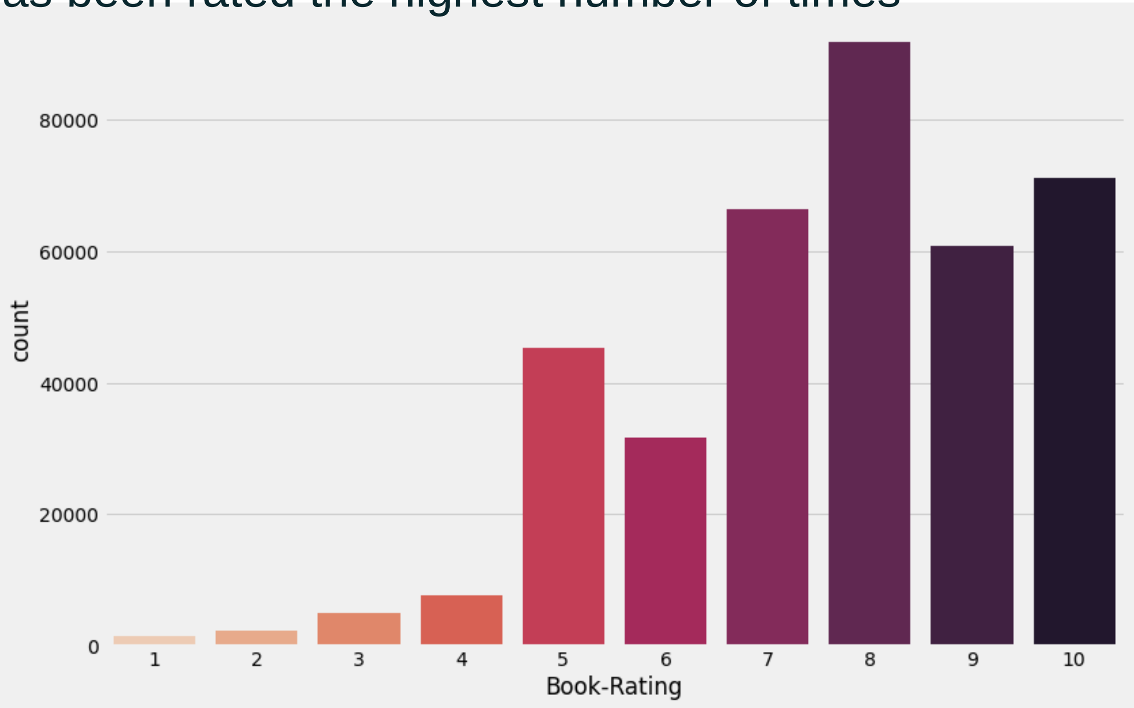
Observations from Book_df (Publishers)

- Harlequin published highest number of books in our given dataset



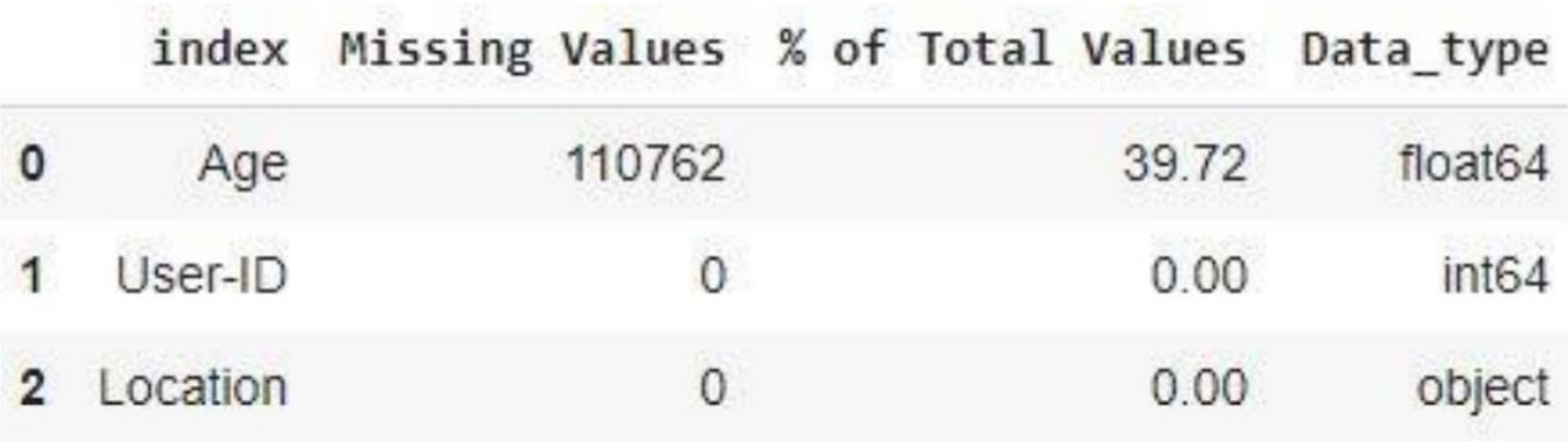
Observations from Ratings_df (Book_Rating)

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



4.Data Cleaning & Outlier Treatment :-

- 1. Null Value Imputation:
  - Age column has 40% missing values

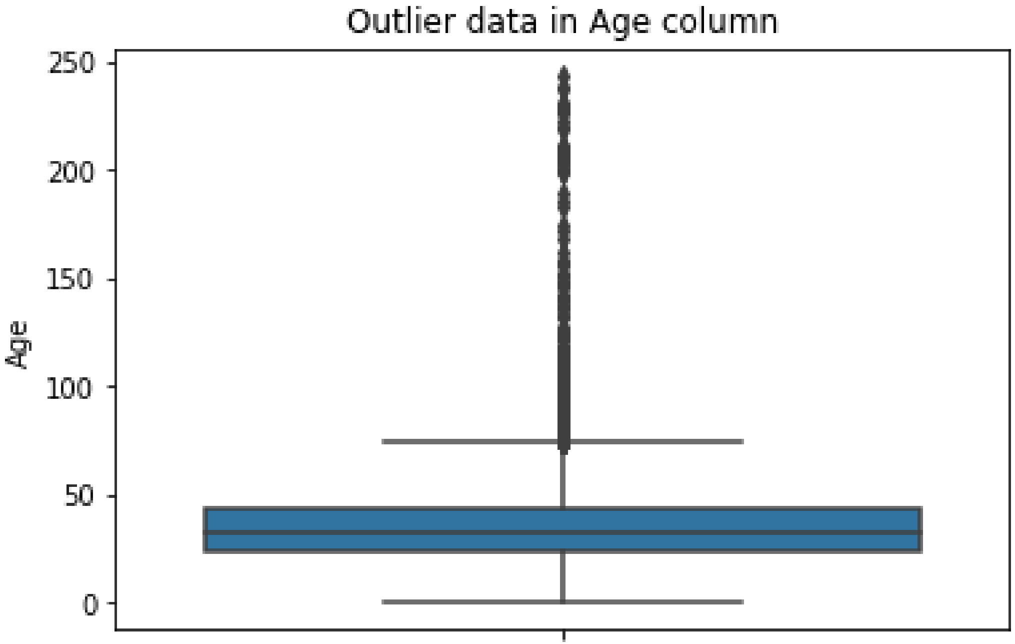


	index	Missing Values	% of Total Values	Data_type
0	Age	110762	39.72	float64
1	User-ID	0	0.00	int64
2	Location	0	0.00	object

Data Cleaning & Outlier Treatment :-

Imputing missing values:-

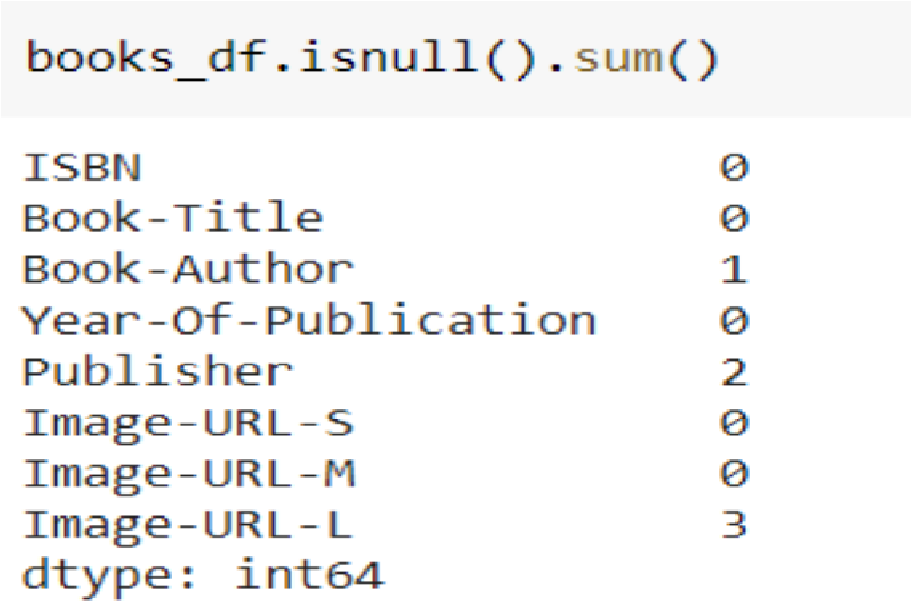
- Previously we found Outliers in Age column
- I used median to fill Nan values



Data Cleaning & Outlier Treatment (continued…)

Null Value Imputation :- (Book_df)

- I found some null values in the books dataset.
- Replaced string by integer values



```
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
```

5. Different Models :-

1 . Popularity Based Recommendation

It is a type of recommendation system which works on the principle of popularity and or anything which is in trend. These systems check about the product or movie which are in trend or are most popular among the users and directly recommend those.

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

Where,

v = number of votes for the books

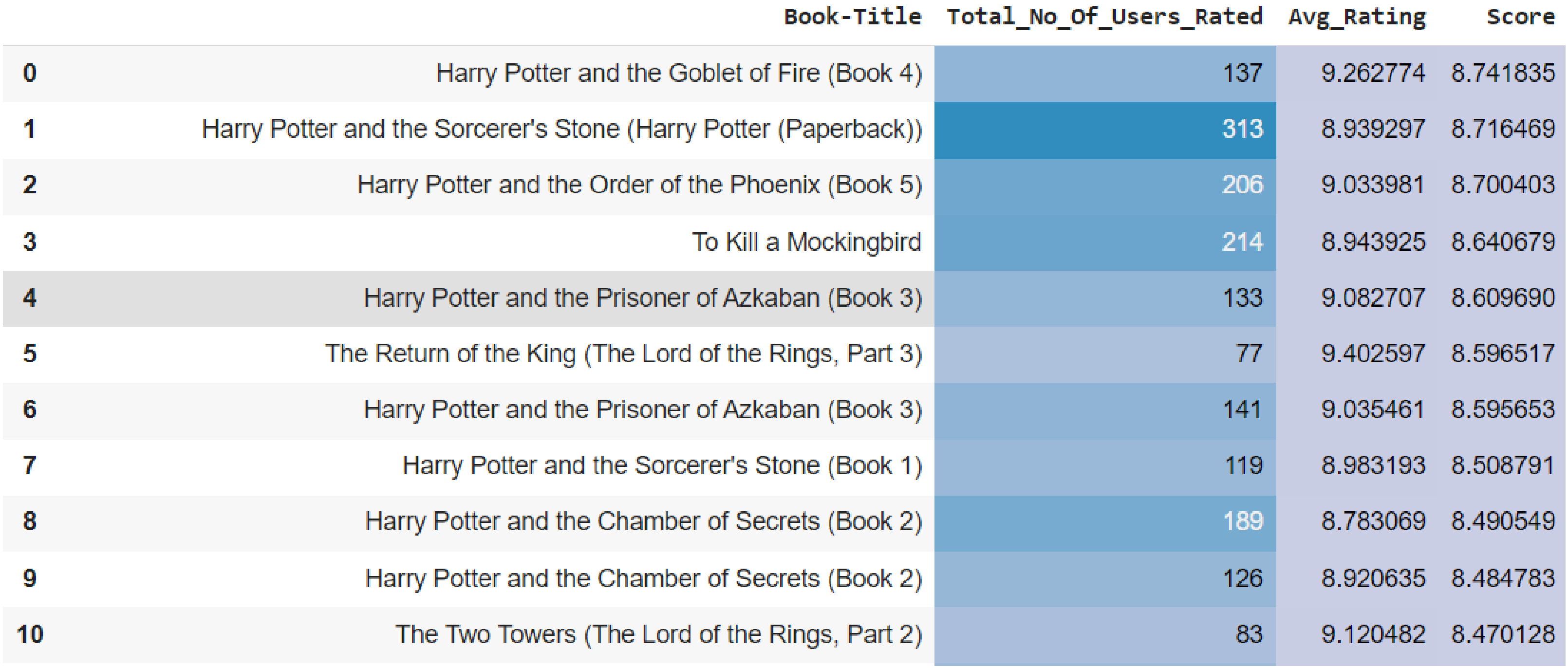
m = minimum votes required to be listed in the chart

R = average rating

C = mean vote

Different Models :- (continued…)

Top 10 books are



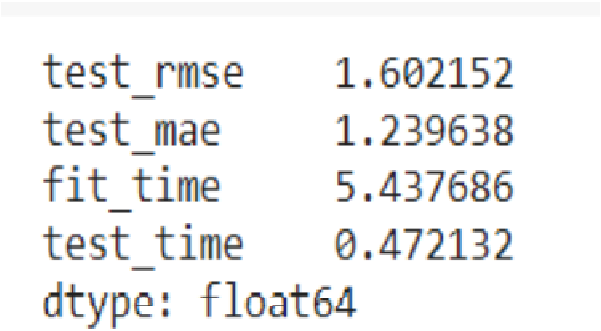
	Book-Title	Total_No_Of_Users_Rated	Avg_Rating	Score
0	Harry Potter and the Goblet of Fire (Book 4)	137	9.262774	8.741835
1	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	313	8.939297	8.716469
2	Harry Potter and the Order of the Phoenix (Book 5)	206	9.033981	8.700403
3	To Kill a Mockingbird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

Different Models :- (continued…)

2. Model based collaborative filtering

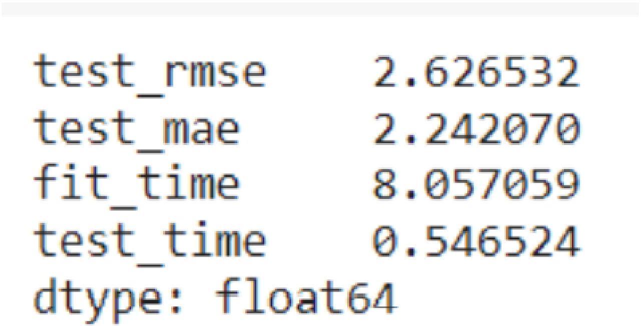
Here I used SVD(Singular value decomposition) & NMF (Non-Negetive Matrix factorization)

SVD



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

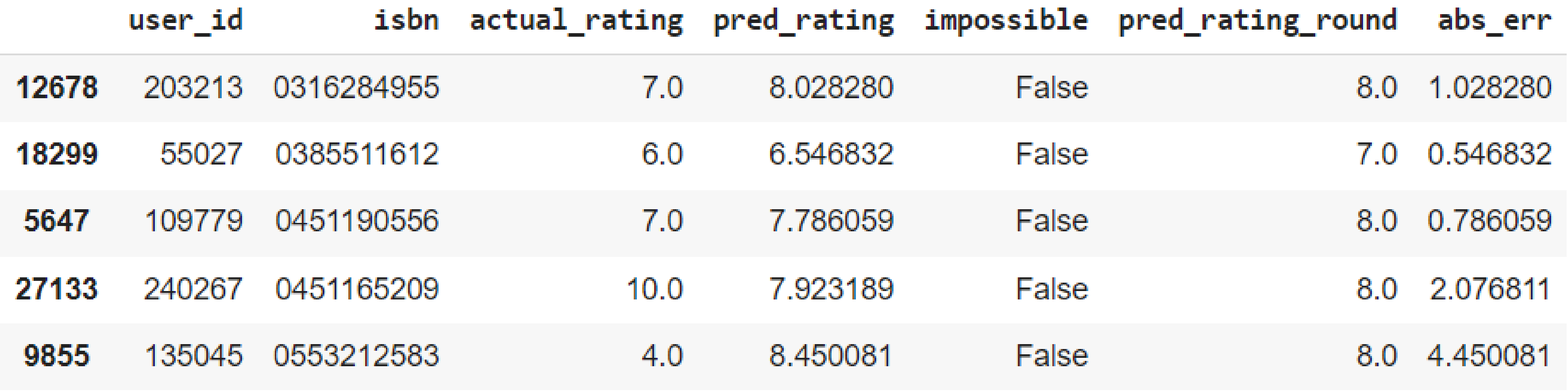
NMF



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

Different Models :- (continued…)

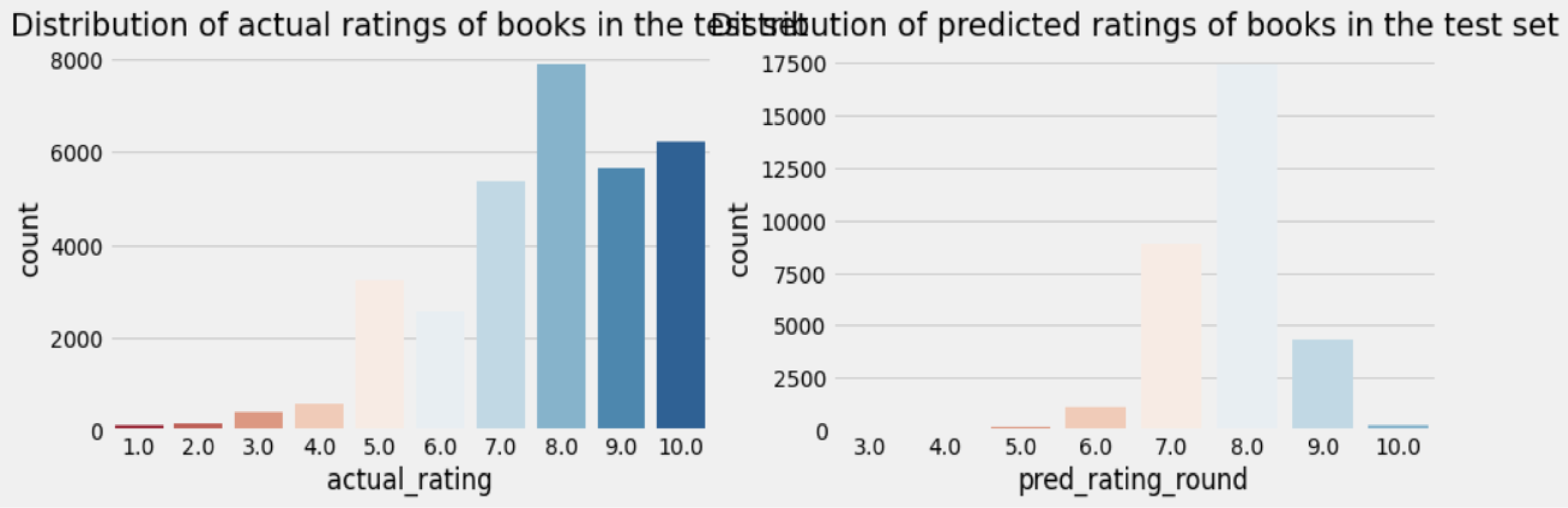
SVD Model Score



	user_id	isbn	actual_rating	pred_rating	impossible	pred_rating_round	abs_err
12678	203213	0316284955	7.0	8.028280	False	8.0	1.028280
18299	55027	0385511612	6.0	6.546832	False	7.0	0.546832
5647	109779	0451190556	7.0	7.786059	False	8.0	0.786059
27133	240267	0451165209	10.0	7.923189	False	8.0	2.076811
9855	135045	0553212583	4.0	8.450081	False	8.0	4.450081

Different Models :- (continued…)

SVD Model Plot

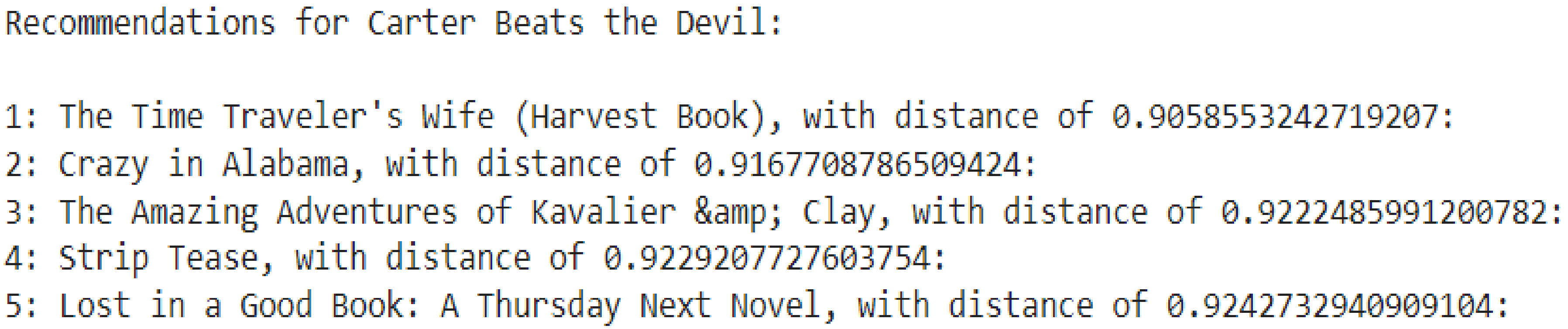


Different Models :- (continued…)

3. Collaborative Filtering-(Item-Item based)

- K-Nearest Neighbour (Using cosine similarity)

Cosine similarity is used as a metric in different machine learning algorithms like the KNN for determining the distance between the neighbours



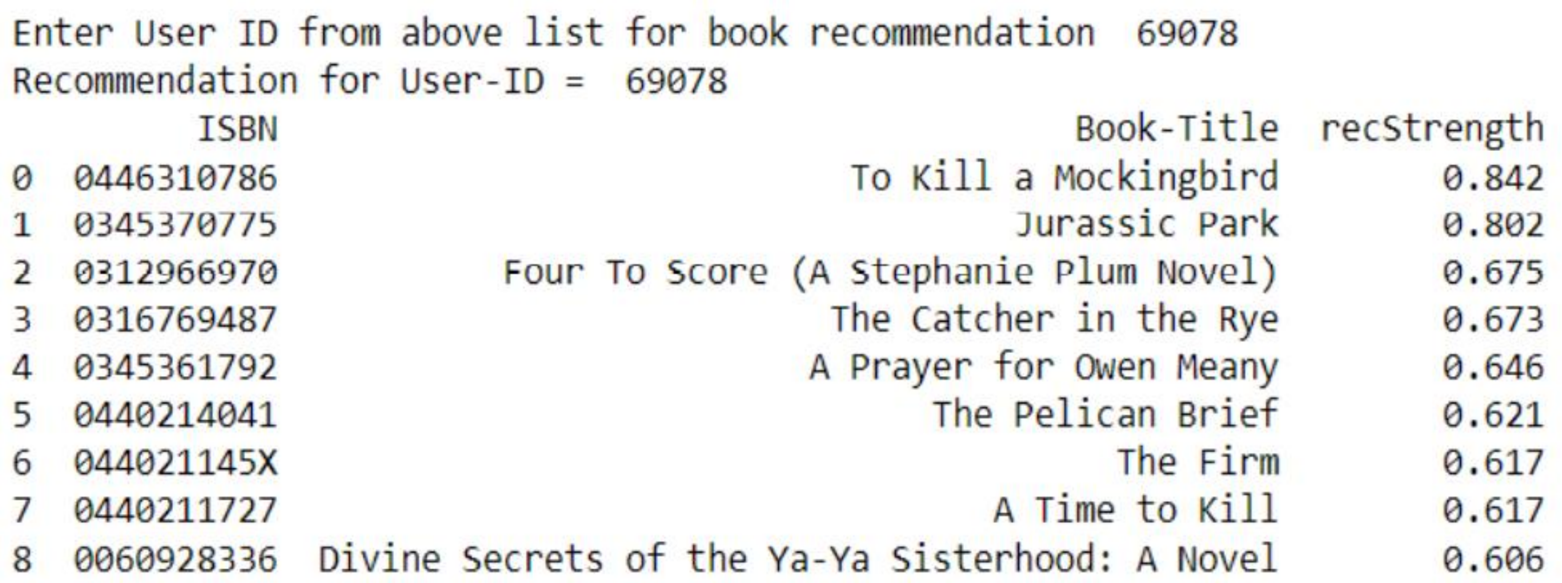
```
Recommendations for Carter Beats the Devil:

1: The Time Traveler's Wife (Harvest Book), with distance of 0.9058553242719207:
2: Crazy in Alabama, with distance of 0.9167708786509424:
3: The Amazing Adventures of Kavalier &amp; Clay, with distance of 0.922248591200782:
4: Strip Tease, with distance of 0.9229207727603754:
5: Lost in a Good Book: A Thursday Next Novel, with distance of 0.9242732940909104:
```

Different Models :- (continued…)

4. Collaborative Filtering-(User-Item based)

Item’s recommendation rating for a user is calculated depending on that items’ ratings by other similar users.



```
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
```

6. 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) .

7. 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.

Thank You