Capstone Project

Book Recommendation System

Presenter Name :

Vikas panchal

Naveen Kumar Batta

Content :

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

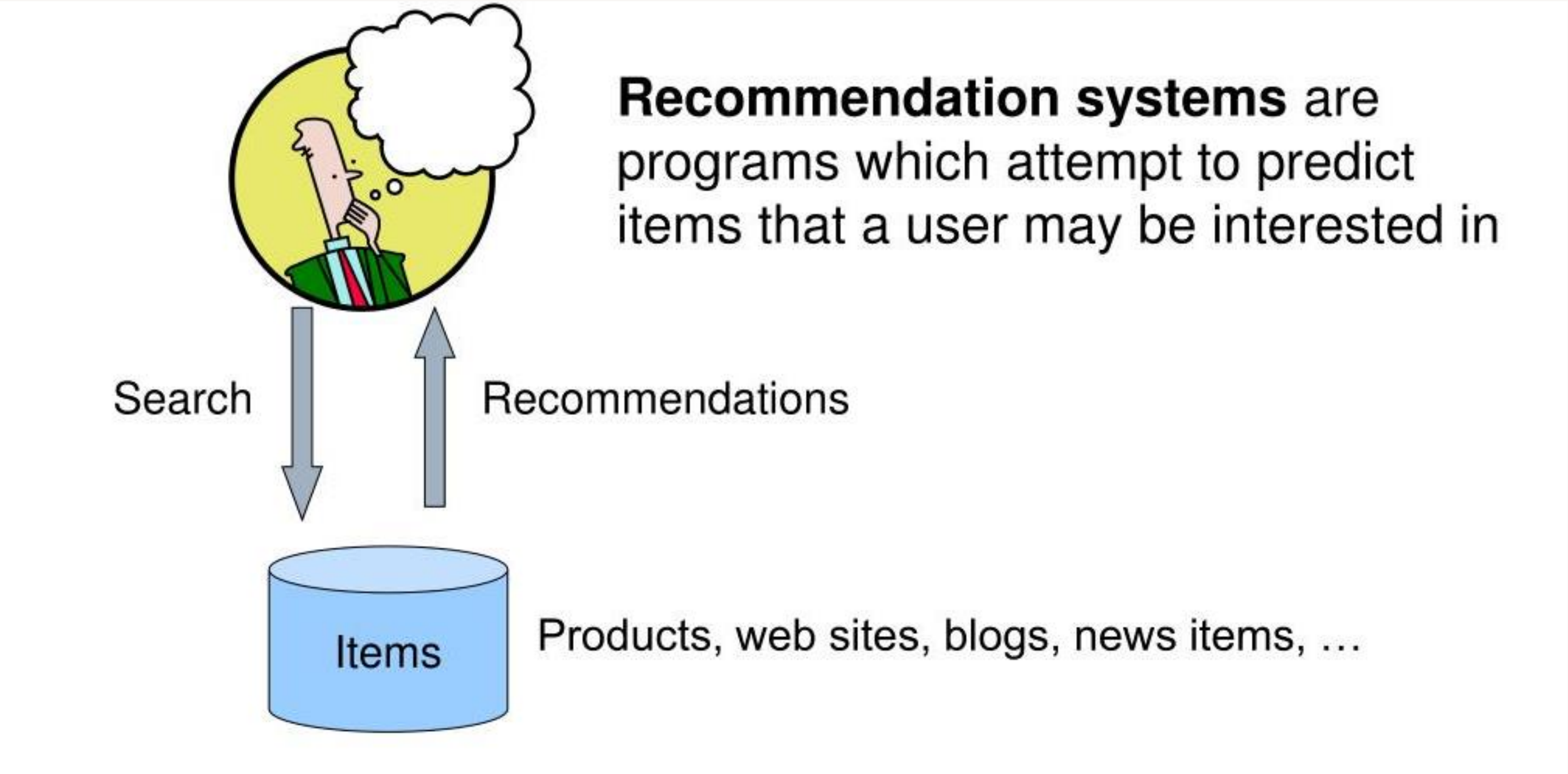
Recommendation System

Recommendation systems produce a ranked list of items on which a user might be interested, in the context of his current choice of an item.

- ❑ Subclass of Information filtering system that seek to predict the ‘rating’ or ‘preference’ that a user would give to them.
- ❑ Applied in variety of applications like movies, books, research articles.
- ❑ Recommendation systems involve predicting user preferences for unseen items • such as movies, songs or books
- ❑ Recommendation systems have become very popular with the increasing availability of millions of products online
- ❑ Recommending relevant products increases the sales

.

What is Recommendation System



Recommendation systems are programs which attempt to predict items that a user may be interested in

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,

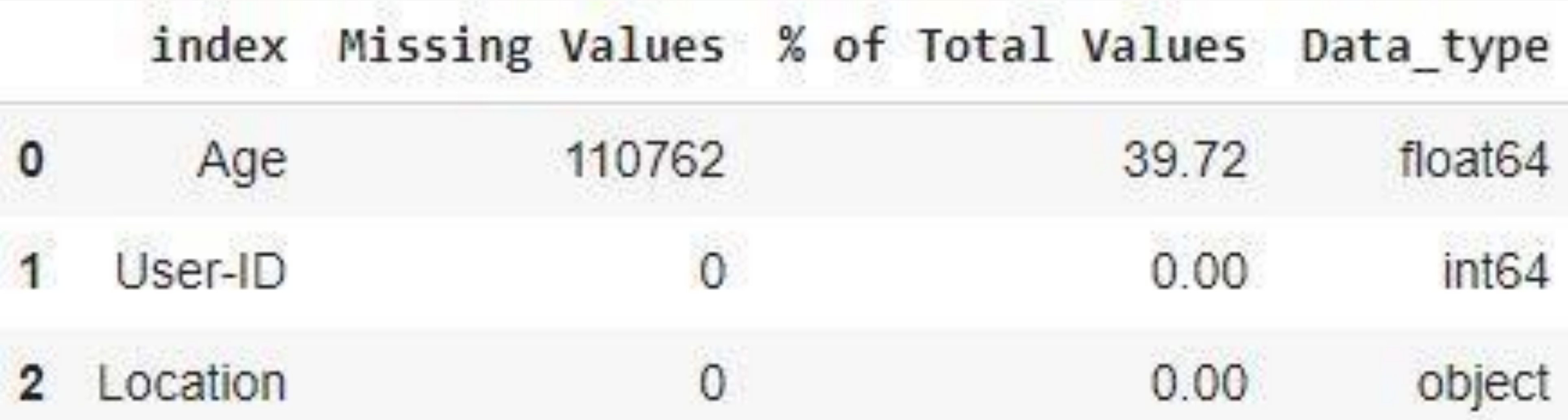
Ratings dataset.

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

Data Cleaning :-

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

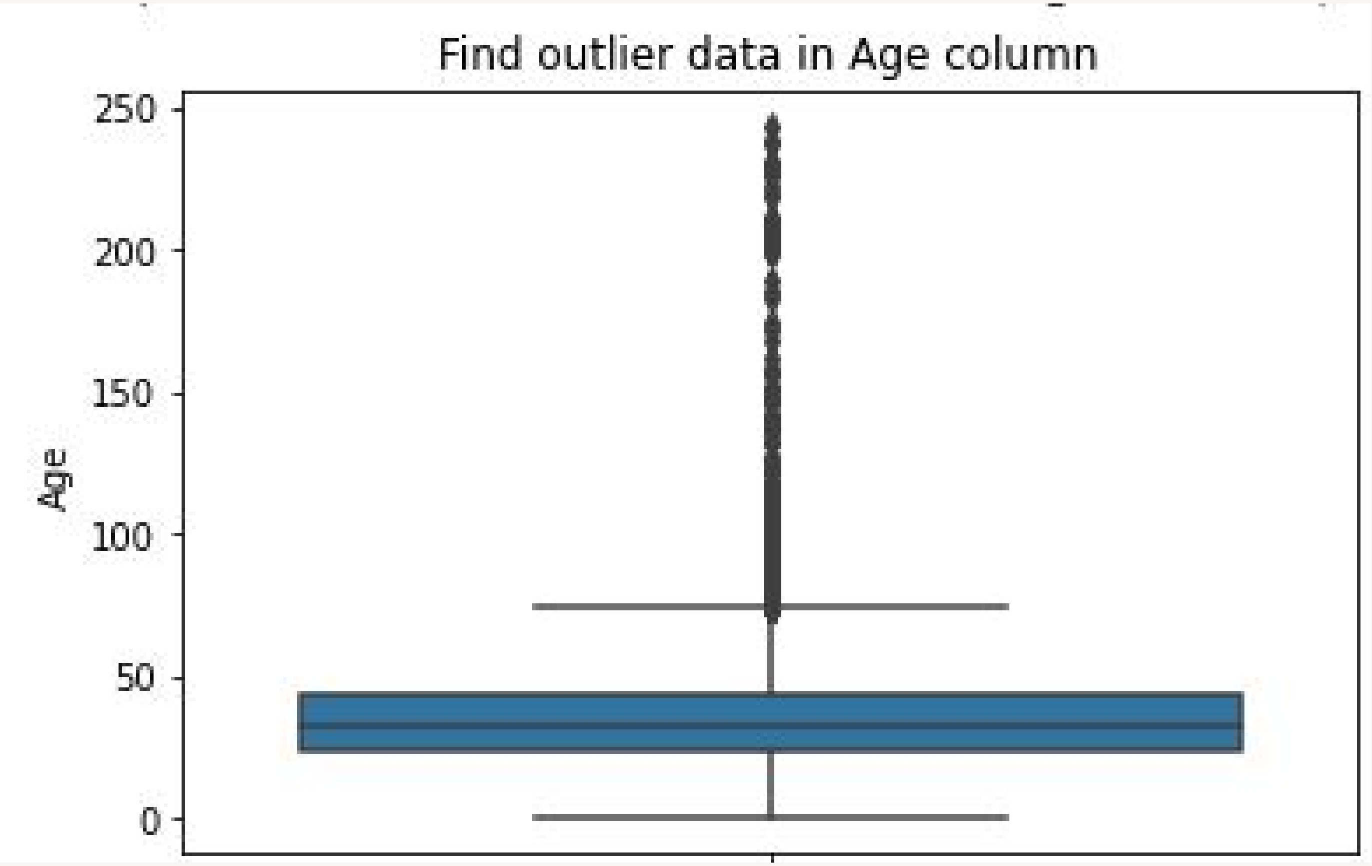
Look for missing data in user dataset



	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

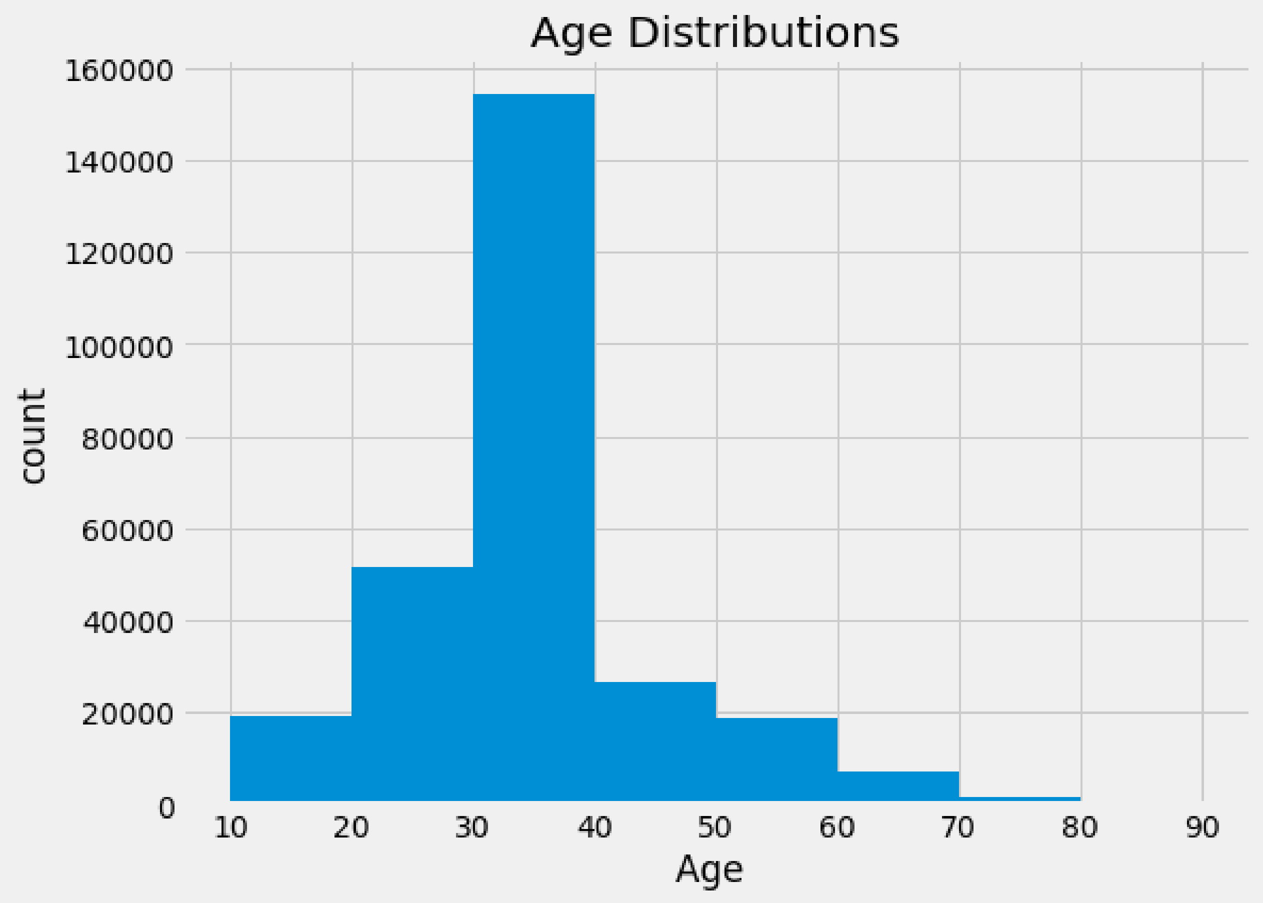
Data Cleaning Checking Outliers (Missing values)

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



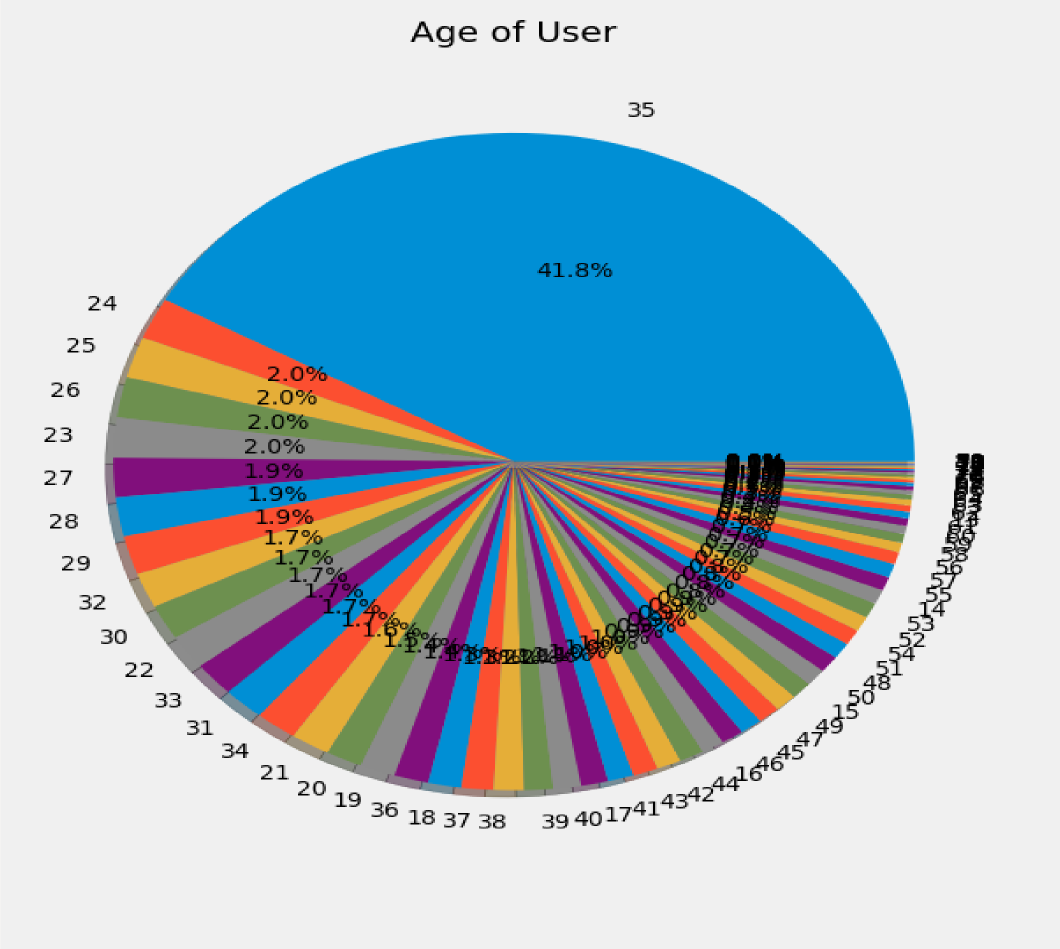
Observations from Users_df (Age)

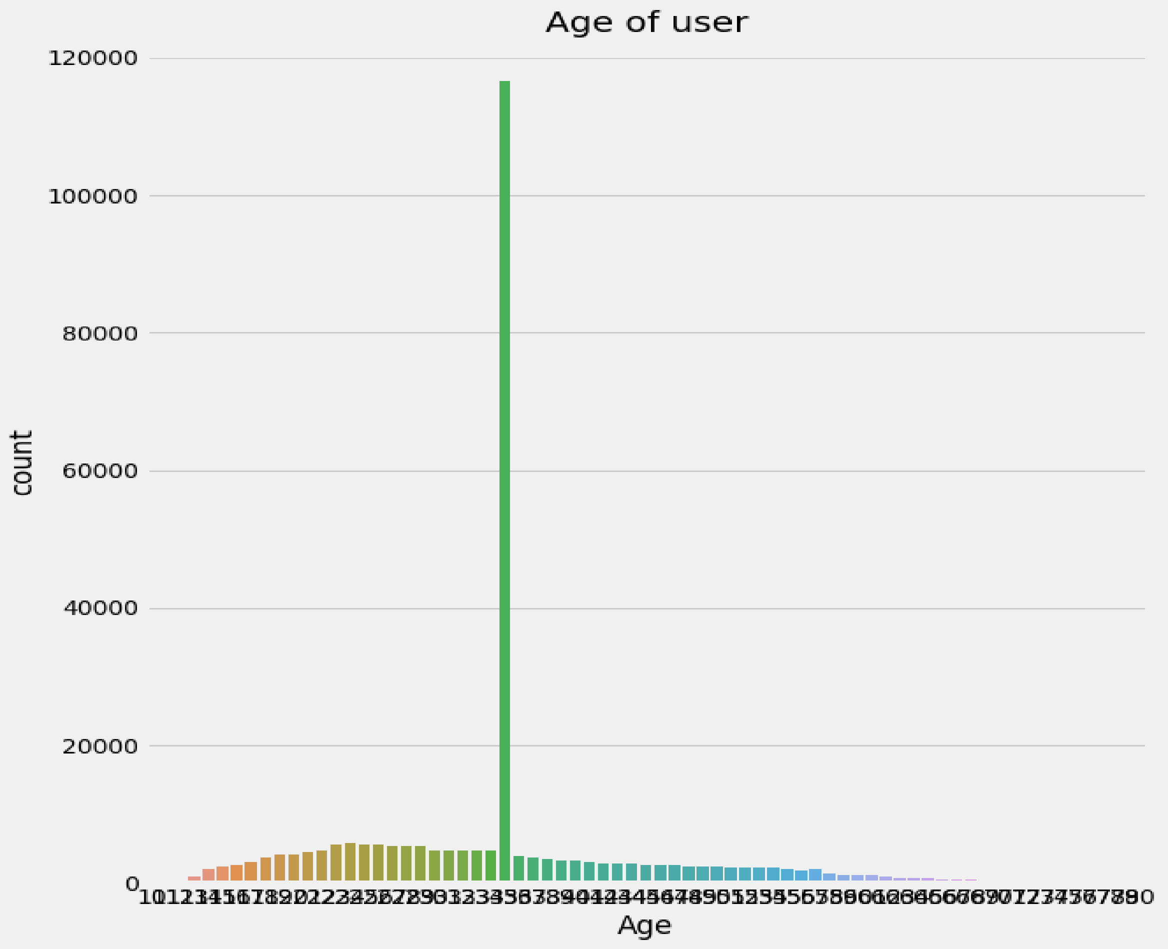
- The Age range distribution is right skewed
- Most active readers lie in age group 20-40



Observation Age of Users (countplot and pie chart):

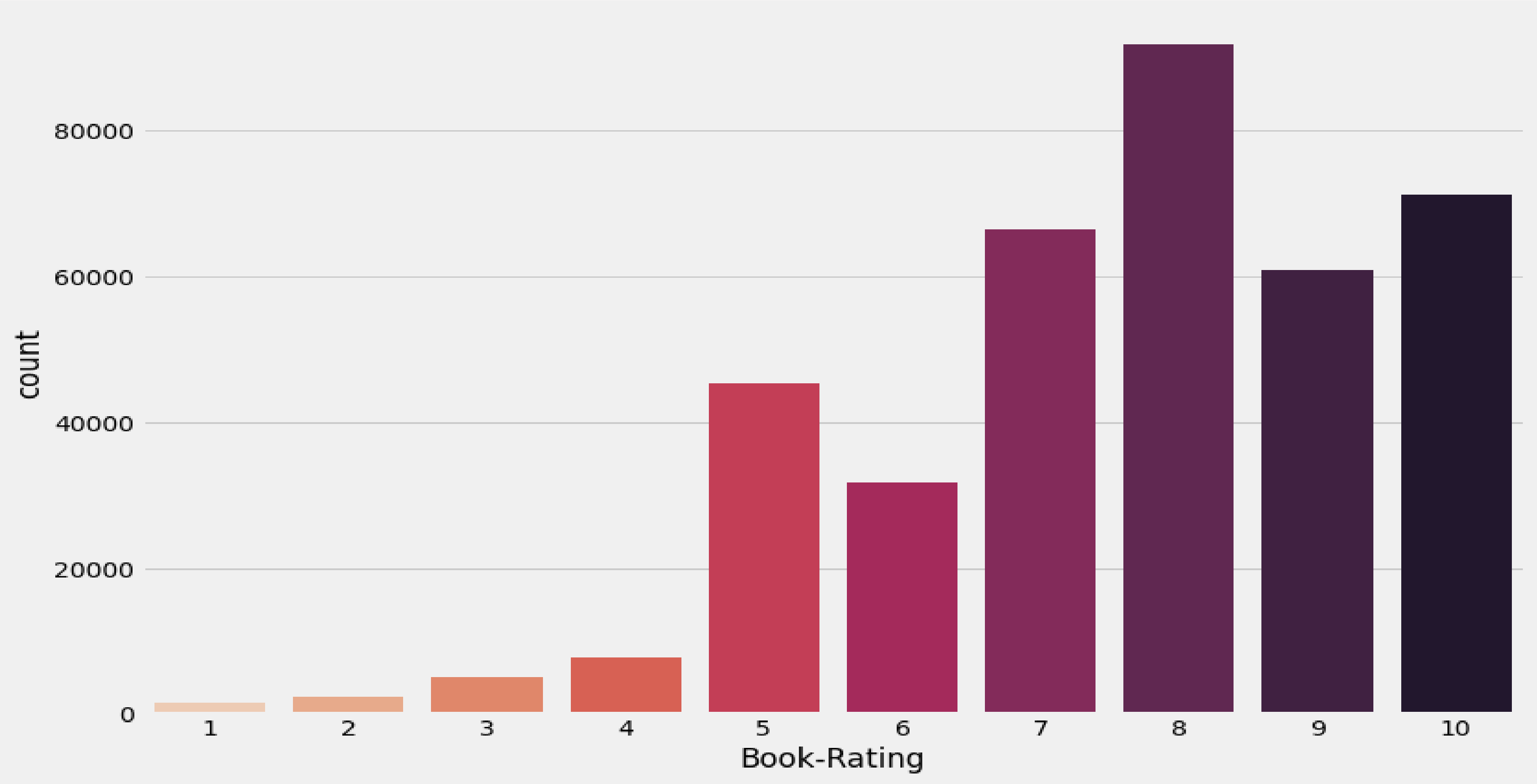
- From above plots we observed that 41.9% of age 34 group read more books compared to other age groups. Also the users with the age 60 and above do not read more books.





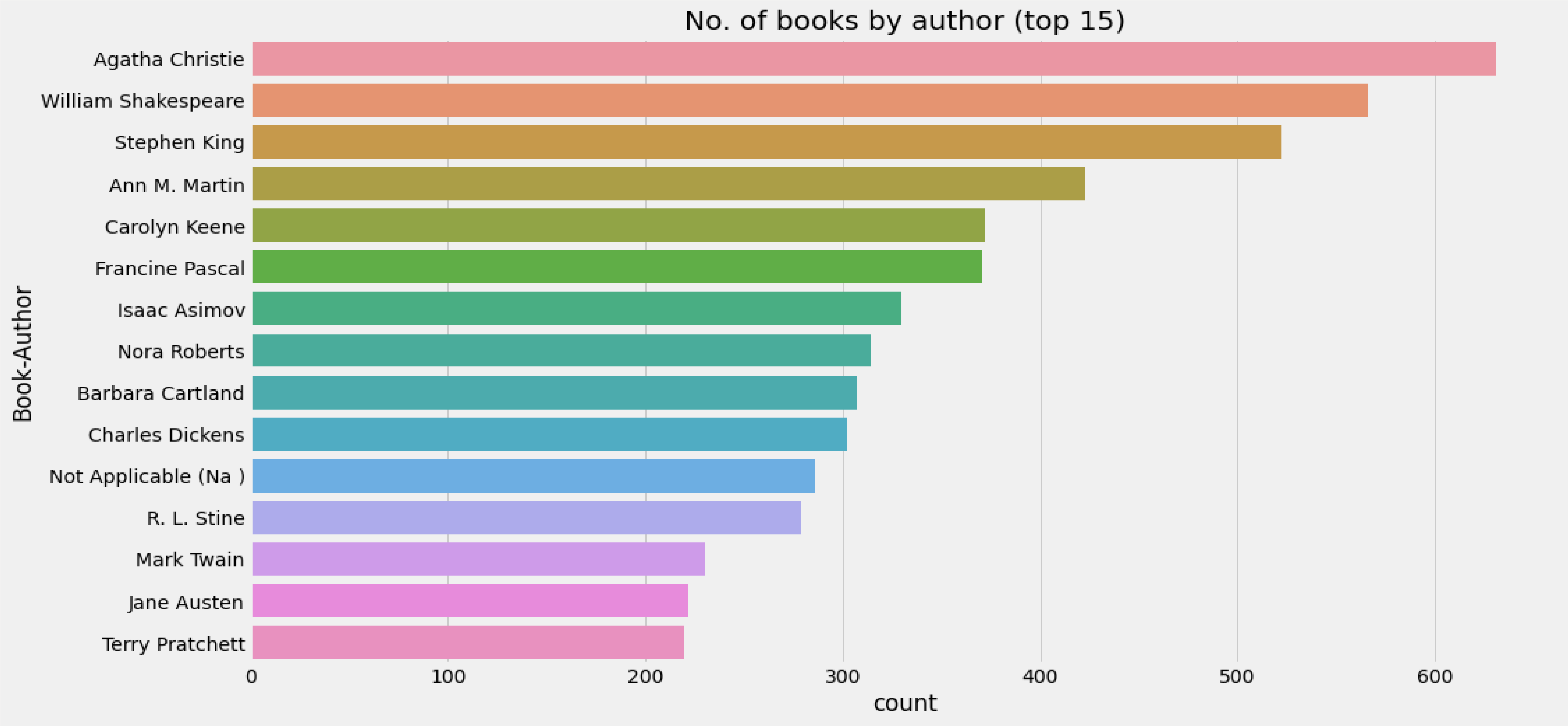
Book Ratings Count:

- It can be observe that higher ratings are more common amongst users and rating 8 has been rated highest number of times
- Higher ratings are more common amongst users



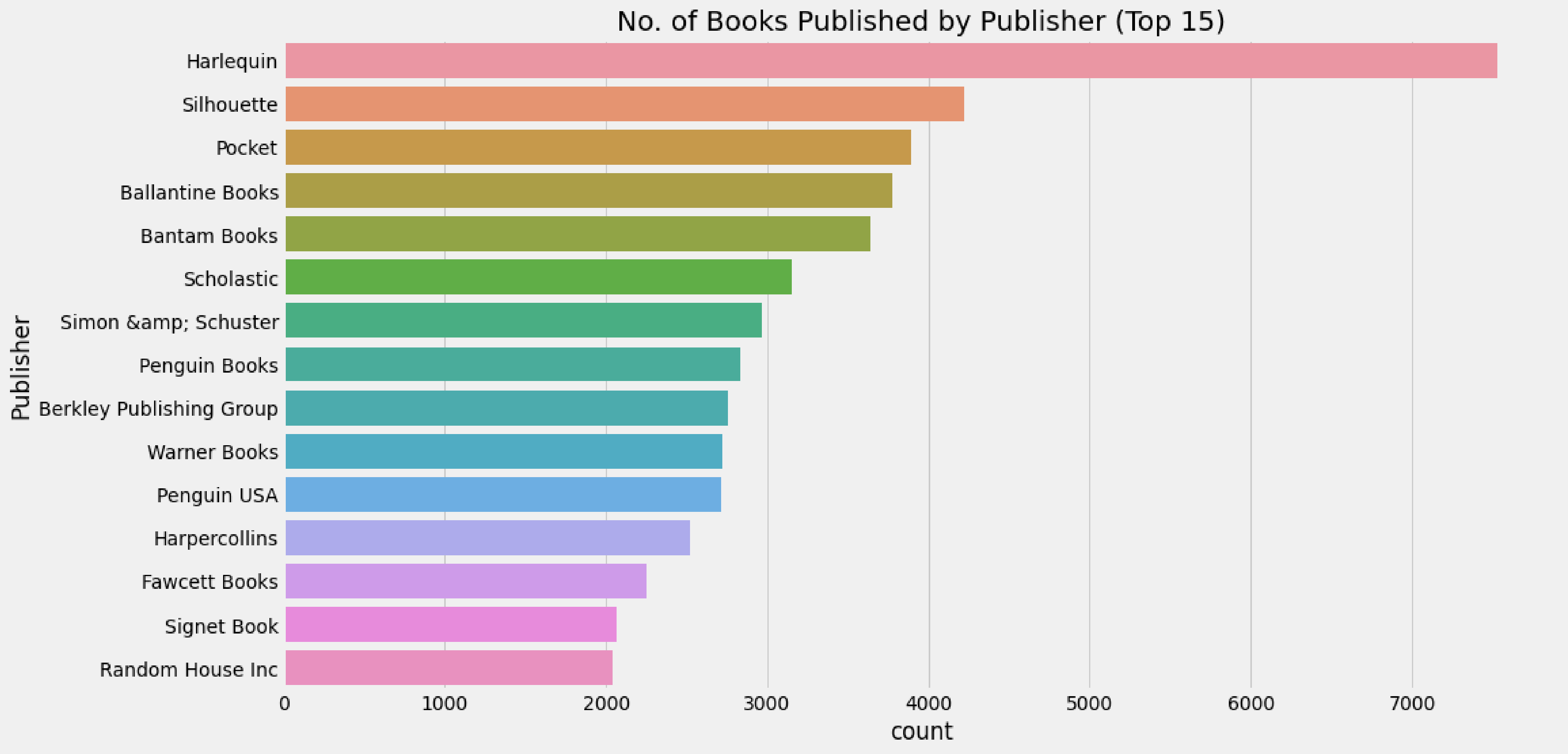
top 15 book authors based on their book count

Agatha christie are highest number of books write in our given dataset



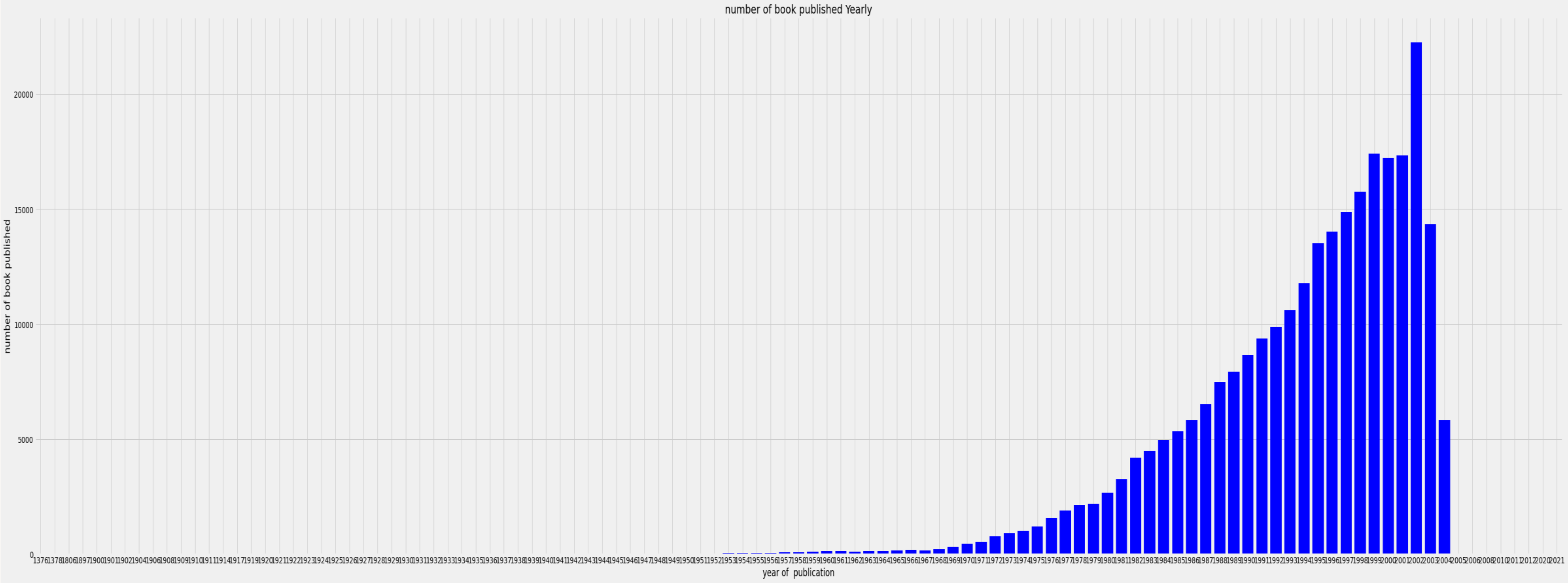
Top 15 Publishers based on their books Published

- Harlequin published highest number of books in our given dataset



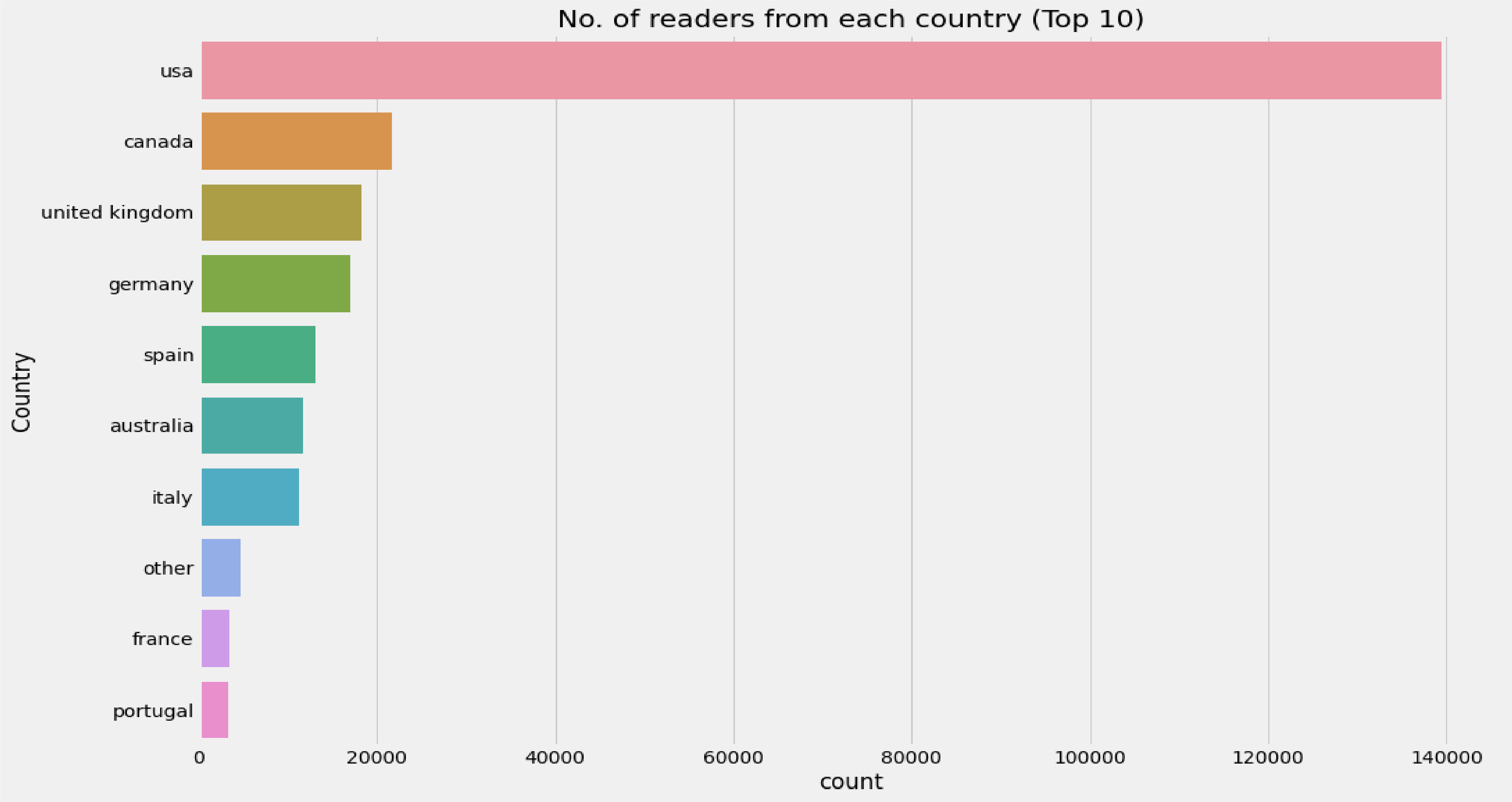
Number of book published by year count:

- So we can observe that publication years are somewhat between 1950 - 2005 here.



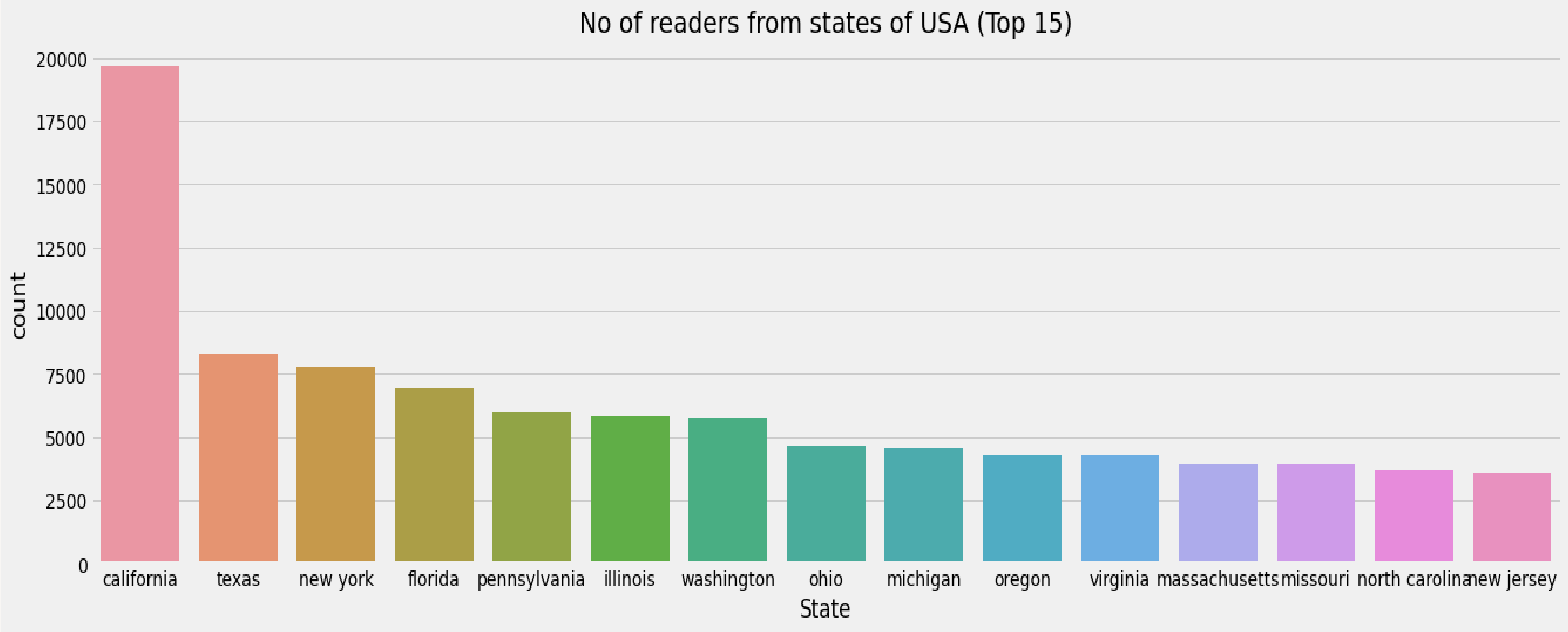
Top 10 countries based on number of readers :

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



Top 15 readers from states of USA :

- Splitting Location column and analyzing State of USA.
- Most active readers are from California



Technology Used

- Google Colab
- python
- seaborn
- Metplotlib
- Numpy
- Pandas
- Flask

Model’s Performed

- Popularity Based Recommendation

  Recommend the top rating books

- Collaborative

  Recommend items those are preferred by similar users Content-based

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

- Content-based

  Recommend items based on similarity between items and user's preferences Hybrid

- Hybrid

  Combines both

Popularity Based Recommendation

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

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

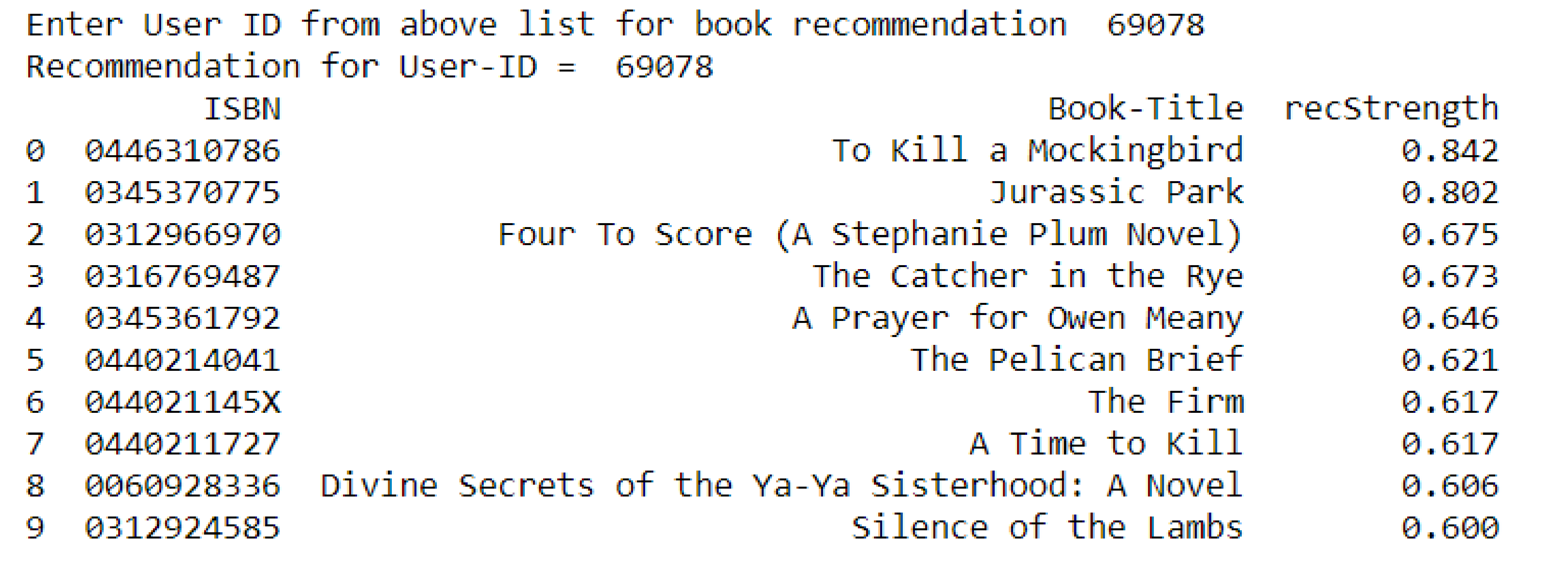
Where,

- ❖ 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.

Collaborative Filtering - (Item-Item based)

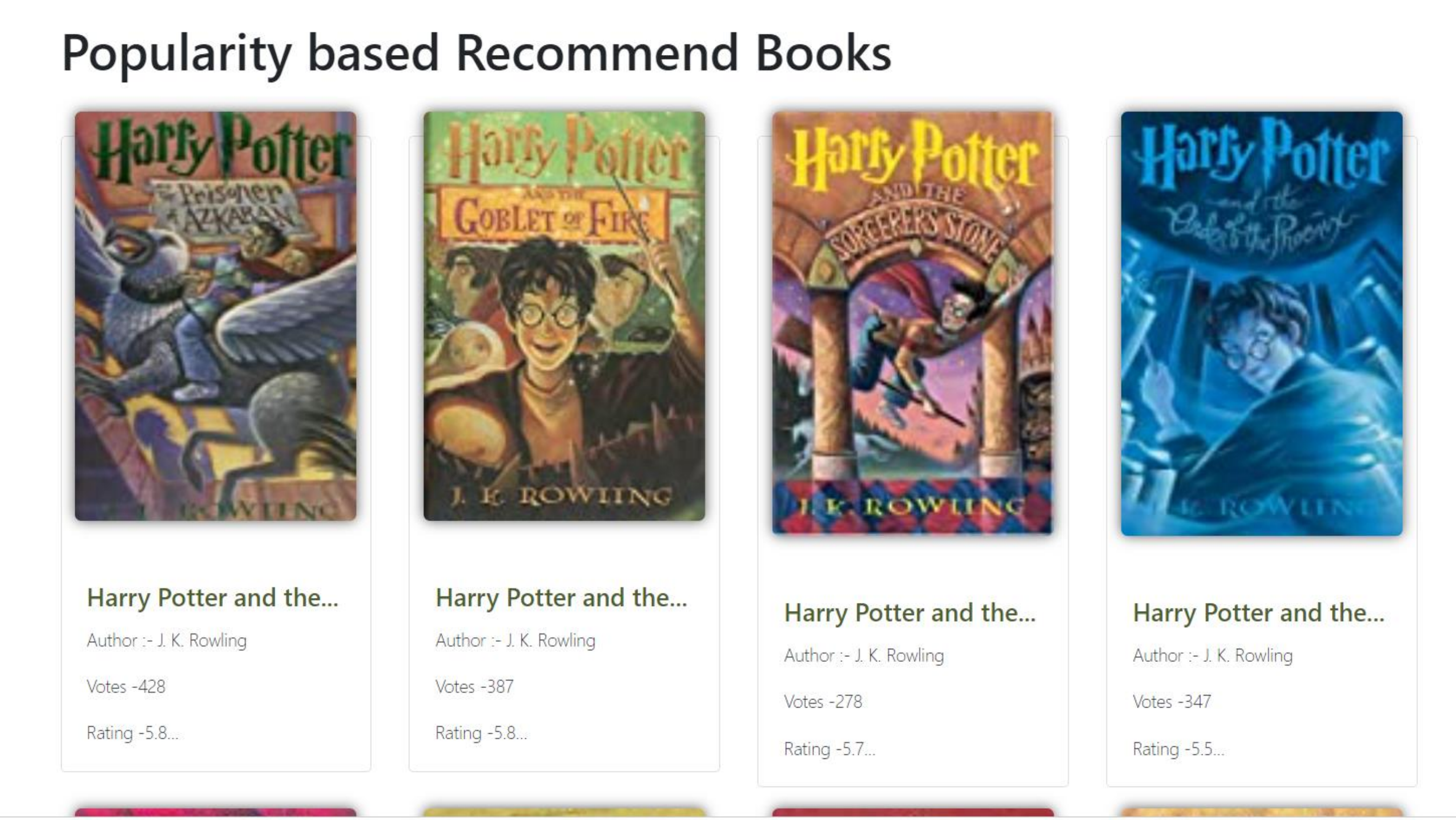
index	Book-Title	num_ratings	avg_ratings
80433	Harry Potter and the Prisoner of Azkaban (Book 3)	428	5.852803738317757
80421	Harry Potter and the Goblet of Fire (Book 4)	387	5.824289405684754
80440	Harry Potter and the Sorcerer's Stone (Book 1)	278	5.737410071942446
80425	Harry Potter and the Order of the Phoenix (Book 5)	347	5.501440922190202
80413	Harry Potter and the Chamber of Secrets (Book 2)	556	5.183453237410072

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
```

Popularity Based Recommendation (user Interface)



Popularity based Recommend Books

Harry Potter and the...

Author :- J. K. Rowling

Votes -428

Rating -5.8...

Harry Potter and the...

Author :- J. K. Rowling

Votes -387

Rating -5.8...

Harry Potter and the...

Author :- J. K. Rowling

Votes -278

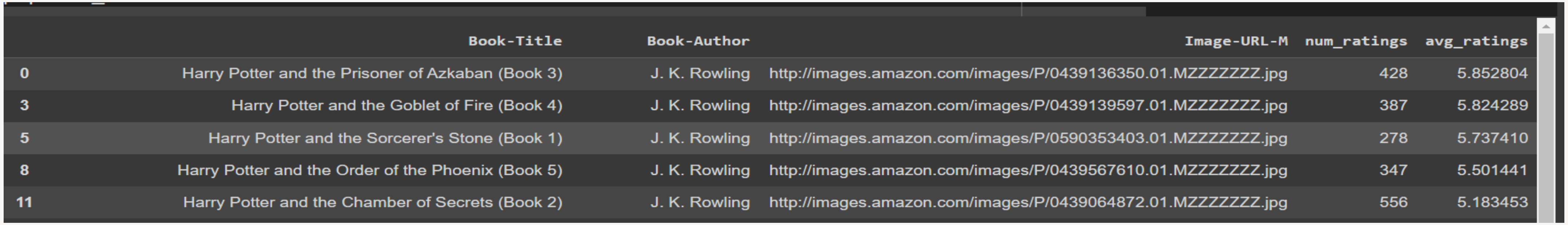
Rating -5.7...

Harry Potter and the...

Author :- J. K. Rowling

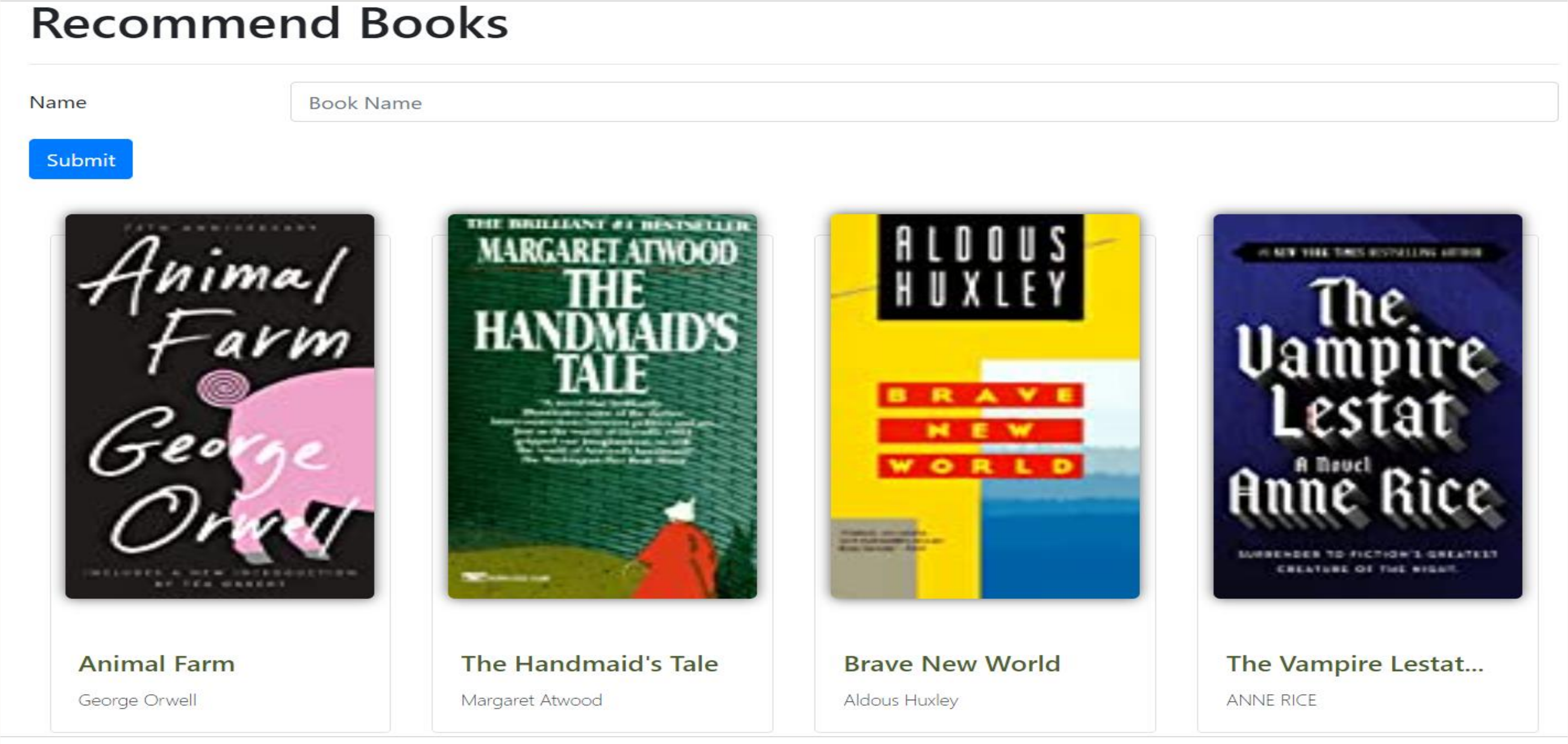
Votes -347

Rating -5.5...



	Book-Title	Book-Author	Image-URL-M	num_ratings	avg_ratings
0	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	http://images.amazon.com/images/P/0439136350.01.MZZZZZZZ.jpg	428	5.852804
3	Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling	http://images.amazon.com/images/P/0439139597.01.MZZZZZZZ.jpg	387	5.824289
5	Harry Potter and the Sorcerer's Stone (Book 1)	J. K. Rowling	http://images.amazon.com/images/P/0590353403.01.MZZZZZZZ.jpg	278	5.737410
8	Harry Potter and the Order of the Phoenix (Book 5)	J. K. Rowling	http://images.amazon.com/images/P/0439567610.01.MZZZZZZZ.jpg	347	5.501441
11	Harry Potter and the Chamber of Secrets (Book 2)	J. K. Rowling	http://images.amazon.com/images/P/0439064872.01.MZZZZZZZ.jpg	556	5.183453

Collaborative Filtering(User Interface): Snapshot



Recommend Books

Name: Book Name

Submit

Animal Farm
George Orwell

The Handmaid's Tale
Margaret Atwood

Brave New World
Aldous Huxley

The Vampire Lestat...
ANNE RICE

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)

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

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