Capstone Project

Book Recommendation System

Unsupervised Machine Learning

INDEX

Book Recommendation System

Problem Description:-

During the last few decades, with the rise of Youtube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys. In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries). Recommendation systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. The main objective is to create a book recommendation system for users.

Concept of Recommender system

AI

A book recommendation system is a type of recommendation system where we have to recommend similar books to the reader based on Content filtering or Collaborative filtering or Hybrid filtering. The books recommendation system is used by online websites which provide ebooks like google play books, open library, good read's, etc. A recommendation system helps an organization to create loyal customers and build trust by them desired products and services for which they came on your site. The recommendation system today are so powerful that they can handle the new customer too who has visited the site for the first time. They recommend the products which are currently trending or highly rated and they can also recommend the products which bring maximum profit to the company.

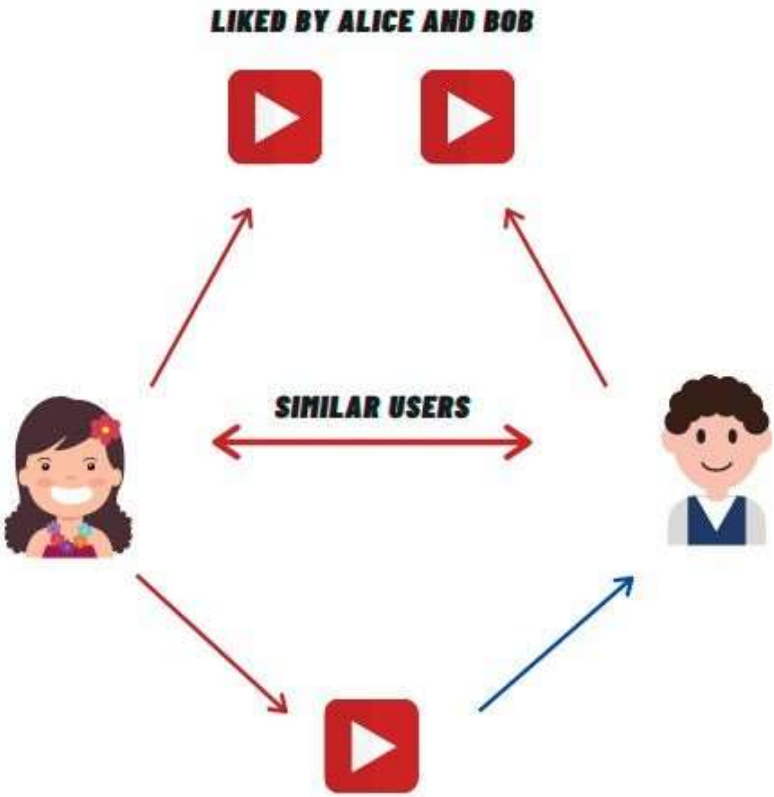
Limitations:-

Users will only get recommendations related to their preferences in their profile, and recommender engine may never recommend any item with other characteristics. As it is based on similarity among items and users, it is not easy to find the neighbour users.

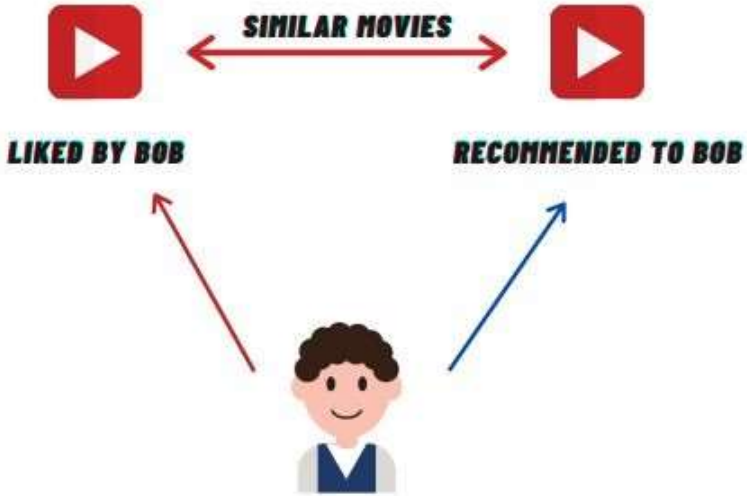
Benefits:-

- Drive Traffic.
- Deliver Relevant Content.
- Engage Shoppers.
- Convert Shoppers to Customers.
- Increase Average Order Value.
- Increase Number of Items per Order.
- Control Merchandising and Inventory Rules.
- Reduce Workload and Overhead.

COLLABORATIVE FILTERING



CONTENT-BASED FILTERING



Data Pipeline

- Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to see the trend.
- Data Processing: In this part we went through each attributes and encoded the categorical features.
- Model Creation: Finally in this part we created the various models. These various models are being analysed and we tried to study various models so as to get the best performing model for our project.

Data Description

The Book-Crossing dataset comprises 3 files:

❑ Users_dataset: Contains the users information

- User-ID (unique for each user)
- Location (contains city, state and country separated by commas)
- Age Shape of Dataset - (278858, 3)

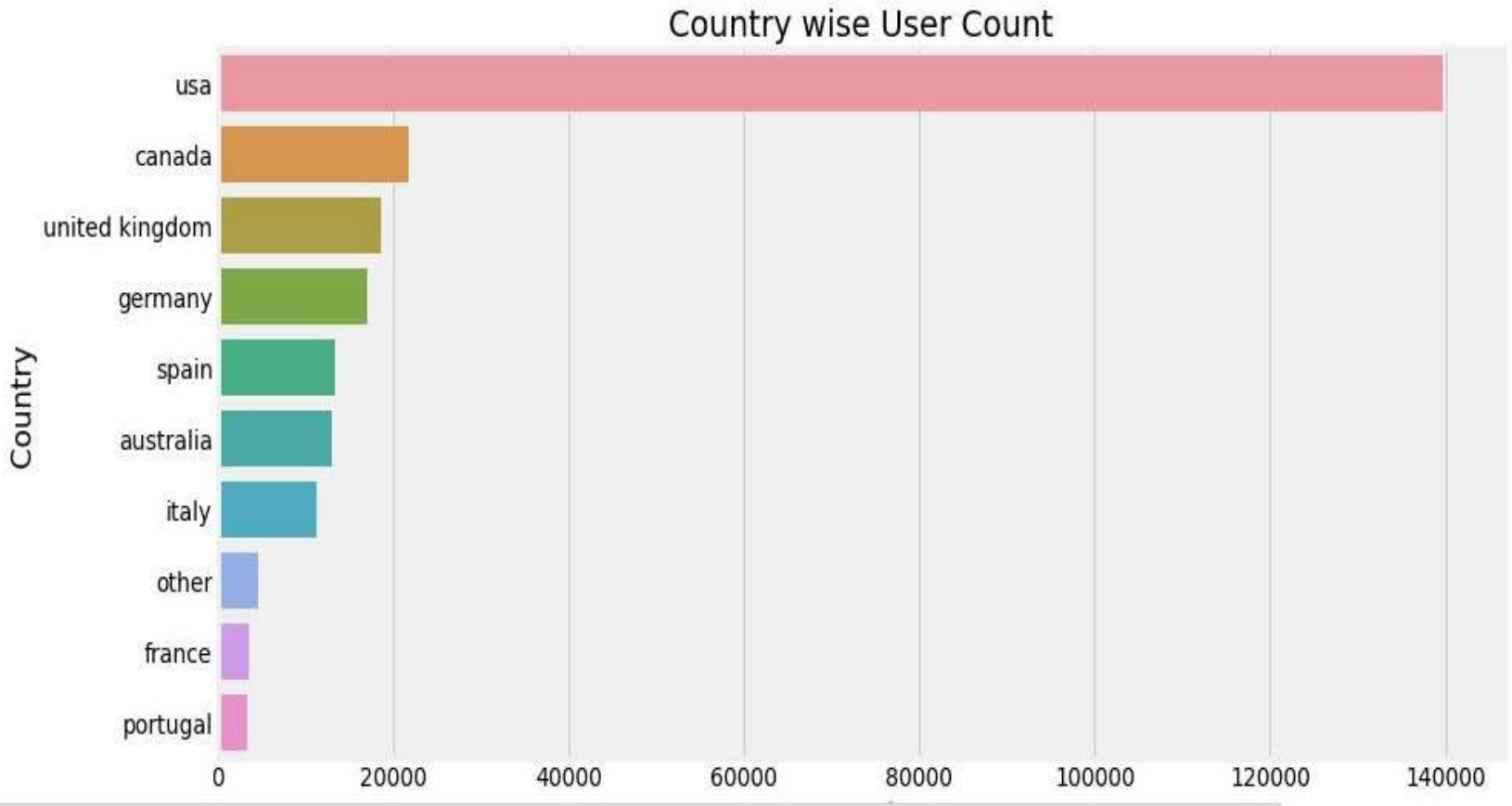
❑ Books_dataset:

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

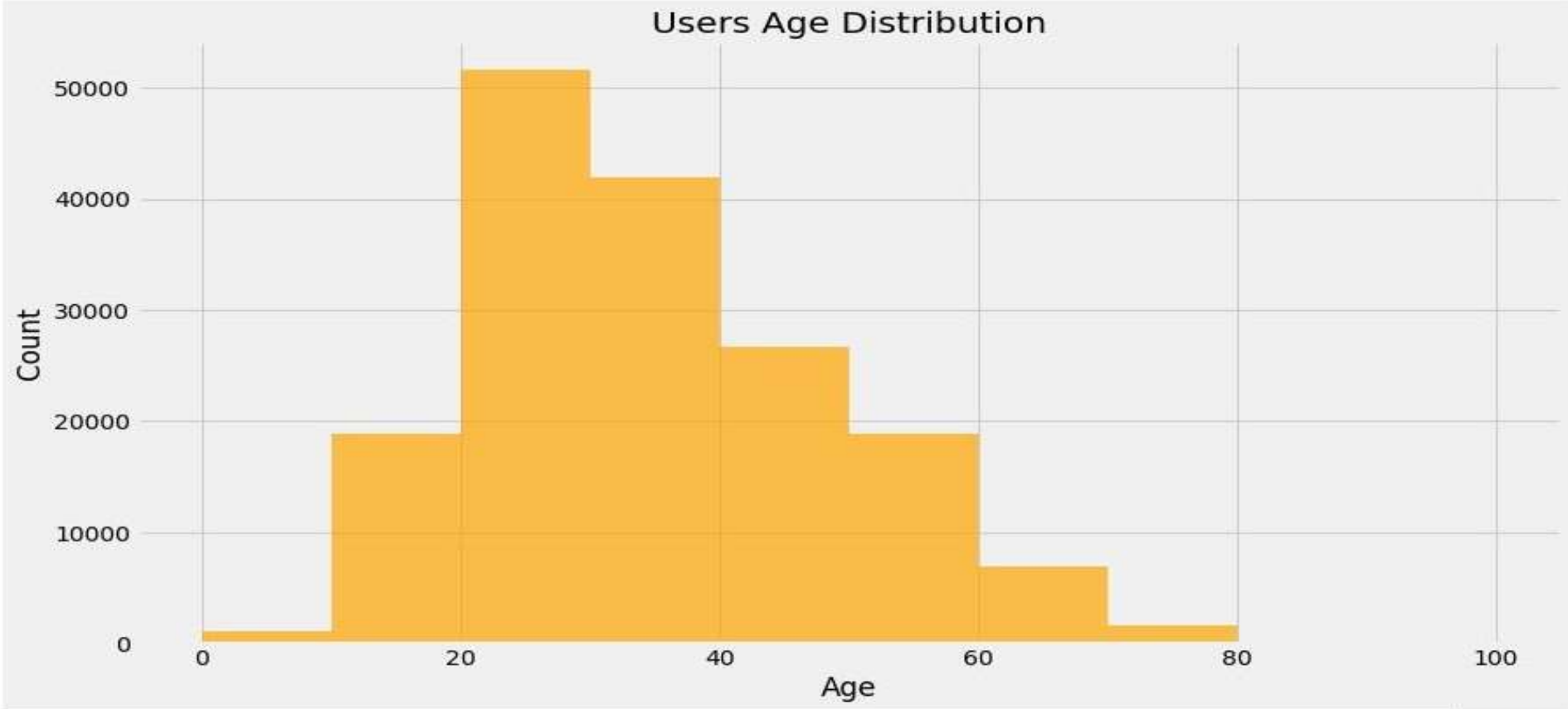
❑ Ratings_dataset: Contains the book rating information.

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

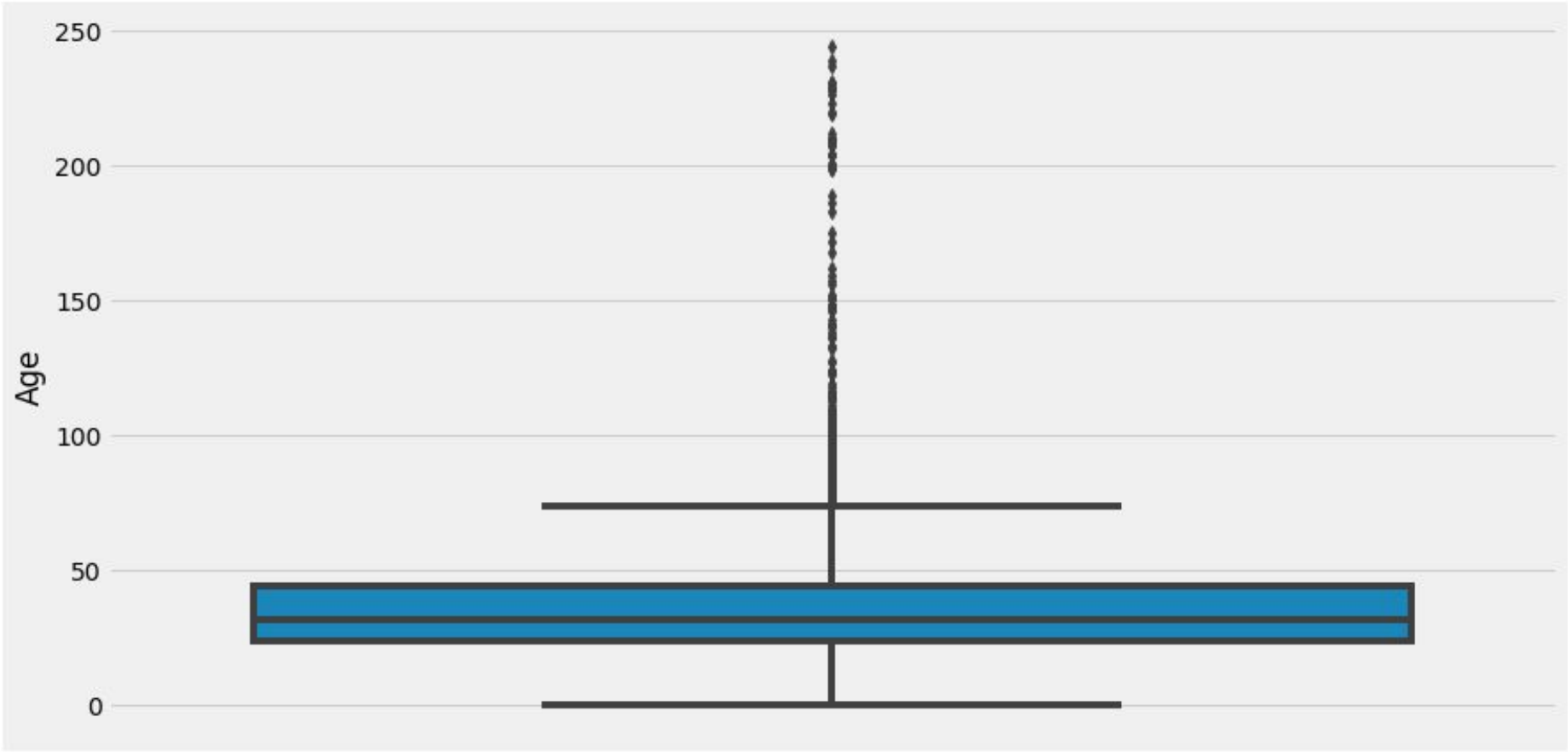
EDA – User Count By Top Countries



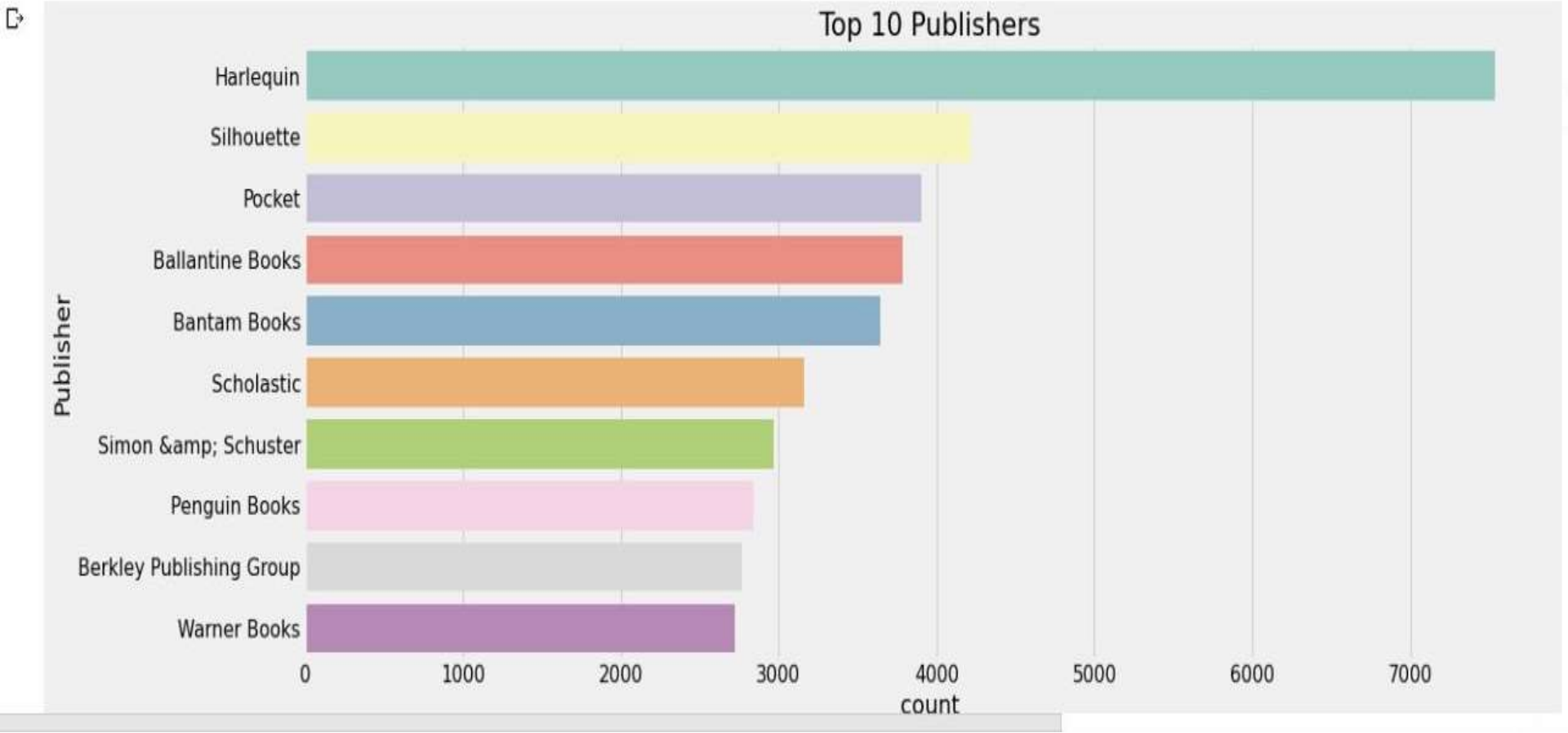
EDA – Age Distribution



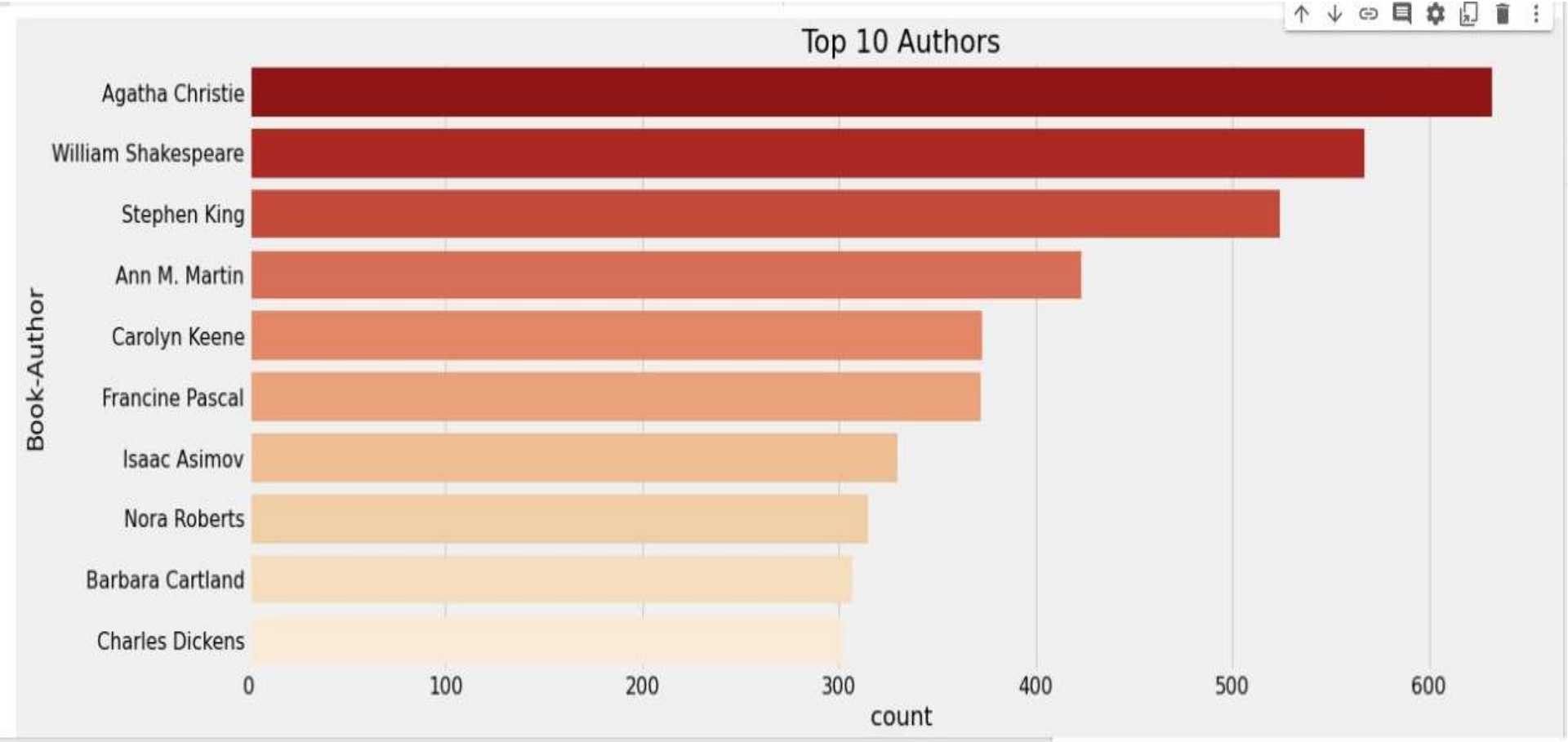
EDA – Outliers In Age Features



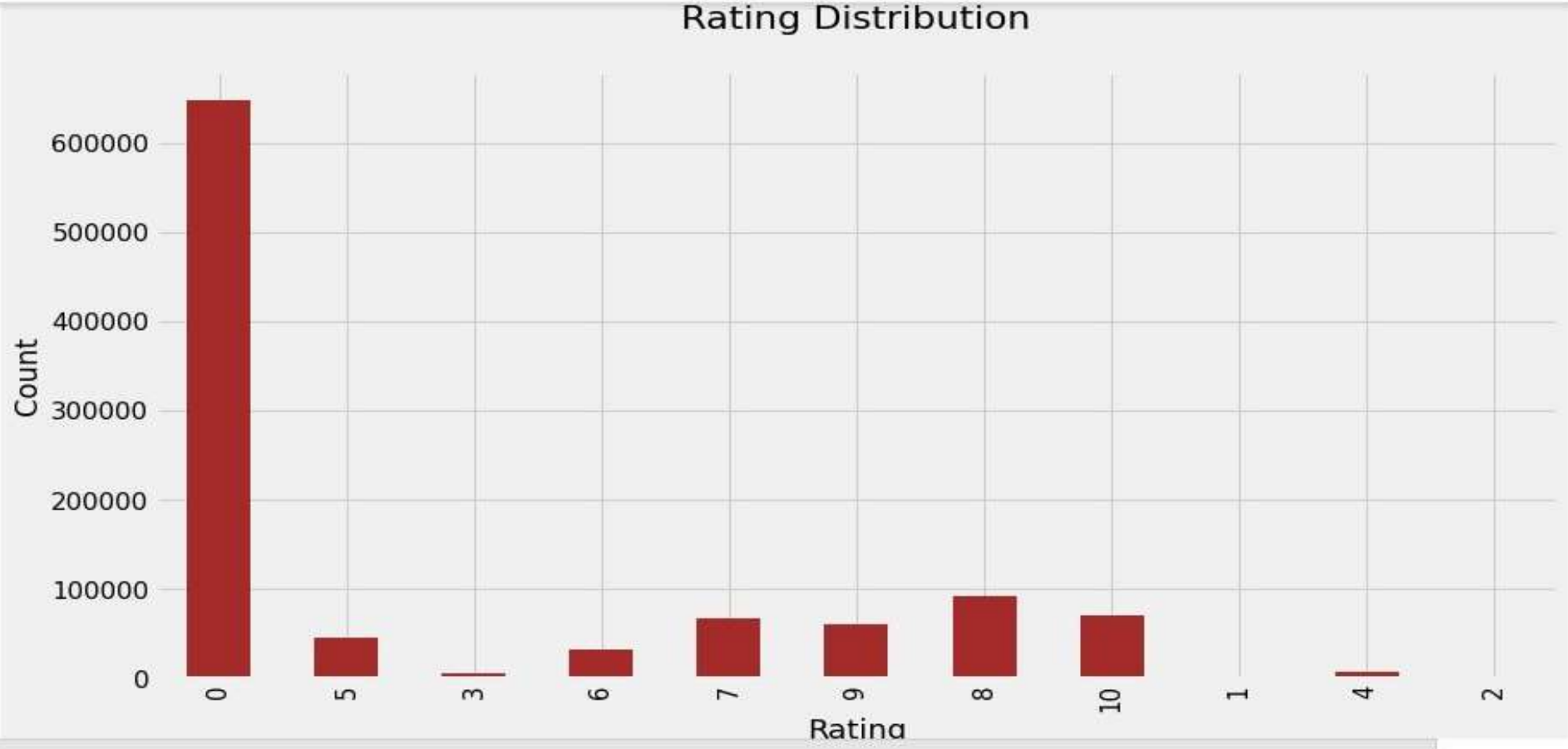
EDA – Top 10 Publisher



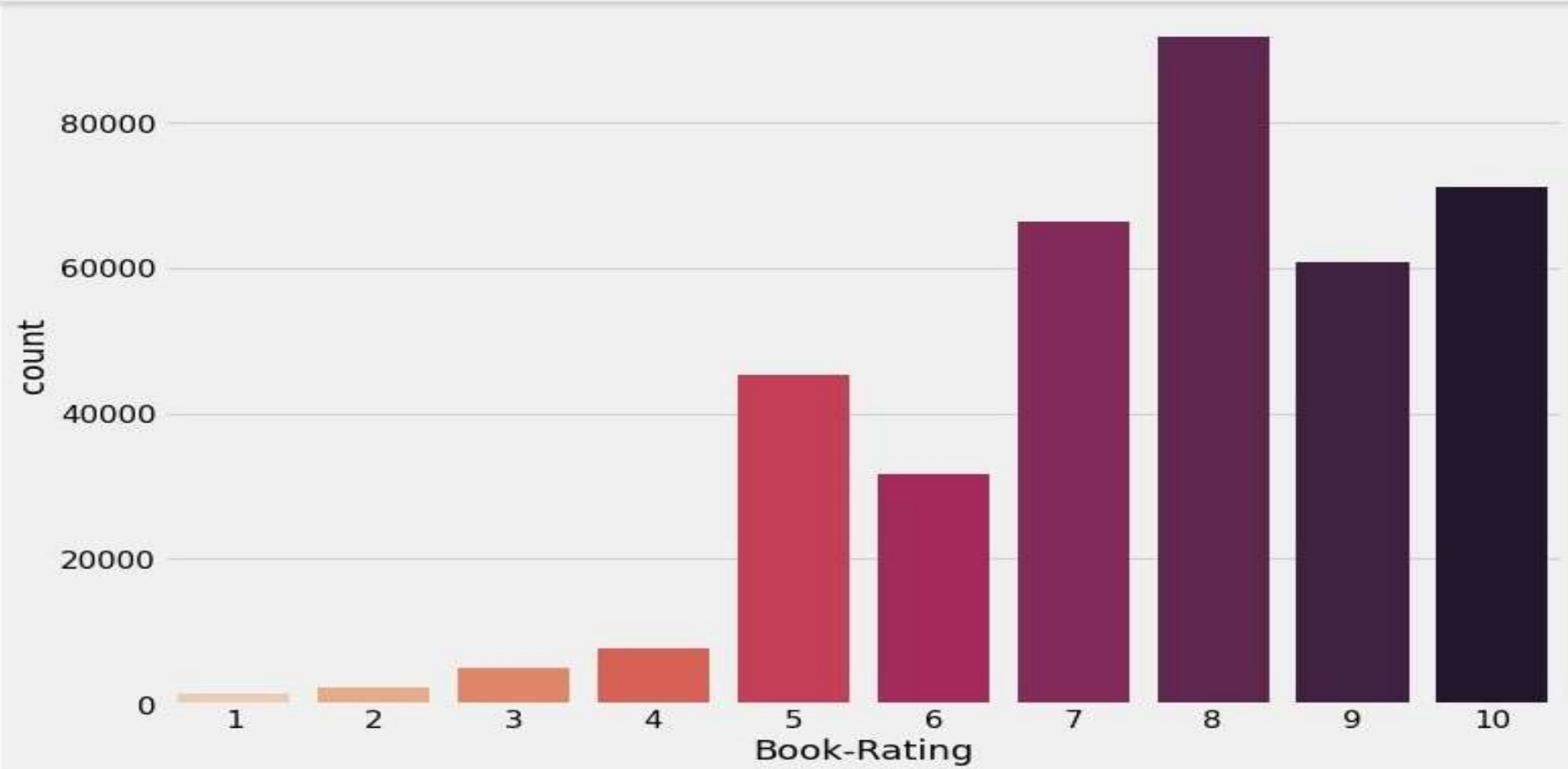
EDA – Top 10 Authors



EDA – Rating Dataset



EDA – Book Rating



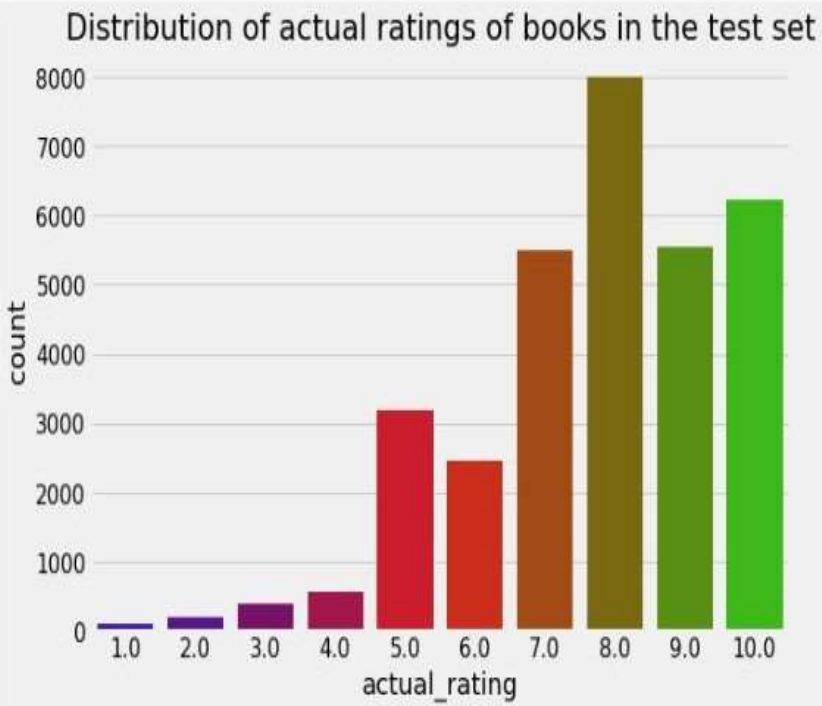
Merging The Dataset & Null Value Imputation

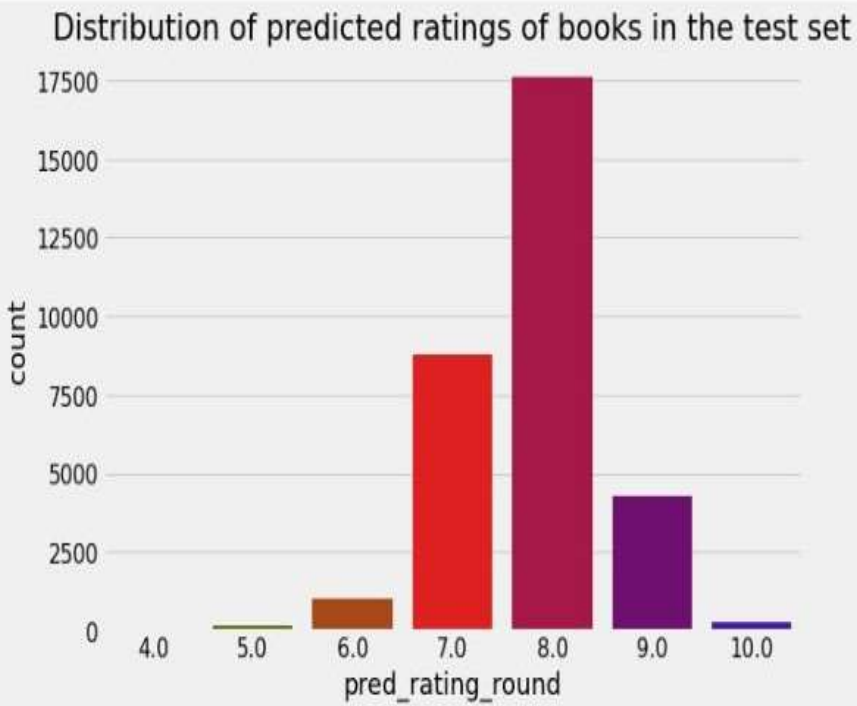
	index	Missing Values	% of Total Values	Data_type
0	User-ID	0	0.0	int64
1	Age	0	0.0	float64
2	Country	0	0.0	object
3	ISBN	0	0.0	object
4	Book-Rating	0	0.0	int64
5	Avg_Rating	0	0.0	float64
6	Total_No_Of_Users_Rated	0	0.0	int64
7	Book-Title	0	0.0	object
8	Book-Author	0	0.0	object
9	Year-Of-Publication	0	0.0	float64
10	Publisher	0	0.0	object

Model’s Performed

- Popularity Based Recommendation.
- Model based collaborative filtering.
- Collaborative Filtering-(Item-Item based).
- Collaborative Filtering-(User-Item based).

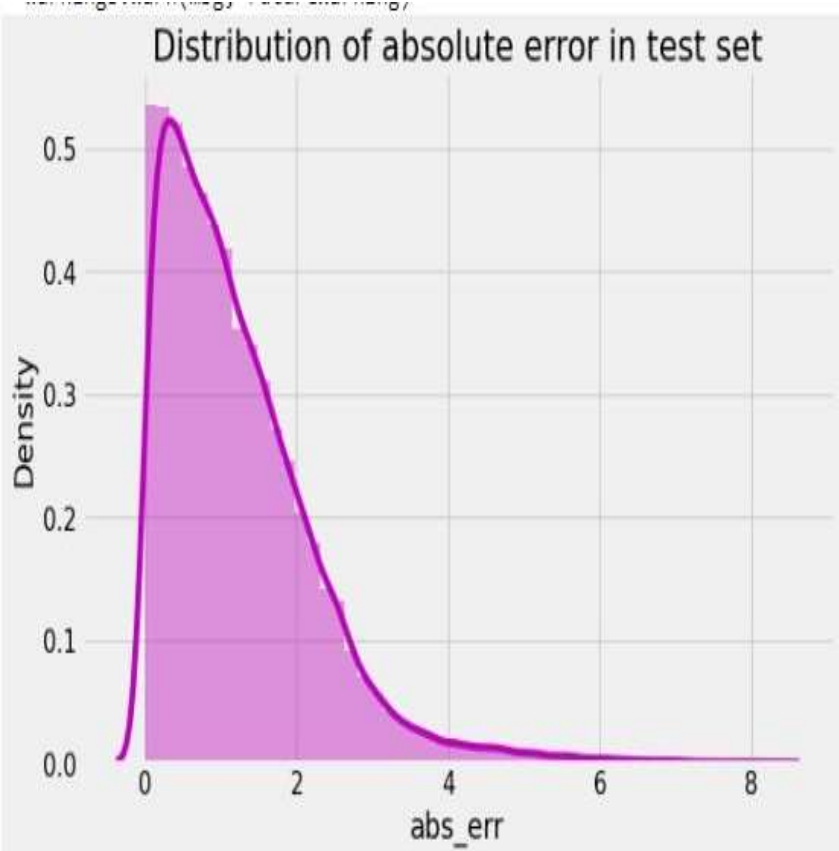
SVD Model Result

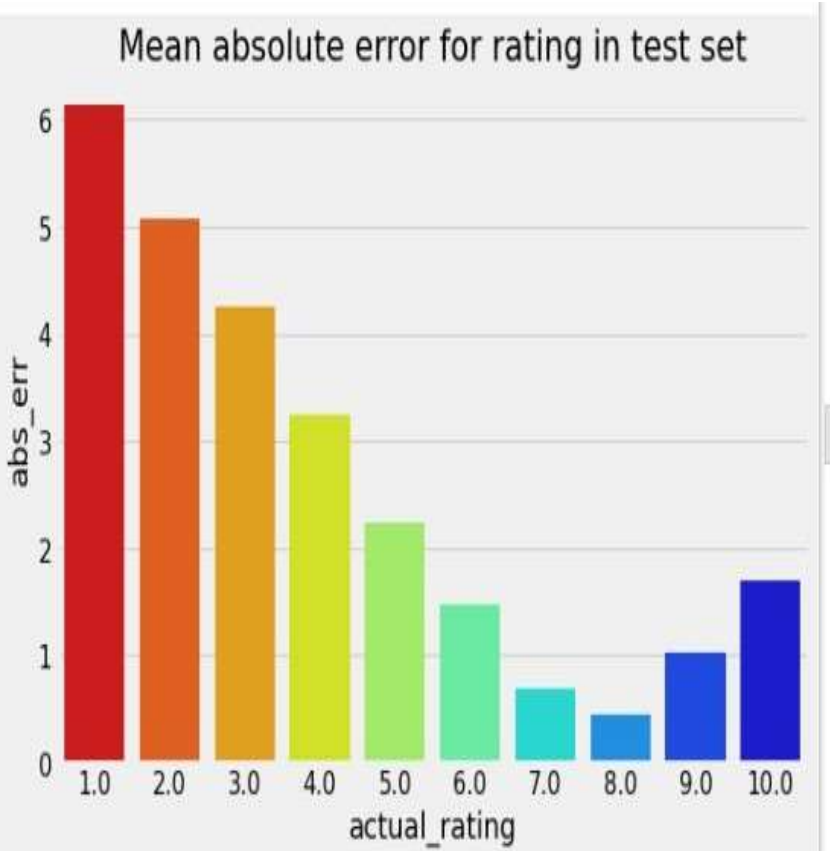




Absolute error of predicted ratings

SVD Model Result





Evaluation

In Recommender Systems, there are a set metrics commonly used for e valuation. We choose to work with TopN accuracy metrics, which evaluates the accuracy of the top recommendations provided to a user, comparing to the items the user has actually interacted in test set.

This evaluation method works as:

- For each user
- For each item the user has interacted in test set
- Sample 100 other items the user has never interacted.
- Ask the recommender model to produce a ranked list of recommended items, from a set composed of one interacted item and the 100 non-interacted items.
- Compute the TopN accuracy metrics for this user and interacted item from the recommendations ranked list.
- Aggregate the global Top-N accuracy metrics.

Model Evaluation

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	User-ID
10	265	335	1389	0.191	0.241	11676
31	187	244	1138	0.164	0.214	98391
45	23	28	380	0.061	0.074	189835
30	85	100	369	0.230	0.271	153662
70	29	35	236	0.123	0.148	23902
7	29	48	204	0.142	0.235	235105
47	25	30	203	0.123	0.148	76499
50	23	36	193	0.119	0.187	171118
42	60	72	192	0.312	0.375	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.
- We can conclude that item-item based collaborative filtering performed better than user-user based collaborative filtering because of lower computation among the memory based approach.
- 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

- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.
- Understanding the metric for evaluation was a challenge as well.
- Decision making on missing value imputations and outlier treatment was quite challenging as well.
- As dataset was quite big enough which led more computation time.

Future Scope

- The future of recommender systems lie in integrating self actualization to do justice to serendipity while recommending which will also support rather than replace human decision-making by understanding preferences.
- Recommender systems can be broadly divided into: Collaborative filtering, Content-based filtering, Hybrid recommender systems, Personality-based recommender systems. Each type of filtering algorithm is used according to the specific need of the application or the product.
- Here we are trying to achieve a recommendation which is not extremely personalised which may feel intrusive to the user and not very generic that it doesn't really account the user's distinct taste. Striking a balance between the two is what needs to be achieved.
- Furthermore there will be plenty of work needed to be done on the famous 'cold start problem' in order to somehow manage collecting just the right amount of implicit information and data to recommend users even if there is little or no direct information available on users.

THANK YOU