CAPSTONE PROJECT

BOOK RECOMMENDATION SYSTEM

BY AMAL KRISHNA N

CONTENT:

- PROBLEM ANALYSIS AND DATA UNDERSTANDING
- UNDERSTANDING BUSINESS PROBLEM
- DATA CLEANING AND PROCESSING
- EDA
- MODELING
- CONCLUSION

PROBLEM STATEMENT

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.

DATA SUMMARY

- Available dataset comprises of 3 files.
- ● Users
- Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.
- ● Books
- › Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover,
- › some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services.
- › Note that in the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website.
- ● Ratings
- Contains the book rating information. Ratings (Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

UNDERSTANDING THE BUSINESS PROBLEM

- ✓ **Increased sales:** increase sales by suggesting books that users are likely to purchase. This can be achieved by making personalized recommendations that match the user's interests and preferences.
- ✓ **Customer satisfaction:** A good recommendation system should be able to provide users with a satisfying and enjoyable experience. This can be achieved by making relevant and accurate recommendations and by providing a user-friendly interface.
- ✓ **User engagement:** A recommendation system can help to keep users engaged with a business's platform by providing a continuous stream of recommendations and new content. This can lead to increased customer loyalty and a longer user lifespan.
- ✓ **Improved user experience:** A recommendation system can help businesses to improve the overall user experience by making it easier for users to discover new books and authors that they may enjoy.
- ✓ **Competitive advantage:** A well-designed recommendation system can provide businesses with a competitive advantage over their rivals, as it can help to attract and retain customers.
- ✓ **Data-driven decision making:** A recommendation system can provide businesses with insights into user preferences and reading habits, which can be used to inform marketing and content strategy decisions.

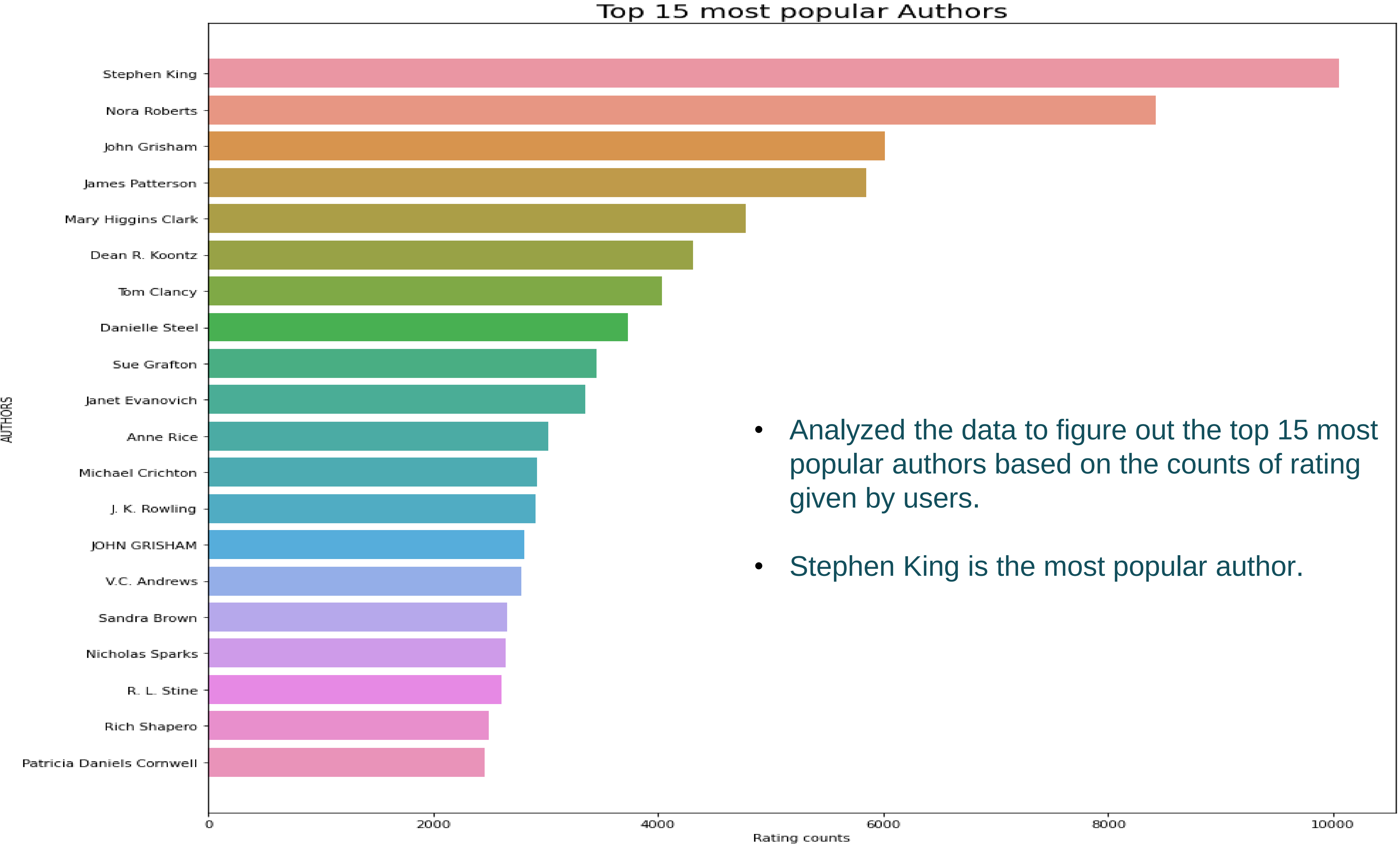
DATA CLEANING AND PROCESSING

- In books dataset books author column have only 1 null value and publisher column have 2 null values
- In users dataset there is a huge amount of null values in age column(110762)
- Basically, missing values of age column will not much more effect on our final aim of recommendation
- Rating dataset does not contain any kind of null values.
- dropping Null value entries, as a best option for recommendation system. And also there is no possibility to fill with approximate values or it is not much worth here.
- dropping 2 of 3 image URL columns and holding large image url column
- dropping user id column which is not needed
- And there are no duplicate entries in all 3 datasets.

Feature engineering

- Merged the books and rating data sets
- Added 2 more columns in the merged data frame i.e., average rating and number of ratings.
- which is mainly useful for EDA part as well as recommendation part

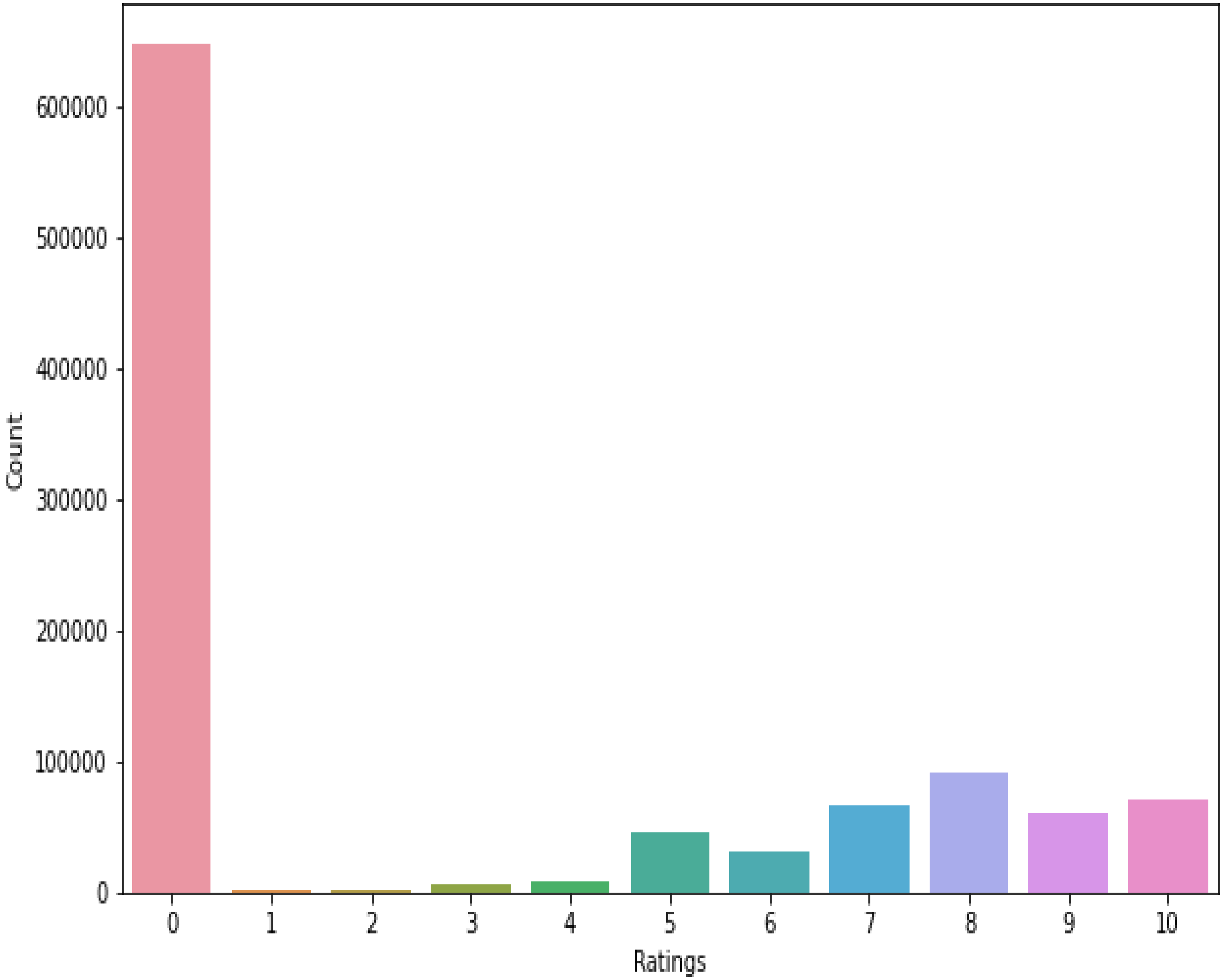
EXPLORATORY DATA ANALYSIS(EDA)



- Analyzed the data to figure out the top 15 most popular authors based on the counts of rating given by users.

- Stephen King is the most popular author.

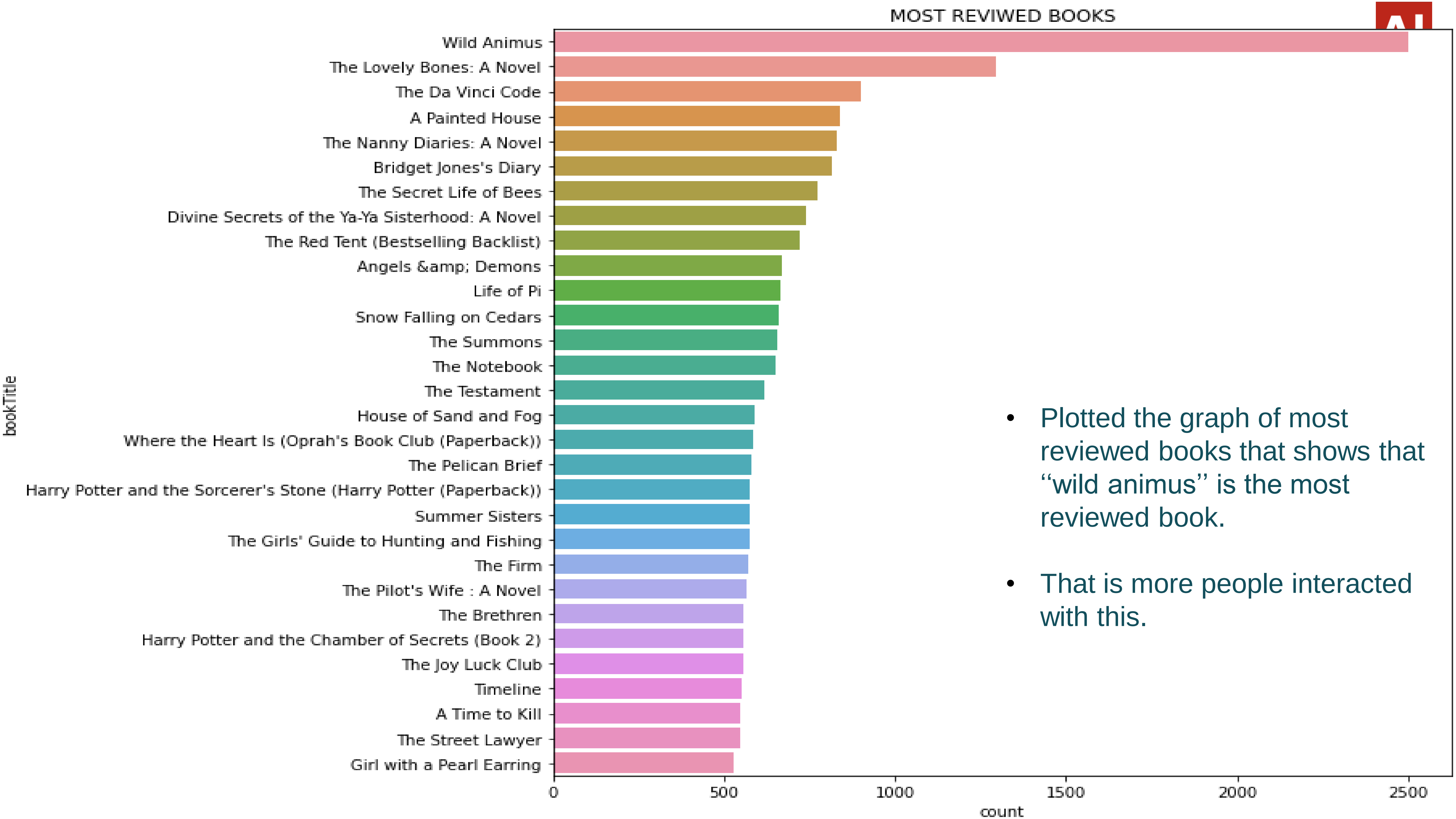
COUNT PLOT OF RATINGS



- Plotted bar chart to find which range of rating most people are given to books.

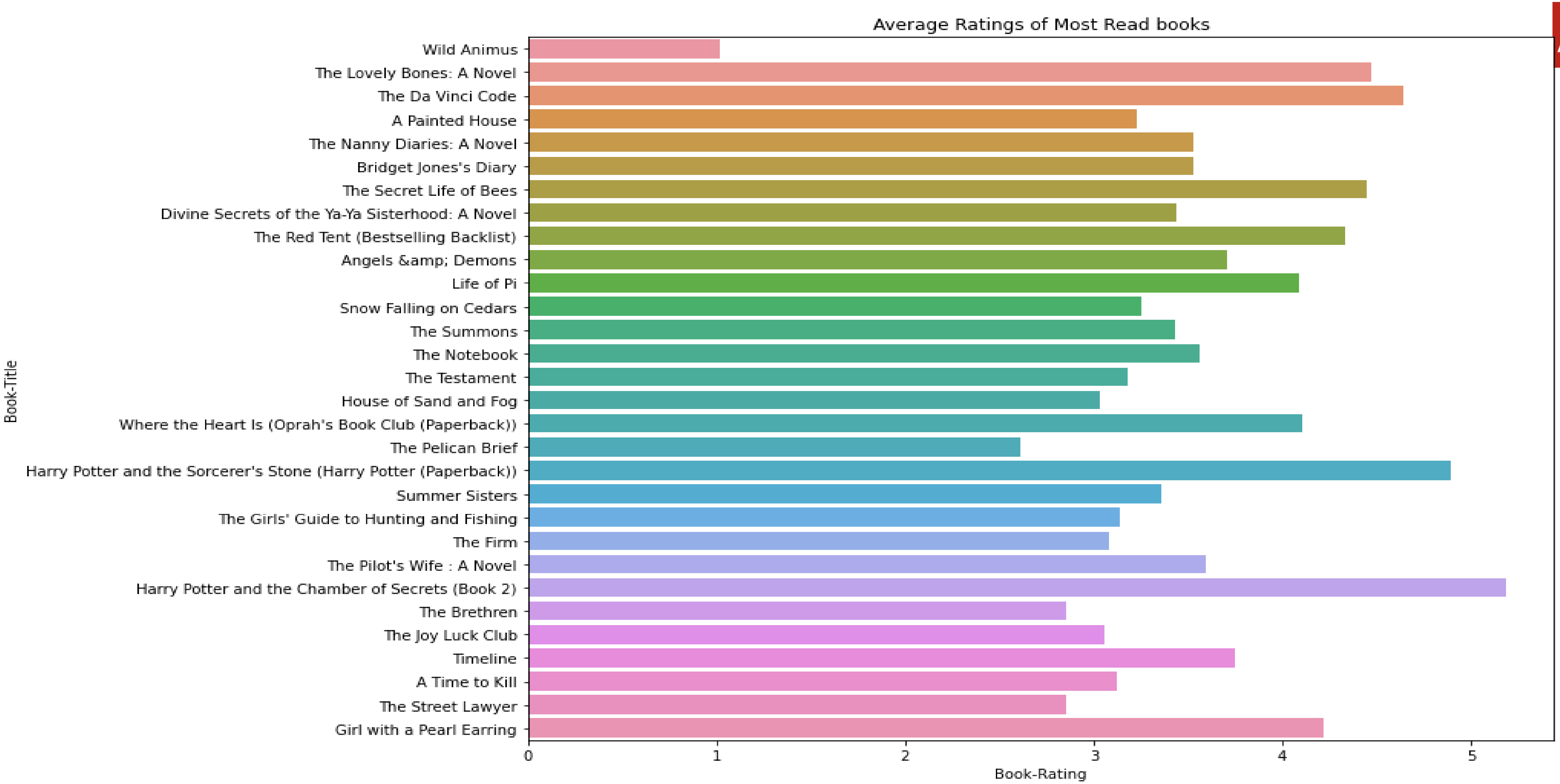
- It showed that the zero rating is more, because most of the people leave without giving any kind of review to books.

- After that, the most given rating is 8. Which means people likes available books



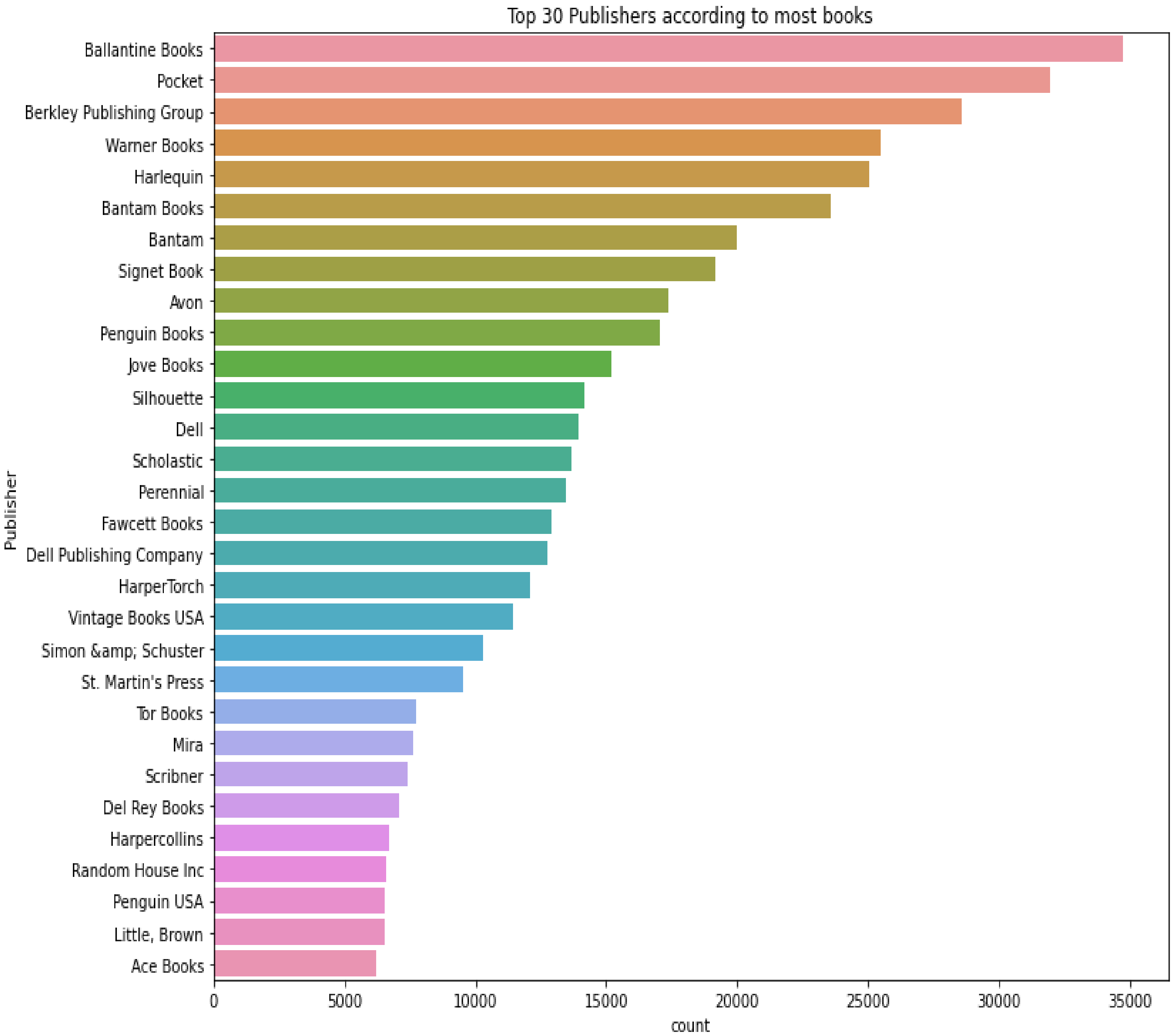
- Plotted the graph of most reviewed books that shows that ‘‘wild animus’’ is the most reviewed book.

- That is more people interacted with this.



- Need to go deeper into the most reviewed books data, so, took the average of ratings of most reviewed books.
- It shows that even if the books have more reviews by number but are less liked by people.

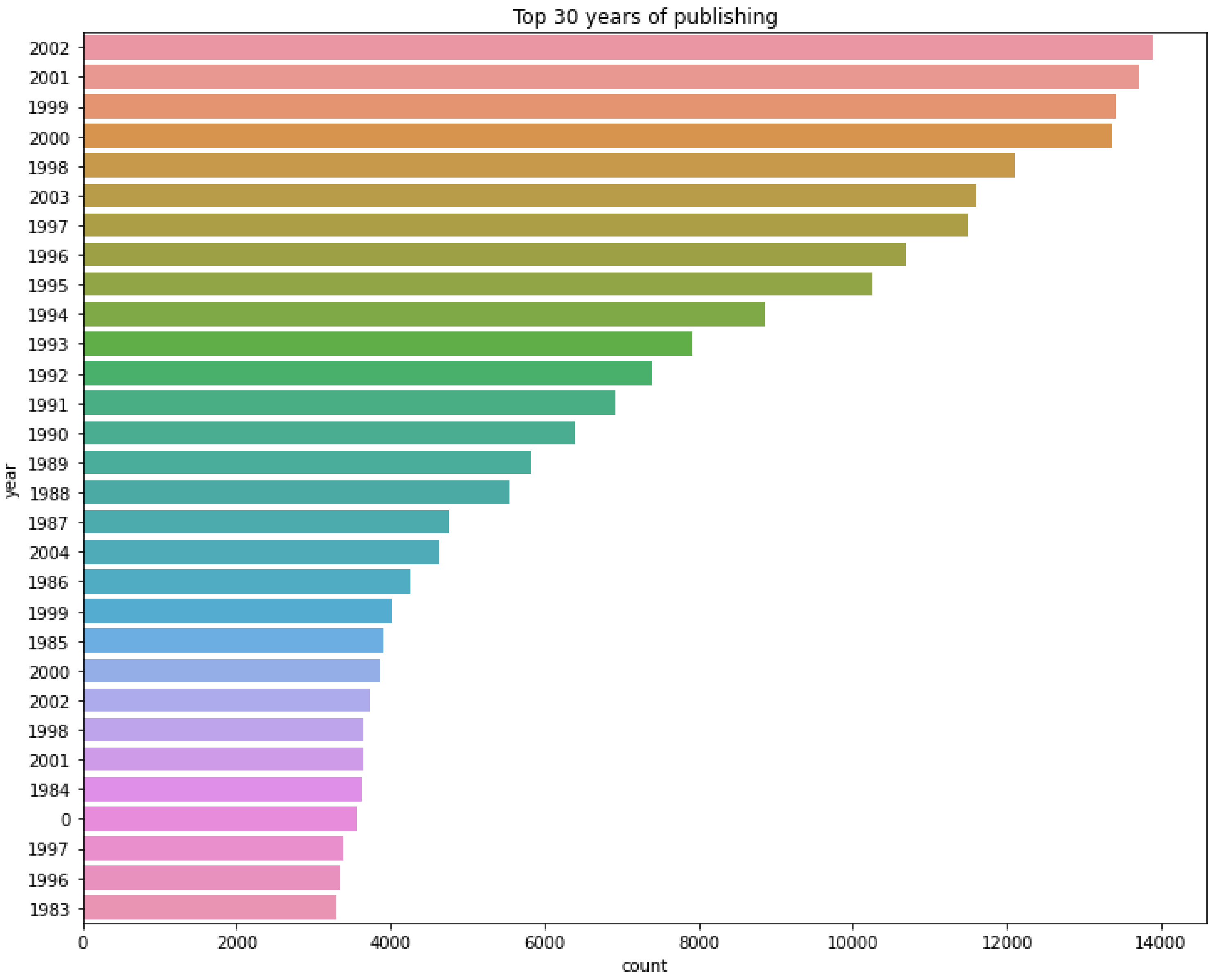
‘Wild animus’ has a very low rating around 1.



- Visualized the top 30 publishers. ‘Ballantine books’ are the top publisher with around 35000 published books.

- Second top publisher is ‘pocket’ also have big number of published books

- Third one is ‘Berkley Publishing Group’ with below 30000 of books



- When we look at the most books published, the year 2002 is the top one year with around 14000 books published.

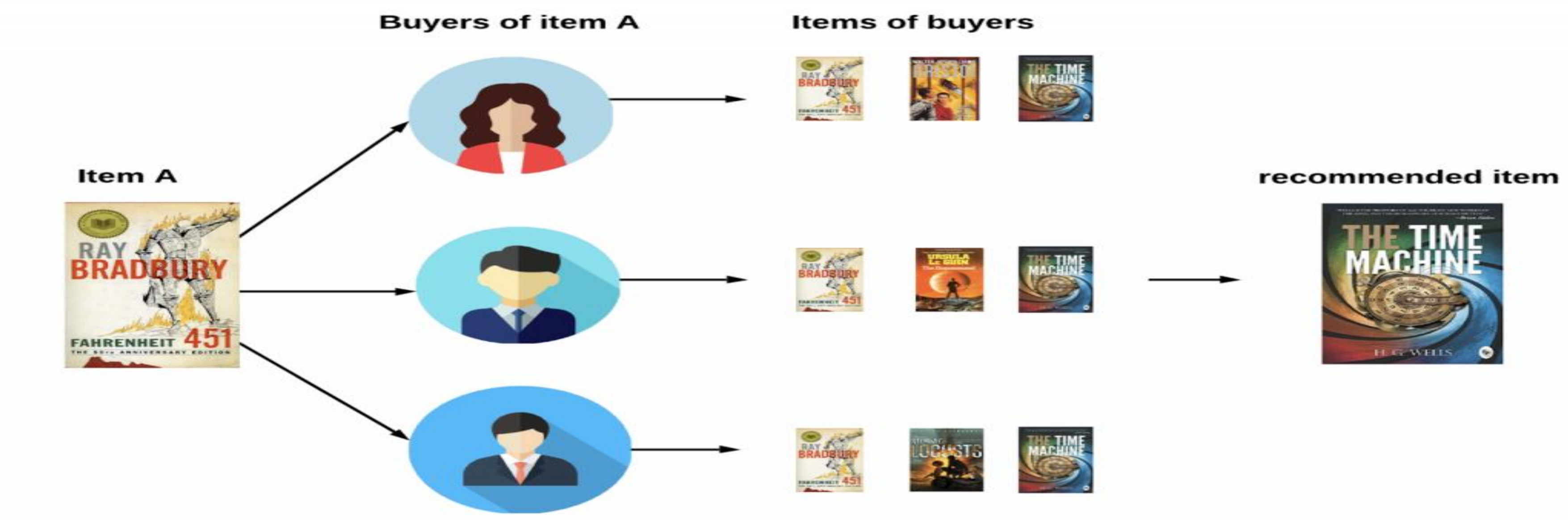
- In 2001 also around 14000 books were published, but less than 2002.

POPULARITY BASED RECMMENDATION SYSTEM

A popularity based recommendation system is a type of recommendation system that suggests items that are popular among the user group.

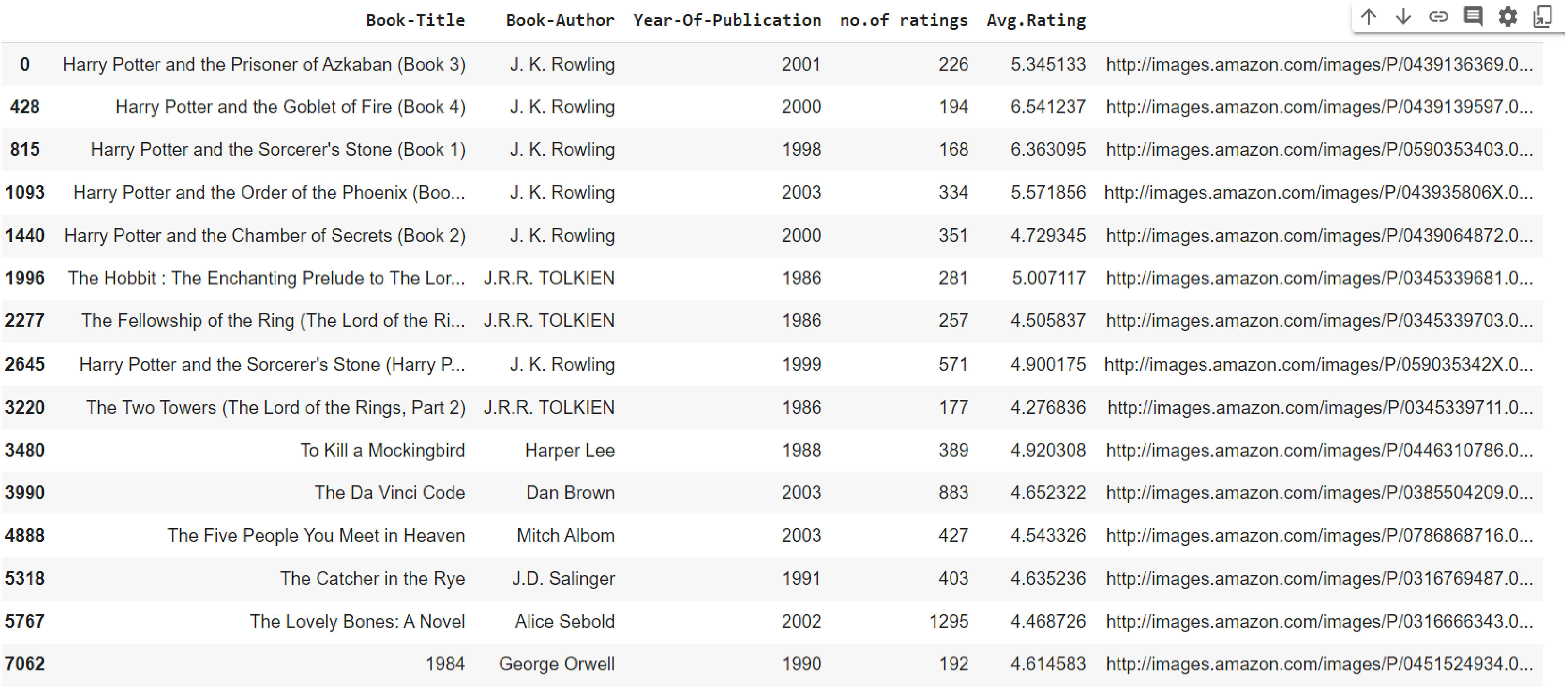
when we have a new user, we will face a cold start issue in recommendation. we can rely on a popularity based recommendation system in this scenario.

It is based on the assumption that items that are popular among many people are more likely to be liked by the user.



One potential drawback of a popularity based recommendation system is that it may not take into account the personal preferences of the user.

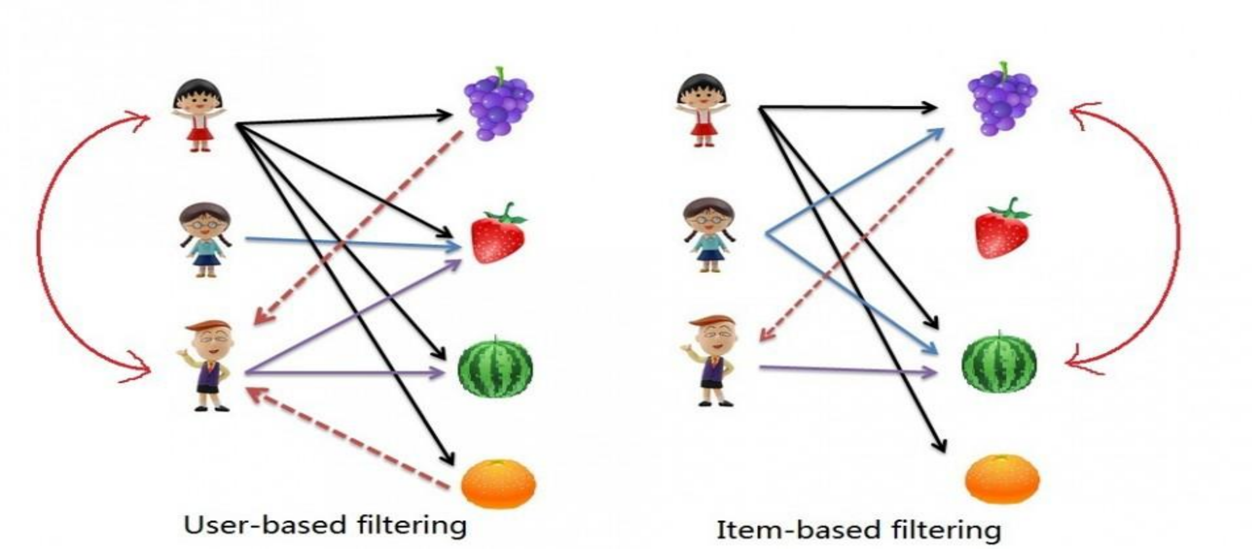
POPULARITY BASED TOP 15 BOOK RECOMMENDATIONS



	Book-Title	Book-Author	Year-Of-Publication	no.of ratings	Avg.Rating	
0	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	2001	226	5.345133	http://images.amazon.com/images/P/0439136369.0...
428	Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling	2000	194	6.541237	http://images.amazon.com/images/P/0439139597.0...
815	Harry Potter and the Sorcerer's Stone (Book 1)	J. K. Rowling	1998	168	6.363095	http://images.amazon.com/images/P/0590353403.0...
1093	Harry Potter and the Order of the Phoenix (Boo...	J. K. Rowling	2003	334	5.571856	http://images.amazon.com/images/P/043935806X.0...
1440	Harry Potter and the Chamber of Secrets (Book 2)	J. K. Rowling	2000	351	4.729345	http://images.amazon.com/images/P/0439064872.0...
1996	The Hobbit : The Enchanting Prelude to The Lor...	J.R.R. TOLKIEN	1986	281	5.007117	http://images.amazon.com/images/P/0345339681.0...
2277	The Fellowship of the Ring (The Lord of the Ri...	J.R.R. TOLKIEN	1986	257	4.505837	http://images.amazon.com/images/P/0345339703.0...
2645	Harry Potter and the Sorcerer's Stone (Harry P...	J. K. Rowling	1999	571	4.900175	http://images.amazon.com/images/P/059035342X.0...
3220	The Two Towers (The Lord of the Rings, Part 2)	J.R.R. TOLKIEN	1986	177	4.276836	http://images.amazon.com/images/P/0345339711.0...
3480	To Kill a Mockingbird	Harper Lee	1988	389	4.920308	http://images.amazon.com/images/P/0446310786.0...
3990	The Da Vinci Code	Dan Brown	2003	883	4.652322	http://images.amazon.com/images/P/0385504209.0...
4888	The Five People You Meet in Heaven	Mitch Albom	2003	427	4.543326	http://images.amazon.com/images/P/0786868716.0...
5318	The Catcher in the Rye	J.D. Salinger	1991	403	4.635236	http://images.amazon.com/images/P/0316769487.0...
5767	The Lovely Bones: A Novel	Alice Sebold	2002	1295	4.468726	http://images.amazon.com/images/P/0316666343.0...
7062	1984	George Orwell	1990	192	4.614583	http://images.amazon.com/images/P/0451524934.0...

COLLABRARIVE FILTERING BASED RECOMMENDATION SYSTEM

Collaborative filtering is a method of building a recommendation system that is based on the preferences and behavior of users. It works by identifying users who have similar preferences and then recommending items that they both like. There are two main types of collaborative filtering: user-based and item-based.



ADVANTAGES

It does not require any information about the content of the items being recommended.

It can also handle large amounts of data effectively and is able to learn from new data as it becomes available.

It is simple to implement

SVD (SINGULAR VALUE DECOMPOSITION)

SVD is a method for decomposing a matrix into the product of three matrices: U, S, and V.
It is often used to reduce the dimensionality of a matrix and to identify patterns in the data.
One way to use SVD for imputing missing values is to decompose the matrix into its U, S, and V components,
and then use the S and V matrices to reconstruct the original matrix with the missing values filled in.
This can be done using the following formula:

$$A = U * S * V^T$$

reasons why singular value decomposition (SVD) might be used in a recommendation system:

- Dimensionality reduction:
- Impute missing values:
- Improve prediction accuracy:
- Handle large-scale data:

COSINE SIMILARITY

- Cosine similarity is a common metric that is used to measure the similarity between two vectors in collaborative filtering.
- It is particularly useful for finding the similarity between users or items based on their ratings of a set of items.
- To calculate the cosine similarity between two vectors, first need to represent each vector as a list of ratings for a set of items.
- In case of building a recommendation system for books, might have a vector for each user that contains their ratings for a set of books.
- Then use the following formula to calculate the cosine similarity between two vectors:

$$\cos(\theta) = (A * B) / (||A|| * ||B||)$$

book name :

Winter Solstice

Top5 Recommendations are

	similarity score
Second Nature	0.942286
A Patchwork Planet	0.936192
Catering to Nobody	0.934048
Echoes	0.933001
The Copper Beech	0.927982

book name :

A Civil Action

Top5 Recommendations are

	similarity score
Into Thin Air : A Personal Account of the Mt. Everest Disaster	0.952487
Secret History	0.940961
Pigs in Heaven	0.937953
The Temple of My Familiar	0.921719
Call of the Wild	0.919866

book name:

1st to Die: A Novel

Top5 Recommendations are:

	similarity score
Pop Goes the Weasel	0.964336
Easy Prey	0.959548
Cat &amp; Mouse (Alex Cross Novels)	0.955471
2nd Chance	0.951208
All That Remains (Kay Scarpetta Mysteries (Paperback))	0.943687

book name :

A Fine Balance

Top5 Recommendations are

	similarity score
The Bonesetter's Daughter	0.920847
Girl in Hyacinth Blue	0.915114
Seabiscuit: An American Legend	0.896713
Prodigal Summer	0.895837
Under the Tuscan Sun	0.892637

book name :

A Bend in the Road

Top5 Recommendations are

	similarity score
The Rescue	0.973764
Small Town Girl	0.944872
That Camden Summer	0.936384
Nights in Rodanthe	0.933510
Suzanne's Diary for Nicholas	0.931360

book name :

Year of Wonders

Top5 Recommendations are

	similarity score
Empire Falls	0.942453
Bel Canto: A Novel	0.929074
The Robber Bride	0.925866
Nickel and Dimed: On (Not) Getting By in America	0.924861
Moo	0.922229

CONCLUSIONS

A book recommendation system can be a valuable tool for businesses that sell or lend books. It can help to increase sales and customer satisfaction, improve the user experience, and provide a competitive advantage. However, implementing a recommendation system can be a challenging process, as it requires the gathering and analysis of a large dataset and the development of machine learning algorithms. It is also important to consider issues such as the cold start problem and user privacy when designing the system.

From EDA:

- When we are considering top 15 authors top most 5 writers are far forward from remaining writers in number of books.
- Stephen king is the top most author
- analysis of rating counts captured that, zero rating is more in dataset, which conveys that most of the people leaving with out give rating to the books
- by the analysis of most reeded / reviewed books and also its average rating figure outed that the most reeded book have very less average rating.
- by Visualizing the top 30 publishers, ‘Ballantine books’ is the top publisher with around 35000 published books.

Overall, a book recommendation system can provide many benefits to businesses, but it is important to carefully plan and execute the project in order to achieve the desired results.

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

Decision making on missing value imputations and outlier treatment was quite challenging as well.

REFERENCES

https://medium.com/

https://www.google.co.in

https://www.kaggle.com/

http://www.librarything.com

THANK YOU