Capstone Project – 4

Book Recommendation System

Presented By:

Aehteshaam Shaikh

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.

The main objective of this project is to create a Book Recommendation system for users.

Data Summary:

1. Books Dataset:

Sr	Column Name	Description	Datatype
1	ISBN	Unique ID for the Book	Object
2	Book-Title	Title of the Book	Object
3	Book-Author	Name of the Author	Object
4	Year-Of-Publication	Year in which book published	Object
5	Publisher	Name of the Publisher	Object
6	Image-URL-S	Url for the image of the book (size – small)	Object
7	Image-URL-M	Url for the image of the book (size – medium)	Object
8	Image-URL-L	Url for the image of the book (size – Large)	Object

Data Summary:

2. Users Dataset:

Sr	Column Name	Description	Datatype
1	User-ID	Unique ID for the User	int
2	Location	Location of the User	Object
3	Age	Age of the User	float

Data Summary:

3. Ratings Dataset:

Sr	Column Name	Description	Datatype
1	User-ID	Unique ID for the User	Int
2	ISBN	Unique ID for the Book	Object
3	Book-Rating	Rating of the Book	int

Data Cleaning:

1. Books Dataset:

- There are 271360 entries and 8 columns with columns Publisher, Book- author and Image-URL-L having some Null Values.

- There was some discrepancy in the Year of Publication column as some entries had the year as "0", "DK Publishing Inc," and "Gallimard," which did not make any sense. Also, the entries whose year is > 2004 have been replaced by the median value, as this dataset itself was published in 2004.

Data Cleaning:

2. Users Dataset:

- This dataset consists of 3 features with 2,78,858 entries with 'Age' column having Null Values.
- The feature Age was Rightly skewed, hence Null values were replaced with Median value
- The Age column contained some values below the age of 5 and above the age of 100 that did not make sense, so they were replaced with the median value.
- A new column ‘Country’ is added to simplify the feature ‘Location’.

Data Cleaning:

3. Ratings Dataset:

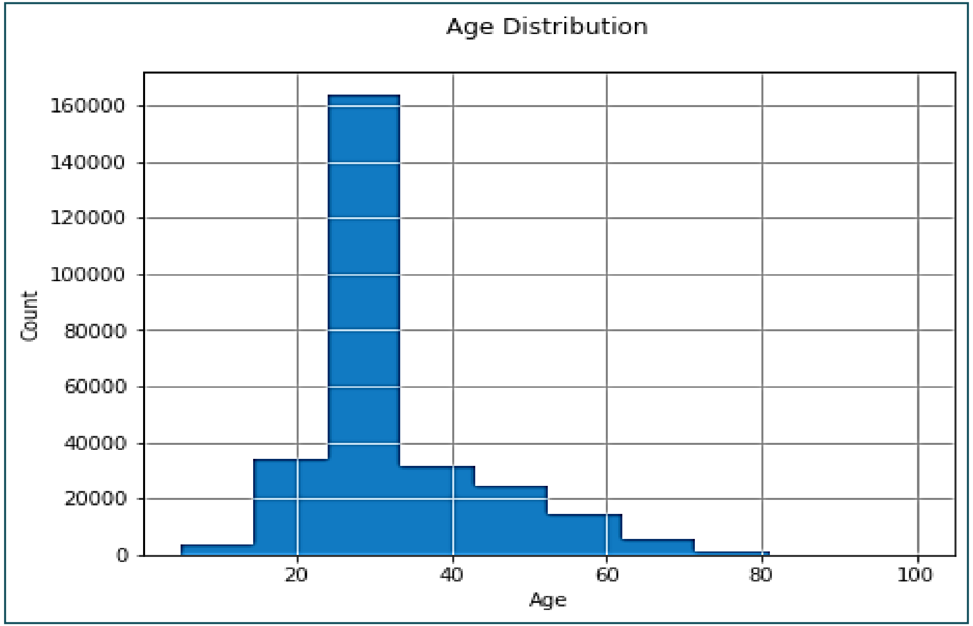
- This Dataset consists 11,49,780 entries with 3 columns and no null values are present in it.

- Created a new dataset that consists of ratings for only those books that are present in our books_df dataset and for only those users who are present in our users dataset

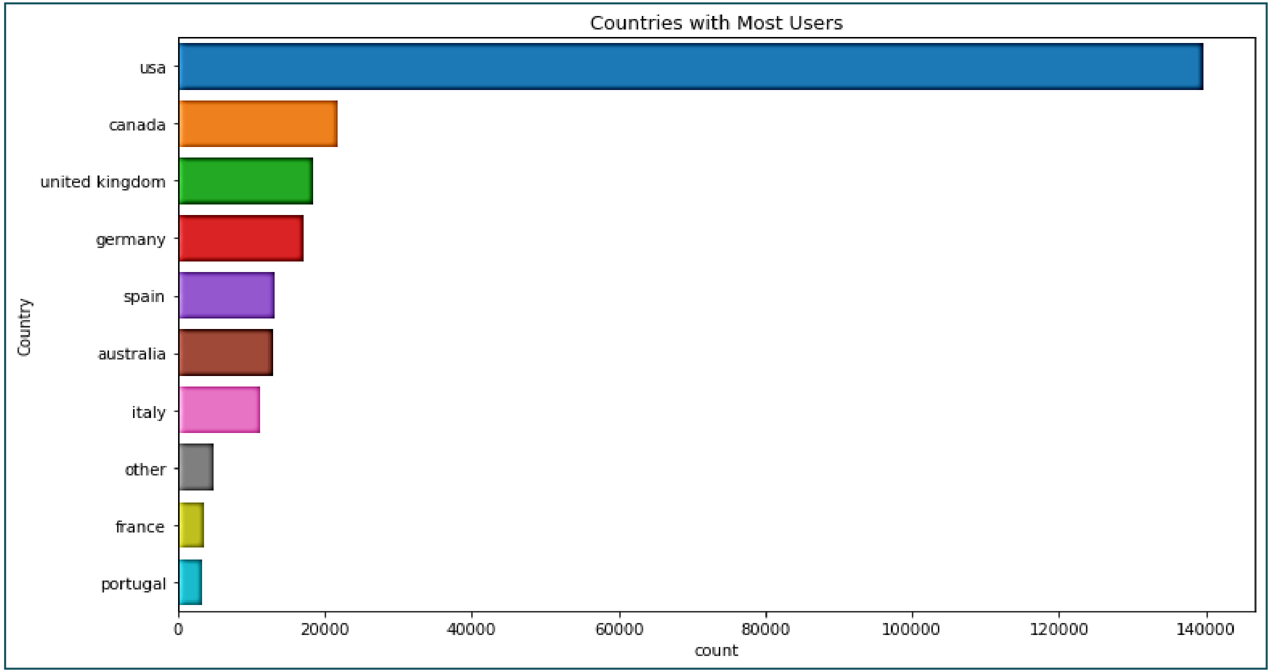
- Since the dataset contains explicit ratings (from 1 to 10) and implicit rating (0), divided the dataset into two parts as ratings_df_explicit and ratings_df_implicit.

Exploratory Data Analysis

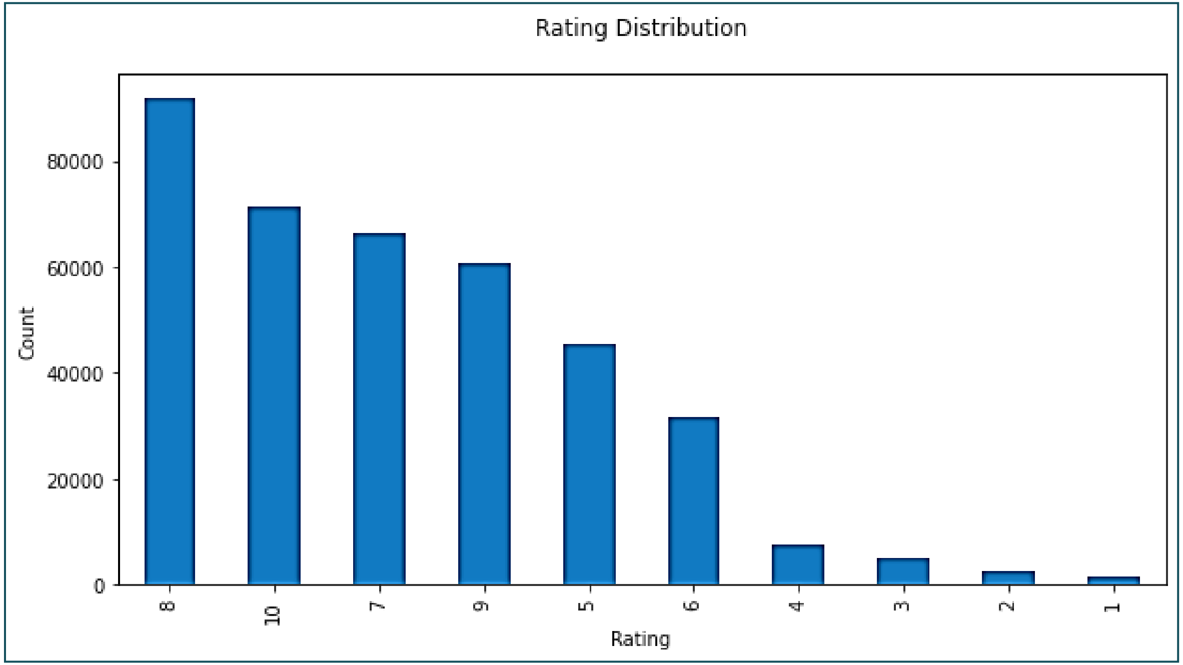
- **Most users are from the age group of 20 to 40**



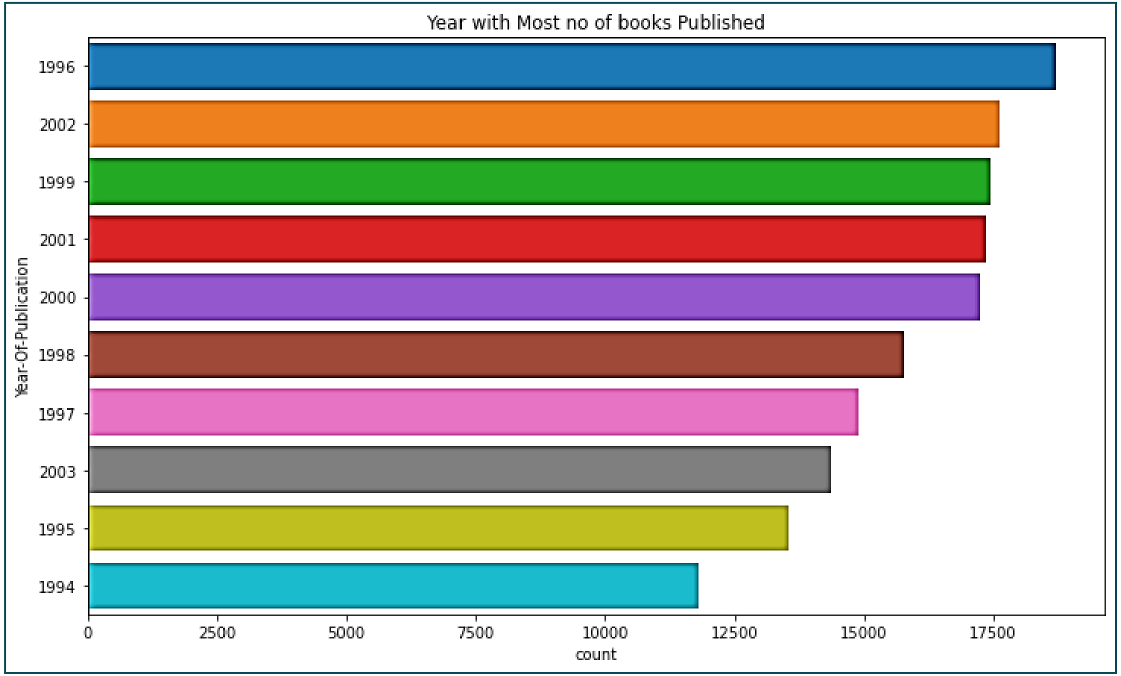
- **Clearly, most of the users are from the USA, followed by Canada.**



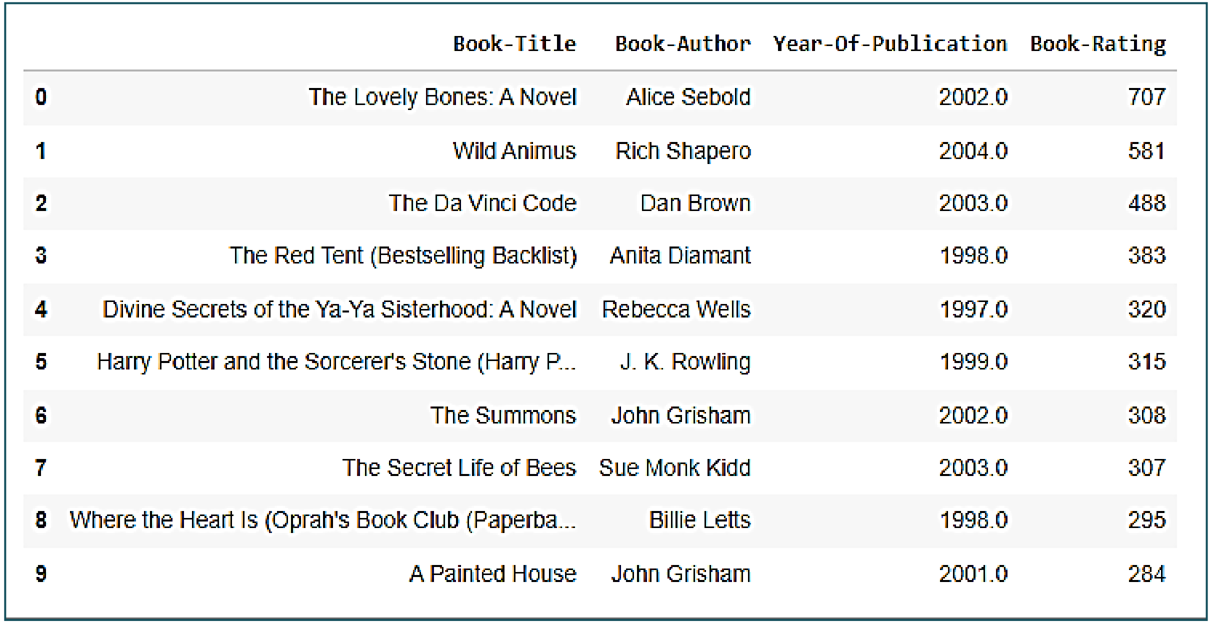
- **The most common rating given by the users is 8, followed by 10.**



- **The year in which most no of books were published in this dataset is 1996 followed by Year 2002.**

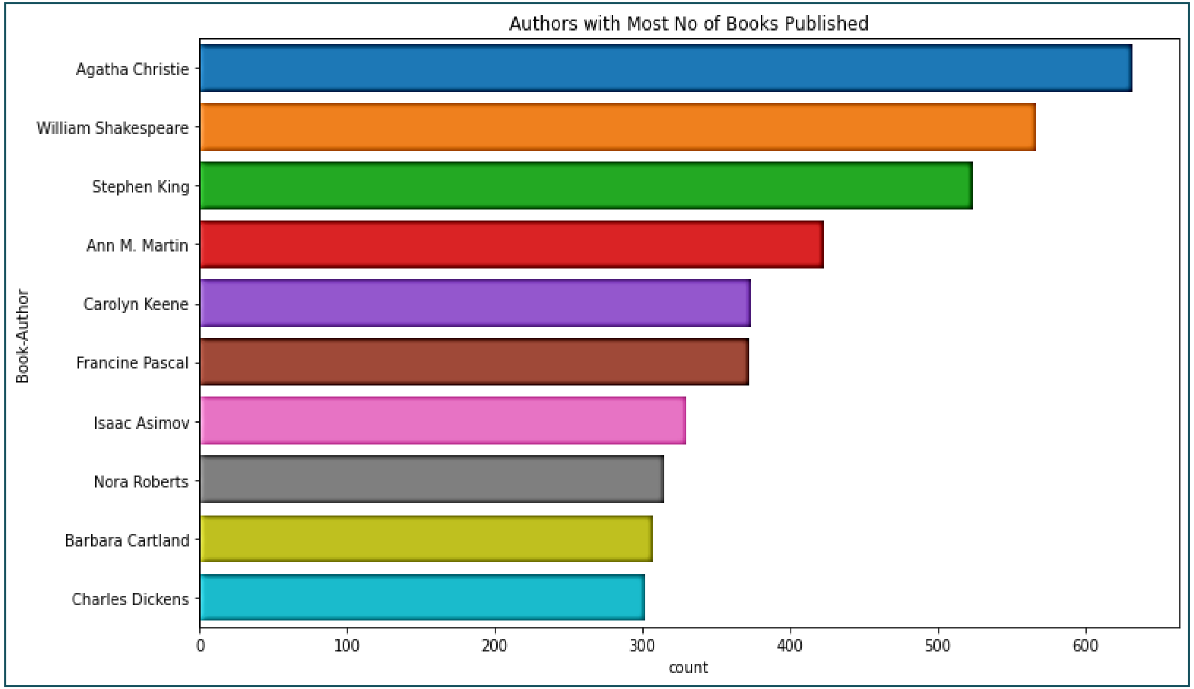


- **The book "The Lovely Bones: A Novel" by Alice Sebold,published in the year 2002, received the highest number of ratings.**

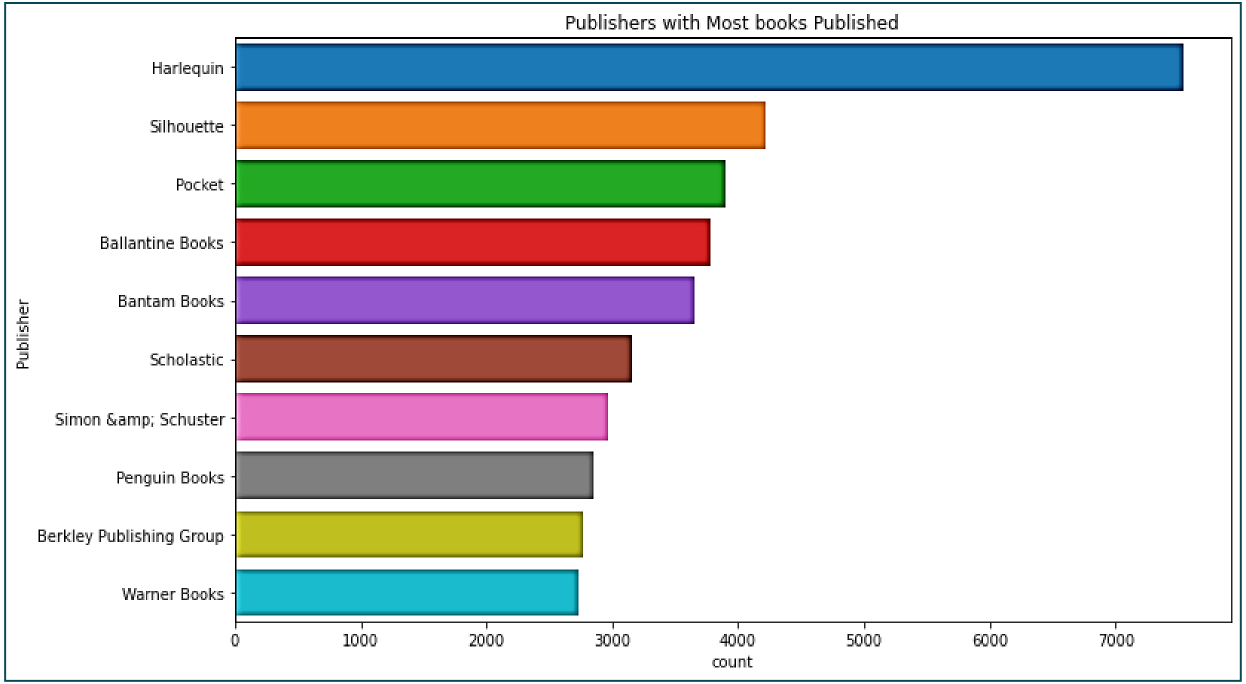


	Book-Title	Book-Author	Year-Of-Publication	Book-Rating
0	The Lovely Bones: A Novel	Alice Sebold	2002.0	707
1	Wild Animus	Rich Shapero	2004.0	581
2	The Da Vinci Code	Dan Brown	2003.0	488
3	The Red Tent (Bestselling Backlist)	Anita Diamant	1998.0	383
4	Divine Secrets of the Ya-Ya Sisterhood: A Novel	Rebecca Wells	1997.0	320
5	Harry Potter and the Sorcerer's Stone (Harry P...	J. K. Rowling	1999.0	315
6	The Summons	John Grisham	2002.0	308
7	The Secret Life of Bees	Sue Monk Kidd	2003.0	307
8	Where the Heart Is (Oprah's Book Club (Paperba...	Billie Letts	1998.0	295
9	A Painted House	John Grisham	2001.0	284

- **Agatha Christie is the Author with most no of books Published followed by William Shakespeare and Stephen king.**



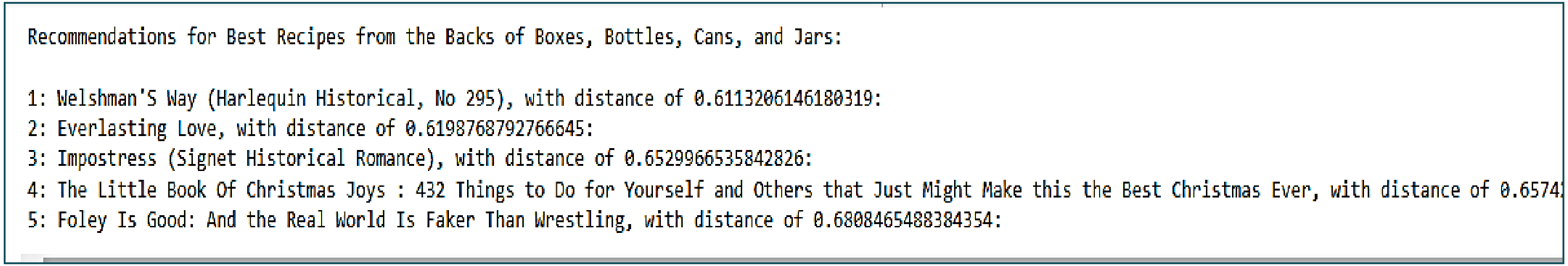
- **Harlequin is the Publisher with most no of books published followed by Silhouette.**



Machine Learning

ML Models Performed :

1. Collaborative Filtering (Item-Item based)

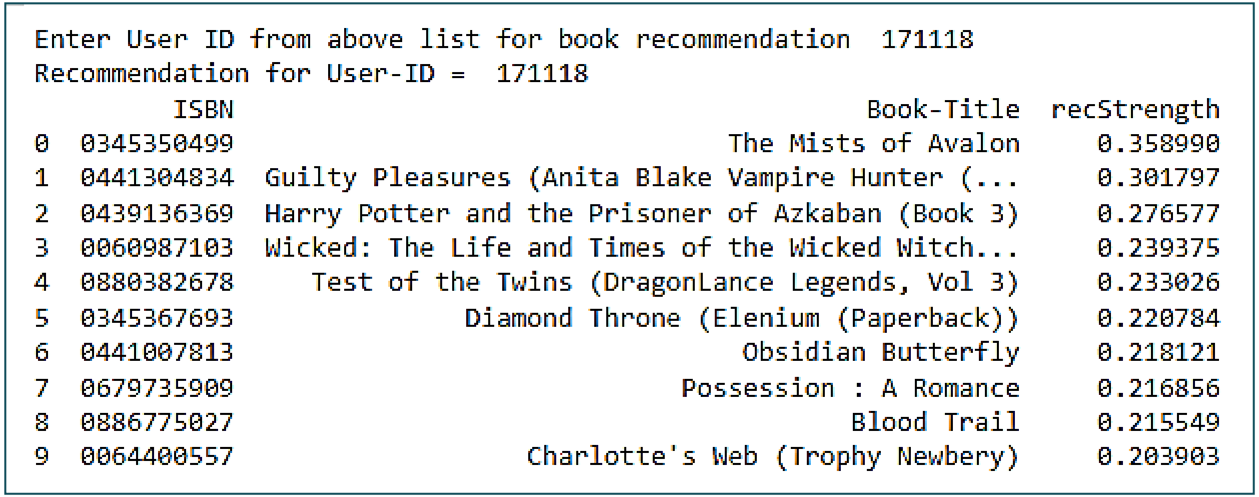


```
Recommendations for Best Recipes from the Backs of Boxes, Bottles, Cans, and Jars:

1: Welshman'S Way (Harlequin Historical, No 295), with distance of 0.6113206146180319:
2: Everlasting Love, with distance of 0.6198768792766645:
3: Impostress (Signet Historical Romance), with distance of 0.6529966535842826:
4: The Little Book Of Christmas Joys : 432 Things to Do for Yourself and Others that Just Might Make this the Best Christmas Ever, with distance of 0.6574
5: Foley Is Good: And the Real World Is Faker Than Wrestling, with distance of 0.6808465488384354:
```

ML Models Performed :

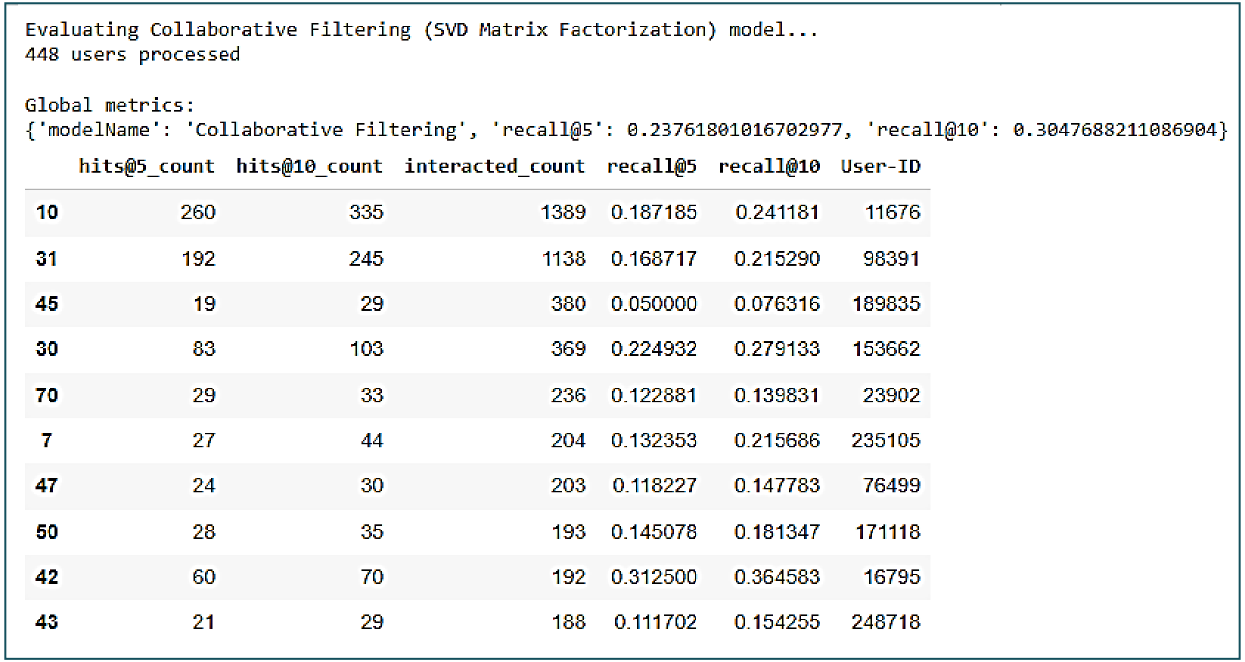
2. Collaborative Filtering (User-Item based)



```
Enter User ID from above list for book recommendation  171118
Recommendation for User-ID =  171118
         ISBN                                         Book-Title  recStrength
0  0345350499                                The Mists of Avalon     0.358990
1  0441304834  Guilty Pleasures (Anita Blake Vampire Hunter (...     0.301797
2  0439136369  Harry Potter and the Prisoner of Azkaban (Book 3)     0.276577
3  0060987103  Wicked: The Life and Times of the Wicked Witch...     0.239375
4  0880382678     Test of the Twins (DragonLance Legends, Vol 3)     0.233026
5  0345367693               Diamond Throne (Elenium (Paperback))     0.220784
6  0441007813                                 Obsidian Butterfly     0.218121
7  0679735909                             Possession : A Romance     0.216856
8  0886775027                                        Blood Trail     0.215549
9  0064400557                   Charlotte's Web (Trophy Newbery)     0.203903
```

Model Evaluation :

After evaluating the Collaborative Filtering model (SVD matrix factorization), got the Recall@5 (23.76 %) and Recall@10 (30.47 %)



```
Evaluating Collaborative Filtering (SVD Matrix Factorization) model...
448 users processed

Global metrics:
{'modelName': 'Collaborative Filtering', 'recall@5': 0.23761801016702977, 'recall@10': 0.3047688211086904}
```

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	User-ID
10	260	335	1389	0.187185	0.241181	11676
31	192	245	1138	0.168717	0.215290	98391
45	19	29	380	0.050000	0.076316	189835
30	83	103	369	0.224932	0.279133	153662
70	29	33	236	0.122881	0.139831	23902
7	27	44	204	0.132353	0.215686	235105
47	24	30	203	0.118227	0.147783	76499
50	28	35	193	0.145078	0.181347	171118
42	60	70	192	0.312500	0.364583	16795
43	21	29	188	0.111702	0.154255	248718

Conclusion:

- After loading the dataset, cleaning the data and performing EDA some important inferences have been made and have been incurred below every visualization.

- We can conclude that item-item based collaborative filtering performed better than user-item based collaborative filtering because of lower computation time and lesser memory usage.

that the column avg views by sp

AI

Thank You