CAPSTONE PROJECT-4

BOOK RECOMMENDATION SYSTEM

KRUSHNAGOPAL BRAHMA

CONTENT

- BUSINESS UNDERSTANDING
- DATA OVERVIEW
- DATA DESCRIPTION
- PREPROCESS DATA
- EXPLORATORY DATA ANALYSIS
- RECOMMENDATION SYSTEM
- CHALLENGES
- CONCLUSIONS

BUSINESS UNDERSTANDING

- **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).**
- **Recommender 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 OVERVIEW

- **The Book-Crossing dataset comprises 3 files.**
- **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 NULLvalues.**
- **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.**
- **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.**

DATA DESCRIPTION

Books Data :

- Data set Contains 'ISBN', 'Book-Title', 'Book-Author', 'Year-Of-Publication', 'Publisher', 'Image-URL-S', 'Image-URL-M', 'Image-URL-L‘ information's.
- The Data set Contains 271354 Rows

Users Data :

- Data set Contains ‘User-ID’ , ‘Location’, ‘Age’ information's.
- The Data set Contains 278858 Rows

Ratings Data :

- Data set Contains ‘User-ID’ , ‘ISBN’, ‘Book-Ratings’ information's.
- The Data set Contains 1149780 Rows

PREPROCESS DATA

Books Dataset Pre-processing

- **We drop those column which contain url of book image because we work on book recommendation no need image and its urls. So we drop 3 columns which contain image urls.**
- **The publication must be int format but there are some error in it at some phase it show yr='DK Publishing Inc', yr='Gallimard' its not posible**

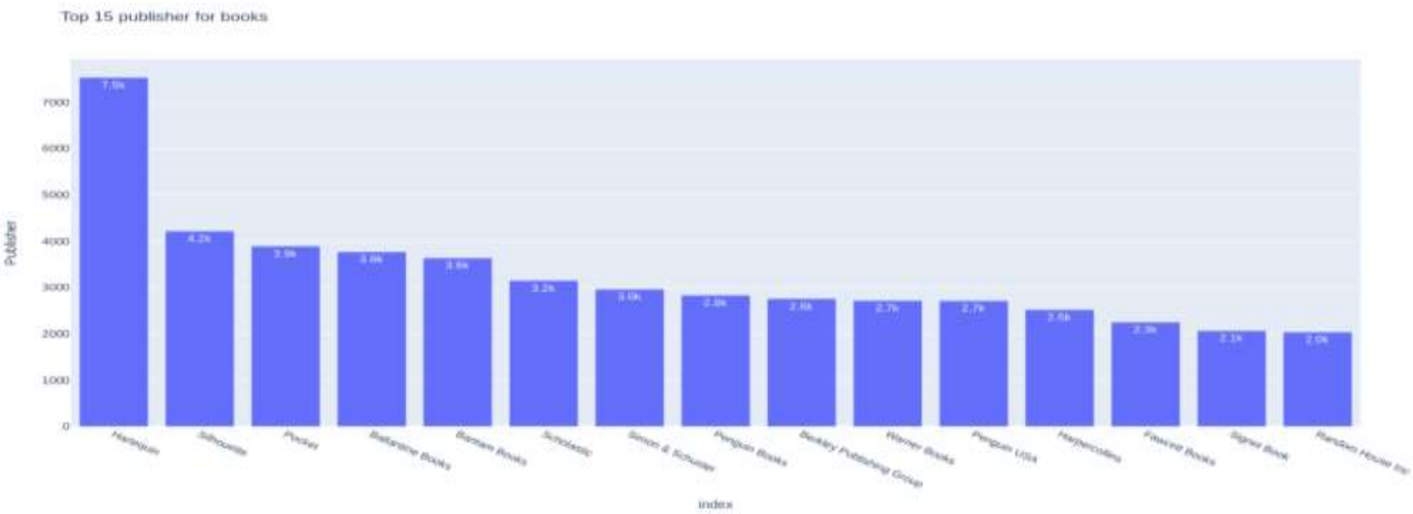
Users Dataset Pre-processing

- **We see that the null valu present in age we try to correct it**
- **There are 11072 valu are null we can't remove these many values so we try to cover it with mean age value.**
- **Ratings Dataset Pre-processing**
- **We se user id 11676 gives rating on 13602 books rating its consider as most impactfull user.**
- **Theare are so many user who are only given rating on 1 book so they might nt proper comaprision rating to our rating.**
- **As we consider if some user gives rating in more than 200 books he might be good impactfull rating user.**

EDA(contd...)

Top 15 publisher of book

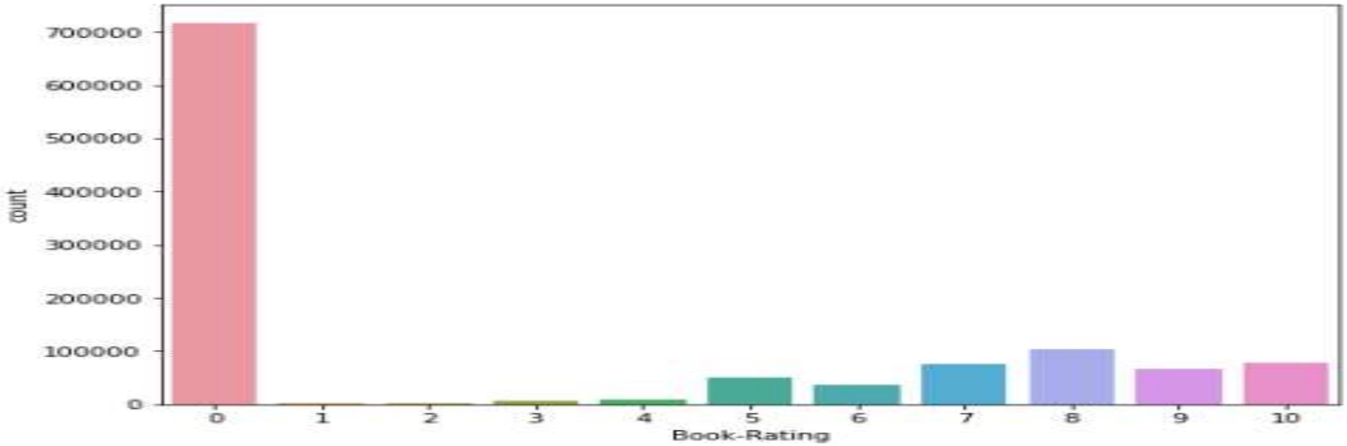
- Here we se that Harlequin publicatation publish most no of book its almost double than Silhouette which is in 2nd place.



EDA(contd...)

Count vs Book Rating

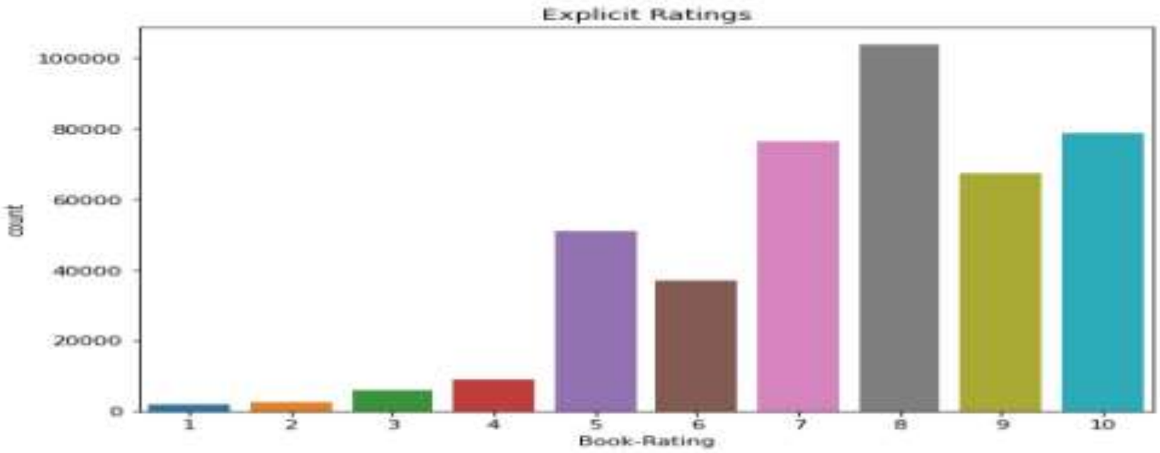
- If we se the rating part most of not given any review here we consider 0.
- It almost 50 to 60 percent of our dataset.



EDA(contd...)

Explicity Rating

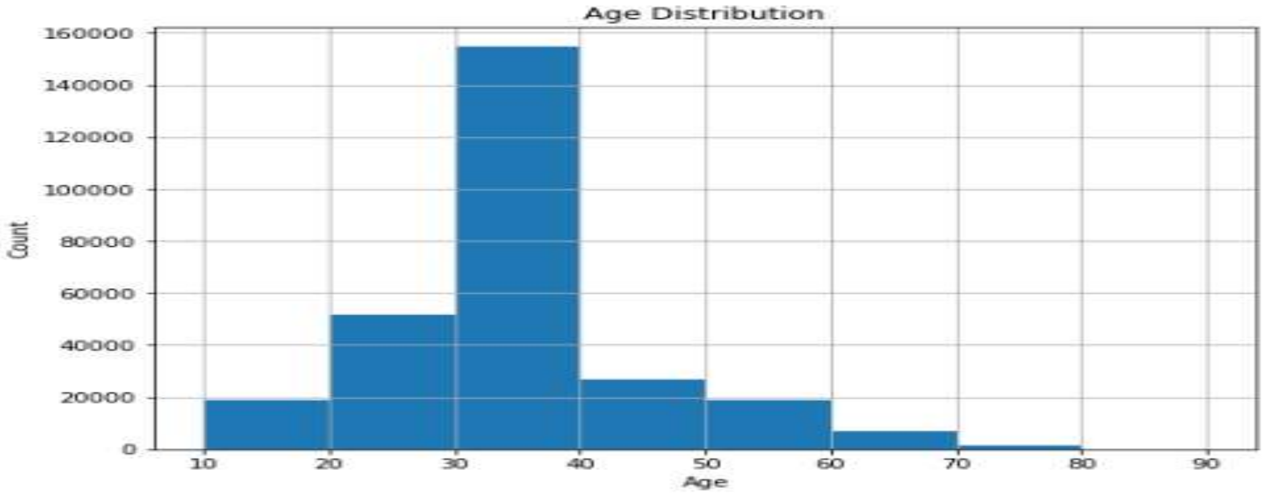
- If we not consider 0 rating user we se highest no of user given rating 8.
- Almost eqal rating given to 7 and 10 then 9.



EDA(contd...)

Age Vs Count

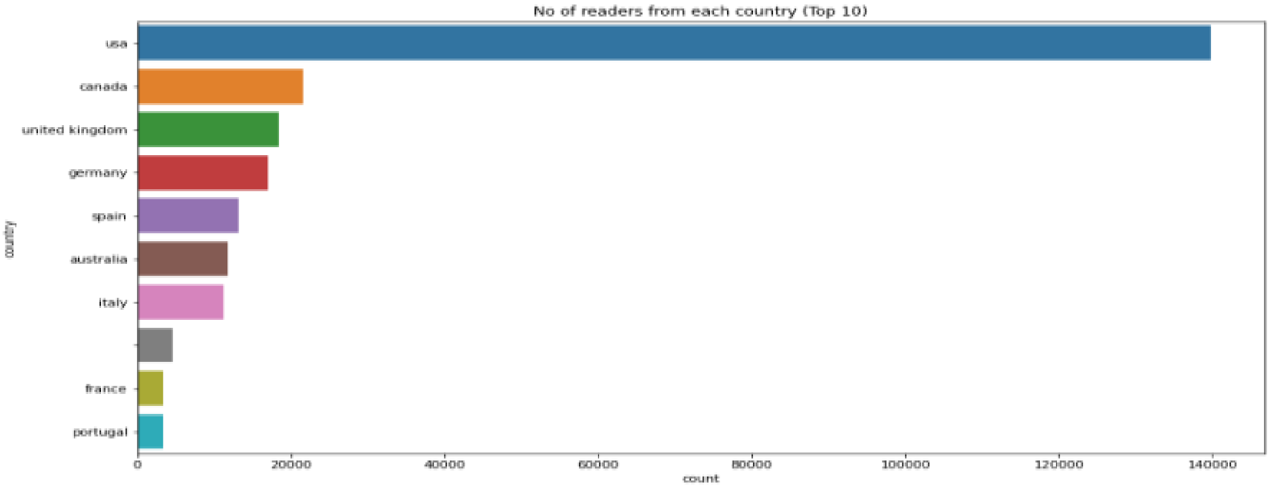
- Most of the user who rated the mv are from 30-40 yrs age then in 2nd its less than half of that catagory is 20-30



EDA(contd...)

No of readers from each country

- Most of the book listed in our daaset are from usa its cover almost 60 to70% of our data



RECOMENTATION SYSTEM

1. Popularity Based (Books reviewed by most of the user)

2.Popular books in country wise

3. Books popular yr wise

4. Popular book of same_author

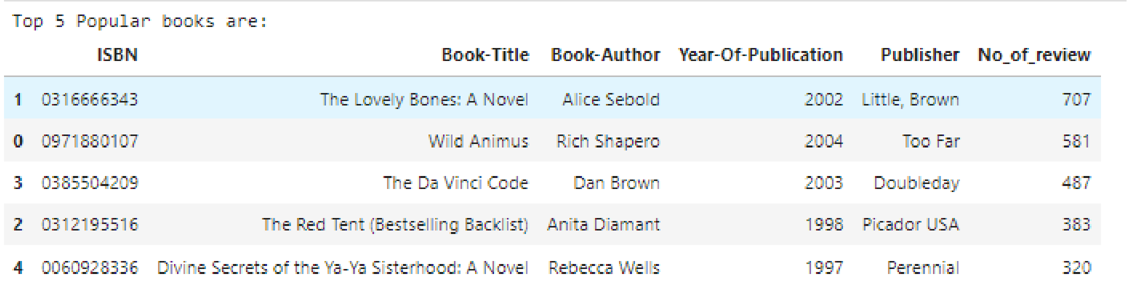
5.Popular Book of same Publsher

6.Top most rating books.

7. Collaborative Filtering (User-Item Filtering)

8.Content Based filtering

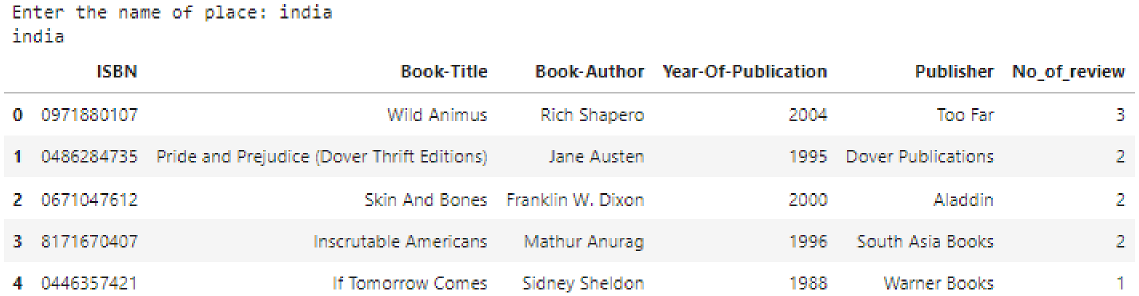
1. Popularity Based (Books reviewed by most of the user)



```
Top 5 Popular books are:
```

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	No_of_review
1	0316666343	The Lovely Bones: A Novel	Alice Sebold	2002	Little, Brown	707
0	0971880107	Wild Animus	Rich Shapero	2004	Too Far	581
3	0385504209	The Da Vinci Code	Dan Brown	2003	Doubleday	487
2	0312195516	The Red Tent (Bestselling Backlist)	Anita Diamant	1998	Picador USA	383
4	0060928336	Divine Secrets of the Ya-Ya Sisterhood: A Novel	Rebecca Wells	1997	Perennial	320

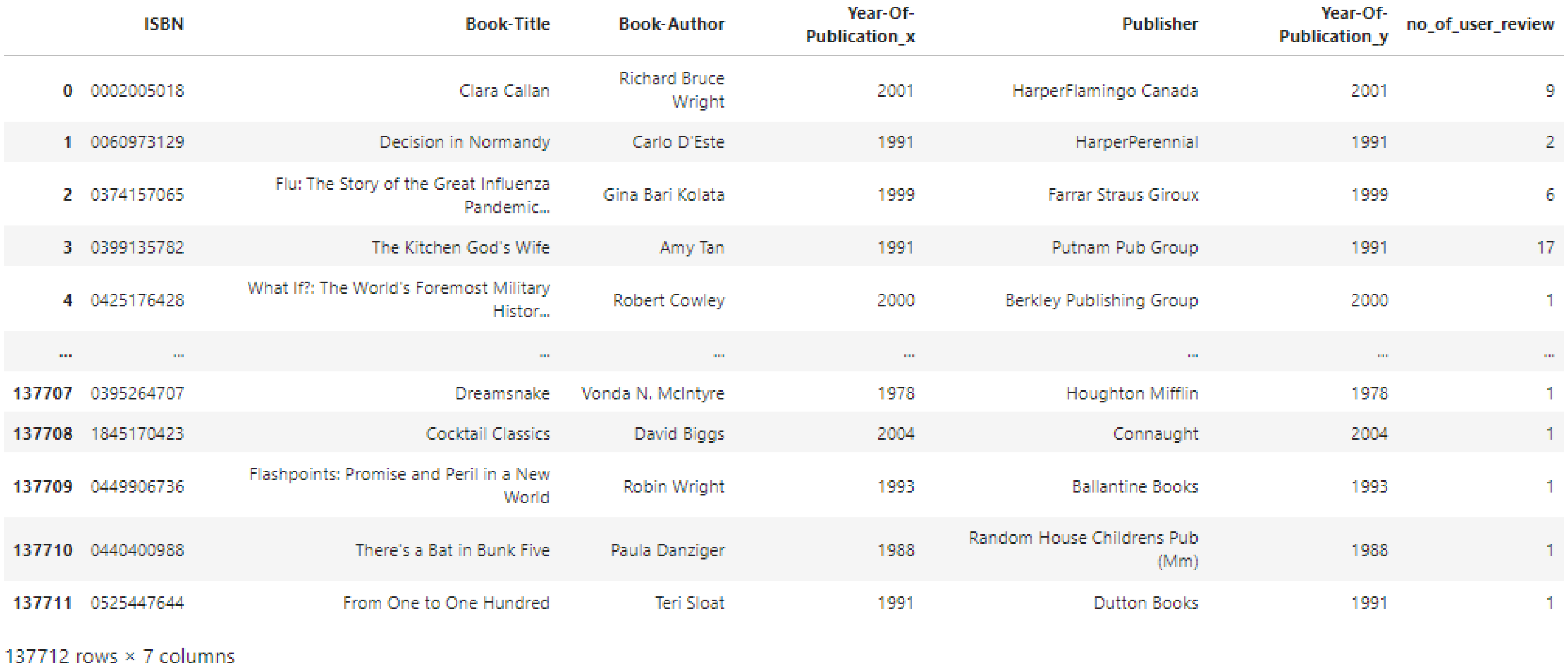
2.Popular books in country wise



```
Enter the name of place: india
india
```

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	No_of_review
0	0971880107	Wild Animus	Rich Shapero	2004	Too Far	3
1	0486284735	Pride and Prejudice (Dover Thrift Editions)	Jane Austen	1995	Dover Publications	2
2	0671047612	Skin And Bones	Franklin W. Dixon	2000	Aladdin	2
3	8171670407	Inscrutable Americans	Mathur Anurag	1996	South Asia Books	2
4	0446357421	If Tomorrow Comes	Sidney Sheldon	1988	Warner Books	1

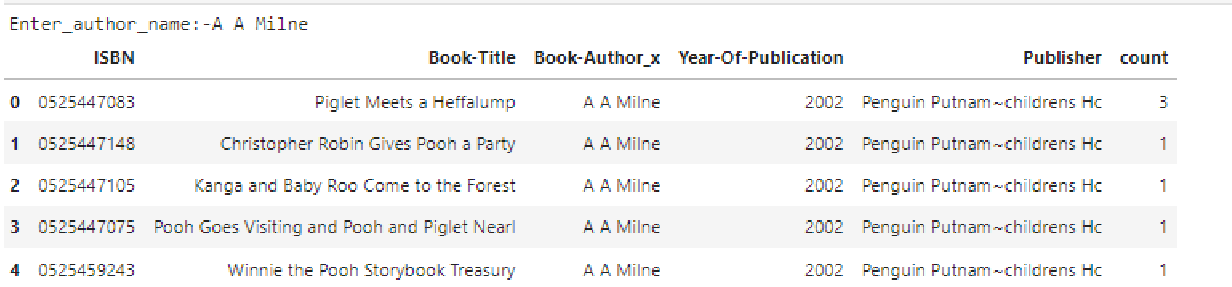
3. Books popular yr wise



	ISBN	Book-Title	Book-Author	Year-Of-Publication_x	Publisher	Year-Of-Publication_y	no_of_user_review
0	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	2001	9
1	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	1991	2
2	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux	1999	6
3	0399135782	The Kitchen God's Wife	Amy Tan	1991	Putnam Pub Group	1991	17
4	0425176428	What If?: The World's Foremost Military Histor...	Robert Cowley	2000	Berkley Publishing Group	2000	1
...	...	...	...	...	...	...	...
137707	0395264707	Dreamsnake	Vonda N. McIntyre	1978	Houghton Mifflin	1978	1
137708	1845170423	Cocktail Classics	David Biggs	2004	Connaught	2004	1
137709	0449906736	Flashpoints: Promise and Peril in a New World	Robin Wright	1993	Ballantine Books	1993	1
137710	0440400988	There's a Bat in Bunk Five	Paula Danziger	1988	Random House Childrens Pub (Mm)	1988	1
137711	0525447644	From One to One Hundred	Teri Sloat	1991	Dutton Books	1991	1

137712 rows × 7 columns

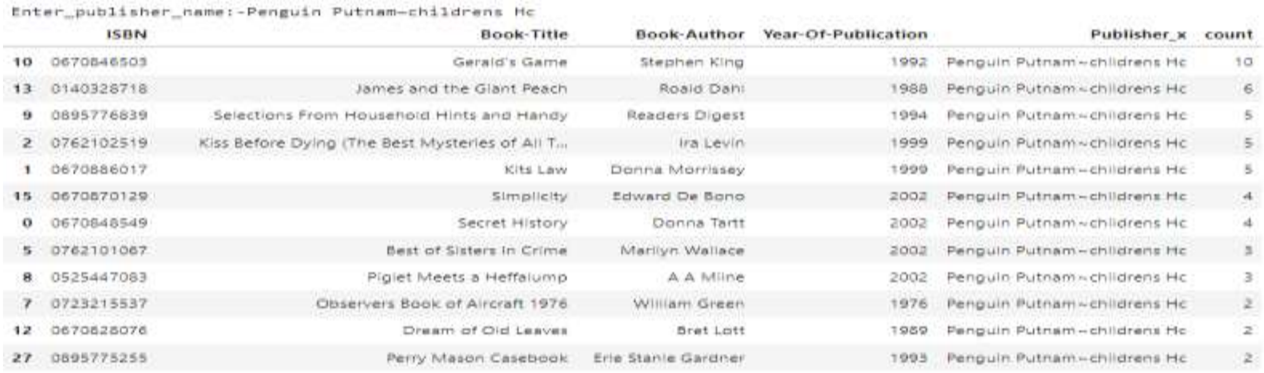
4. Popular book of same_author



```
Enter_author_name:-A A Milne
```

	ISBN	Book-Title	Book-Author_x	Year-Of-Publication	Publisher	count
0	0525447083	Piglet Meets a Heffalump	A A Milne	2002	Penguin Putnam~childrens Hc	3
1	0525447148	Christopher Robin Gives Pooh a Party	A A Milne	2002	Penguin Putnam~childrens Hc	1
2	0525447105	Kanga and Baby Roo Come to the Forest	A A Milne	2002	Penguin Putnam~childrens Hc	1
3	0525447075	Pooh Goes Visiting and Pooh and Piglet Nearl	A A Milne	2002	Penguin Putnam~childrens Hc	1
4	0525459243	Winnie the Pooh Storybook Treasury	A A Milne	2002	Penguin Putnam~childrens Hc	1

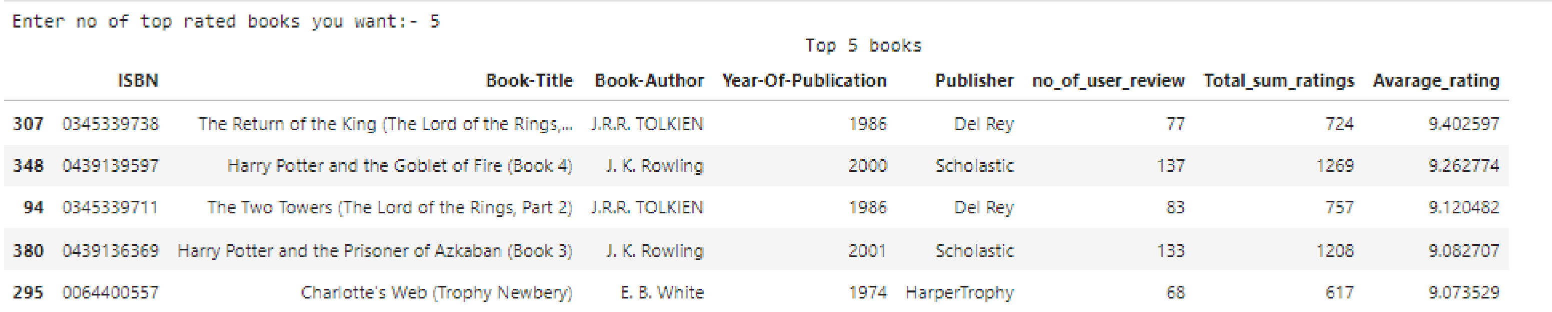
5.Popular Book of same Publsher



```
Enter_publisher_name:-Penguin Putnam~childrens Hc
```

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher_x	count
10	0670846503	Gerald's Game	Stephen King	1992	Penguin Putnam~childrens Hc	10
13	0140328718	James and the Giant Peach	Roald Dahl	1988	Penguin Putnam~childrens Hc	6
9	0895776839	Selections From Household Hints and Handy	Readers Digest	1994	Penguin Putnam~childrens Hc	5
2	0762102519	Kiss Before Dying (The Best Mysteries of All T...	Ira Levin	1999	Penguin Putnam~childrens Hc	5
1	0670886017	Kits Law	Donna Morrissey	1999	Penguin Putnam~childrens Hc	5
15	0670870129	Simplicity	Edward De Bono	2002	Penguin Putnam~childrens Hc	4
0	0670848549	Secret History	Donna Tartt	2002	Penguin Putnam~childrens Hc	4
5	0762101067	Best of Sisters in Crime	Marilyn Wallace	2002	Penguin Putnam~childrens Hc	3
8	0525447083	Piglet Meets a Heffalump	A A Milne	2002	Penguin Putnam~childrens Hc	3
7	0723215537	Observers Book of Aircraft 1976	William Green	1976	Penguin Putnam~childrens Hc	2
12	0670828076	Dream of Old Leaves	Bret Lott	1989	Penguin Putnam~childrens Hc	2
27	0895775255	Perry Mason Casebook	Erle Stanle Gardner	1993	Penguin Putnam~childrens Hc	2

6.Top most rating books.

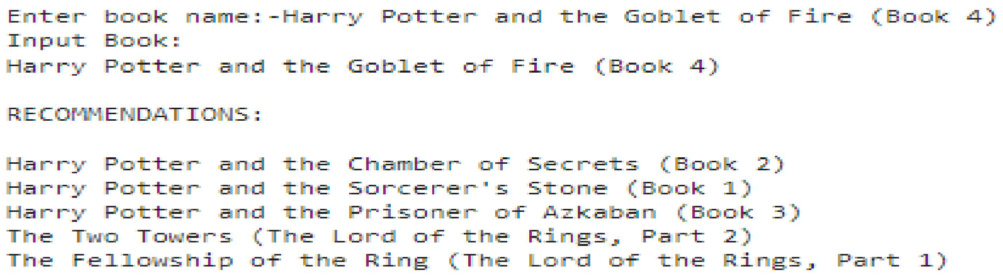


```
Enter no of top rated books you want:- 5
```

Top 5 books

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	no_of_user_review	Total_sum_ratings	Avarage_rating
307	0345339738	The Return of the King (The Lord of the Rings,...	J.R.R. TOLKIEN	1986	Del Rey	77	724	9.402597
348	0439139597	Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling	2000	Scholastic	137	1269	9.262774
94	0345339711	The Two Towers (The Lord of the Rings, Part 2)	J.R.R. TOLKIEN	1986	Del Rey	83	757	9.120482
380	0439136369	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	2001	Scholastic	133	1208	9.082707
295	0064400557	Charlotte's Web (Trophy Newbery)	E. B. White	1974	HarperTrophy	68	617	9.073529

7. Collaborative Filtering (User-Item Filtering)

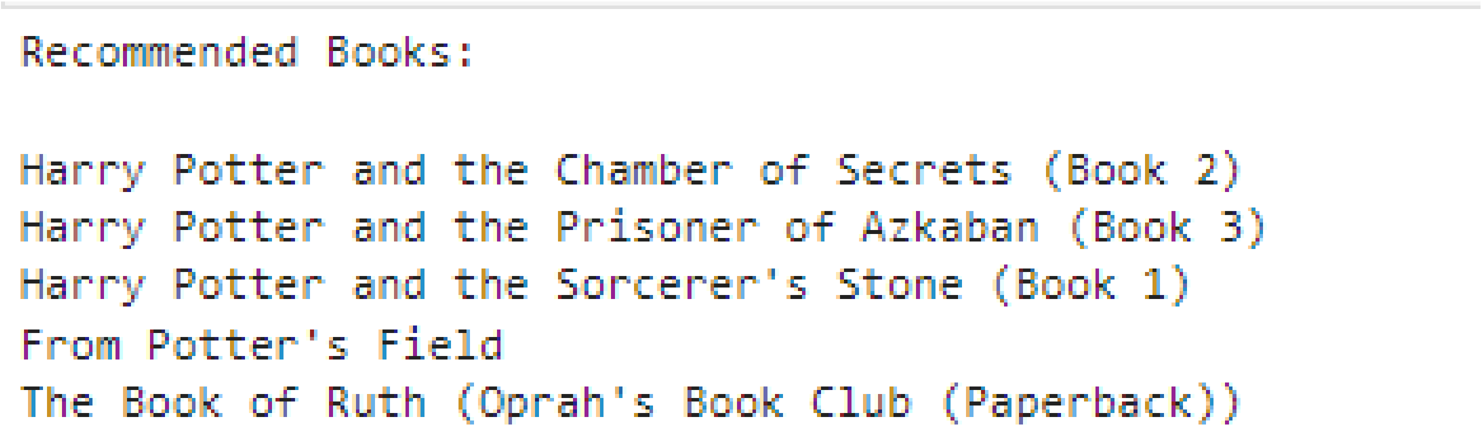


```
Enter book name:-Harry Potter and the Goblet of Fire (Book 4)
Input Book:
Harry Potter and the Goblet of Fire (Book 4)

RECOMMENDATIONS:

Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Prisoner of Azkaban (Book 3)
The Two Towers (The Lord of the Rings, Part 2)
The Fellowship of the Ring (The Lord of the Rings, Part 1)
```

8.Content Based filtering



```
Recommended Books:

Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Sorcerer's Stone (Book 1)
From Potter's Field
The Book of Ruth (Oprah's Book Club (Paperback))
```

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.
- 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.

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 30-40 and most of them came from NorthAmerican and European countries namely USA, Canada, UK, Germany and Spain. Among them USA have highest no book.
- 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.
- In book Wild Animus user gives a high no of rating which is 2.5k its so high in compare to other books.
- Author with the most books was Agatha Christie, William Shakespeare and Stephen King.
- For modelling, it was observed that for model based collaborative filtering and content based filtering doing best .
- In other aproach the avg rating The Return of the King and hary poter have most liked books .

THANK YOU