Capstone Project

Book Recommendation System

By

Shivani Thakur

CONTENT

- Problem Statement
- Data Understanding & description
- Data Information
- Data Processing
- Checking Outliers
- Data Visualization
- EDA
- EDA Conclusion
- Recommender System
- Model Evaluation
- Conclusion

PROBLEM STATMENT

A book recommendation system is a type of recommendation system where we have to recommend similar books to the reader based on his interest. The books recommendation system is used by online websites which provide ebooks like google play books, open library, good Read's, etc.

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to u sers (items being movies to watch, text to read, products to buy, or anything else depending on indu stries).

Recommendation systems are used in hundreds of different services everywhere from online shoppi ng to music to movies. Recommender systems are really critical in some industries as they can gen erate a huge amount of income when they are efficient or also be a way to stand out significantly fro m competitors.

The main objective of our project is to create book recommendation systems for users on various ap proaches .

DATA UNDERSTANDING & DESCRIPTION

We have 3 files in our dataset which is extracted from some books selling websites.

Books

ISBN - We have a unique ISBN number for all the books.

Book Title – Book title correspond to the ISBN number.

Book Author - Name of book author

Year of Publication - In which year book published

Publisher - Publishing company/house name

URL Links (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large

Users Data

User ID - A unique user id of all the users.

Location - City, state and country of the user.

Age - Age of the user

Ratings

Book Rating - Rating provide by the user for a particular book between 0-10

User ID & ISBN – Basically this to map book rating with other 2 data

DATA INFORMATION

```
1 books_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271047 entries, 0 to 271046
Data columns (total 5 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   ISBN                 271047 non-null  object
 1   Book-Title           271047 non-null  object
 2   Book-Author          271047 non-null  object
 3   Year-Of-Publication  271047 non-null  float64
 4   Publisher            271047 non-null  object
dtypes: float64(1), object(4)
memory usage: 10.3+ MB
```

```
1 rating_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1149780 entries, 0 to 1149779
Data columns (total 3 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   User-ID      1149780 non-null  int64
 1   ISBN         1149780 non-null  object
 2   Book-Rating  1149780 non-null  int64
dtypes: int64(2), object(1)
memory usage: 26.3+ MB
```

```
1 #inspecting the columns in users_df
2 users_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 278858 entries, 0 to 278857
Data columns (total 3 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   User-ID   278858 non-null  int64
 1   Location  278858 non-null  object
 2   Age       168096 non-null  float64
dtypes: float64(1), int64(1), object(1)
memory usage: 6.4+ MB
```

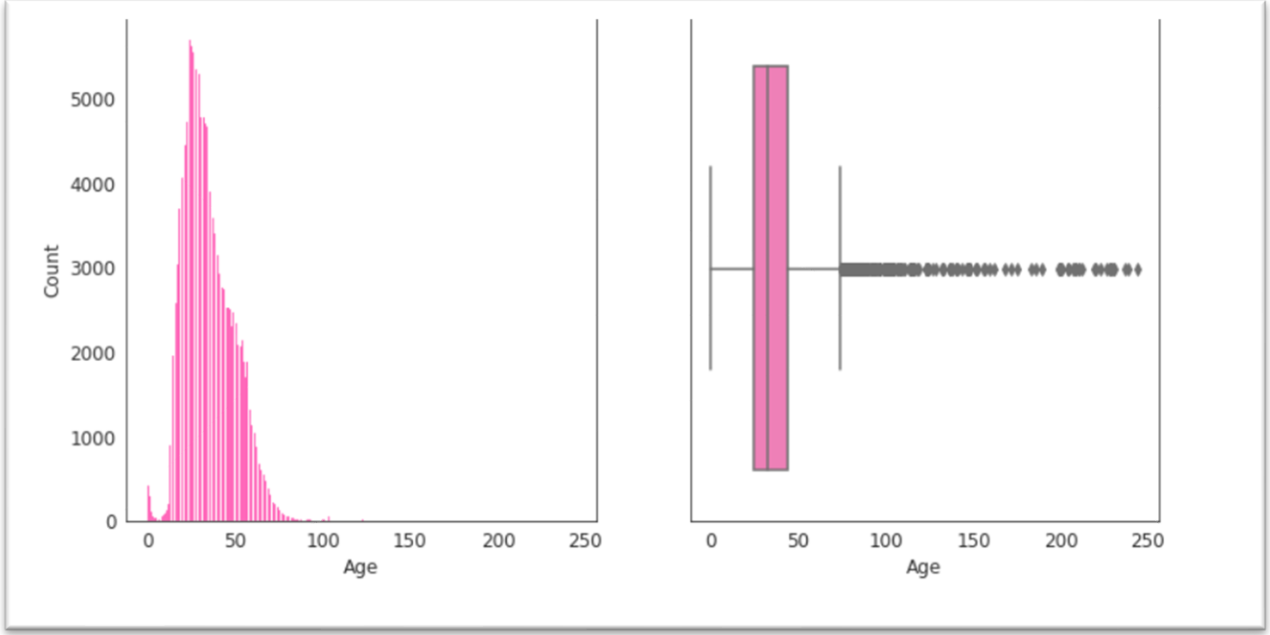
As you can see , rating df has no null objects while books and users df has null objects.

Data Processing

As we understand our data so before going further we will try to figure:

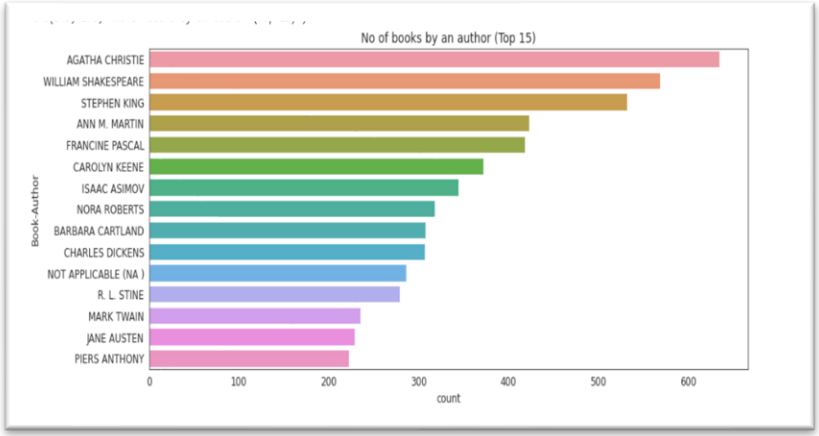
- Checked duplicate:-There are no duplicates.
- Null values and outliers .
- Replaced null values and treated outliers.
- There are some incorrect entries in data which was replaced with correct values

Checking Outliers

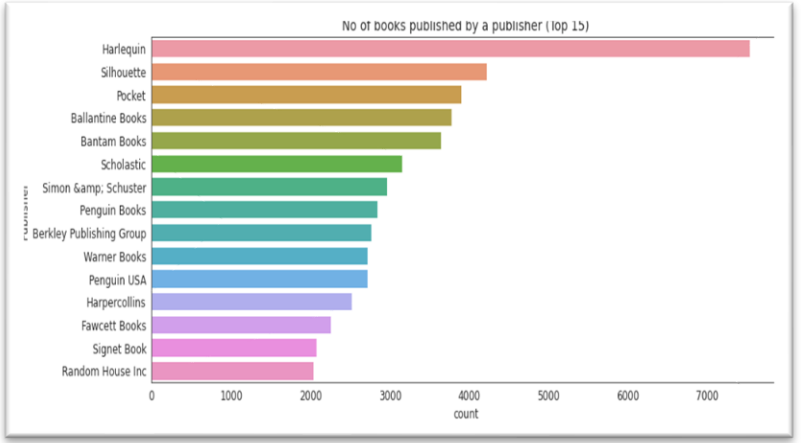


- ✓ Most of the users are from the age group 25-50.
- ✓ It is highly unlikely to have users under the age of 4 and above 100.The peaks near 0 and 100 in the kdeplot indicates that there are some outlier values in the 'Age' column
- ✓ It is highly unlikely to have users of age above 95 and below 4 in this case.Let's replace these values with np. nan

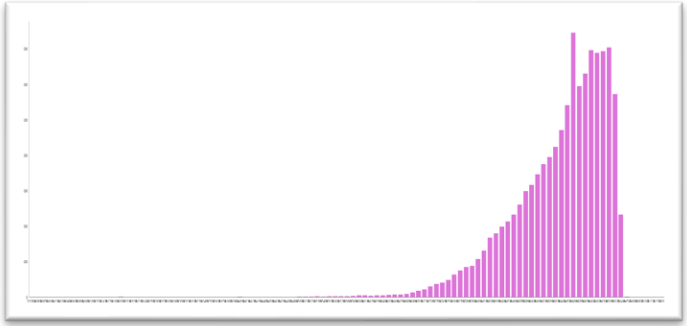
Data Visualization



Count of books by author



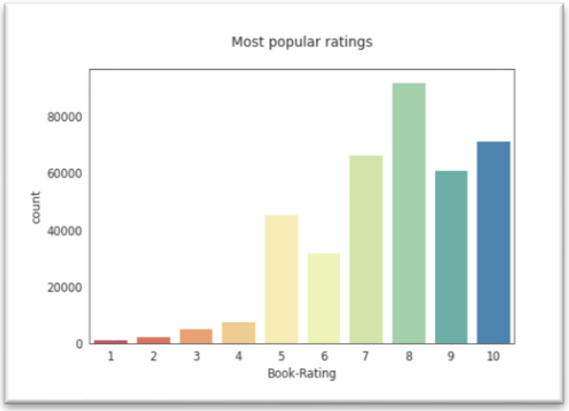
Count of top 15 books published by publisher



Count of books published yearly

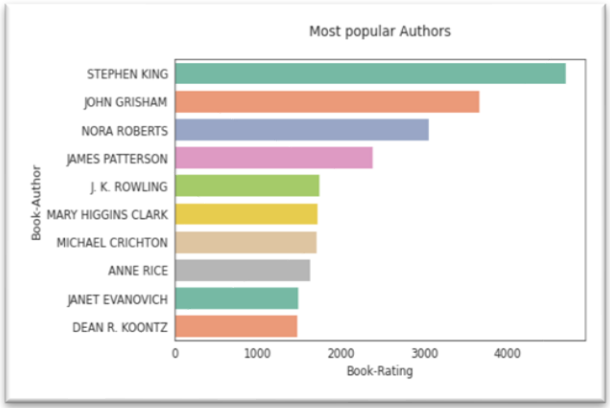
Exploratory Data Analysis

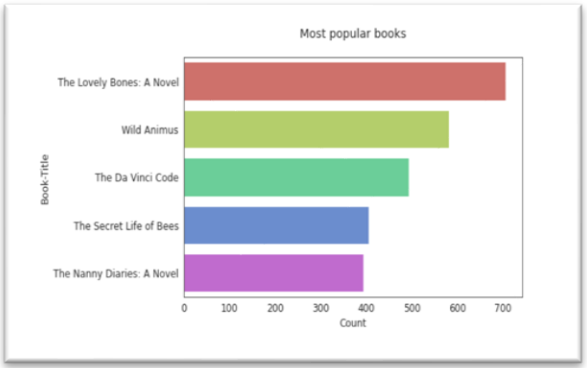
Book Data



Most of users have given above 4 ratings to books 8 is the most common rating given by users.

The most popular author is Stephen King followed by John Grisham and Nora Roberts.

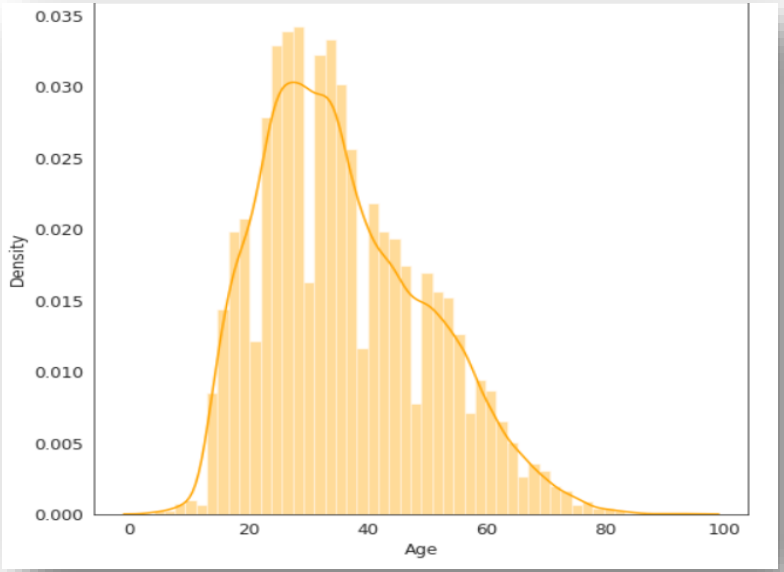




The book which has been rated by most number of users is 'The Lovely Bones'

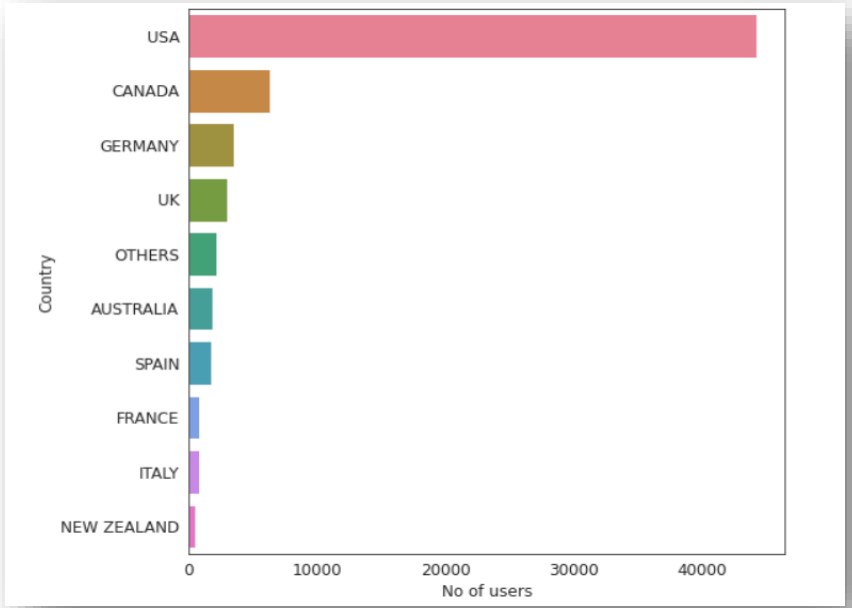
EDA : User Data

Age distribution of users



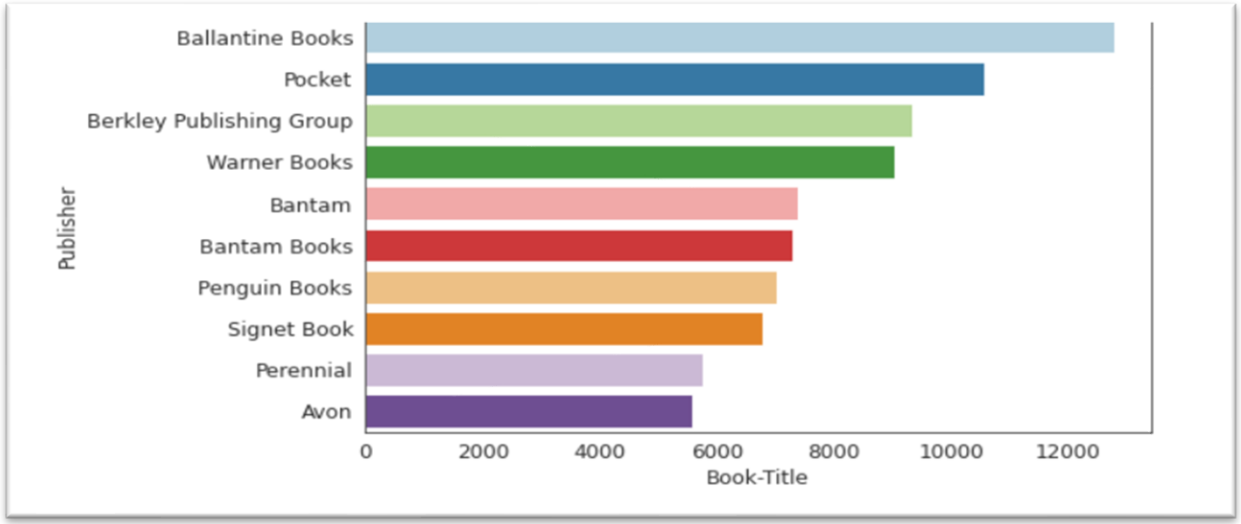
- The majority of readers are between the ages of 25 and 40.
- Readers who are 80 to 100 years old make up a tiny minority.

Countries with most readers



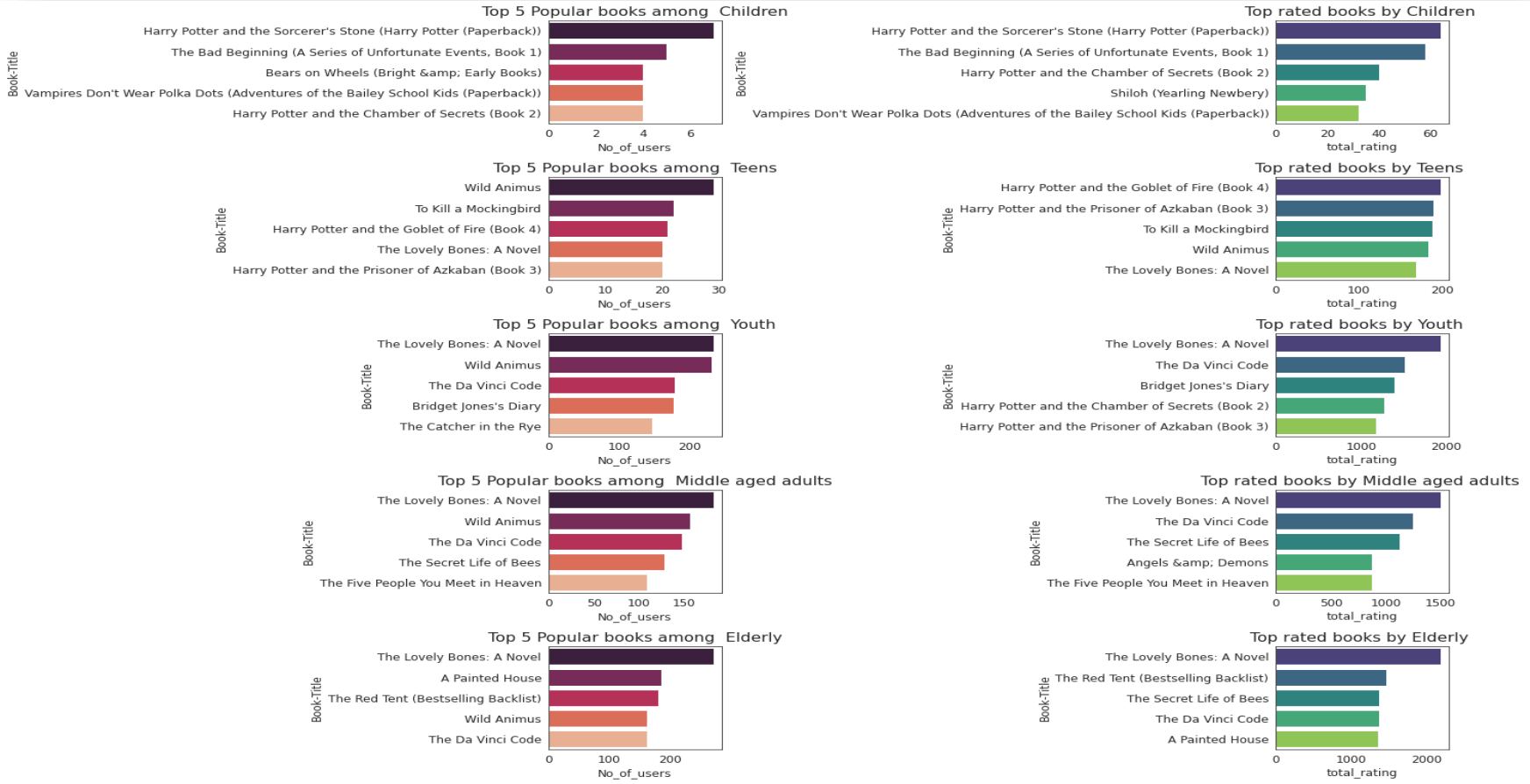
Most of the readers are from the United States. While other countries have very less users.

Publisher with most books



- ✓ Here we have some top most rated publishers.
- ✓ Ballantine Books is most popular publisher followed by Pocket and Berkley Publishing Group based on the number of users who have rated their books.
- ✓ Perennial and Avon have the least ratings

Top 5 Genres/Categories in the list



EDA CONCLUSIONS

- The Lovely Bones: A Novel and Wild Animus are the two most read books.
- Most popular book author based on the number of ratings is Stephan King .
- Ballantine Books and Pocket are the top publishers based on the number of ratings that their books have received.
- The majority of readers are between the ages of 20 and 40.
- The majority of readers who have given the books ratings are from the United States and Canada.
- Regardless of the age group, The Lovely Bones and Wild animus appear on lists of the top-rated books.

Recommender System

Popularity-Based Recommendation System

Collaborative filtering

Memory Based Approach

KNN with cosine metric

Collaborative filtering

- SVD (Singular Value Decomposition)

Popularity Based Recommender System

It is a type of recommendation system that bases choices on factors like popularity and/or current trends

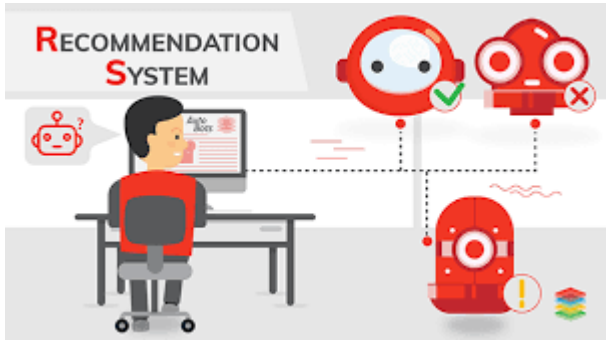
	ISBN	Book-Rating	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	0316666343	707	The Lovely Bones: A Novel	ALICE SEBOLD	2002.0	Little, Brown
1	0971880107	581	Wild Animus	RICH SHAPERO	2004.0	Too Far
2	0385504209	487	The Da Vinci Code	DAN BROWN	2003.0	Doubleday
3	0312195516	383	The Red Tent (Bestselling Backlist)	ANITA DIAMANT	1998.0	Picador USA
4	0060928336	320	Divine Secrets of the Ya-Ya Sisterhood: A Novel	REBECCA WELLS	1997.0	Perennial
5	059035342X	313	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	J. K. ROWLING	1999.0	Arthur A. Levine Books
6	0142001740	307	The Secret Life of Bees	SUE MONK KIDD	2003.0	Penguin Books
7	0446672211	295	Where the Heart Is (Oprah's Book Club (Paperback))	BILLIE LETTS	1998.0	Warner Books
8	044023722X	281	A Painted House	JOHN GRISHAM	2001.0	Dell Publishing Company
9	0452282152	278	Girl with a Pearl Earring	TRACY CHEVALIER	2001.0	Plume Books

This table indicates top 10 books based on rating given by maximum users.

Country-based book recommendation

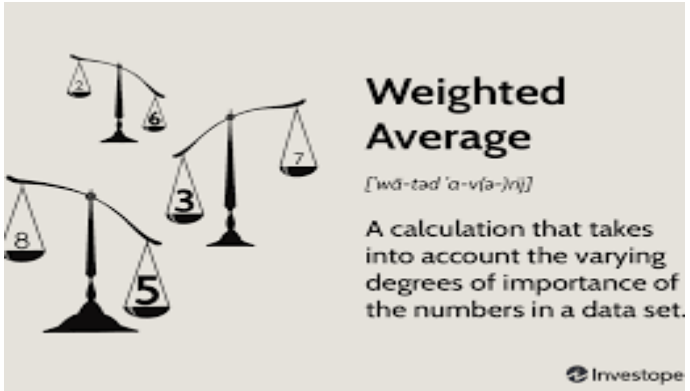
most popular recommendations country wise
E.g: India

	ISBN	Book-Rating	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	0971880107	3	Wild Animus	RICH SHAPERO	2004.0	Too Far
1	0671047612	2	Skin And Bones	FRANKLIN W. DIXON	2000.0	Aladdin
2	0486284735	2	Pride and Prejudice (Dover Thrift Editions)	JANE AUSTEN	1995.0	Dover Publications
3	8171670407	2	Inscrutable Americans	MATHUR ANURAG	1996.0	South Asia Books
4	0006944035	1	Secret Island / Secret Mountain (Two-in-ones)	ENID BLYTON	1994.0	HarperCollins Publishers



Weighted average rating method

- Using Weighted average for each Book’s Average Rating
- W = (Rv + Cm)/(v + m)
- where
- W= Weighted Rating
- R = Average of the Books rating
- v = No of people who have rated the books(number of votes)
- m = minimum no of votes to be listed
- C = the mean rating across all the books



Weighted average rating of the books

	Book-Title	Book-Author	avg_rating	ratings_count	weighted_average
46516	Harry Potter and the Chamber of Secrets Postcard Book	J. K. ROWLING	9.869565	23	9.52
122145	The Two Towers (The Lord of the Rings, Part 2)	J. R. R. TOLKIEN	9.653846	52	9.50
30142	Dilbert: A Book of Postcards	SCOTT ADAMS	9.923077	13	9.36
81784	Postmarked Yesteryear: 30 Rare Holiday Postcards	PAMELA E. APKARIAN-RUSSELL	10.000000	11	9.34
118127	The Return of the King (The Lord of the Rings, Part 3)	J.R.R. TOLKIEN	9.397436	78	9.31
17713	Calvin and Hobbes	BILL WATTERSON	9.583333	24	9.29
100902	The Authoritative Calvin and Hobbes (Calvin and Hobbes)	BILL WATTERSON	9.600000	20	9.25
72637	My Sister's Keeper : A Novel (Picoult, Jodi)	JODI PICOULT	9.545455	22	9.23
118123	The Return of the King (The Lord of The Rings, Part 3)	J. R. R. TOLKIEN	9.625000	16	9.20
120090	The Sneetches and Other Stories	DR. SEUSS	10.000000	8	9.17

This is the list of most favourite books based on the weighted rating scores. The book 'Harry Potter and the Chamber of Secrets Postcard Book' seems to have top this chart.

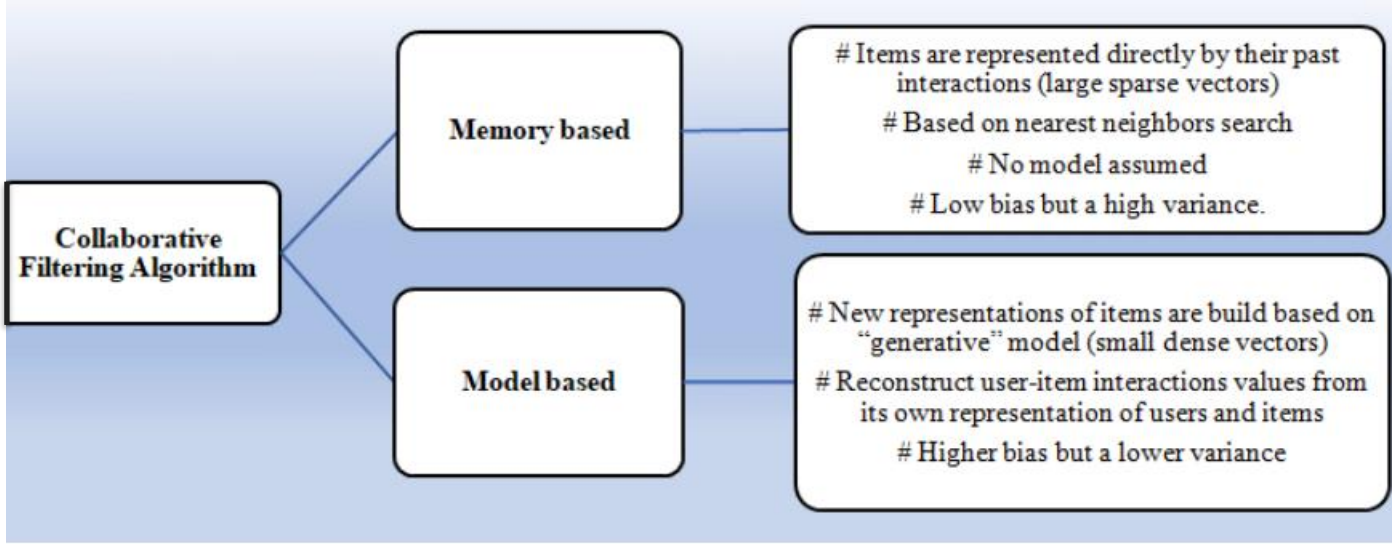
Author based recommender system

	Book-Title	weighted_average
46516	Harry Potter and the Chamber of Secrets Postcard Book	9.52
46520	Harry Potter and the Goblet of Fire (Book 4)	9.10
46532	Harry Potter and the Prisoner of Azkaban (Book 3)	9.02
46539	Harry Potter and the Sorcerer's Stone (Book 1)	9.02
46524	Harry Potter and the Order of the Phoenix (Book 5)	9.01

- This recommender system recommends books from the same author as entered by the user.
- The author of the book Harry Potter and the Chamber of Secrets is J. K. ROWLING
- Here are the top 5 books from the same author

Collaborative filtering:

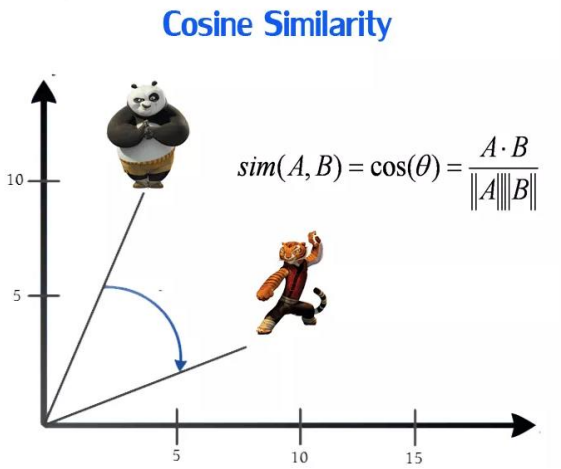
- **Collaborative Filtering** is a Machine Learning technique used to identify relationships between pieces of data. This technique is frequently used in recommender systems to identify similarities between user data and items.
- This means that if *Users A* and *B* both like *Product A*, and *User B* also likes *Product B*, then *Product B* could be recommended to *User A* by the system.



Memory Based Approach

KNN Based Algorithm

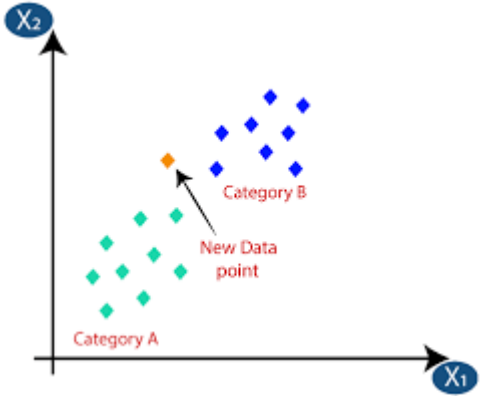
This algorithm takes into consideration up-to 'K' nearest users (in user based collaborative filtering) or 'K' nearest items (in item based collaborative filtering) for making recommendations. By default, the algorithm is 'user-based', and k is 40 (km in is 1). This means ratings of 40 nearest users are considered while recommending an item to a user. Some variants of this algorithm include with Means, with Zscore & Baseline wherein the average rating of users, or the normalized ZScore of ratings or the baseline rating are also considered as the system generates recommendations



```
The top 10 Recommended books for Harry Potter and the Chamber of Secrets (Book 2) are:

Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Dr. Seuss's A B C (I Can Read It All by Myself Beginner Books)
The Second Generation
Lover Beware
J. K. Rowling: The Wizard Behind Harry Potter
A Dash of Death
So Much to Tell You
Dragonquest Achille Cover
```

The top 10 Recommended books

KNN Based



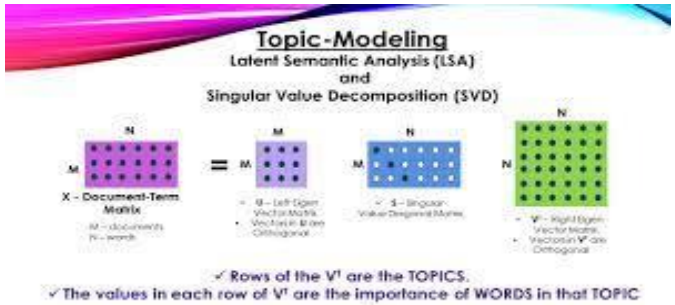

```
The top 10 Recommended books for Harry Potter and the Chamber of Secrets (Book 2) are:

Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Order of the Phoenix (Book 5)
Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
The Fellowship of the Ring (The Lord of the Rings, Part 1)
The Hobbit: or There and Back Again
The Two Towers (The Lord of the Rings, Part 2)
Dr. Seuss's A B C (I Can Read It All by Myself Beginner Books)
The Second Generation
```

KNN with cosine metric

Model Based Approach

SVD (Singular Value Decomposition)

This algorithm takes a matrix factorization approach. The user-item rating matrix is factorized into smaller dimension user & item matrices consisting of latent factors (hidden characteristics). By default, number of latent factors is 100. These latent factors are able to capture the known user-item rating preference & in the process are able to predict an estimated rating for all user-item pair where user has not yet rated an item



These are books that the user ID 254 has already rated

```
These are  books that the user ID 254 has already rated

274                                  The Golden Compass (His Dark Materials, Book 1)
282                                                                Making Minty Malone
285                                                                        Animal Farm
356                                                                The Secret Life of Bees
485                                            She's Come Undone (Oprah's Book Club)
1175                                                               American Gods
2785                                                 The Hobbit: or There and Back Again
2809                                   Harry Potter and the Sorcerer's Stone (Book 1)
2969                                                               The Bonesetter's Daughter
3459                                 Harry Potter and the Chamber of Secrets (Book 2)
3839                                Harry Potter and the Prisoner of Azkaban (Book 3)
4241                                                               American Gods: A Novel
5431                                     Harry Potter and the Goblet of Fire (Book 4)
5432                                  Harry Potter and the Chamber of Secrets (Book 2)
6096                                                                       The Dark Half
6330                                Harry Potter and the Prisoner of Azkaban (Book 3)
9253                                  The Golden Compass (His Dark Materials, Book 1)
11033                Familiar Lullaby (Fear Familiar) (Harlequin Intrigue, No 614)
12761                  The Fellowship of the Ring (The Lord of the Rings, Part 1)
14779                                                                      The Duke
15509                                                  Complete Chronicles of Narnia
15511                                                                      Stardust
15536     Amazing Grace : Lives of Children and the Conscience of a Nation, The
15537                                                  Something Wicked This Way Comes
Name: Book-Title, dtype: object
```

Recommending books for User ID: 254

Recommending books for User ID: 254

	ISBN	Book-Title	Book-Author	Publisher
0	043935806X	Harry Potter and the Order of the Phoenix (Book 5)	J. K. ROWLING	Scholastic
1	059035342X	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	J. K. ROWLING	Arthur A. Levine Books
2	0439139600	Harry Potter and the Goblet of Fire (Book 4)	J. K. ROWLING	Scholastic Paperbacks
3	0446310786	To Kill a Mockingbird	HARPER LEE	Little Brown & Company
4	0385504209	The Da Vinci Code	DAN BROWN	Doubleday
5	0345339681	The Hobbit : The Enchanting Prelude to The Lord of the Rings	J.R.R. TOLKIEN	Del Rey
6	0385484518	Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	MITCH ALBOM	Doubleday
7	0316769487	The Catcher in the Rye	J.D. SALINGER	Little, Brown
8	0345339703	The Fellowship of the Ring (The Lord of the Rings, Part 1)	J.R.R. TOLKIEN	Del Rey
9	0345339711	The Two Towers (The Lord of the Rings, Part 2)	J.R.R. TOLKIEN	Del Rey

SVD and NMF models comparison Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF) are matrix factorization techniques used for dimensionality reduction. Surprise package provides implementation of those algorithms.

Model Evaluation

In Recommender Systems, there are a set metrics commonly used for evaluation. We choose to work with Top-N 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 follows:

- 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 Top-N accuracy metrics for this user and interacted item from the recommendations ranked list
- Aggregate the global Top-N accuracy metrics

Evaluating Collaborative Filtering (SVD Matrix Factorization) model

Global metrics

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	_person_id
36	35	89	545	0.064220	0.163303	11676
202	62	81	139	0.446043	0.582734	98391
271	27	35	93	0.290323	0.376344	153662
60	22	36	88	0.250000	0.409091	16795
474	16	28	73	0.219178	0.383562	95359
485	52	60	72	0.722222	0.833333	114368
390	32	32	61	0.524590	0.524590	104636
456	17	24	54	0.314815	0.444444	158295
660	34	43	54	0.629630	0.796296	123883
659	7	13	53	0.132075	0.245283	35859

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 evaluated the performance of Singular Value Decomposition based recommender and obtained a Global Recall@5 of 30 % and Recall@10 of 41%

Thank You!