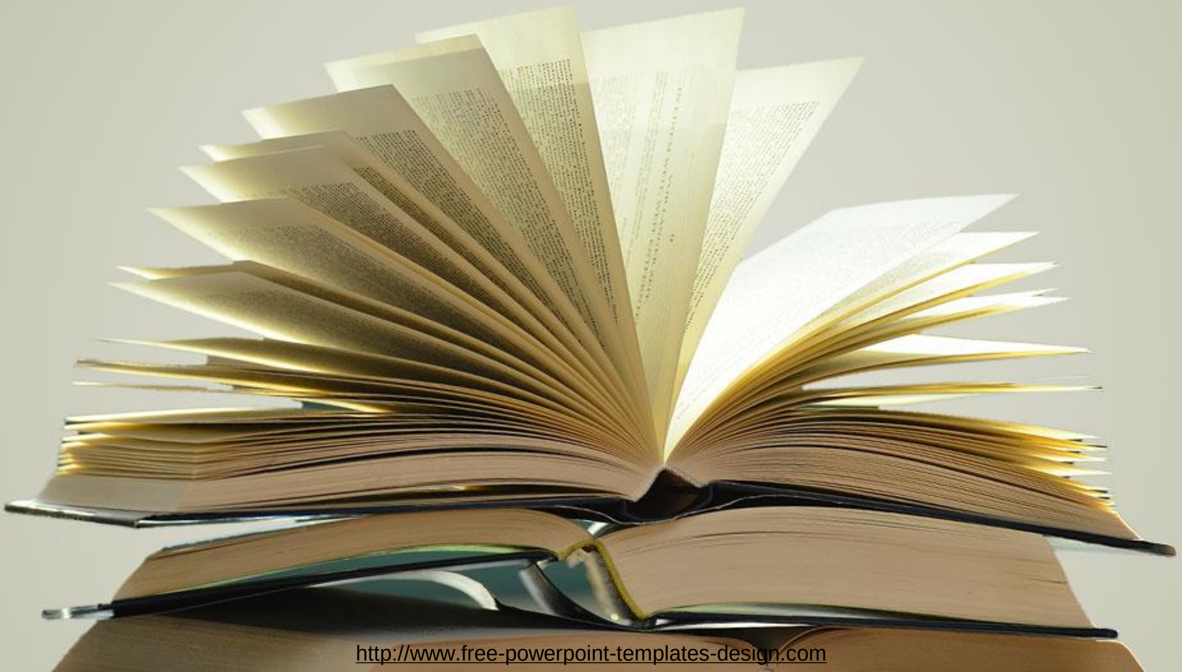
BOOK RECOMMENDATION SYSTEMS

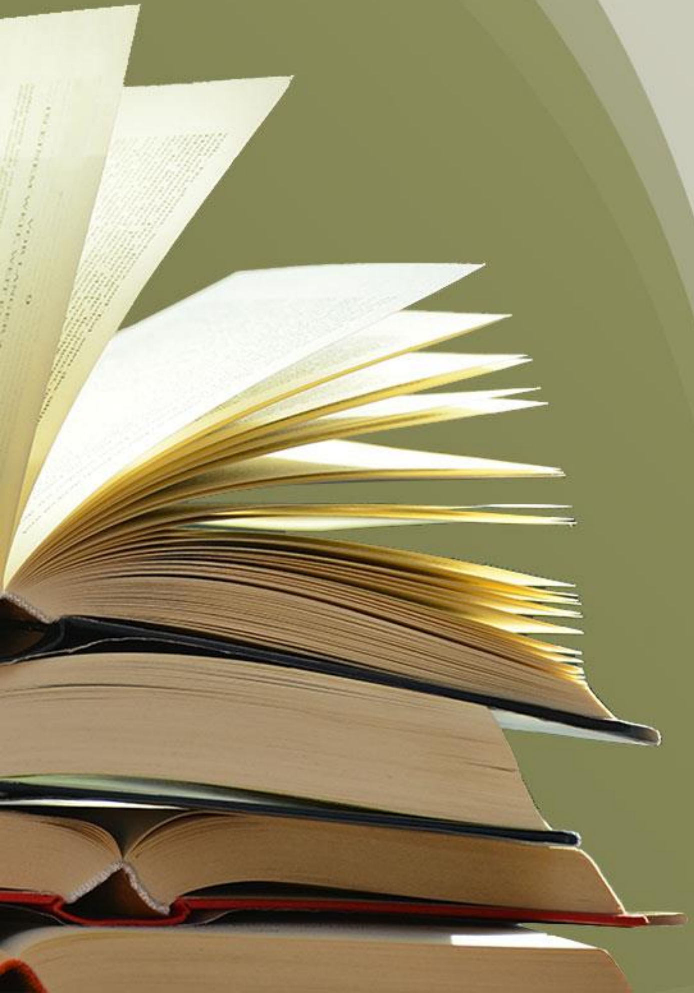
Lestari Aprina - Final Project

Data Science Batch 22

December 2023



http://www.free-powerpoint-templates-design.com



Content

1 Project Background

2 Data Understanding & Data Preprocessing

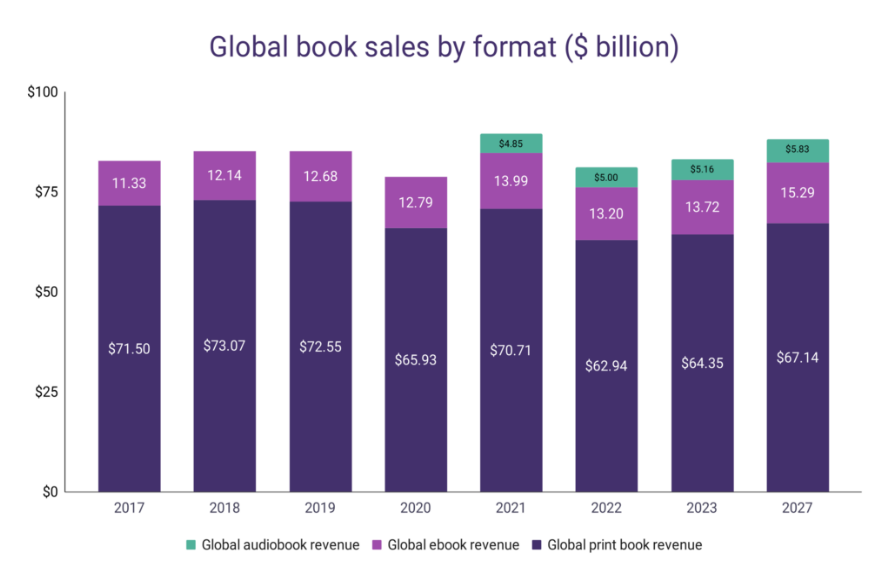
3 Exploratory Data Analysis

4 Modelling

5 Summary & Recommendation

Project Background

Project Background



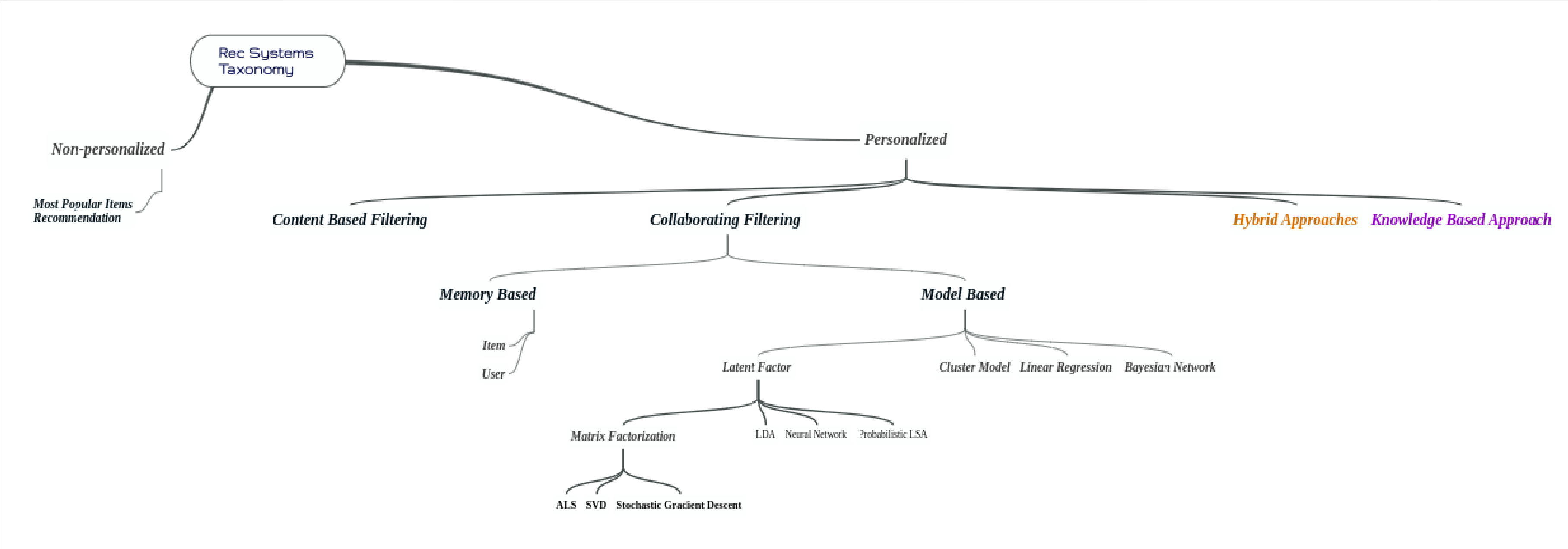
Source: https://wordsrated.com/global-book-sales-statistics/#:~:text=As%20of%202023%2C%20the%20global,to%202021%2C%20with%20%2476.14%20billion.

The emergence of several web services over the past few decades has made recommender systems increasingly prevalent in our daily lives.

The main objective is to create a book recommendation system for the users.

In certain businesses, recommender systems play a vital role because, when implemented well, they can yield big profits or serve as a means of differentiating oneself from the competition.

Project Background



Rec Systems Taxonomy

Non-personalized

Most Popular Items Recommendation

Personalized

Content Based Filtering

Collaborating Filtering

Hybrid Approaches

Knowledge Based Approach

Memory Based

Item

User

Model Based

Latent Factor

Cluster Model

Linear Regression

Bayesian Network

Matrix Factorization

LDA

Neural Network

Probabilistic LSA

ALS

SVD

Stochastic Gradient Descent

Source: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/data

Data Understanding & Data Preprocessing

Dataset

COLLECTION METHODOLOGY

Collected by Cai-Nicolas Ziegler in a 4-week crawl (2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems.

Contains 278,858 users(anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

Books

Identified by their respective ISBN

Shape of Dataset (271360, 8)

Users

User-ID (unique for each user)

Shape of Dataset (278858, 3)

Ratings

Expressed on a scale from 1-10

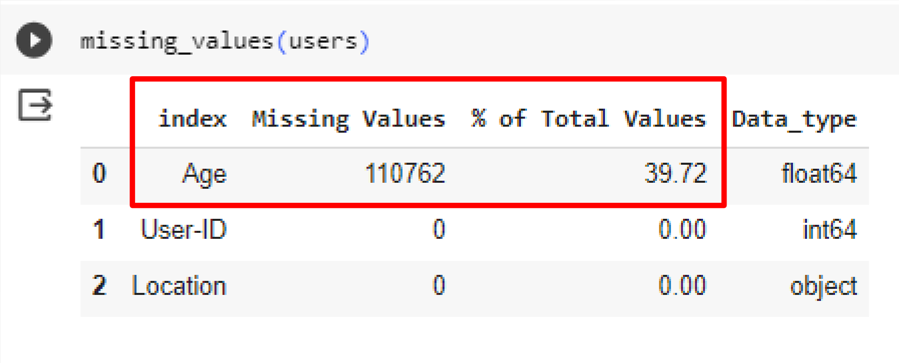
Shape of Dataset (1149780, 3)

Source: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/data

Data Preprocessing

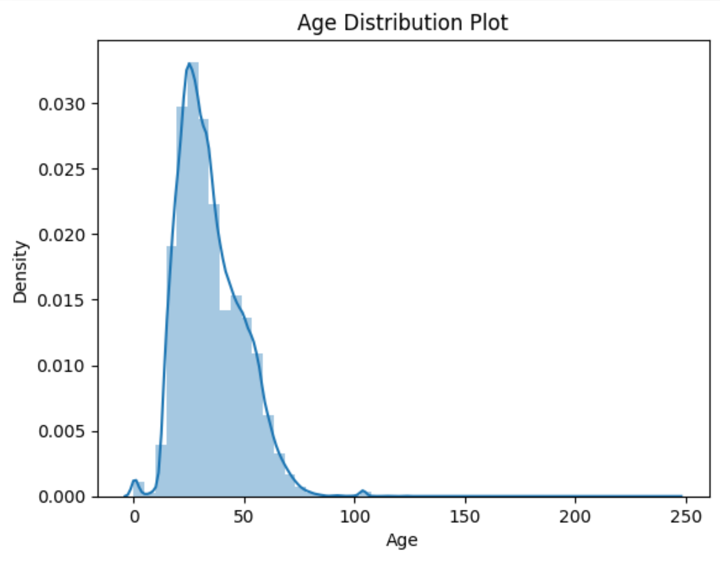
1. Users Dataset

Age column - Missing values and outliers handling



```
missing_values(users)
```

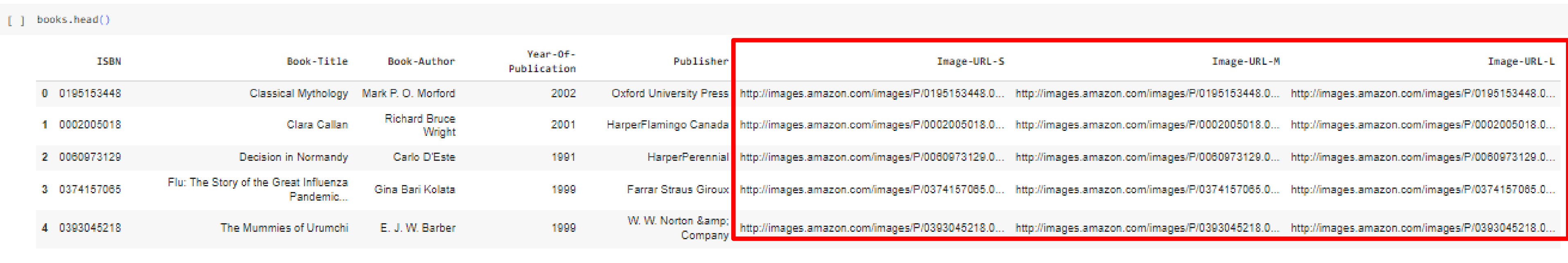
	index	Missing Values	% of Total Values	Data_type
0	Age	110762	39.72	float64
1	User-ID	0	0.00	int64
2	Location	0	0.00	object



Data Preprocessing

2. Books Dataset

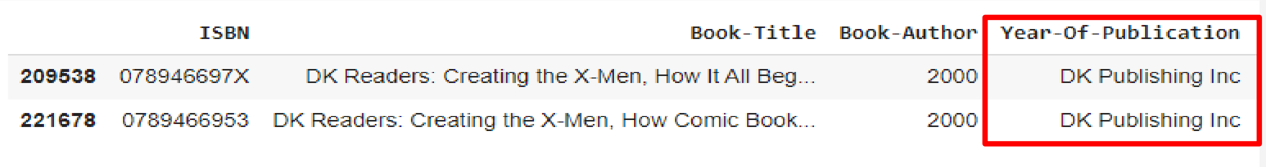
Drop unnecessary columns



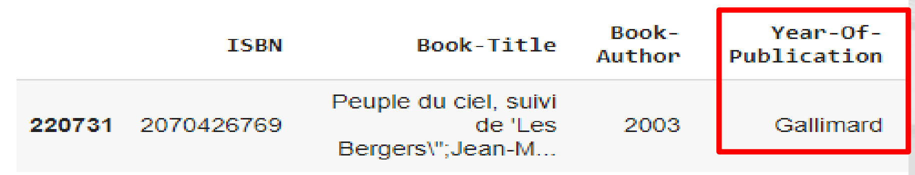
```
[ ] books.head()
```

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Image-URL-M	Image-URL-L
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...
3	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...
4	0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton &amp; Company	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...

Replace strings by int values to do some corrections due to error in the dataset



	ISBN	Book-Title	Book-Author	Year-Of-Publication
209538	078946697X	DK Readers: Creating the X-Men, How It All Beg...	2000	DK Publishing Inc
221678	0789466953	DK Readers: Creating the X-Men, How Comic Book...	2000	DK Publishing Inc

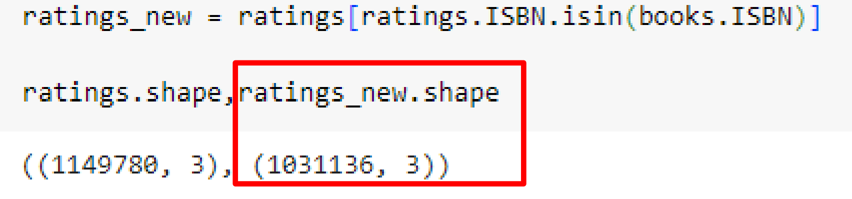


	ISBN	Book-Title	Book-Author	Year-Of-Publication
220731	2070426769	Peuple du ciel, suivi de 'Les Bergers\";Jean-M...	2003	Gallimard

Data Preprocessing

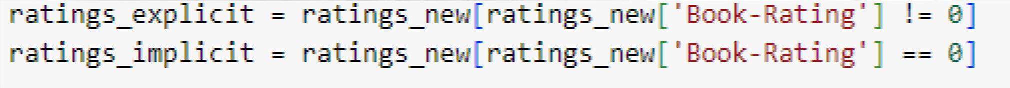
3. Ratings Dataset

Drop rows having book ISBN which are not part of books dataset

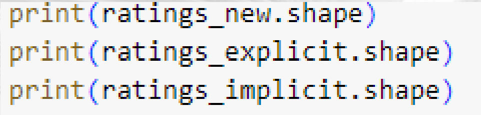


```
ratings_new = ratings[ratings.ISBN.isin(books.ISBN)]

ratings.shape,ratings_new.shape

((1149780, 3), (1031136, 3))
```

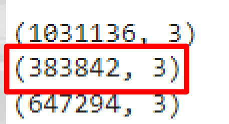
Segregating implicit and explicit ratings datasets



```
ratings_explicit = ratings_new[ratings_new['Book-Rating'] != 0]
ratings_implicit = ratings_new[ratings_new['Book-Rating'] == 0]
```



```
print(ratings_new.shape)
print(ratings_explicit.shape)
print(ratings_implicit.shape)
```

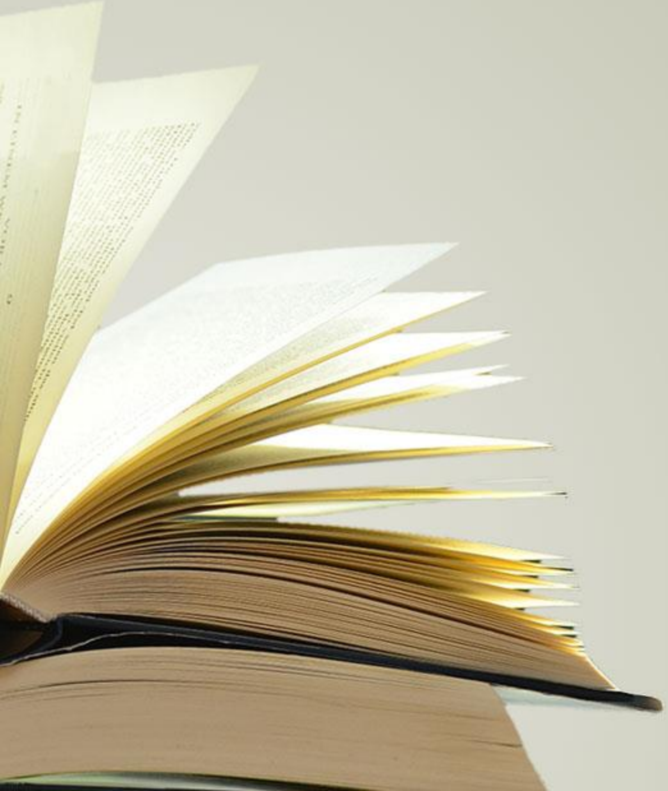


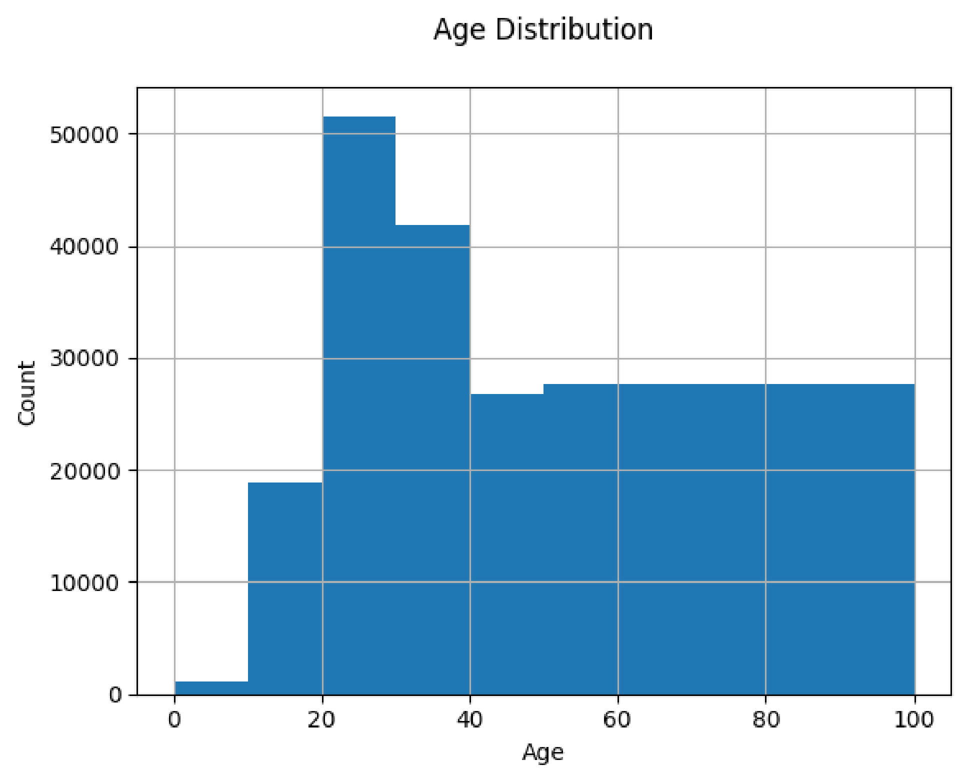
```
(1031136, 3)
(383842, 3)
(647294, 3)
```

Exploratory Data Analysis

Users Age Distribution

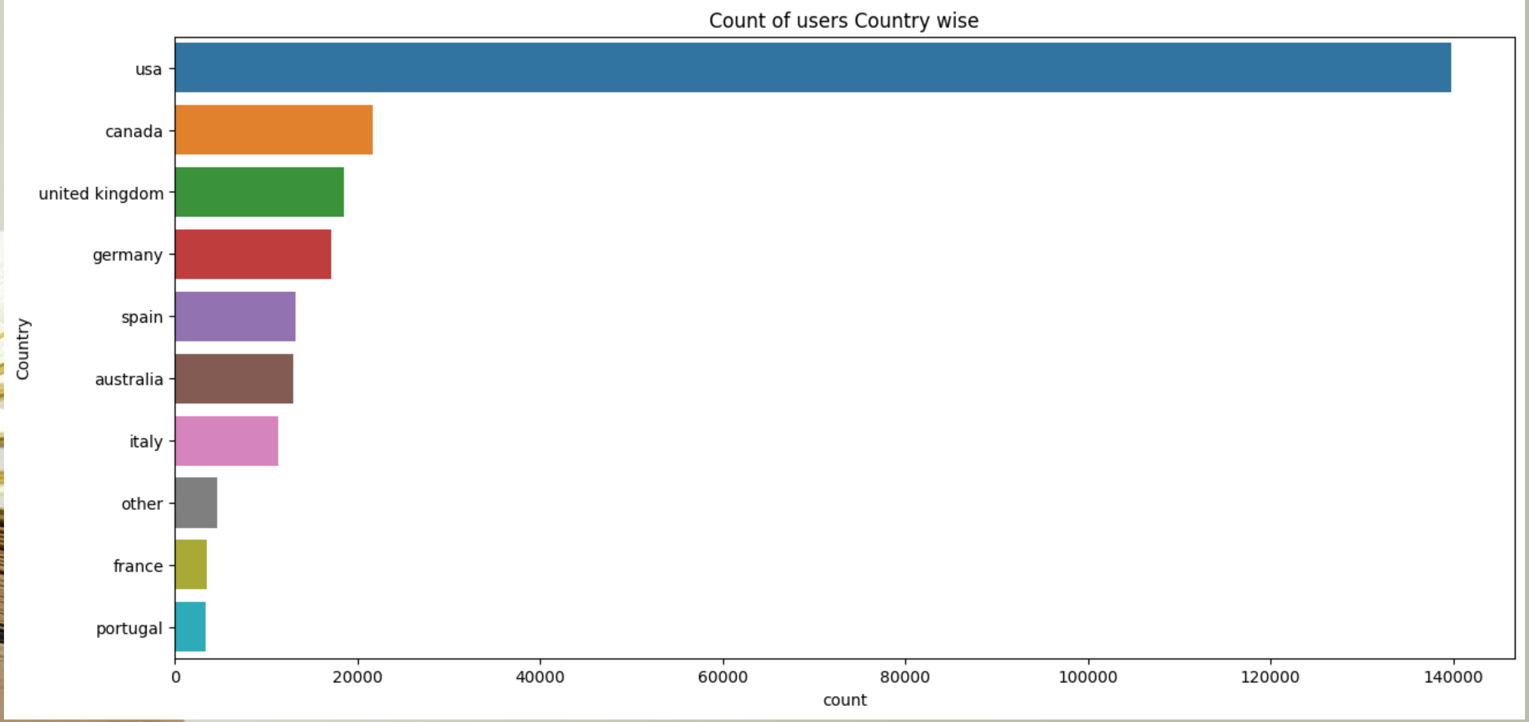
What is the age range of the most active users?





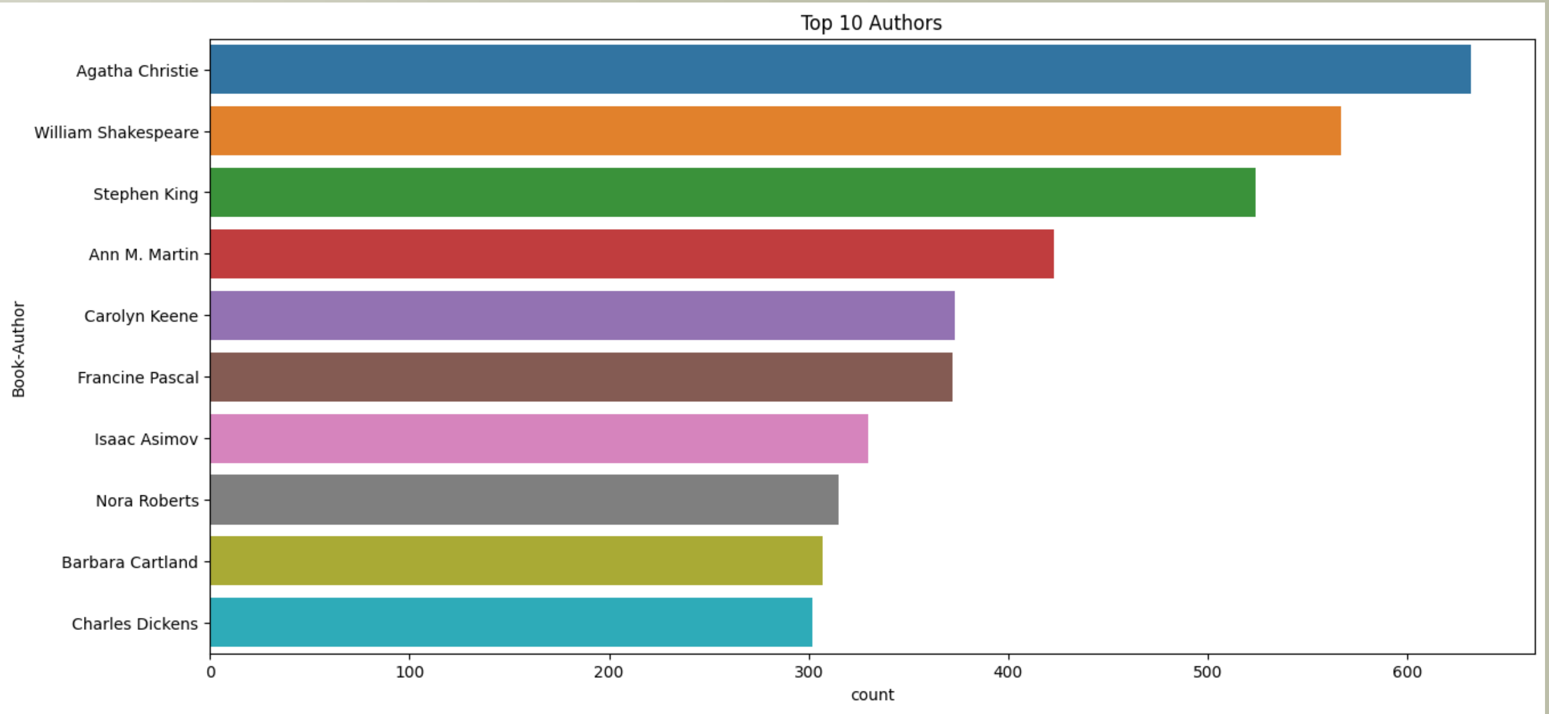
Users Location Distribution

Where is the country of the most active users come from?



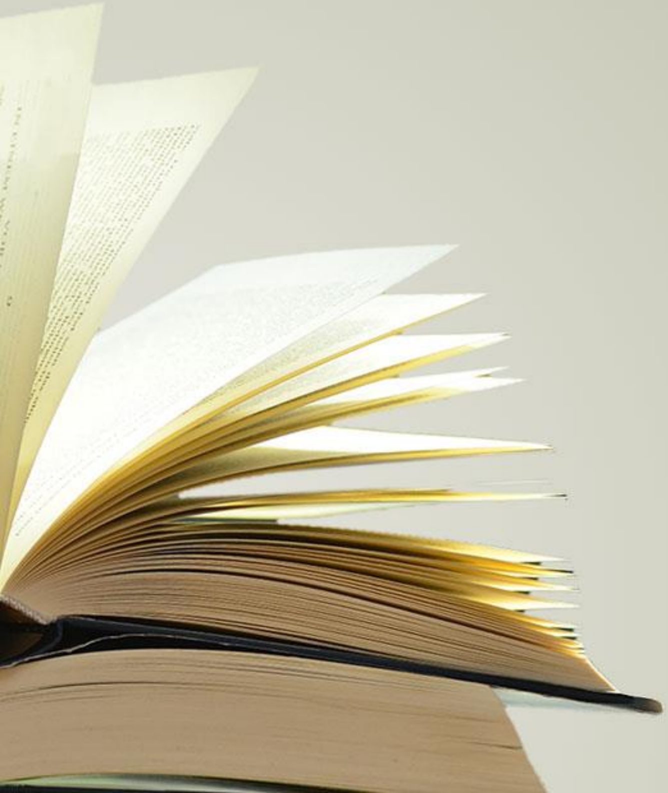
Book Authors

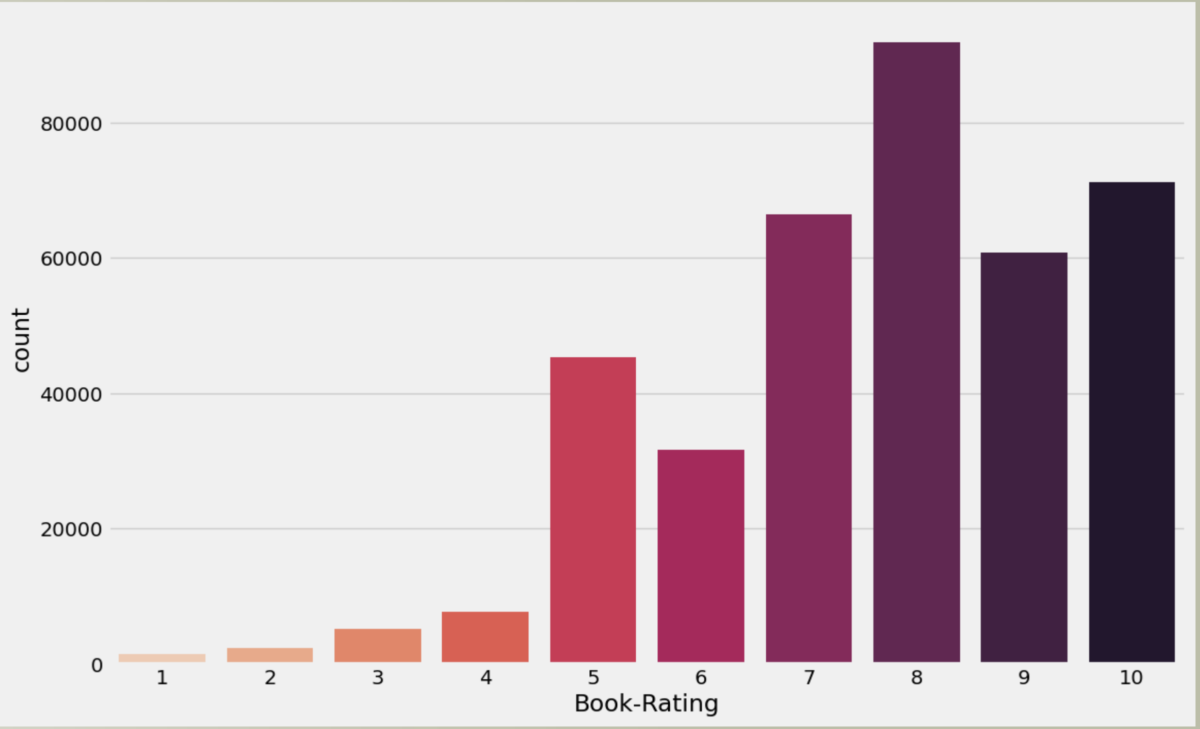
Who is the author which have written the most books?



Ratings Distribution

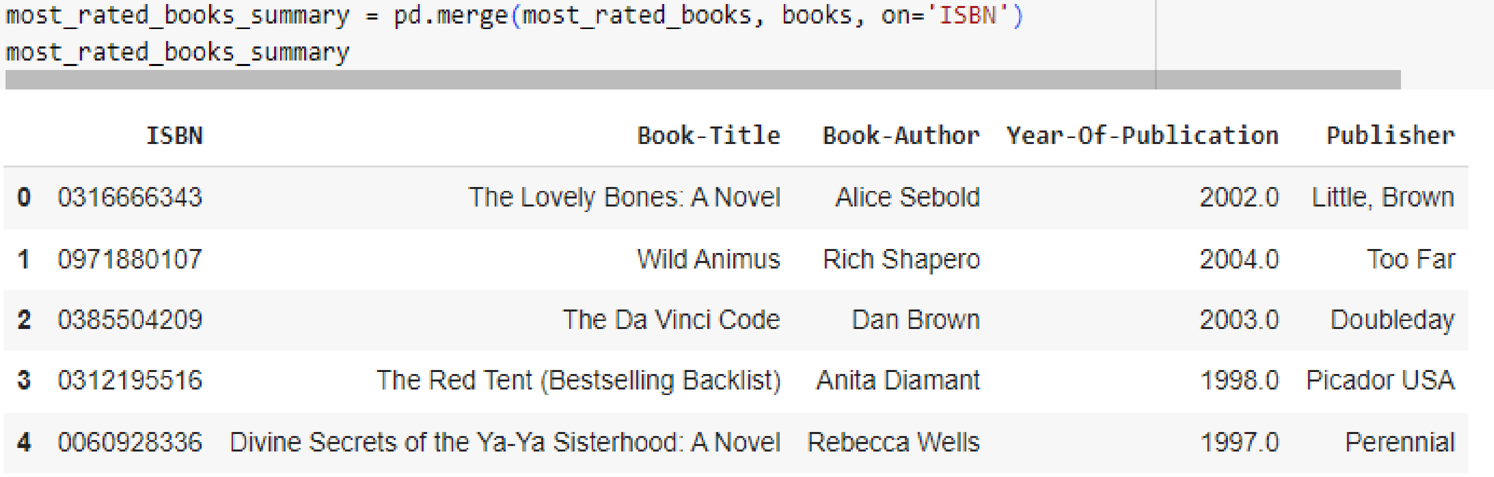
What is the highest rating for the most of the books?





Ratings Distribution

What is the type/genre of most rated books?



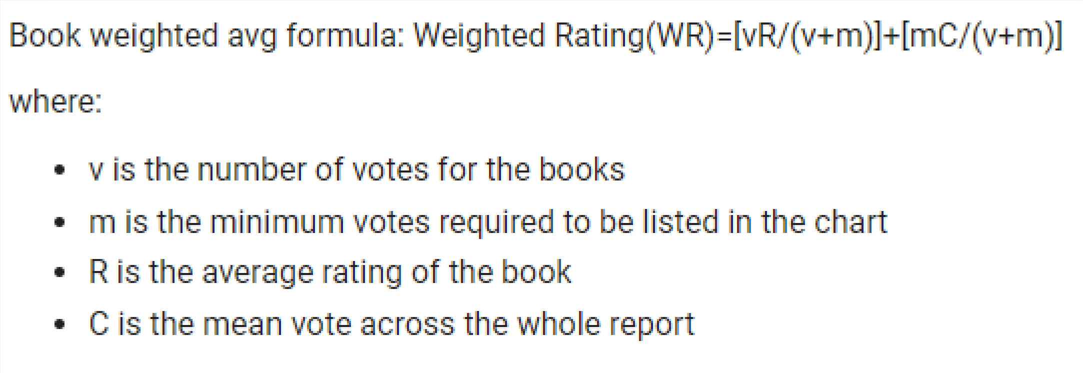
```
most_rated_books_summary = pd.merge(most_rated_books, books, on='ISBN')
most_rated_books_summary
```

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	0316666343	The Lovely Bones: A Novel	Alice Sebold	2002.0	Little, Brown
1	0971880107	Wild Animus	Rich Shapero	2004.0	Too Far
2	0385504209	The Da Vinci Code	Dan Brown	2003.0	Doubleday
3	0312195516	The Red Tent (Bestselling Backlist)	Anita Diamant	1998.0	Picador USA
4	0060928336	Divine Secrets of the Ya-Ya Sisterhood: A Novel	Rebecca Wells	1997.0	Perennial

Modelling

Modelling

1. Popularity Based Filtering



Book weighted avg formula: Weighted Rating(WR)=[vR/(v+m)]+[mC/(v+m)]

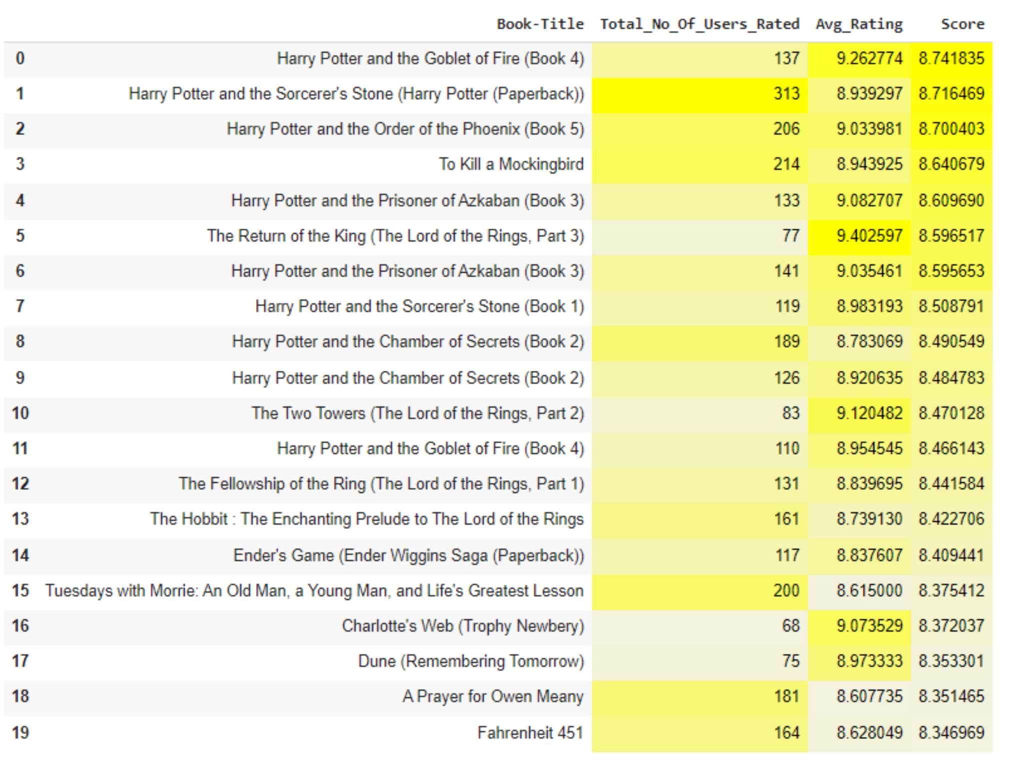
where:

- v is the number of votes for the books
- m is the minimum votes required to be listed in the chart
- R is the average rating of the book
- C is the mean vote across the whole report

Source: https://medium.com/the-owl/recommender-systems-f62ad843f70c

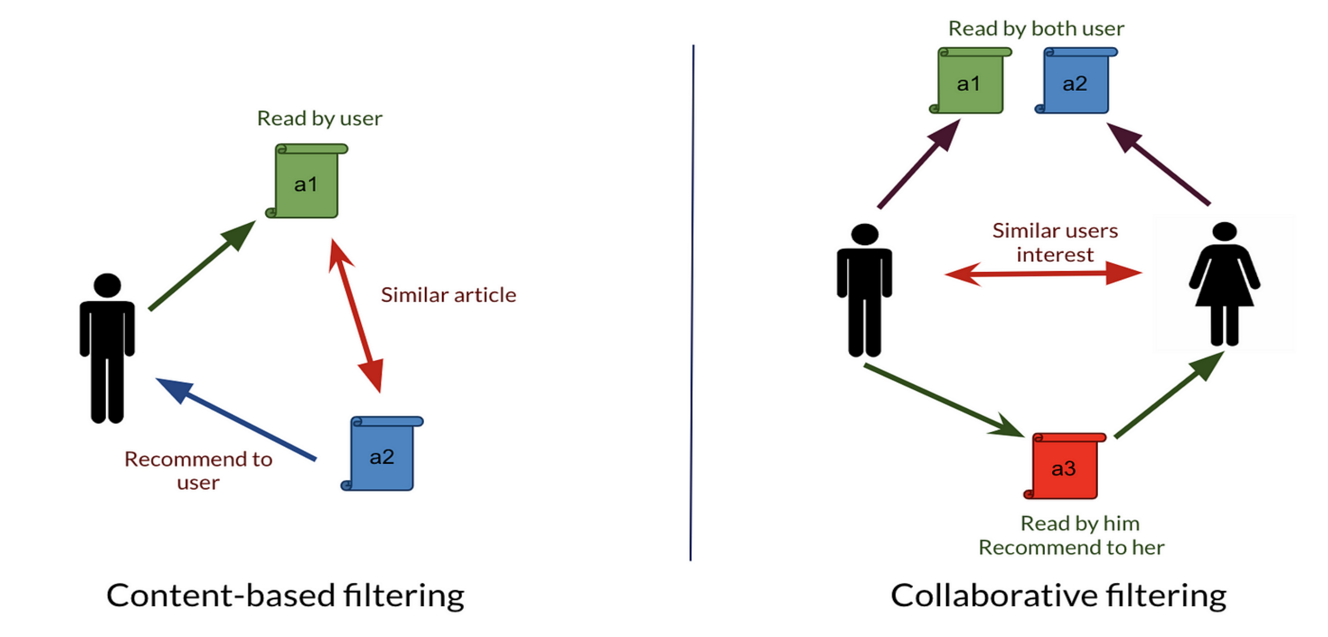
Modelling

1. Popularity Based Filtering



	Book-Title	Total_No_Of_Users_Rated	Avg_Rating	Score
0	Harry Potter and the Goblet of Fire (Book 4)	137	9.262774	8.741835
1	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	313	8.939297	8.716469
2	Harry Potter and the Order of the Phoenix (Book 5)	206	9.033981	8.700403
3	To Kill a Mockingbird	214	8.943925	8.640679
4	Harry Potter and the Prisoner of Azkaban (Book 3)	133	9.082707	8.609690
5	The Return of the King (The Lord of the Rings, Part 3)	77	9.402597	8.596517
6	Harry Potter and the Prisoner of Azkaban (Book 3)	141	9.035461	8.595653
7	Harry Potter and the Sorcerer's Stone (Book 1)	119	8.983193	8.508791
8	Harry Potter and the Chamber of Secrets (Book 2)	189	8.783069	8.490549
9	Harry Potter and the Chamber of Secrets (Book 2)	126	8.920635	8.484783
10	The Two Towers (The Lord of the Rings, Part 2)	83	9.120482	8.470128
11	Harry Potter and the Goblet of Fire (Book 4)	110	8.954545	8.466143
12	The Fellowship of the Ring (The Lord of the Rings, Part 1)	131	8.839695	8.441584
13	The Hobbit : The Enchanting Prelude to The Lord of the Rings	161	8.739130	8.422706
14	Ender's Game (Ender Wiggins Saga (Paperback))	117	8.837607	8.409441
15	Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	200	8.615000	8.375412
16	Charlotte's Web (Trophy Newbery)	68	9.073529	8.372037
17	Dune (Remembering Tomorrow)	75	8.973333	8.353301
18	A Prayer for Owen Meany	181	8.607735	8.351465
19	Fahrenheit 451	164	8.628049	8.346969

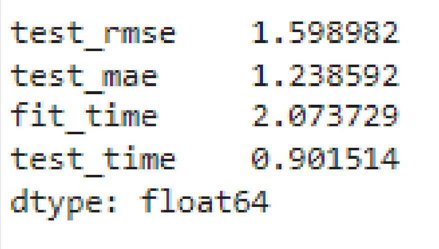
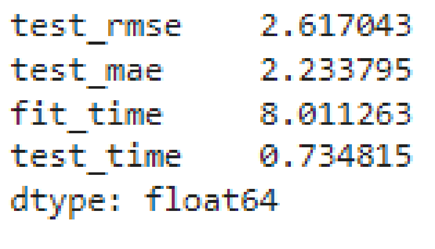
Modelling



Source: https://tech.timesinternet.in/recommendation-system-from-economic-times-3ad53871af40

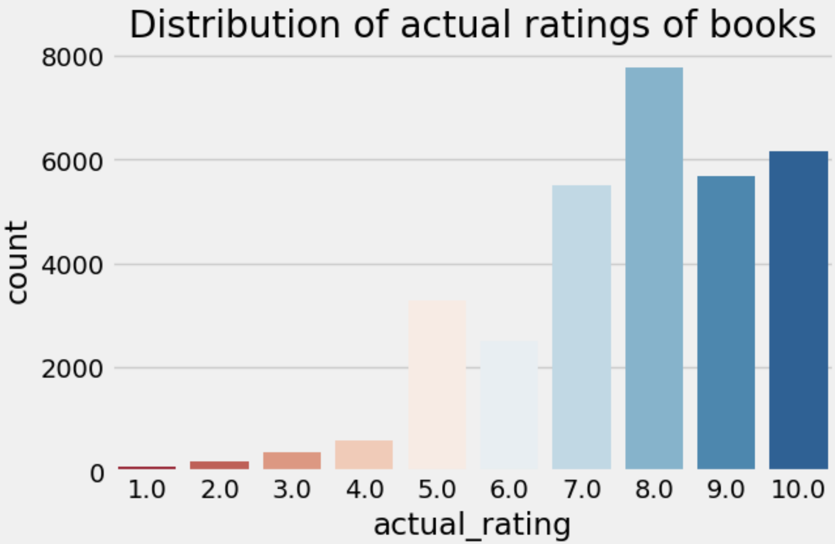
Modelling

2. Model Based Collaborative Filtering

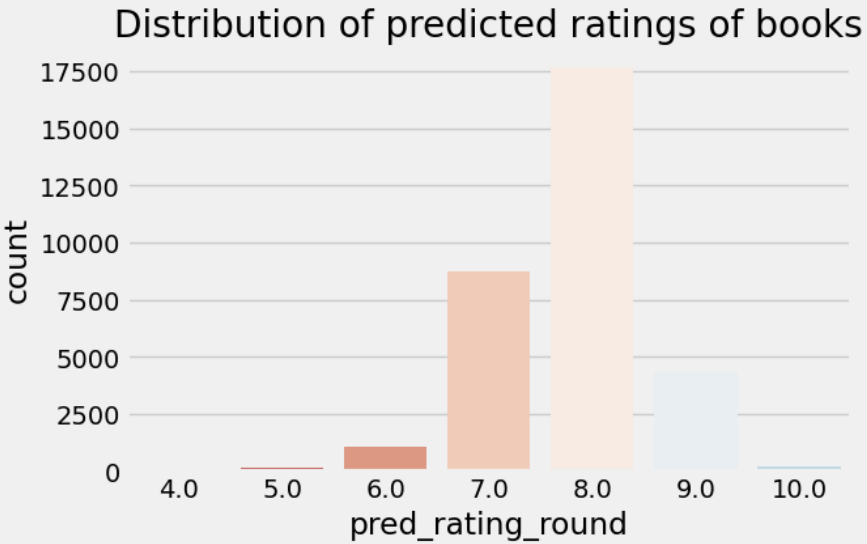
<table>
<tr><th>SVD</th><th>NMF</th></tr>
<tr><td>
<pre>
test_rmse   1.598982
test_mae   1.238592
fit_time   2.073729
test_time   0.901514
dtype: float64
</pre></td><td>
<pre>
test_rmse   2.617043
test_mae   2.233795
fit_time   8.011263
test_time   0.734815
dtype: float64
</pre></td></tr>
</table>

Modelling

2. Model Based Collaborative Filtering – SVD Model Results



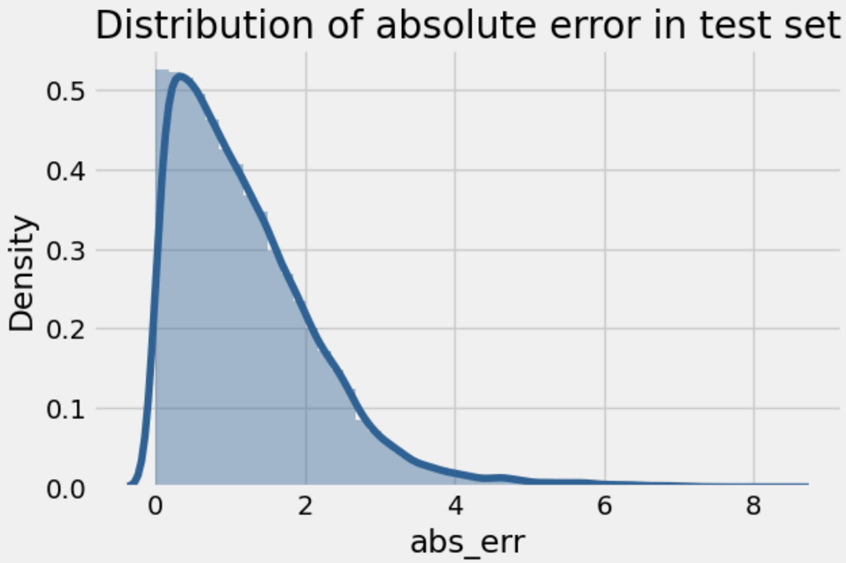
Distribution of actual ratings of books

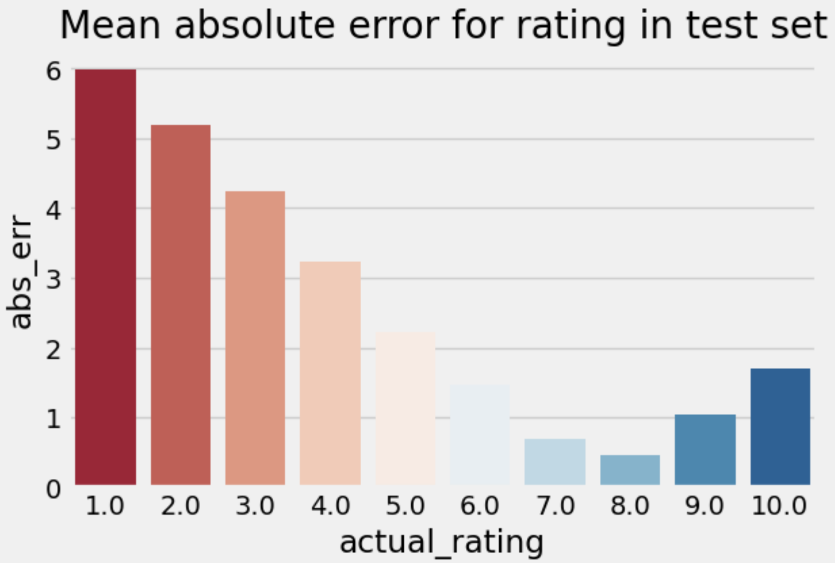


Distribution of predicted ratings of books

Modelling

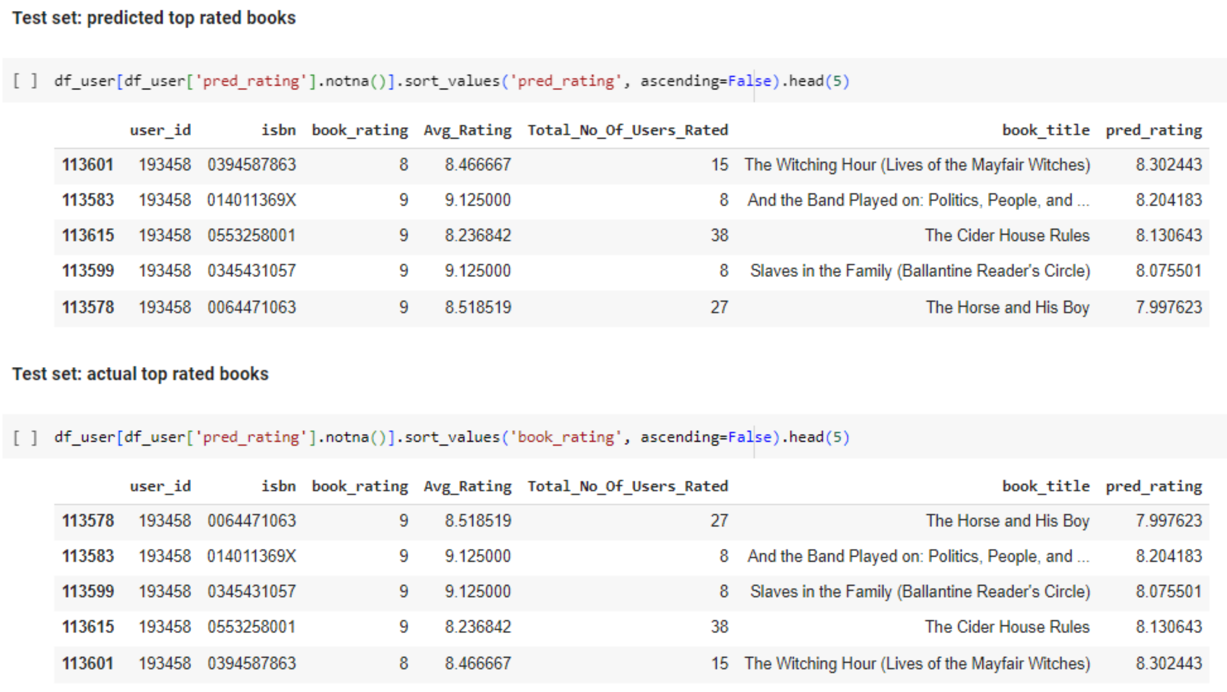
2. Model Based Collaborative Filtering – SVD Model Results


Distribution of absolute error in test set


Mean absolute error for rating in test set

Modelling

2. Model Based Collaborative Filtering – SVD Model Results (user_id 193458)



Test set: predicted top rated books

```
[ ] df_user[df_user['pred_rating'].notna()].sort_values('pred_rating', ascending=False).head(5)
```

	user_id	isbn	book_rating	Avg_Rating	Total_No_Of_Users_Rated	book_title	pred_rating
113601	193458	0394587863	8	8.466667	15	The Witching Hour (Lives of the Mayfair Witches)	8.302443
113583	193458	014011369X	9	9.125000	8	And the Band Played on: Politics, People, and ...	8.204183
113615	193458	0553258001	9	8.236842	38	The Cider House Rules	8.130643
113599	193458	0345431057	9	9.125000	8	Slaves in the Family (Ballantine Reader's Circle)	8.075501
113578	193458	0064471063	9	8.518519	27	The Horse and His Boy	7.997623

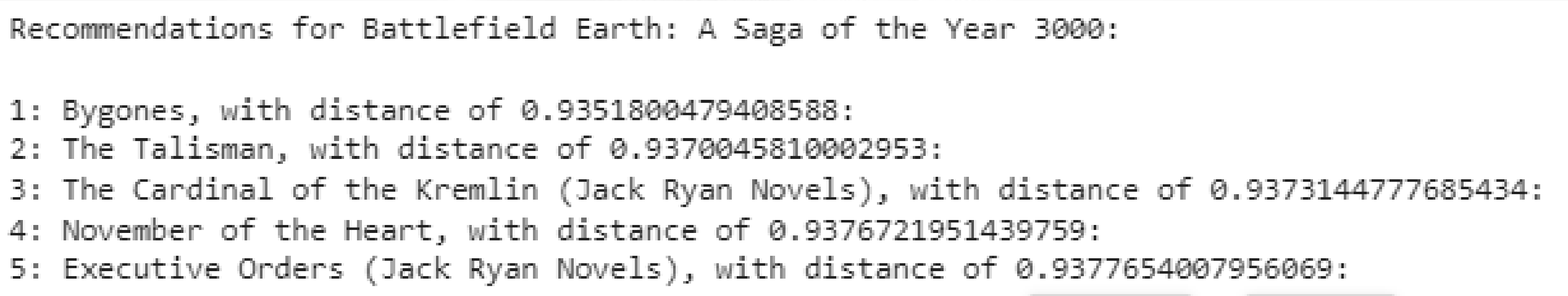
Test set: actual top rated books

```
[ ] df_user[df_user['pred_rating'].notna()].sort_values('book_rating', ascending=False).head(5)
```

	user_id	isbn	book_rating	Avg_Rating	Total_No_Of_Users_Rated	book_title	pred_rating
113578	193458	0064471063	9	8.518519	27	The Horse and His Boy	7.997623
113583	193458	014011369X	9	9.125000	8	And the Band Played on: Politics, People, and ...	8.204183
113599	193458	0345431057	9	9.125000	8	Slaves in the Family (Ballantine Reader's Circle)	8.075501
113615	193458	0553258001	9	8.236842	38	The Cider House Rules	8.130643
113601	193458	0394587863	8	8.466667	15	The Witching Hour (Lives of the Mayfair Witches)	8.302443

Modelling

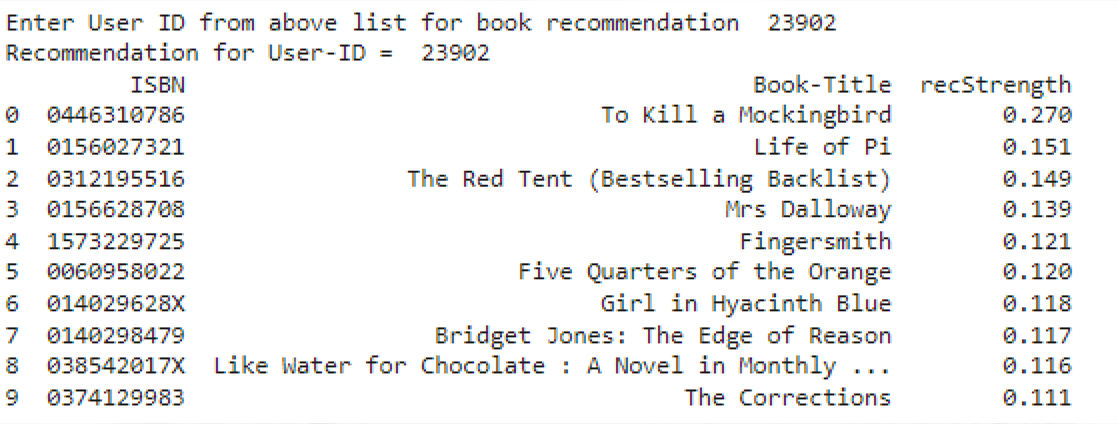
3. Memory Based Collaborative Filtering (Item-Item Based Collaborative Filtering)



```
Recommendations for Battlefield Earth: A Saga of the Year 3000:

1: Bygones, with distance of 0.9351800479408588:
2: The Talisman, with distance of 0.9370045810002953:
3: The Cardinal of the Kremlin (Jack Ryan Novels), with distance of 0.9373144777685434:
4: November of the Heart, with distance of 0.9376721951439759:
5: Executive Orders (Jack Ryan Novels), with distance of 0.9377654007956069:
```

Modelling

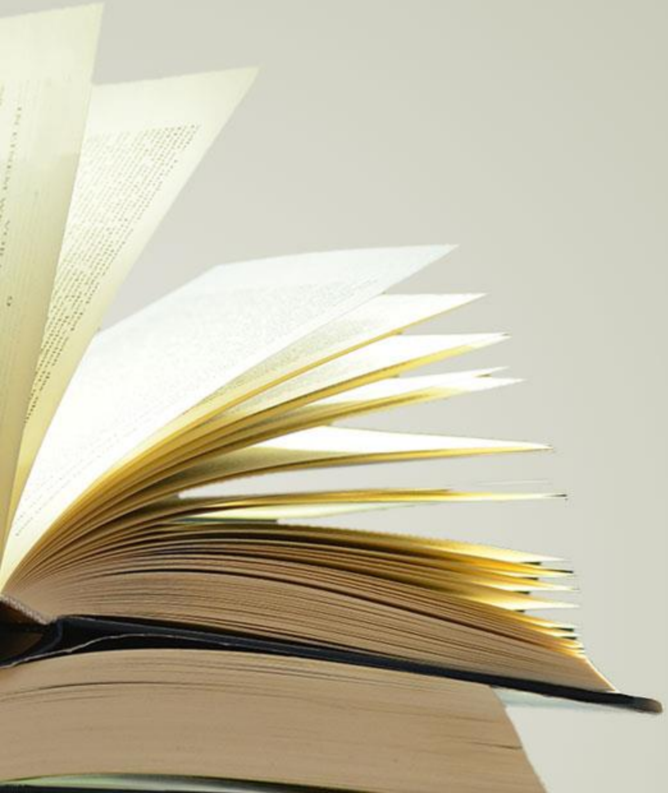
3. Memory Based Collaborative Filtering (User-Item Based Collaborative Filtering)



```
Enter User ID from above list for book recommendation  23902
Recommendation for User-ID =  23902
         ISBN                                         Book-Title  recStrength
0  0446310786                              To Kill a Mockingbird        0.270
1  0156027321                                         Life of Pi        0.151
2  0312195516                The Red Tent (Bestselling Backlist)        0.149
3  0156628708                                       Mrs Dalloway        0.139
4  1573229725                                        Fingersmith        0.121
5  0060958022                        Five Quarters of the Orange        0.120
6  014029628X                              Girl in Hyacinth Blue        0.118
7  0140298479                  Bridget Jones: The Edge of Reason        0.117
8  038542017X  Like Water for Chocolate : A Novel in Monthly ...        0.116
9  0374129983                                    The Corrections        0.111
```

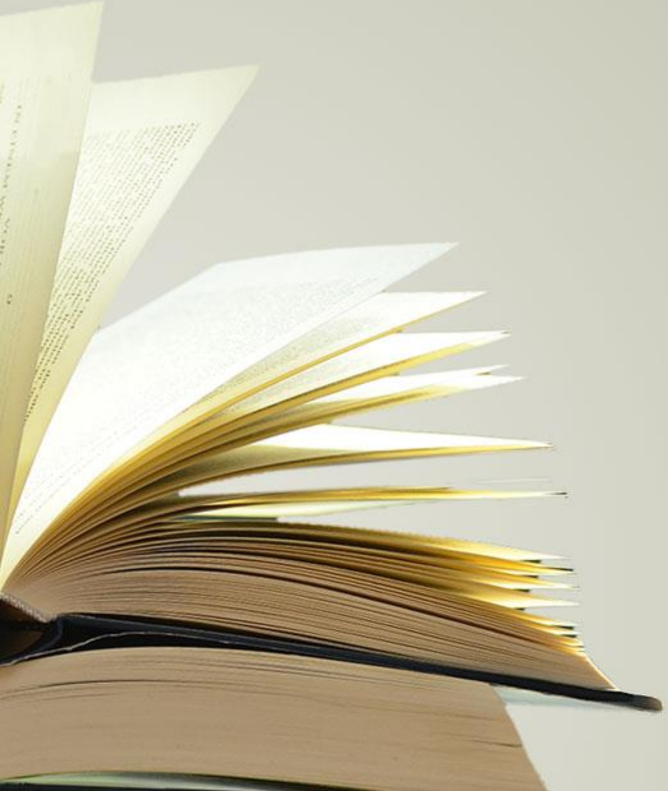
Summary & Recommendation

Summary



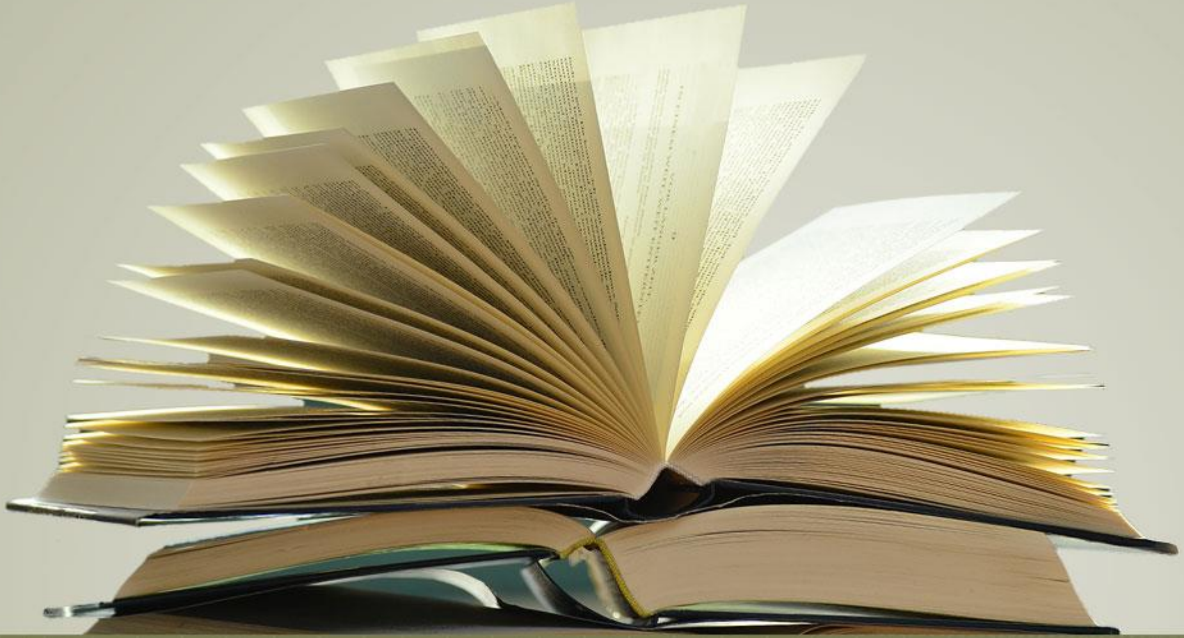
- The Top 5 most rated books were essentially novels
- Majority of the users were of the age range 20-30s with most of them came from USA, Canada, UK, Germany and Spain
- Author with the most books was Agatha Christie, William Shakespeare and Stephen King
- Most of the books have high ratings with maximum books being rated 8. Ratings below 5 are few in number
- For modelling, the model based collaborative filtering SVD technique worked way better than NMF with lower Mean Absolute Error (MAE)
- The memory based collaborative filtering, item-item based performed better than user-user based because of lower computation

Recommendation



- Make a hybrid recommendation system, which combines content-based filtering and collaborative filtering method

- Given more information regarding the books dataset, namely features like Genre, Description, etc., we could implement a content filtering based system and compare the results with the collaborative filtering based system



Thank you