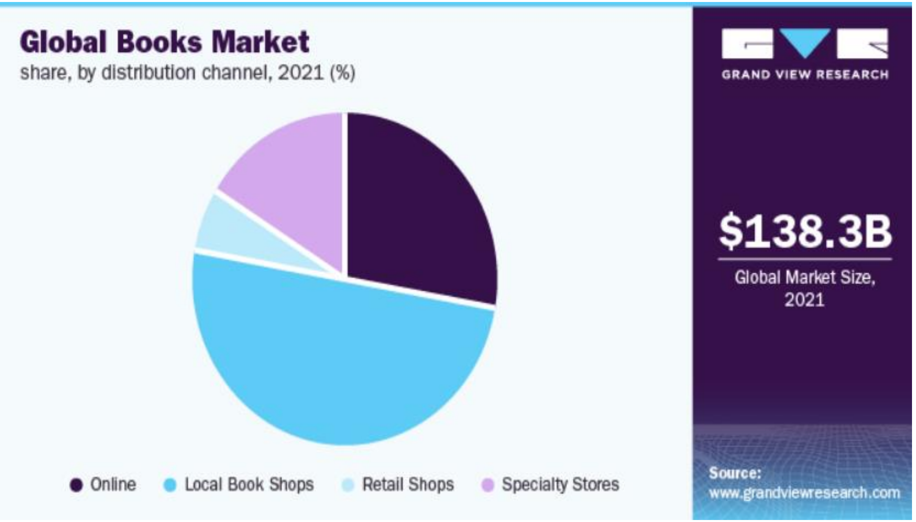
Book Recommendation System

Submitted By
Vikas Chaudhary

Flow of the Presentation

- Introduction
- Data Preparation
- EDA
- Algorithm Implementation
- Challenges
- Conclusion

Introduction



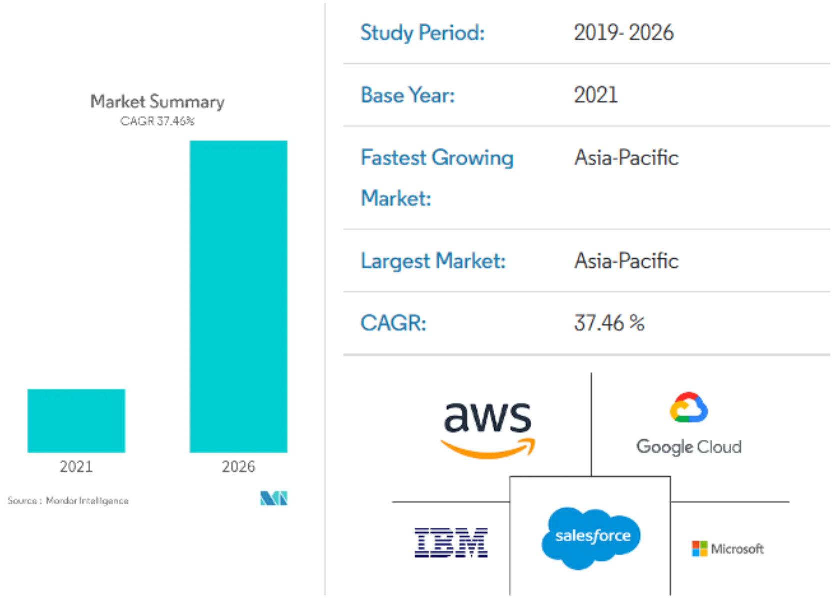
Source: www.grandviewresearch.com

As per the Market Analysis Report by Grand View Research, The global books market size was valued at USD 138.35 billion in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 1.9% from 2022 to USD 164.22 billion in 2030.

- Hard copy segment accounted for the largest market revenue share of around 78.7% in 2021.
- The online channel is anticipated to register faster growth during forecast years with a CAGR of 2.9% from 2022 to 2030.

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.

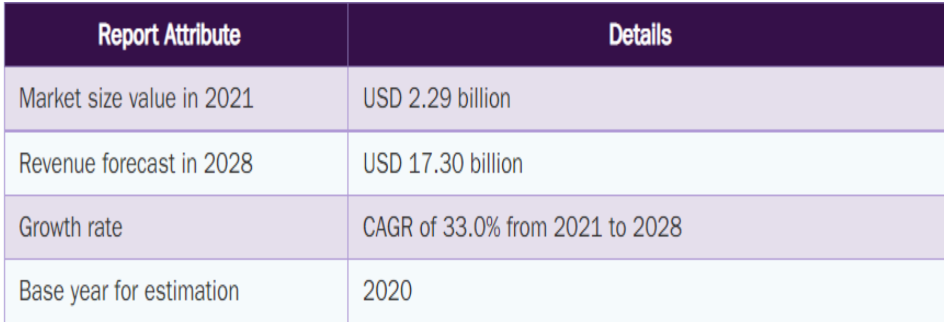
Recommendation Systems – Present and Future



Study Period:	2019- 2026
Base Year:	2021
Fastest Growing Market:	Asia-Pacific
Largest Market:	Asia-Pacific
CAGR:	37.46 %

An effectively build recommendation system has the potential to change the business in its entirety.

- As per Mordor Intelligence, the Recommendation Engine market was valued at USD 2.12 billion in 2020, and it is expected to reach USD 15.13 billion by 2026, registering a CAGR of 37.46% during the period of 2021-2026.
- Similar trends are also shown in a report by Grand View Research (given in the table below).



Report Attribute	Details
Market size value in 2021	USD 2.29 billion
Revenue forecast in 2028	USD 17.30 billion
Growth rate	CAGR of 33.0% from 2021 to 2028
Base year for estimation	2020

Objective: On the basis of the given datasets that contain the required records, we need to build a Machine Learning (ML) model to recommend book(s).

Methodology: Unsupervised Machine Learning (ML)

Database Summary:

Three datasets are being provided:

1. Books: with 271360 rows and 8 columns it contains details about book.
2. Users: with 278858 rows and 3 columns it contains details about users.
3. Ratings: with 1149780 and 3 columns it contains details about the ratings given to a book by users.

Data Preparation

About Books Dataset: It contains the given 8 columns

1. ISBN – International Standard Book Number, an identification number of book.
2. Book-Title – Name of the book
3. Book-Author – Author of the book
4. Year-of-Publication – Year when the book was published
5. Publisher – Name of the Publisher
6. Image-URL-S
7. Image-URL-M
8. Image-URL-L

6, 7, 8 contain the link to the image of the cover of the book

About Users Dataset: It contains the given 3 columns

1. User-ID – ID number of the user
2. Location – Location (City, Province/State, Country) of the user
3. Age – Age of the user

About Rating Dataset: It contains the given 3 columns

1. User-ID – ID number of the user
2. ISBN – International Standard Book Number, an identification number of book.
3. Book-Rating – Rating is given by the user to the book

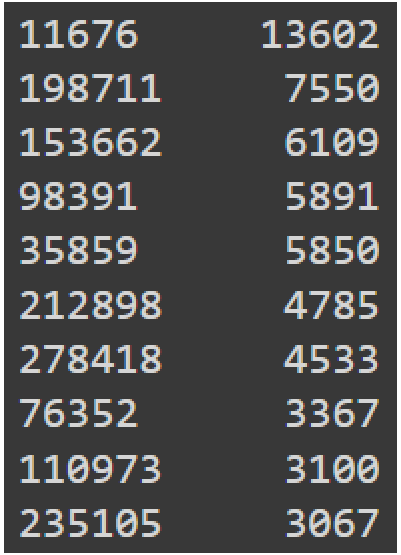
Overview of Datasets

1. Books Dataset:
- Some columns have 1 or 2 missing values so those entries were deleted.
- Columns with Image URLs were deleted.

2. Users Dataset:
- It has some null values in the 'Age' column

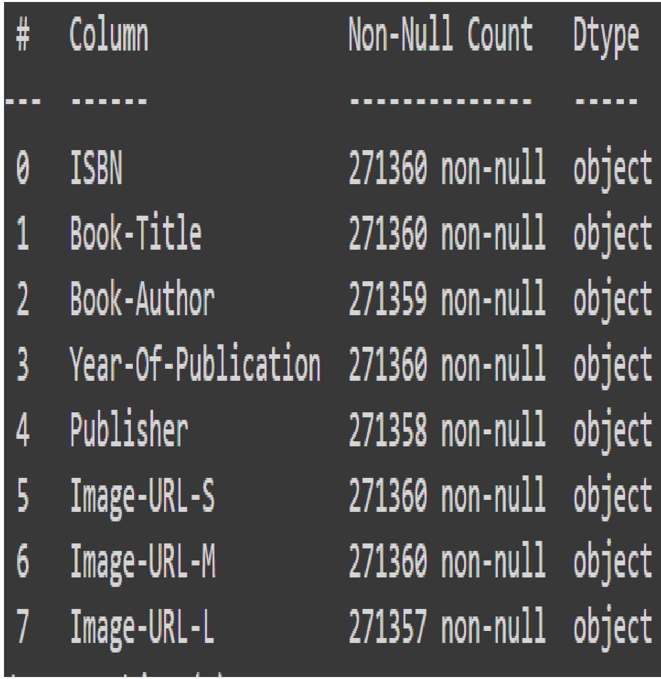
3. Rating Dataset:
- It doesn't have any null values in any column

Top 10 Users (ID) to Rate Books



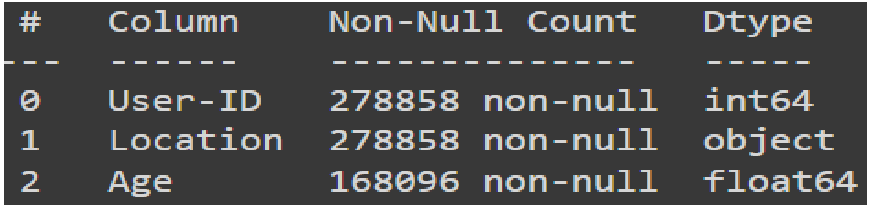
```
11676      13602
198711      7550
153662      6109
98391       5891
35859       5850
212898      4785
278418      4533
76352       3367
110973      3100
235105      3067
```

Books Dataset



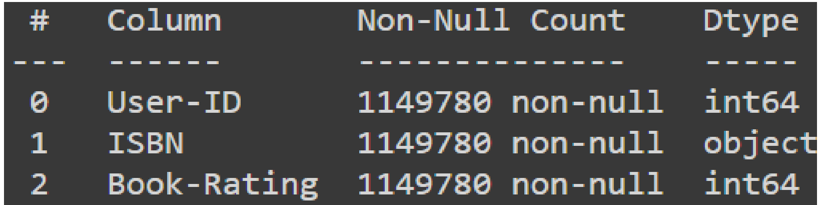
```
#   Column               Non-Null Count   Dtype
---  ------               --------------   -----
0   ISBN                 271360 non-null  object
1   Book-Title           271360 non-null  object
2   Book-Author          271359 non-null  object
3   Year-Of-Publication  271360 non-null  object
4   Publisher            271358 non-null  object
5   Image-URL-S          271360 non-null  object
6   Image-URL-M          271360 non-null  object
7   Image-URL-L          271357 non-null  object
```

Users Dataset



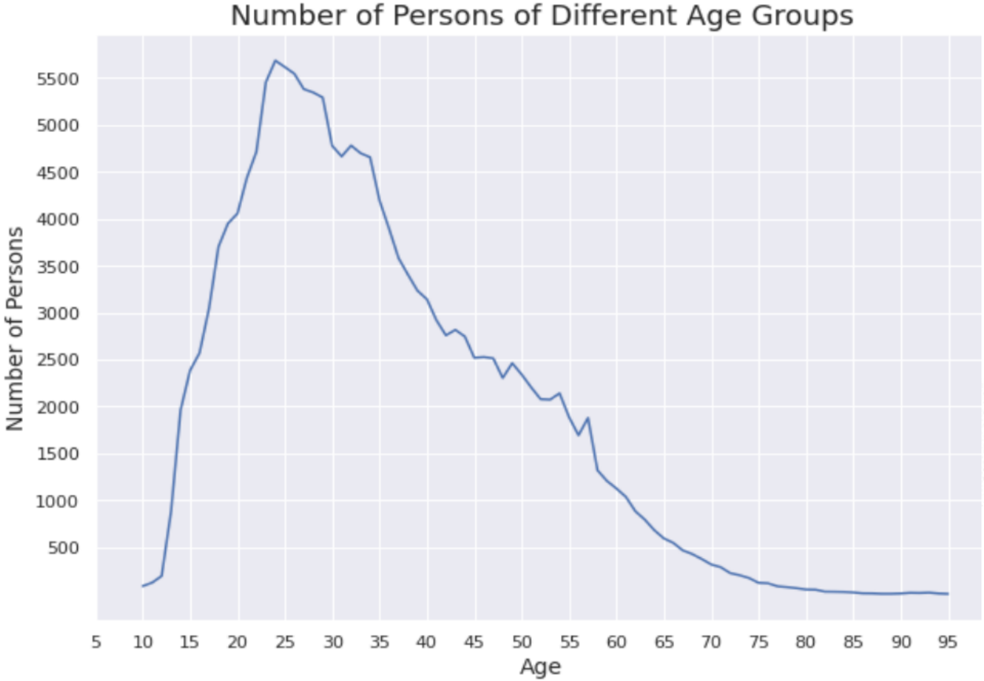
```
#    Column     Non-Null Count   Dtype
---  ------     --------------   -----
0    User-ID    278858 non-null  int64
1    Location   278858 non-null  object
2    Age        168096 non-null  float64
```

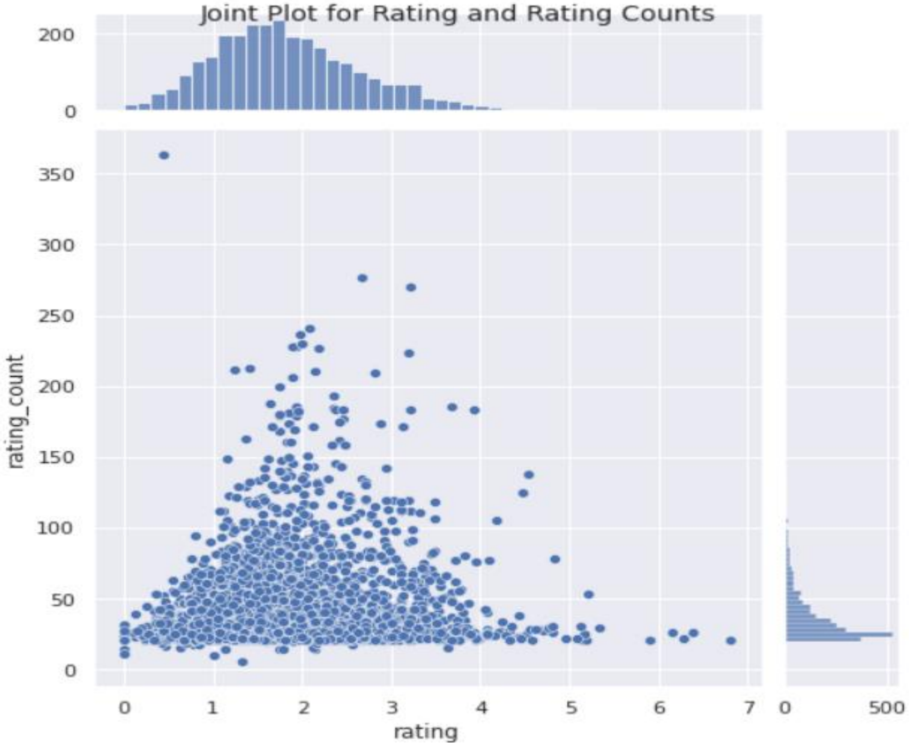
Ratings Dataset



```
#    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
```

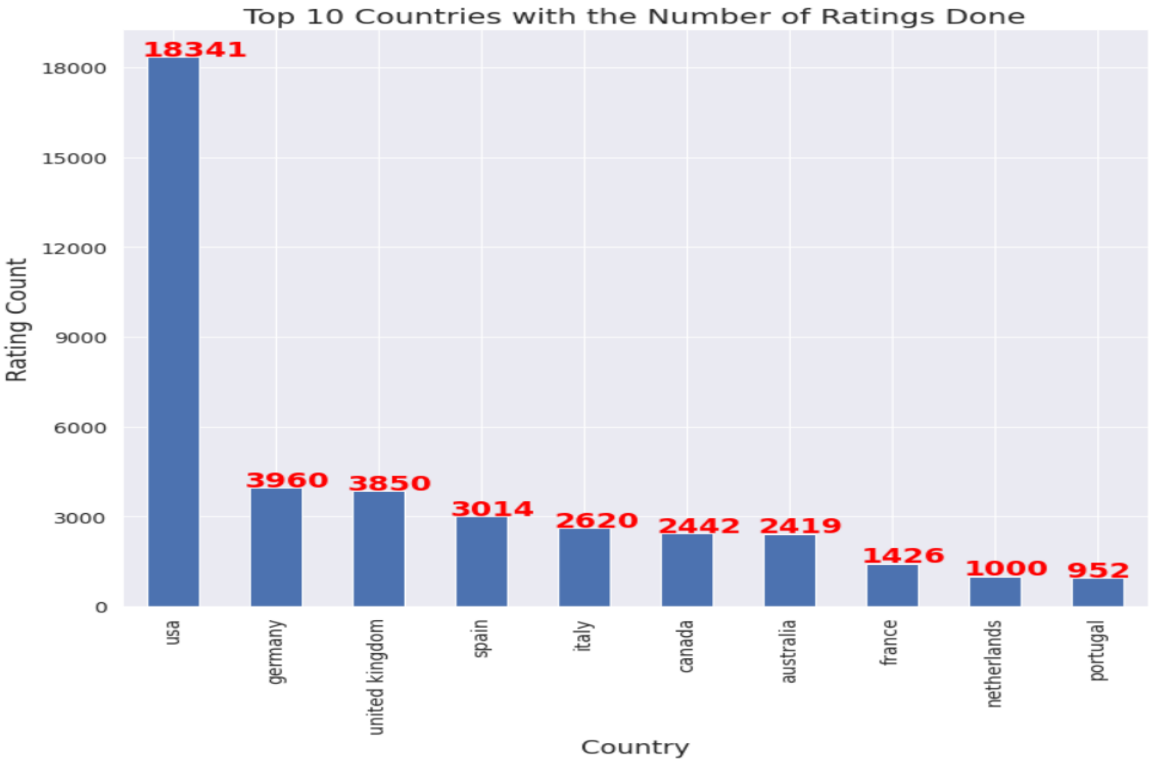
EDA

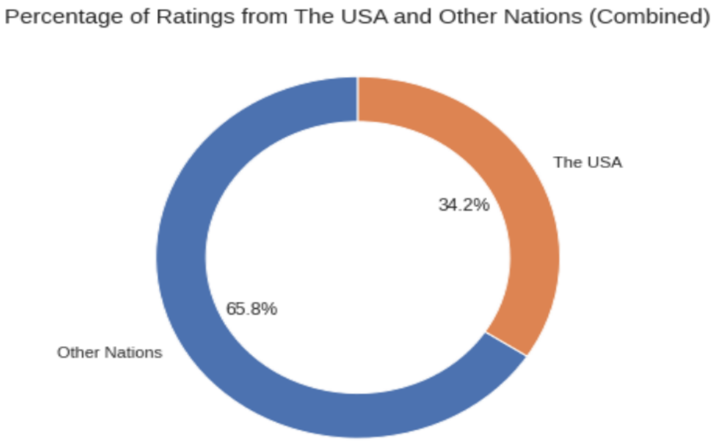


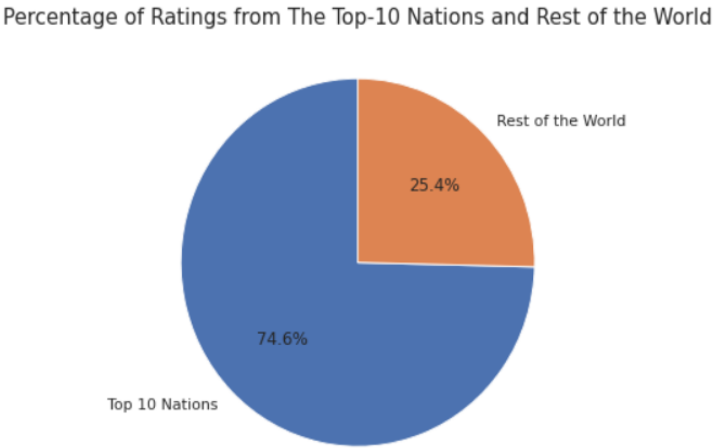


- The maximum number of users to rate the books are of the age of 24.
- Joint Plot of ‘rating’ and ‘rating_count’ after removing ‘rating_count’ below 20 shows that there are so many books below 50 ‘rating_count’.

Ratings Origin (Country)

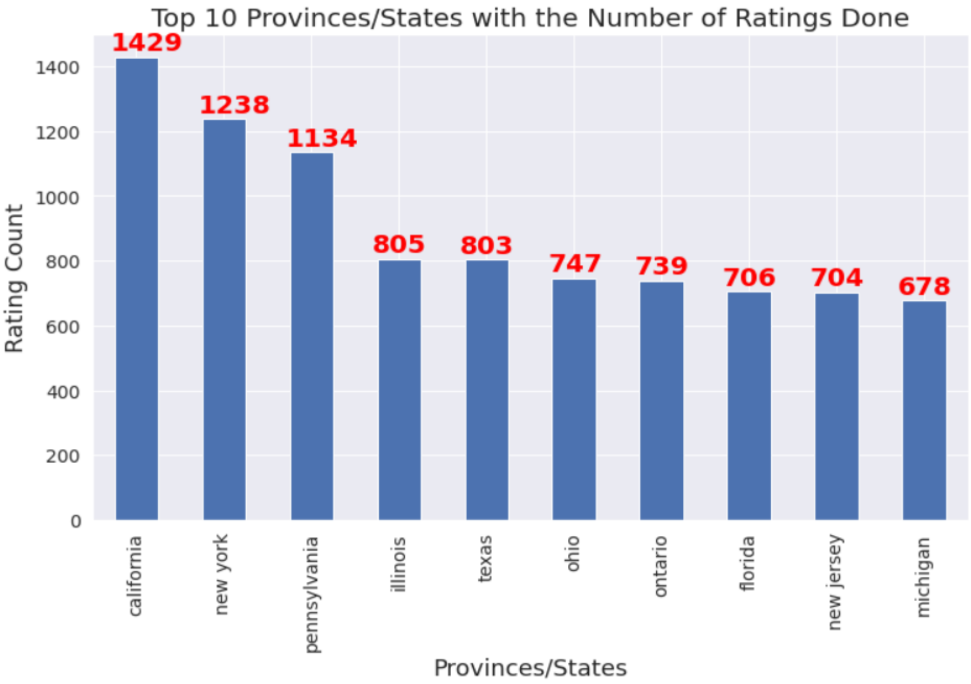


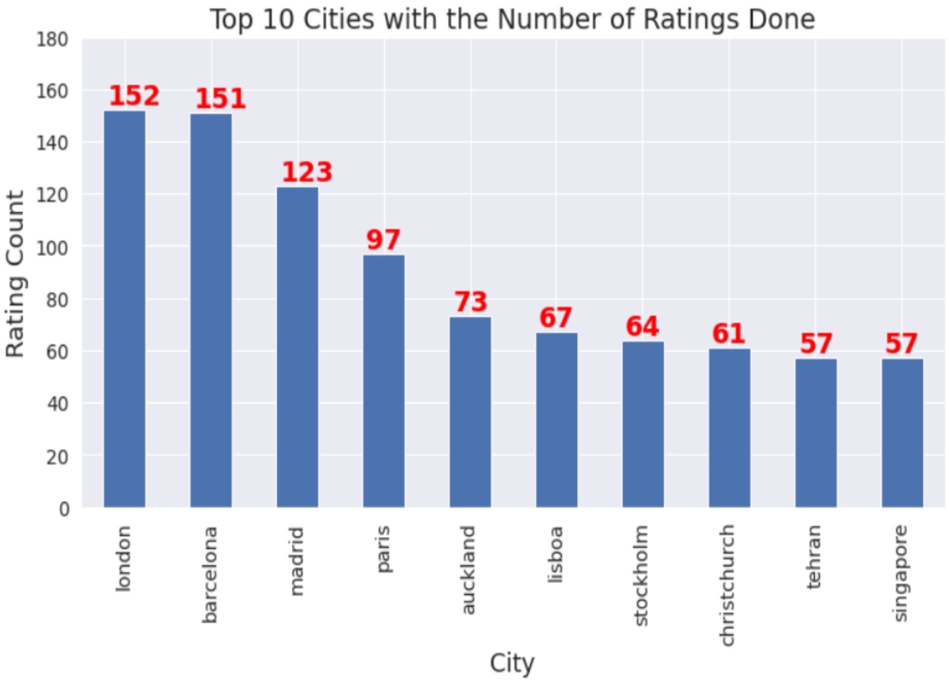




- The USA is the dominating origin of ratings.
- Nearly 1 in 3 ratings were done from The USA.
- 3 in 4 ratings among top 10 nations are American and European nations.

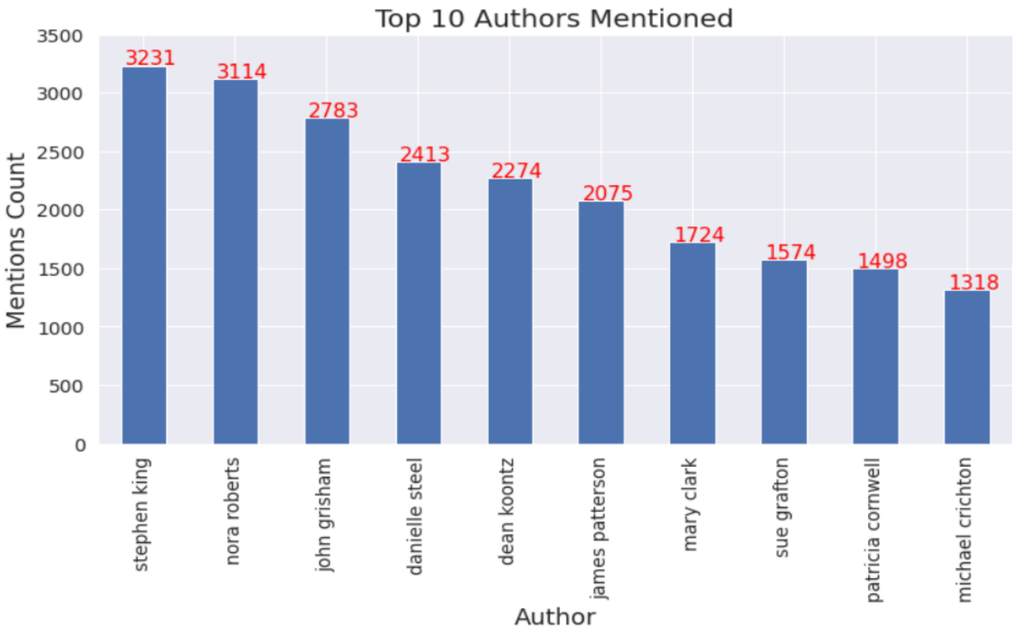
Ratings Origin (Province/State and City)



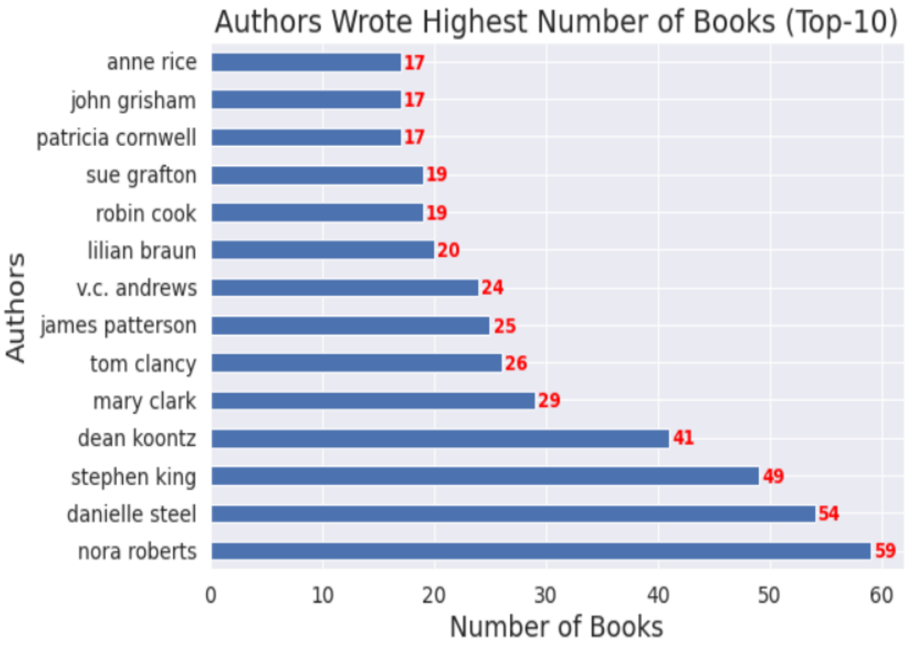


- In provinces/states The USA is also dominating here, most of the states in the top 10 are from The USA.
- When it comes to cities in the top 20 they are from the entire globe but still, most of them are European and capital cities.

About Authors



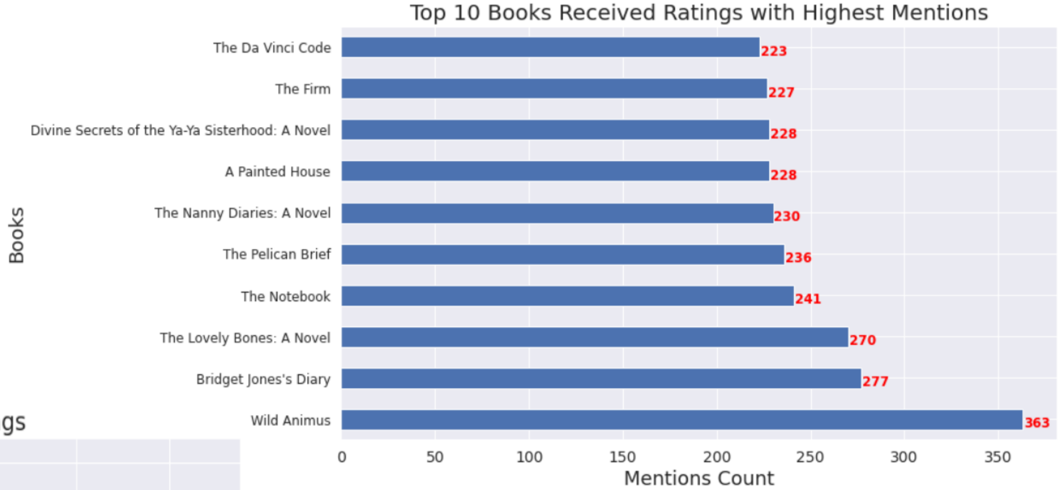
Nora Roberts has written the highest number of books, while Danielle Steel and Stephen King are 2nd and 3rd respectively.

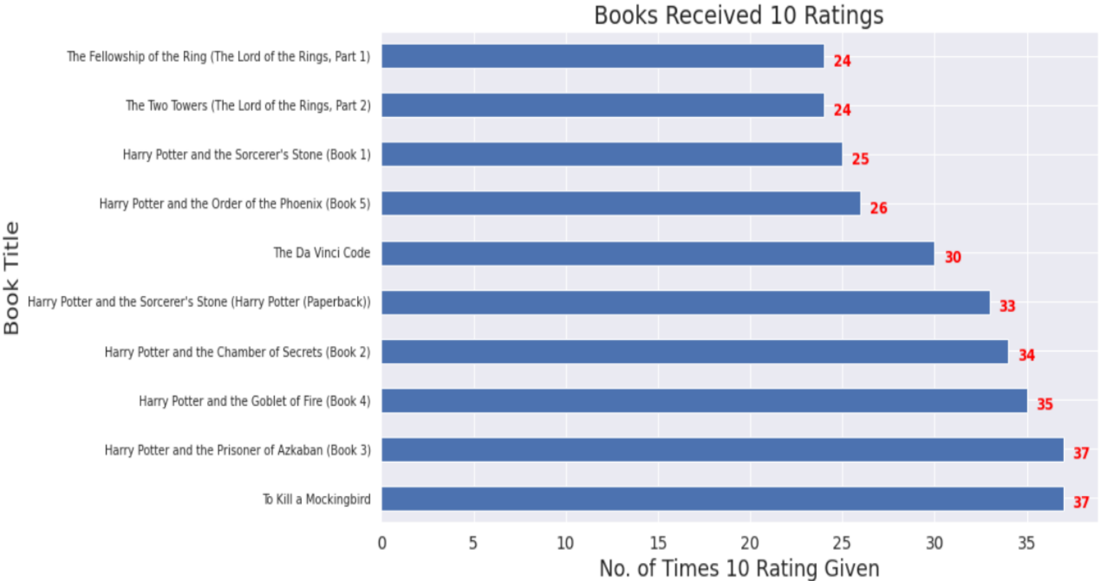


Stephen King has the highest mentions, Nora Roberts and John Grisham are 2nd and 3rd respectively.

About Books

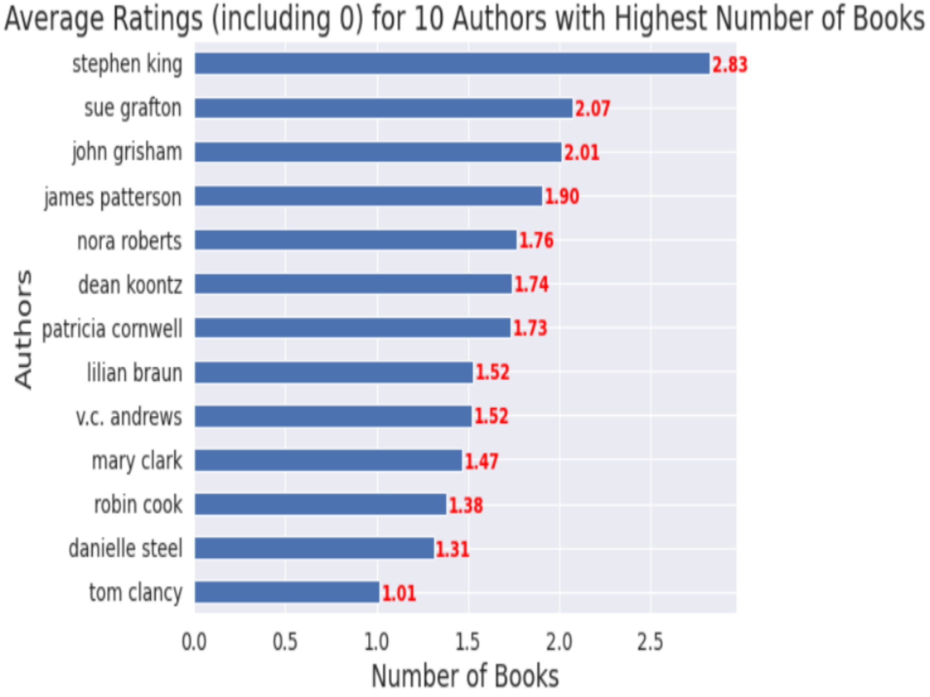
‘Wild Animus’ has the highest mentions, ‘Bridget Jones’s Diary’ and ‘The Lovely Bones: A Novel’ are 2nd and 3rd highest mentioned books.



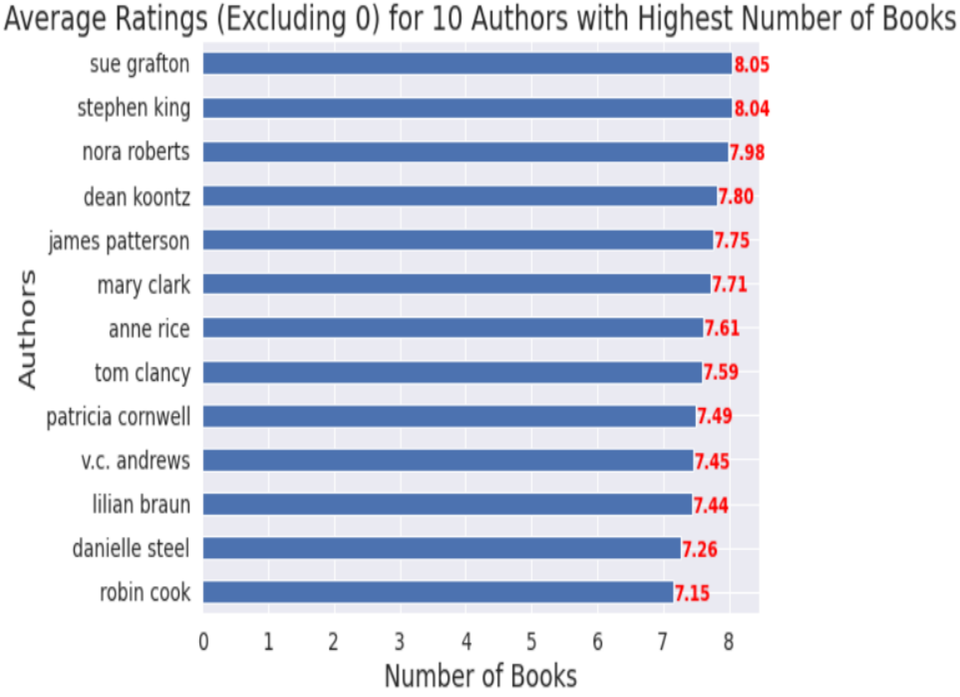


- ‘To Kill a Mockingbird’ has received 10 ratings highest number of times.
- Harry Potter and The Lord of the Rings series books have received the highest number of 10 ratings.

Average Ratings (Authors)



Average Ratings (including 0) for 10 Authors with Highest Number of Books

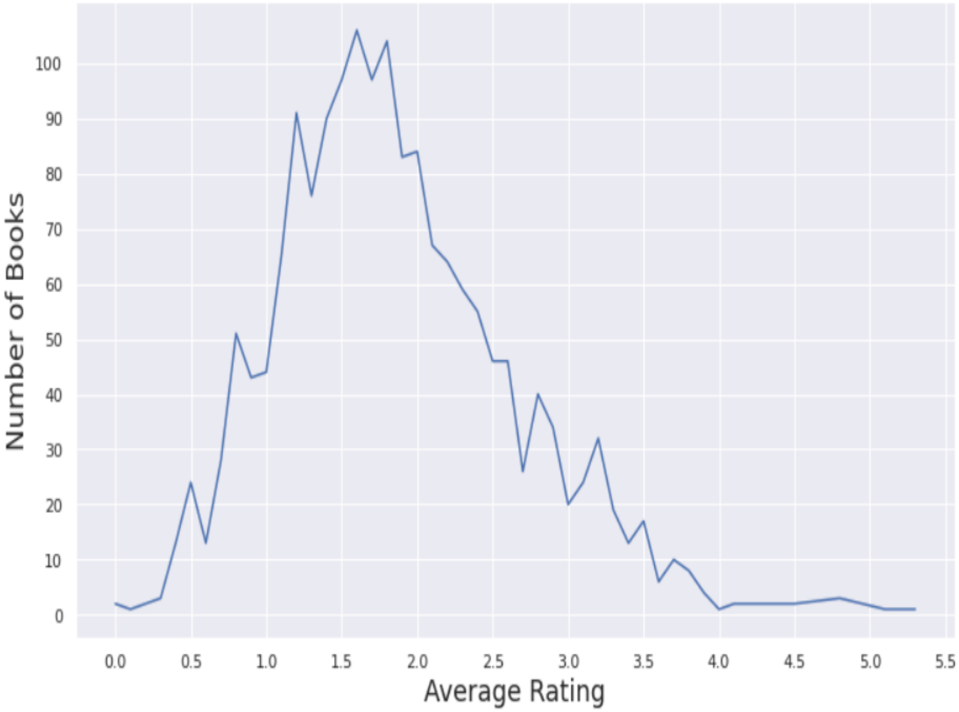


Average Ratings (Excluding 0) for 10 Authors with Highest Number of Books

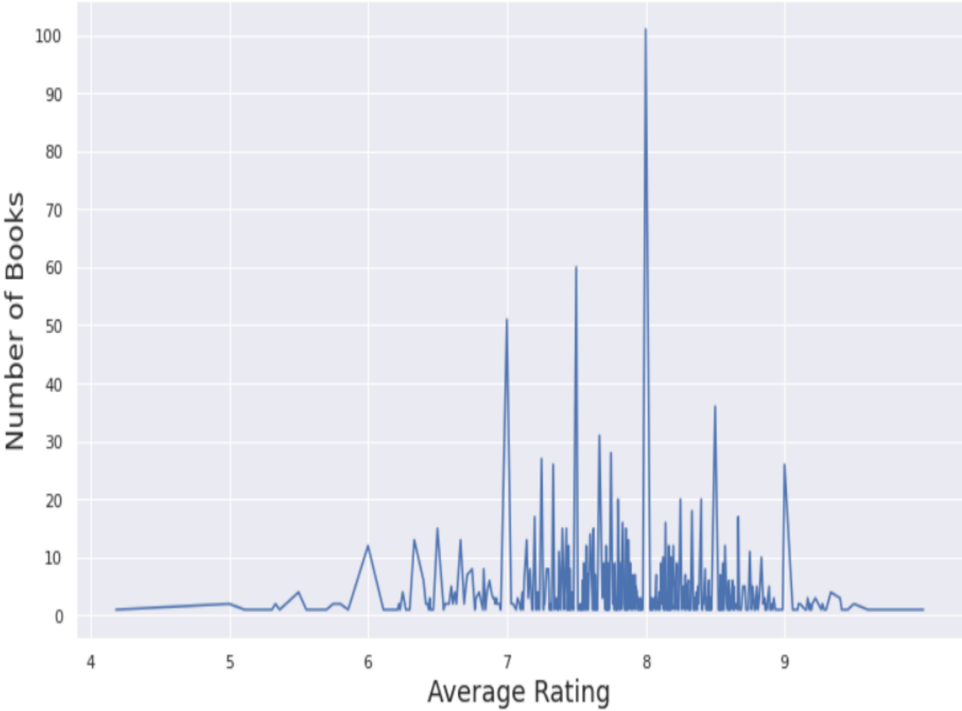
Among the top-10 authors who wrote the highest number of books, Stephen King has the highest average ratings when considering 0 as a rating. When we exclude 0 from the rating Sue Grafton had the highest average rating.

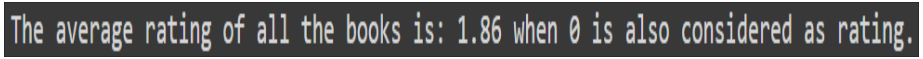
Average Ratings (Books)

Number of Books with Average Rating (including 0 as rating)

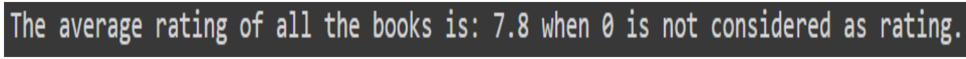


Number of Books with Average Rating (exluding 0 as rating)



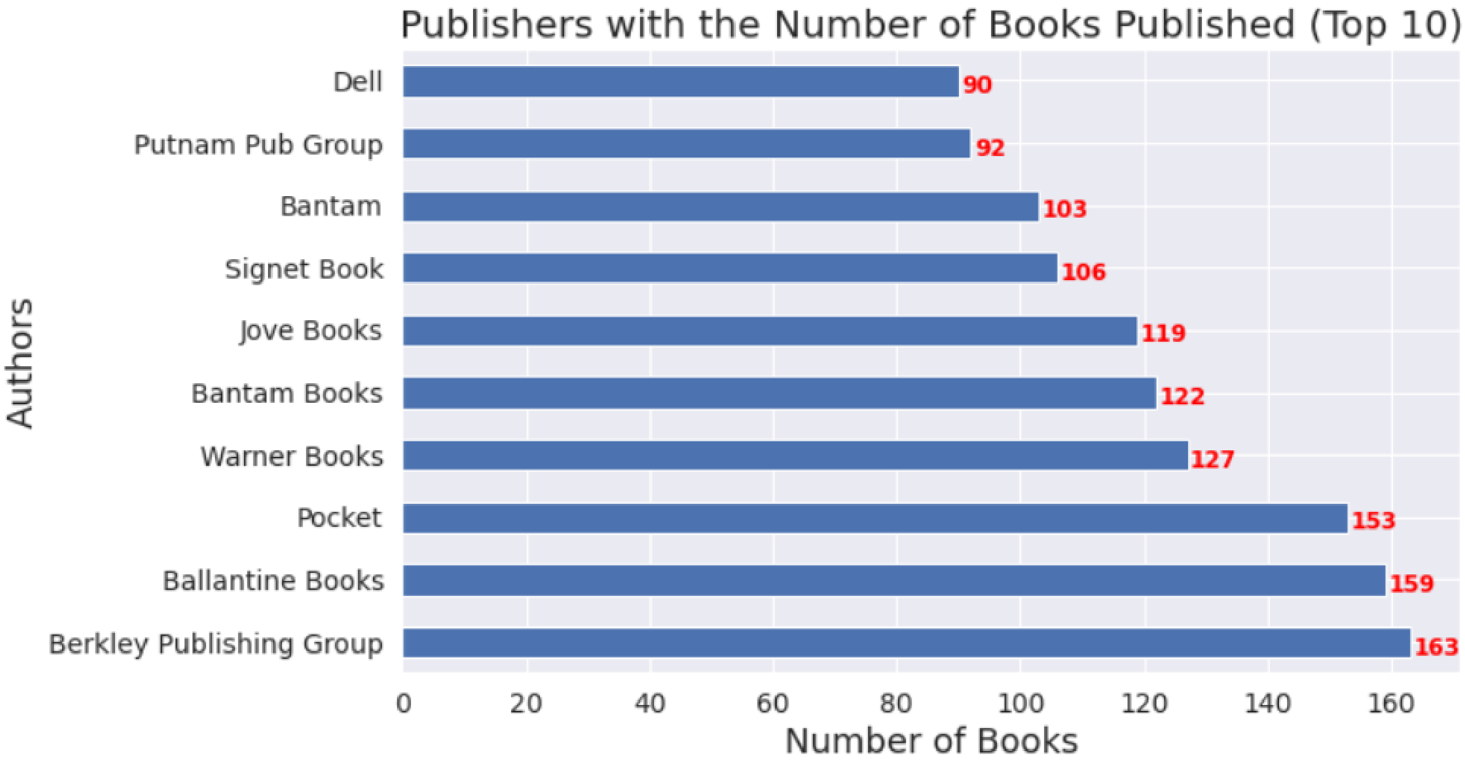


```
The average rating of all the books is: 1.86 when 0 is also considered as rating.
```



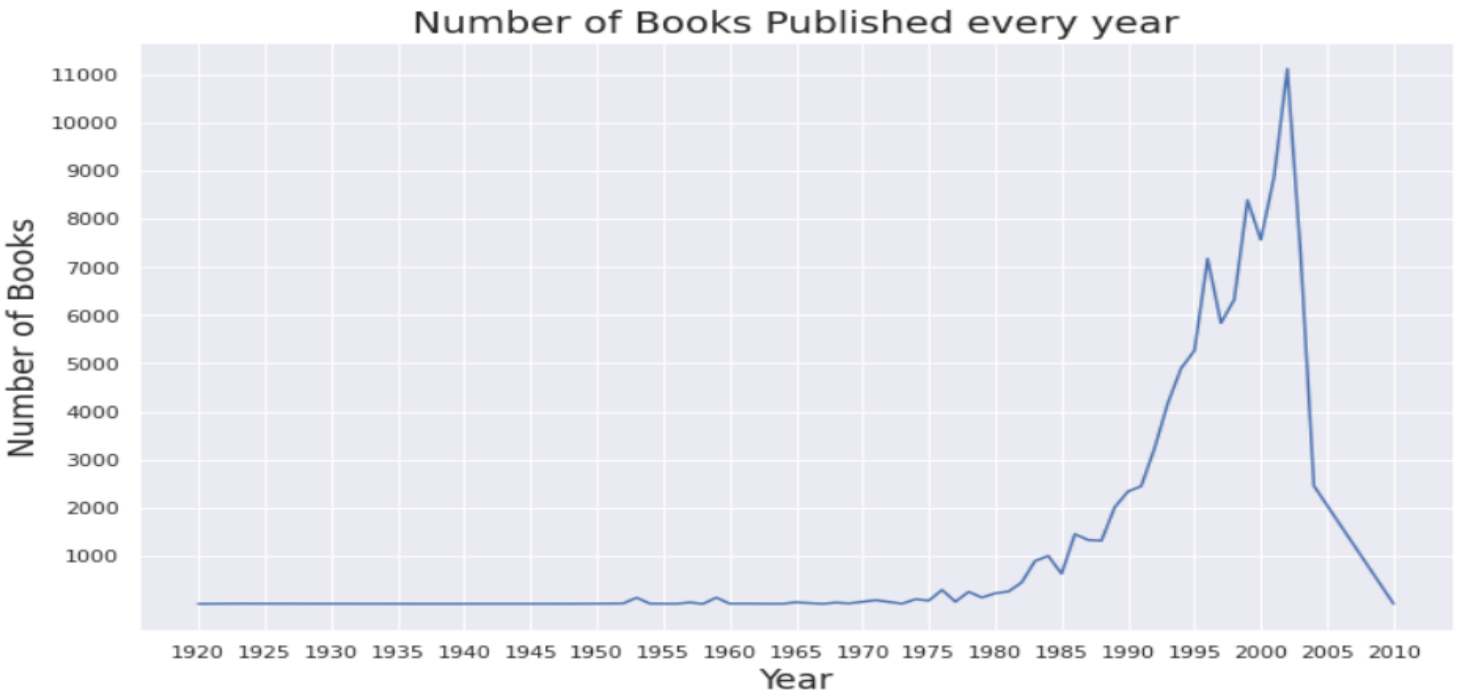
```
The average rating of all the books is: 7.8 when 0 is not considered as rating.
```

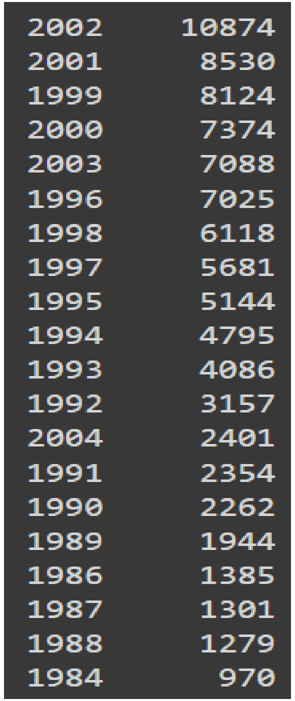
Top Publishers



Berkley Publishing Group has published the highest number of books.

Number of Books Published





```
2002     10874
2001      8530
1999      8124
2000      7374
2003      7088
1996      7025
1998      6118
1997      5681
1995      5144
1994      4795
1993      4086
1992      3157
2004      2401
1991      2354
1990      2262
1989      1944
1986      1385
1987      1301
1988      1279
1984       970
```

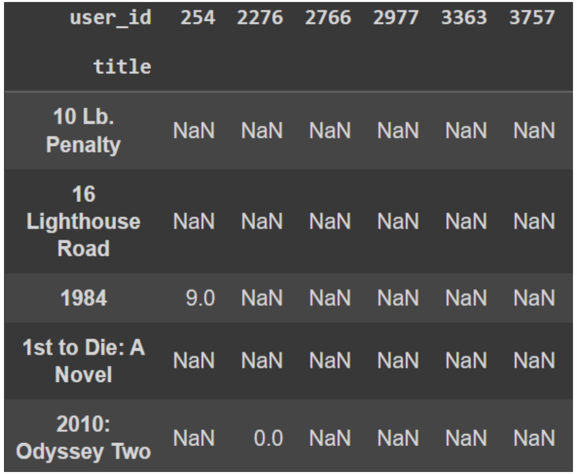
- The database has books published from 1920 to 2010.
- The highest number of books are published in 2002.

Algorithm Implementation

‘user_id’, representing the ID of the user and ‘title’ the book title are required to build the Recommendation System.

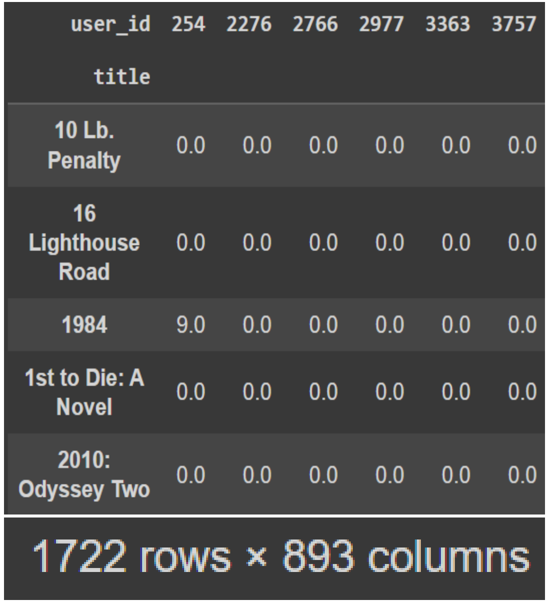
Steps Involved after finalizing dataset:

STEP 1: Making Pivot Table



user_id	254	2276	2766	2977	3363	3757
title						
10 Lb. Penalty	NaN	NaN	NaN	NaN	NaN	NaN
16 Lighthouse Road	NaN	NaN	NaN	NaN	NaN	NaN
1984	9.0	NaN	NaN	NaN	NaN	NaN
1st to Die: A Novel	NaN	NaN	NaN	NaN	NaN	NaN
2010: Odyssey Two	NaN	0.0	NaN	NaN	NaN	NaN

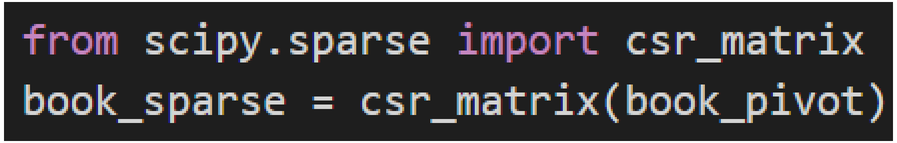
STEP 2: Filling NaN Value of Pivot Table with ‘0’



user_id	254	2276	2766	2977	3363	3757
title						
10 Lb. Penalty	0.0	0.0	0.0	0.0	0.0	0.0
16 Lighthouse Road	0.0	0.0	0.0	0.0	0.0	0.0
1984	9.0	0.0	0.0	0.0	0.0	0.0
1st to Die: A Novel	0.0	0.0	0.0	0.0	0.0	0.0
2010: Odyssey Two	0.0	0.0	0.0	0.0	0.0	0.0

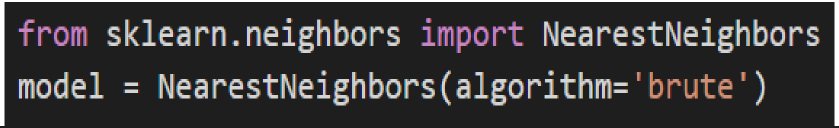
1722 rows × 893 columns

STEP 3: Making Sparse Matrix

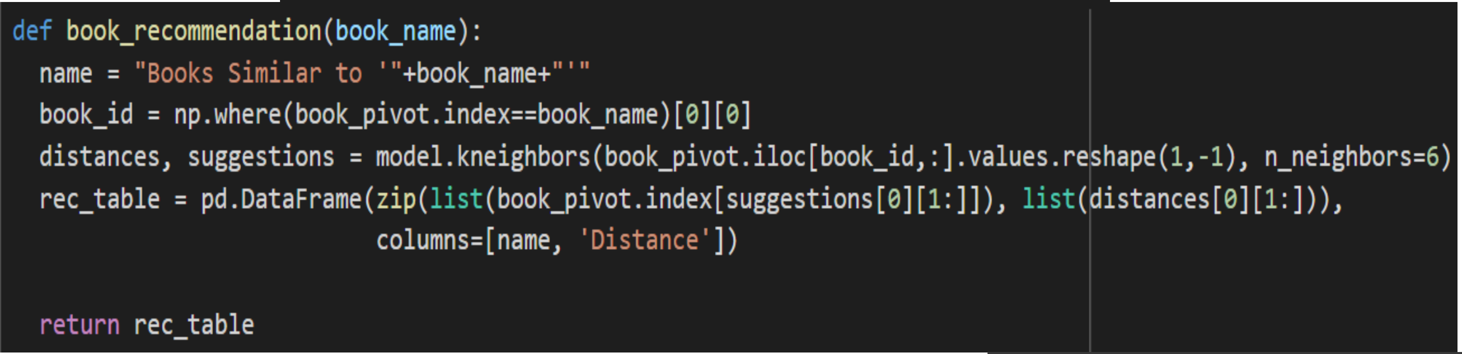


```
from scipy.sparse import csr_matrix
book_sparse = csr_matrix(book_pivot)
```

STEP 4: Implementing Algorithm

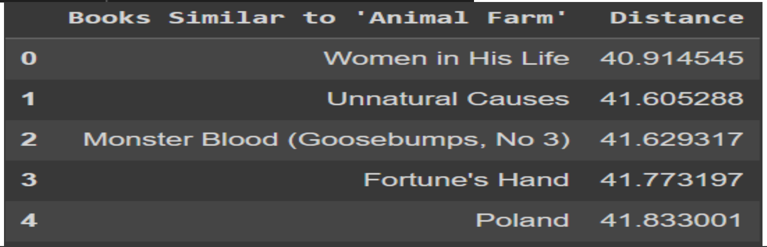


```python
from sklearn.neighbors import NearestNeighbors
model = NearestNeighbors(algorithm='brute')
```



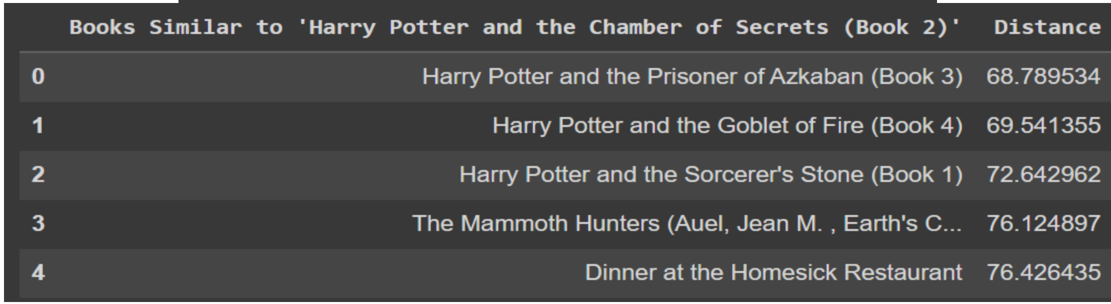
```python
def book_recommendation(book_name):
  name = "Books Similar to '"+book_name+"'"
  book_id = np.where(book_pivot.index==book_name)[0][0]
  distances, suggestions = model.kneighbors(book_pivot.iloc[book_id,:].values.reshape(1,-1), n_neighbors=6)
  rec_table = pd.DataFrame(zip(list(book_pivot.index[suggestions[0][1:]]), list(distances[0][1:])),
                           columns=[name, 'Distance'])

  return rec_table
```



	Books Similar to 'Animal Farm'	Distance
0	Women in His Life	40.914545
1	Unnatural Causes	41.605288
2	Monster Blood (Goosebumps, No 3)	41.629317
3	Fortune's Hand	41.773197
4	Poland	41.833001

STEP 5: Testing Result



	Books Similar to 'Harry Potter and the Chamber of Secrets (Book 2)'	Distance
0	Harry Potter and the Prisoner of Azkaban (Book 3)	68.789534
1	Harry Potter and the Goblet of Fire (Book 4)	69.541355
2	Harry Potter and the Sorcerer's Stone (Book 1)	72.642962
3	The Mammoth Hunters (Auel, Jean M. , Earth's C...	76.124897
4	Dinner at the Homesick Restaurant	76.426435

Challenges

- Three different datasets were required to make the final dataset. All three were large datasets.
- There are several books with multiple authors but they are different books with common book titles.
- The name of some authors with different publishers has the different format as some entries have a middle name or short form of name.
- There are many 0 in rating columns that reduces the average ratings when considered as actual rating.
- Some books have less number of ratings resulting in unfair rating distribution.
- Contrary to the above point the books with the higher number of ratings will have an unfair advantage in the final recommendation system as compared to those with less number of ratings.

Conclusion

The following are the important findings from the final dataframe after considering 30 as the minimum number of ratings received by any book and 200 is the minimum number of ratings done by any user.

- It has 1722 books
- Contains 1028 authors
- And 528 publishers
- North America and Europe are the dominant markets for all including readers, authors and publishers.
- The readers from The USA have the highest presence as compared to any country or region.
- The majority of the population in the database is young.
- There are many ways to make or deploy this project, our objective must be explicit before deciding on the final dataset for training.

Thank You