UNSUPERVISED LEARNING CAPSTONE PROJECT

BOOK RECOMMENDATION SYSTEM

Presented By:

- **Raj Vijay Sonar (Cohort Vindhya)**

RECOMMENDATION SYSTEM

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).

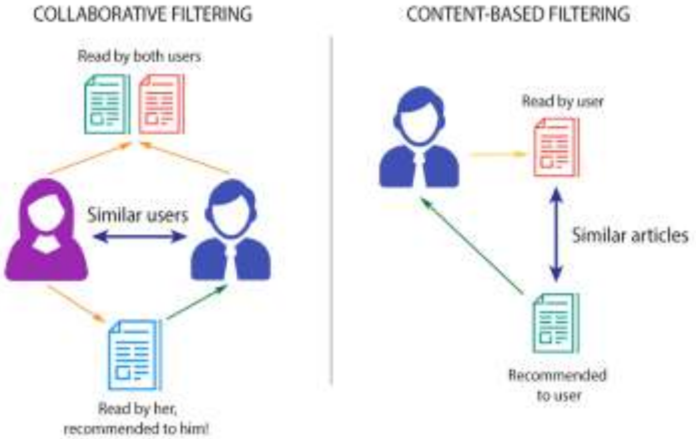
During the last few decades, with the rise of YouTube, Amazon, Netflix and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys.

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.

TYPES OF RECOMMENDER SYSTEMS

COLLABORATIVE FILTERING

This method makes automatic predictions about the interests of a user by collecting preferences or taste information from many users. The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on a set of items, A is more likely to have B's opinion for a given item than that of a randomly chosen person.



CONTENT BASED FILTERING

This method uses only information about the description and attributes of the items users has previously consumed to model user's preferences. In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended.

HYBRID MODEL APPROACH

It is combining collaborative filtering and content-based filtering to recommend more effectively. These methods can also be used to overcome some of the common problems in recommender systems such as cold start and the sparsity problem.

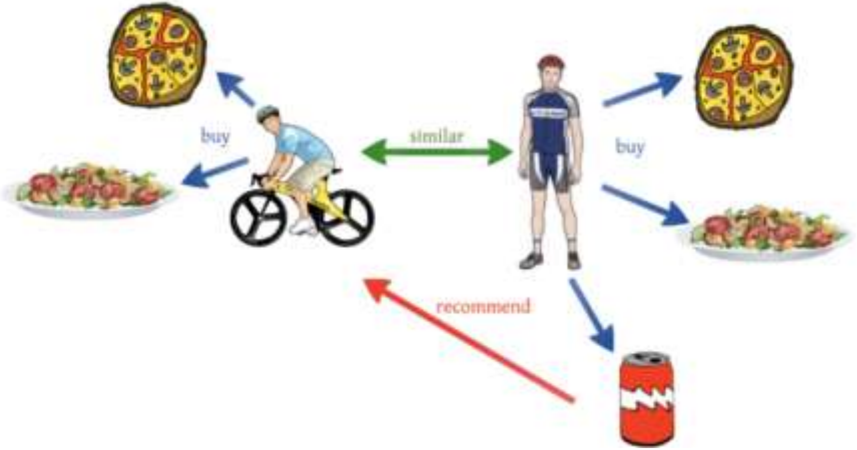
COLLABORATIVE FILTERING

MEMORY BASED APPROACH

This approach uses the memory of previous users interactions to compute users similarities based on items they've interacted (user-based approach) or compute items similarities based on the users that have interacted with them (item-based approach).

MODEL BASED APPROACH

In this approach, models are developed using different machine learning algorithms to recommend items to users. There are many model-based CF algorithms, like Neural Networks, Bayesian Networks, Clustering Techniques, and Latent Factor Models such as Singular Value Decomposition (SVD) and Probabilistic Latent Semantic Analysis.



DATASETS

BOOKS DATAFRAME

This dataframe contains the information regarding the books. Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset.

COLUMNS IN DATAFRAME:

ISBN: Unique ID to identify a book
Book-Title: Title of the book
Book-Author: Author of the book
Year-Of-Publication: The year in which the book was published
Publisher: Publisher of the book
Image-URL-S: Image URL of the book (Small Size)
Image-URL-M: Image URL of the book (Medium Size)
Image-URL-L: Image URL of the book (Large Size)

- The dataframe had 271360 rows and 8 columns.
- No duplicate entries were present in the dataset.
- Dropped the columns of Image URL.
- Replaced the null values of the ‘Author’ column with the highest repeated author i.e., ‘Agatha Christie’ and similarly for ‘Publisher’ column with ‘Harlequin’

USERS DATAFRAME

This dataframe contains the information related to the users who read the books.

COLUMNS IN DATAFRAME:

User-ID: Unique ID of the user
Location: Location of the user
Age: Age of the user

- The dataframe had 278858 rows and 3 columns.
- No duplicate entries were present in the dataset.
- The ‘Age’ column had 110762 null values which was around 40% of the data, so did not impute the missing data of the ‘Age’ column.
- Created a function ‘country’ to extract the country name from the provided location of the user.

RATINGS DATAFRAME

This dataframe contains the book rating information.
Book-Ratings are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

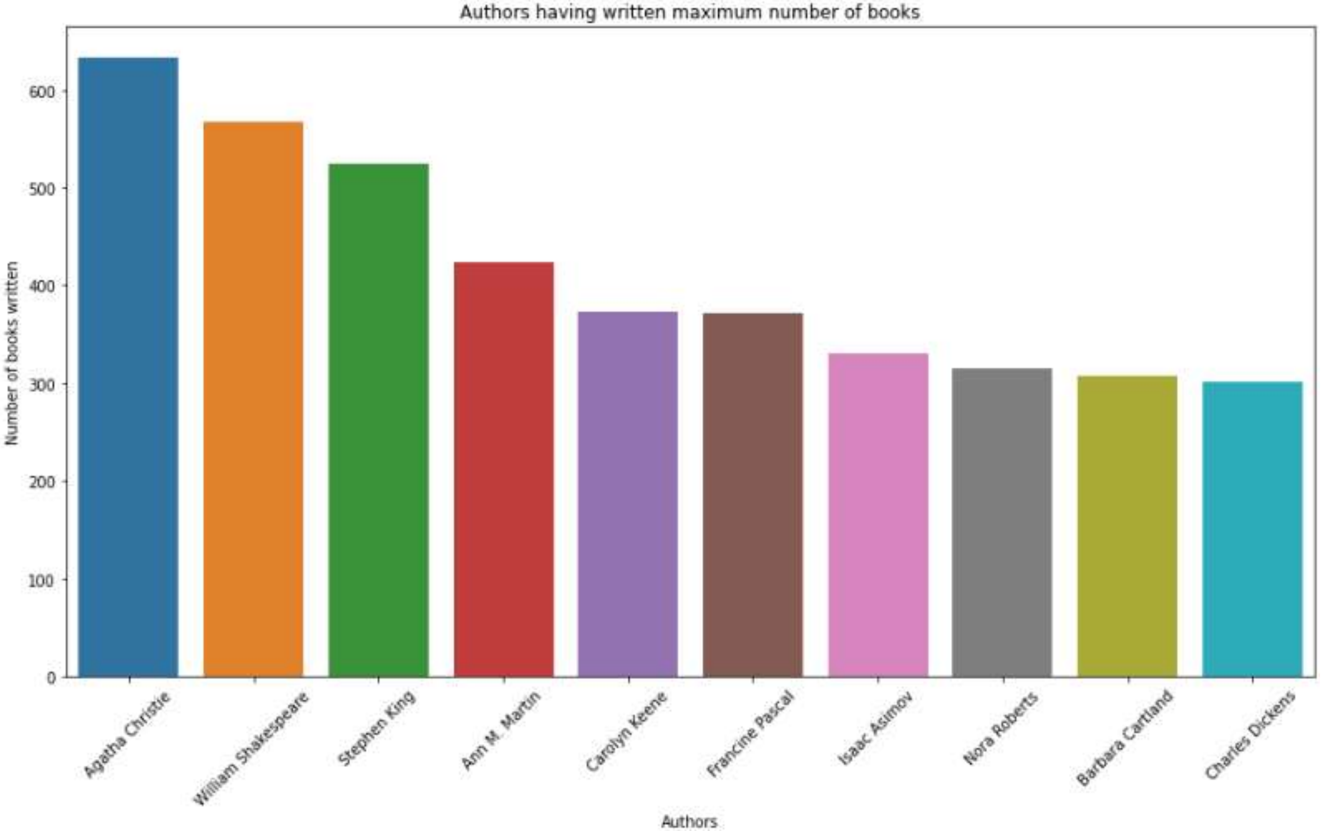
COLUMNS IN DATAFRAME:

User-ID: Unique ID of the user
ISBN: Unique ID to identify a book
Book-Rating: Ratings of the book (In the range of 0-10)

- The dataframe had 1149780 rows and 3 columns.
- No duplicate entries were present in the dataset.
- No null values were present in the dataset.

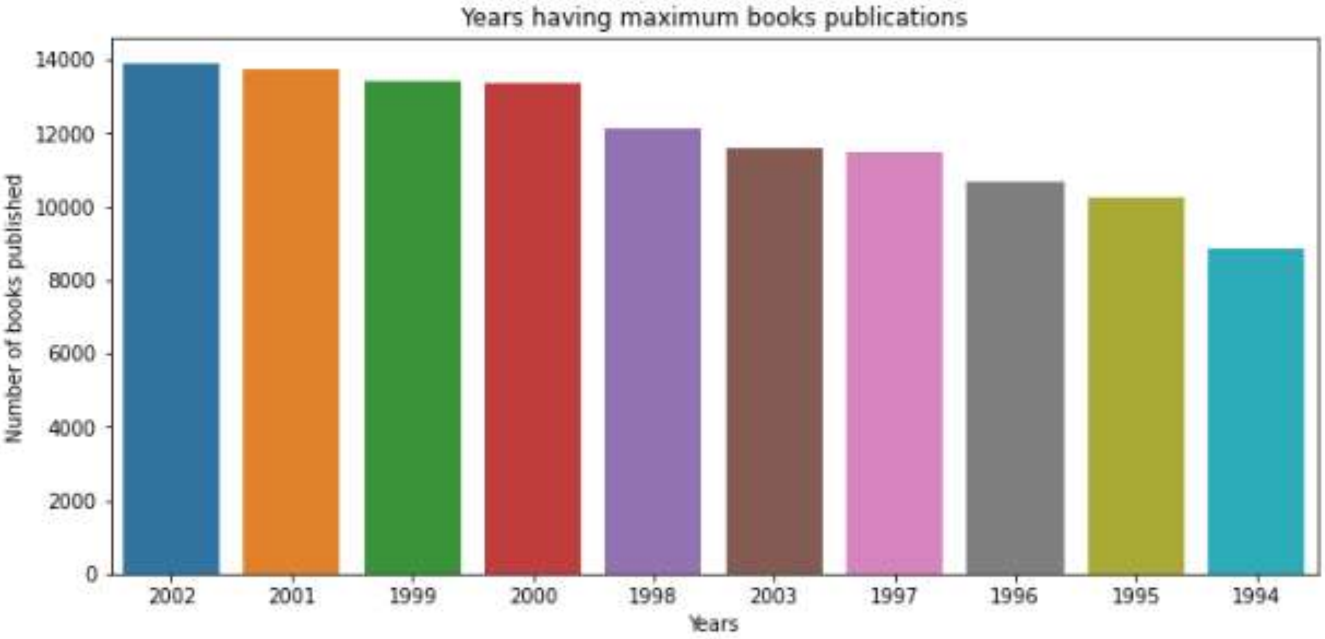
EXPLORATORY DATA ANALYSIS

Top 10 authors who have written maximum number of the books



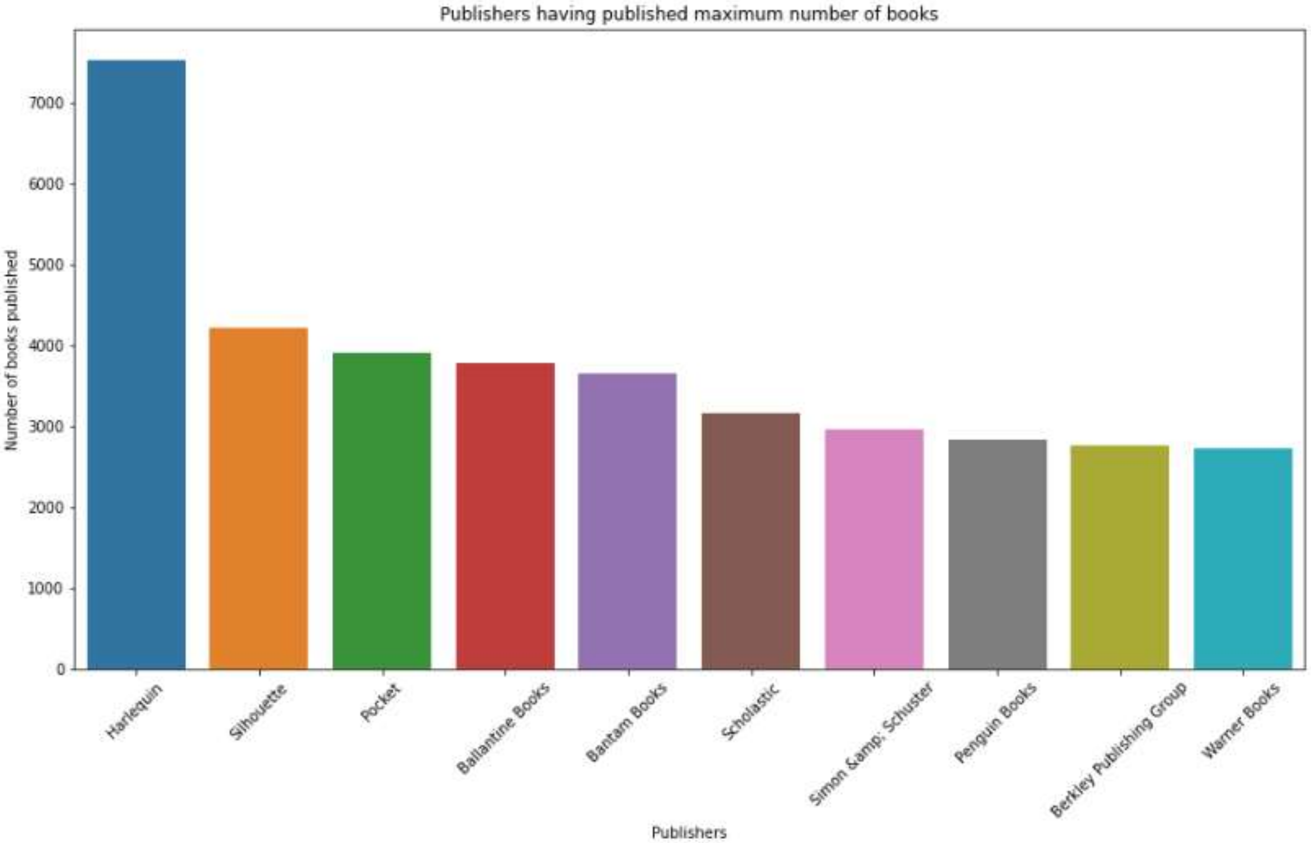
The author ‘Agatha Christie’ has written maximum number of books i.e., around 650, followed by William Shakespeare and Stephen King

Top 10 years in which maximum number of books were published



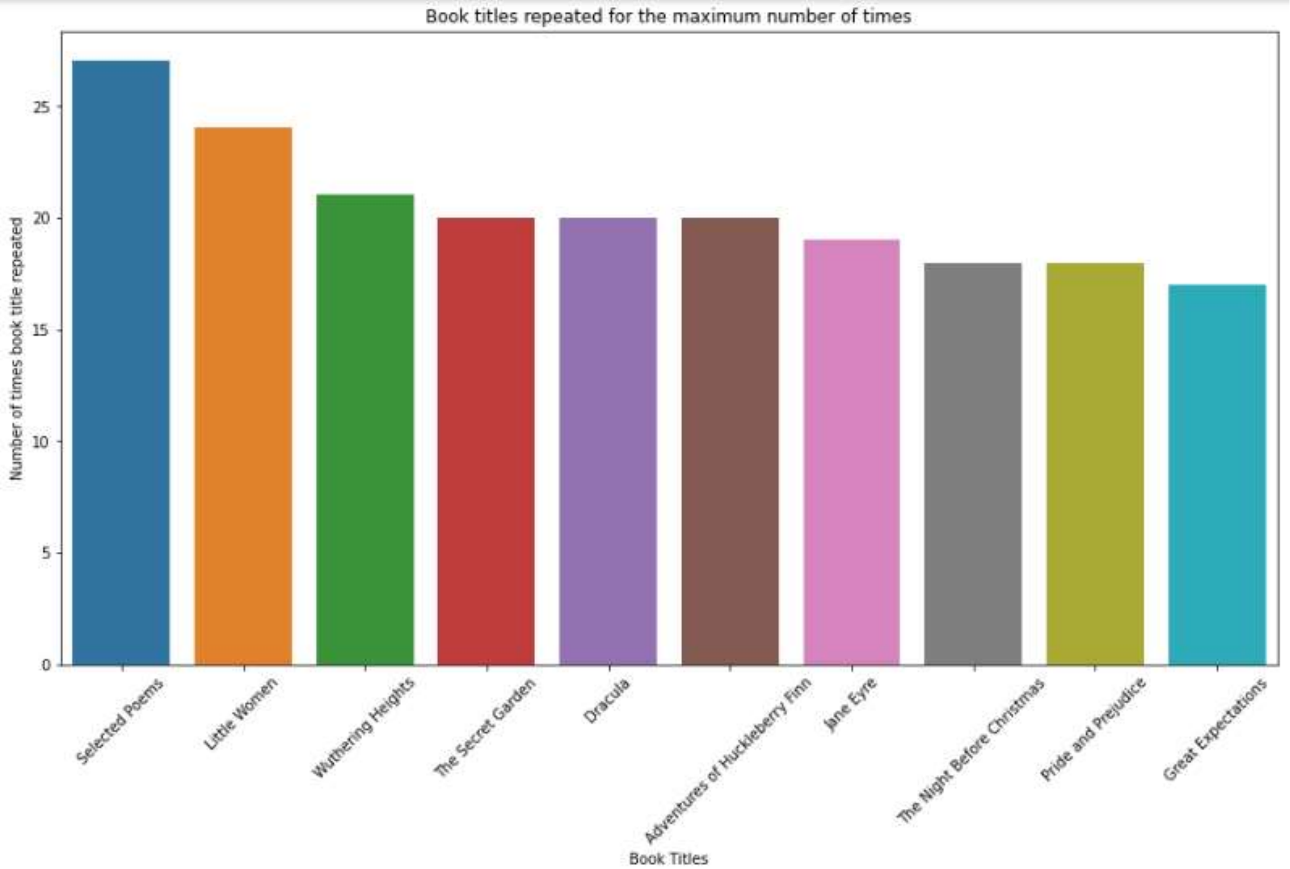
Maximum number of books were published in the year 2002 which was around 1380 followed by the years 2001 and 1999

Top 10 publishers who have published maximum number of books



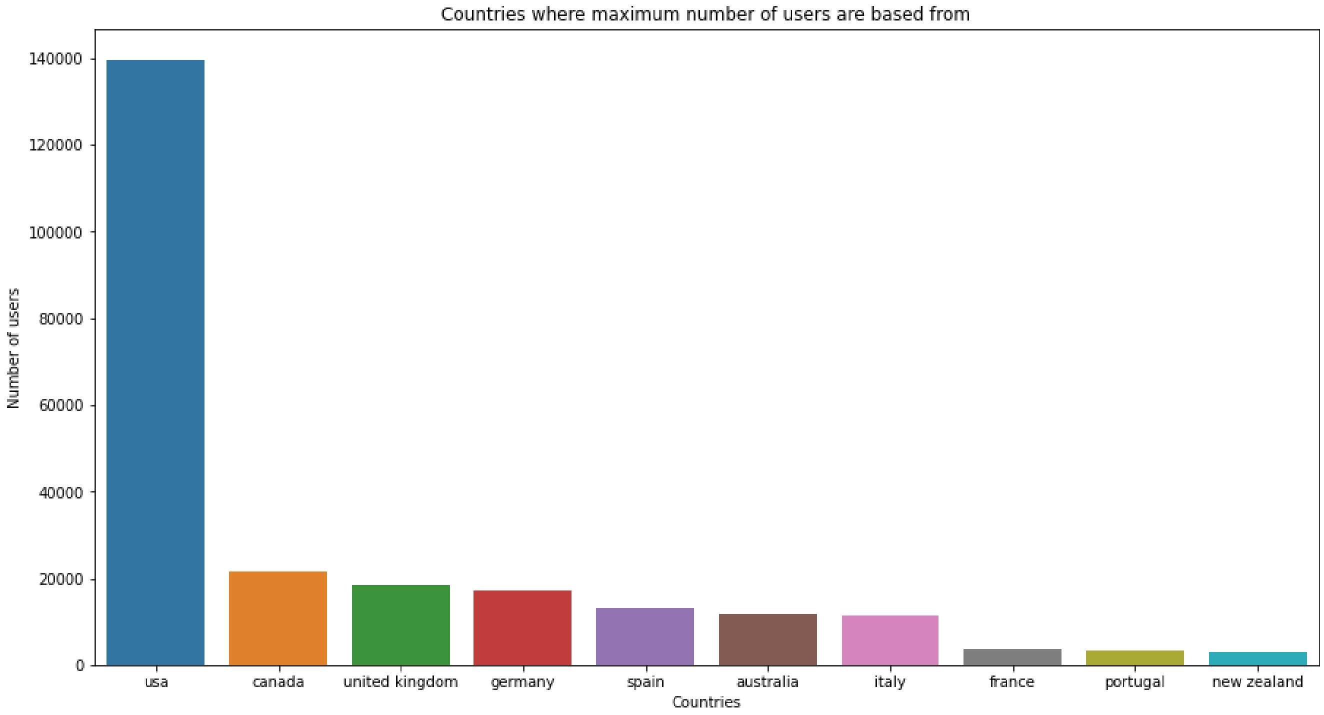
Maximum number of books were published by the publisher ‘Harlequin’ i.e., around 7600, followed by ‘Silhouette’ and ‘Pocket’

Top 10 book titles repeated for the maximum number of times



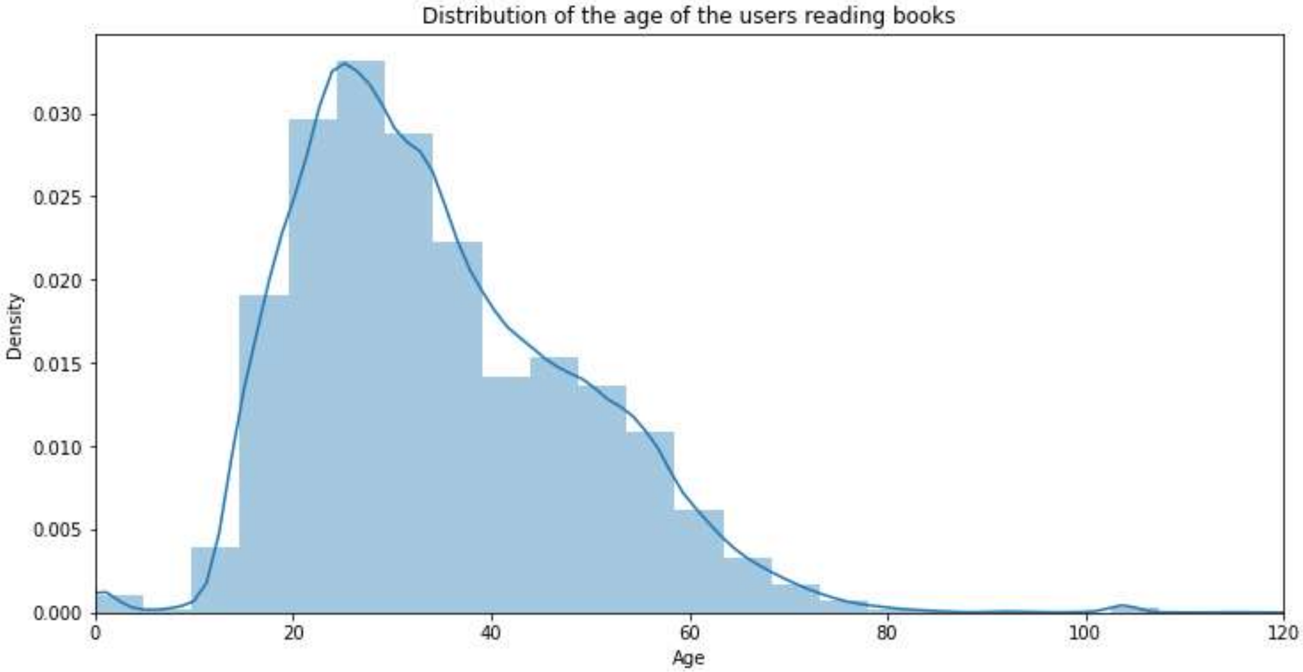
The book title that has been repeated maximum number of times is ‘Selected Poems’ i.e., around 28 times followed by ‘Little Women’ and ‘Wuthering Heights’

Top 10 countries where maximum number of users are based from



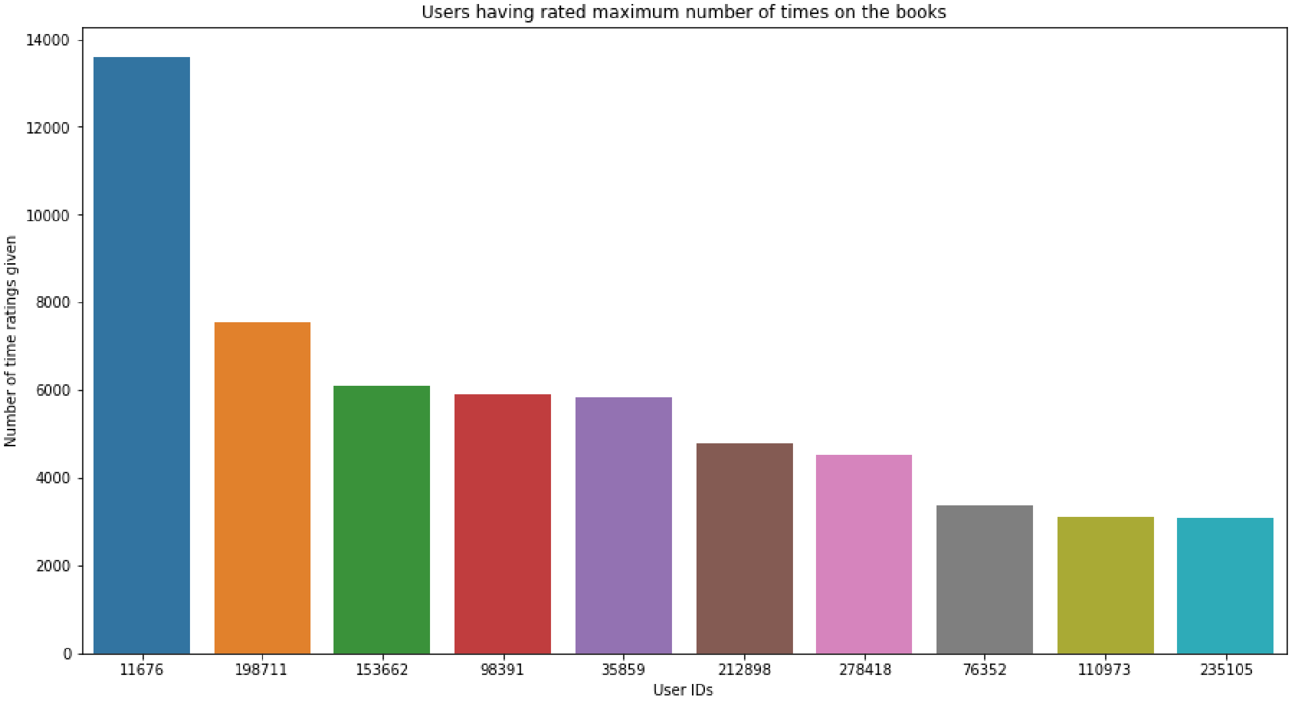
Maximum number of users are based in the United States i.e., about 14000 followed by Canada and the United Kingdom

Distribution of the reading age of the users



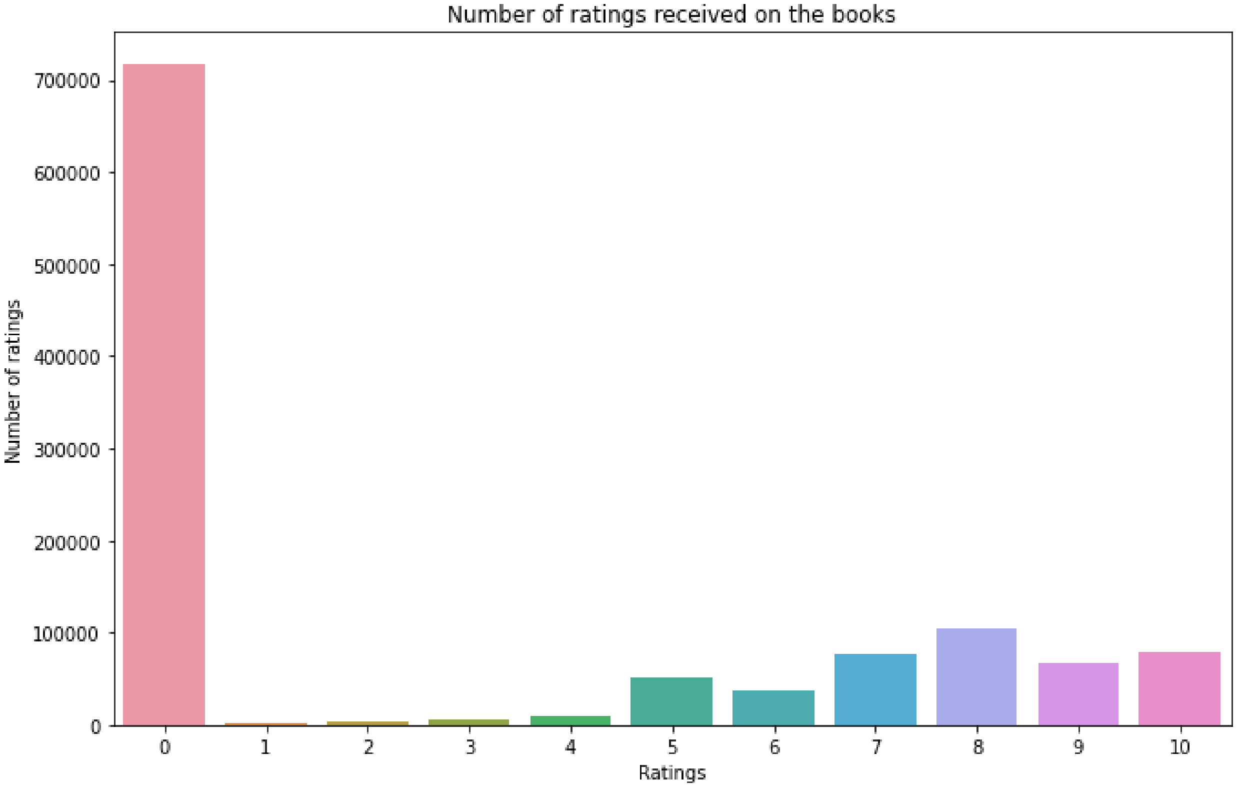
The average reading age of the user is around 20-40. The age at which the users read the maximum is about 28-30 years of age.

Top 10 users who have rated maximum number of times on the books



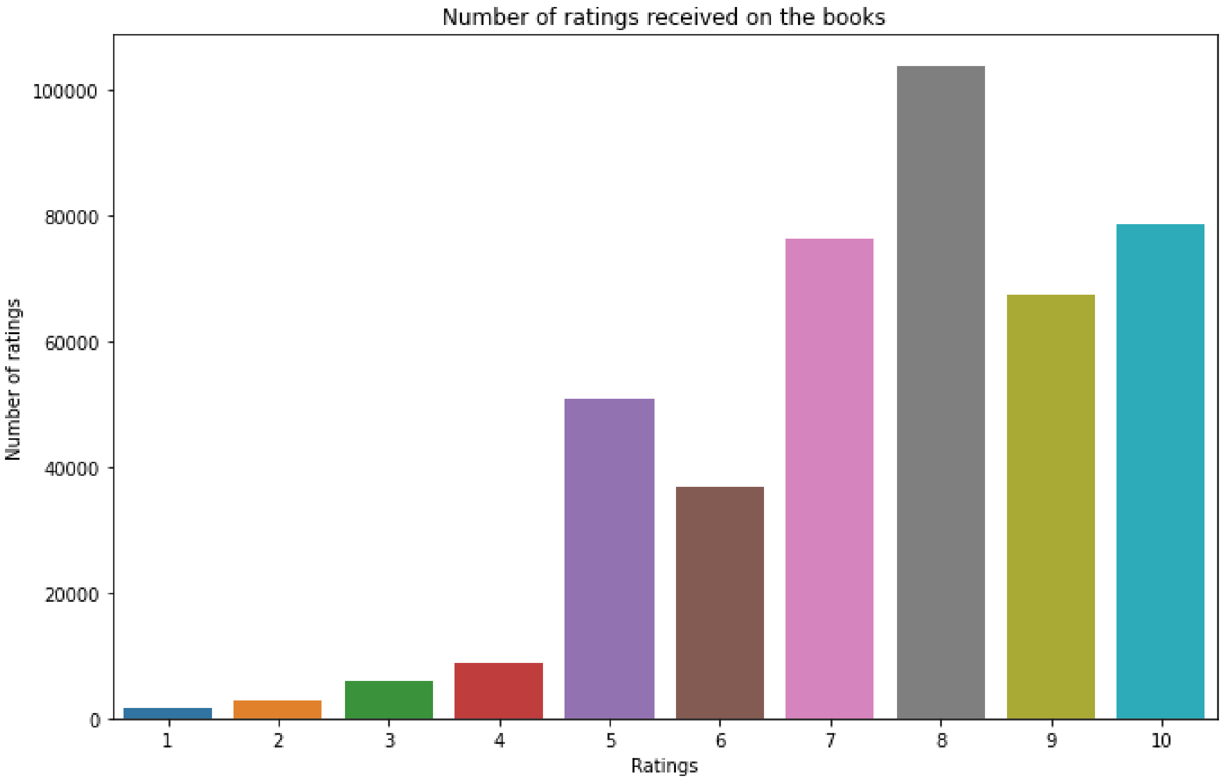
The user id number 11676 has rated for the maximum number of times on the books followed by user id number 198711 and 153662

Number of different ratings received on the books



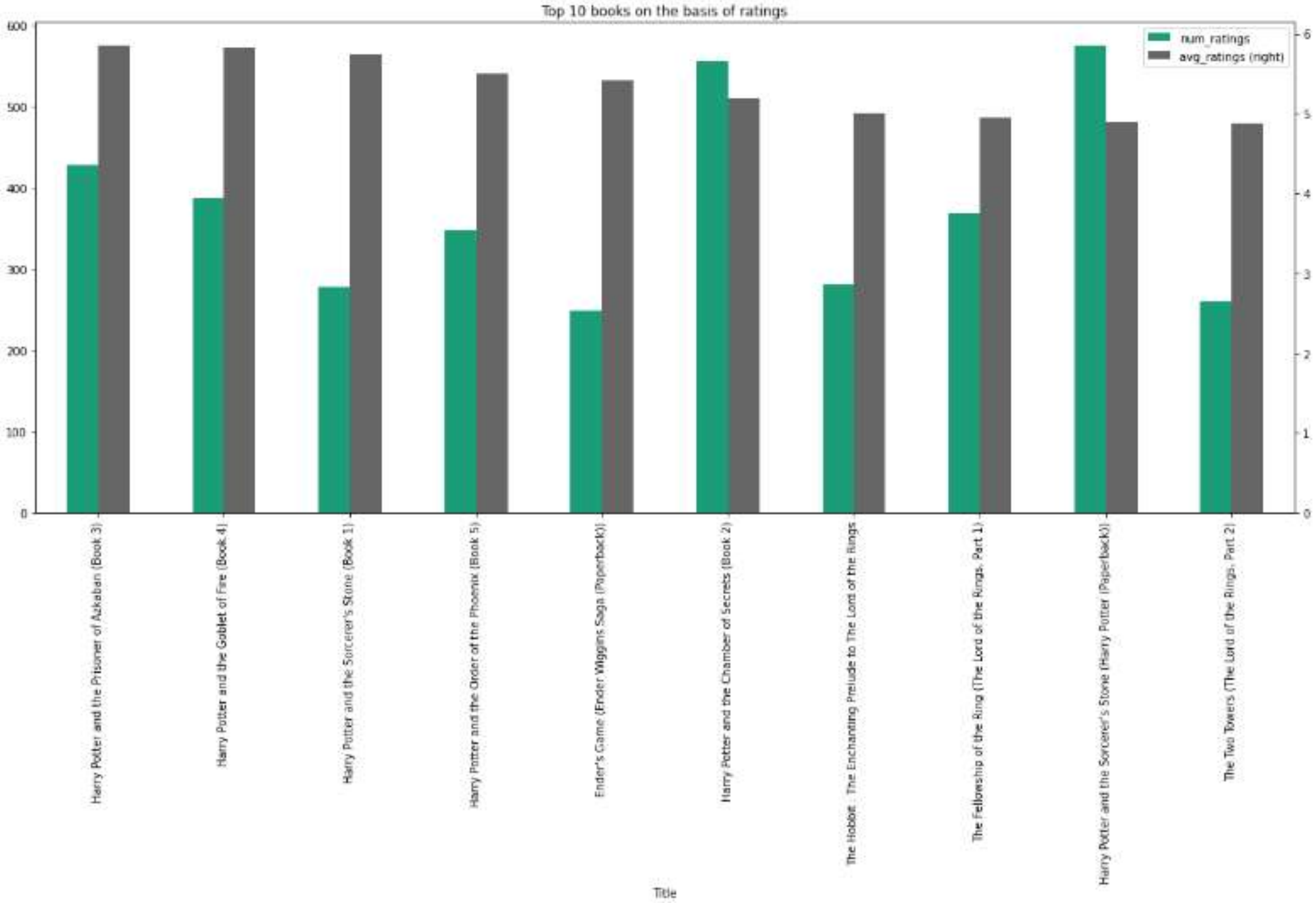
Maximum number of books have received the ratings as 0

Number of different ratings received on the books excluding 0 rating



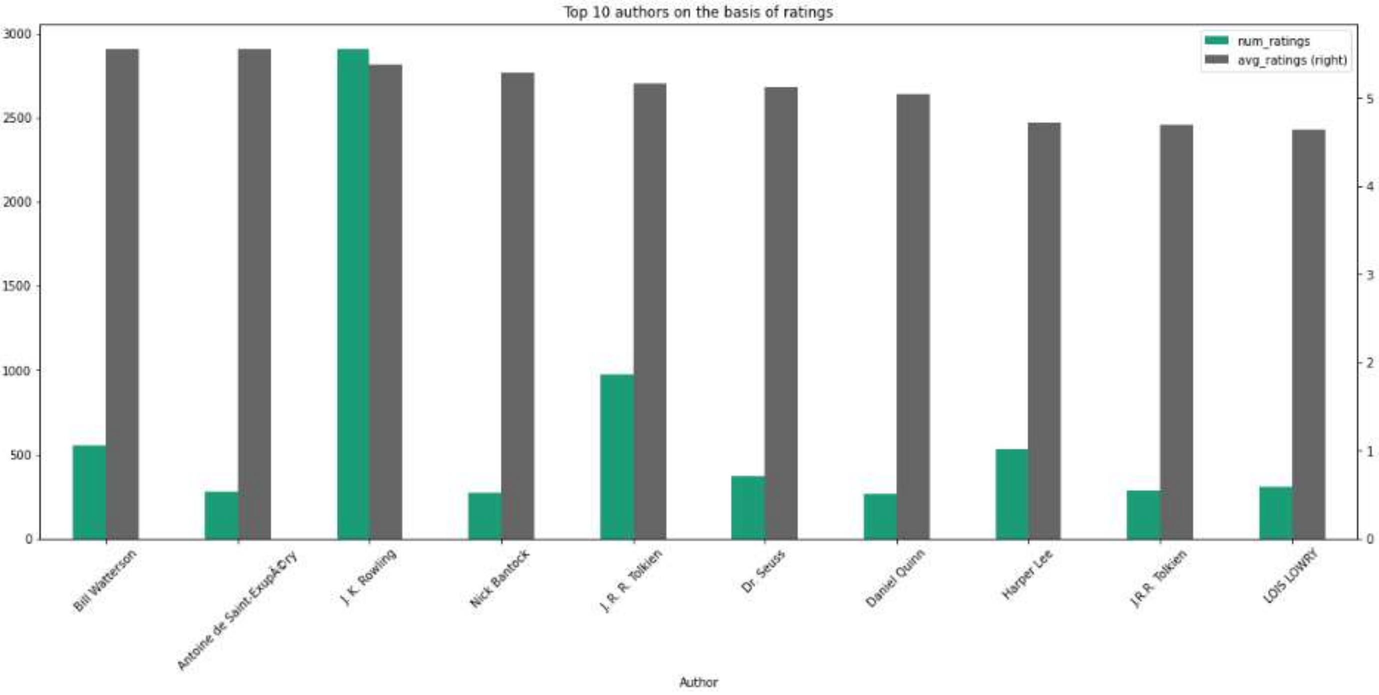
After excluding 0 ratings it can be seen that maximum number of books have received the ratings as 8 and followed by 10 and 7

Top 10 most famous books on the basis of number of ratings and average rating



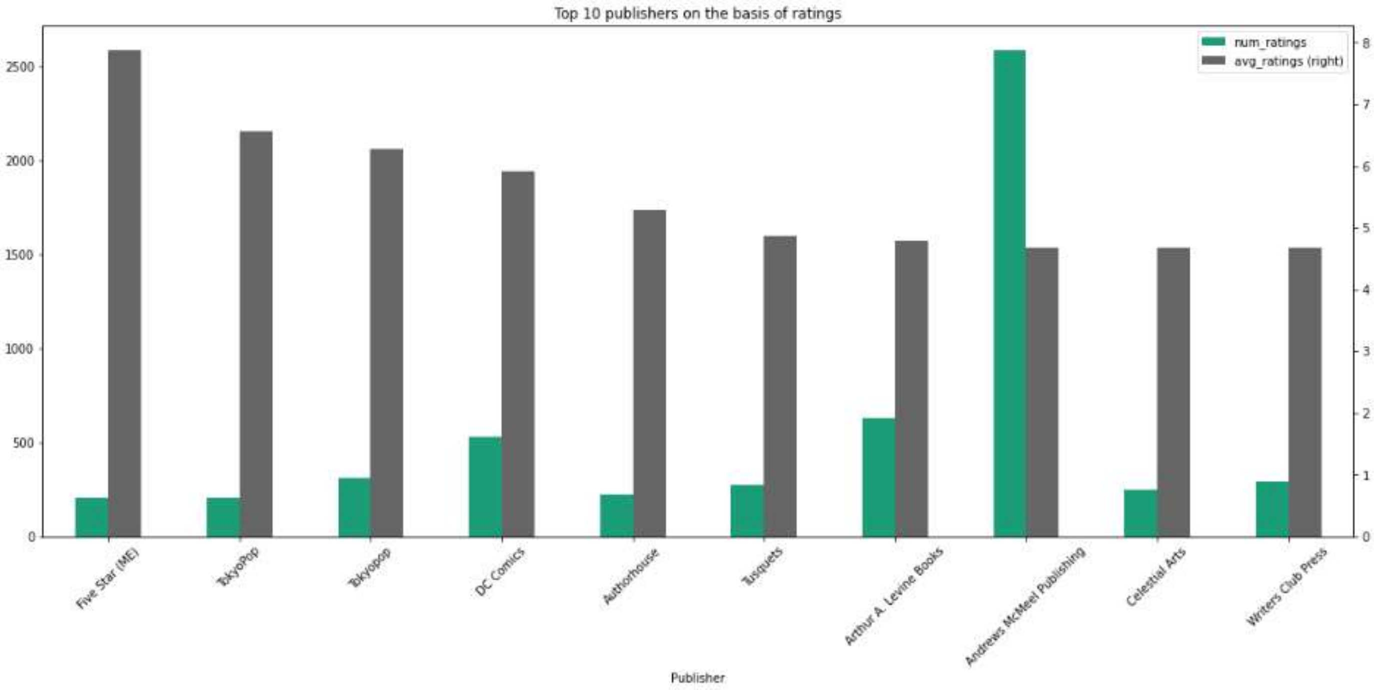
The books 'Harry Potter and the Chamber of Secrets (Book 2)' and 'Harry Potter and Sorcerer's Stone' have the highest number of ratings, which means a large number of people have read those books.

Top 10 most famous authors on the basis of number of ratings and average rating



The author ‘J.K. Rowling’ (writer of Harry Potter) has the highest number of ratings, which means the author has reached a large number of audience.

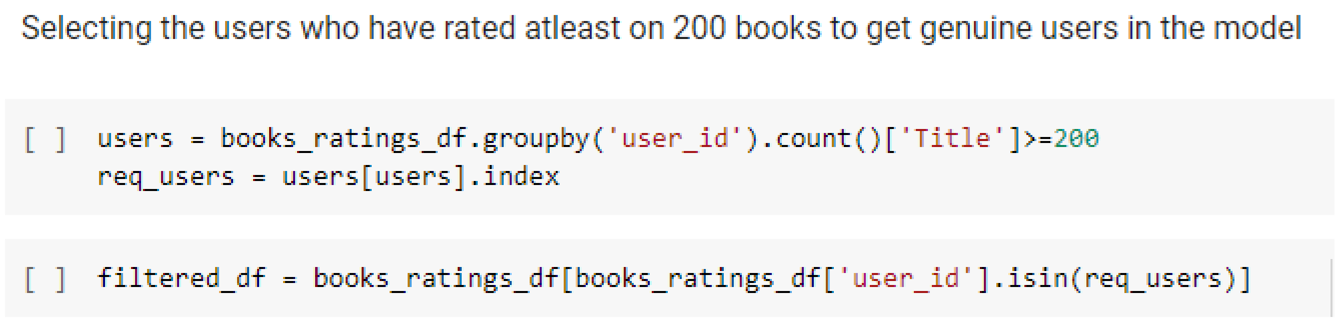
Top 10 most famous publishers on the basis of number of ratings and average rating



The publisher ‘Andrews McMeel Publishing’ has received the highest number of ratings amongst all and thus it has maximum reach to the audience.

COLLABORATIVE FILTERING APPROACH

1. Selected the users who have atleast rated on 200 or more books to get a group of intellectual and genuine users.

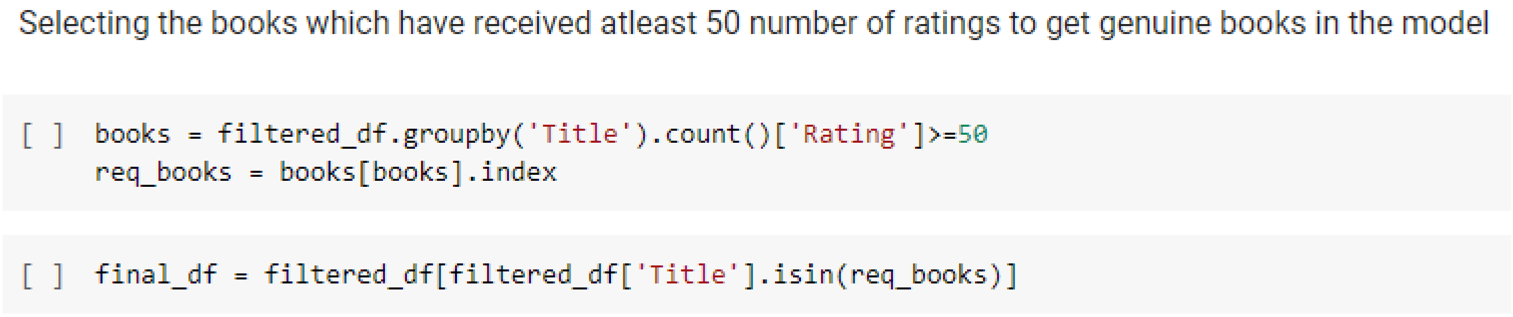


Selecting the users who have rated atleast on 200 books to get genuine users in the model

```
[ ]  users = books_ratings_df.groupby('user_id').count()['Title']>=200
     req_users = users[users].index
```

```
[ ]  filtered_df = books_ratings_df[books_ratings_df['user_id'].isin(req_users)]
```

2. Selected the books which have received atleast 50 ratings or more to get a collection of all the good and genuine books.

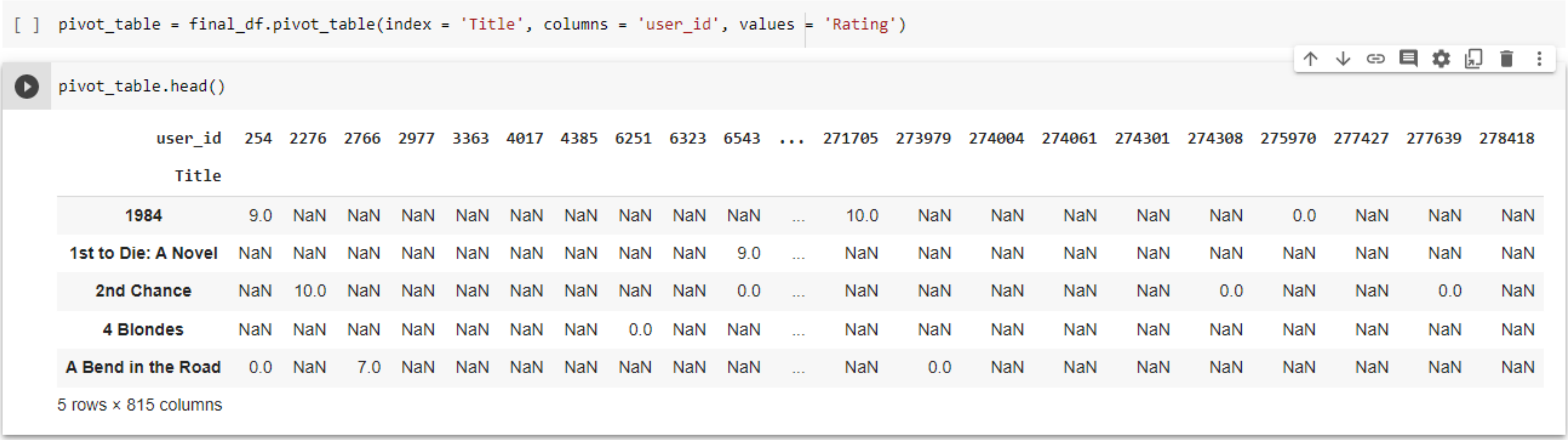


Selecting the books which have received atleast 50 number of ratings to get genuine books in the model

```
[ ]  books = filtered_df.groupby('Title').count()['Rating']>=50
     req_books = books[books].index
```

```
[ ]  final_df = filtered_df[filtered_df['Title'].isin(req_books)]
```

3. Created a pivot table with the index as the ‘Book Title’, attributes as the ‘User-ID’ and the values of the pivot table as the ‘rating’ given by a particular user to a particular book.



```
[ ] pivot_table = final_df.pivot_table(index = 'Title', columns = 'user_id', values = 'Rating')
```

```
pivot_table.head()
```

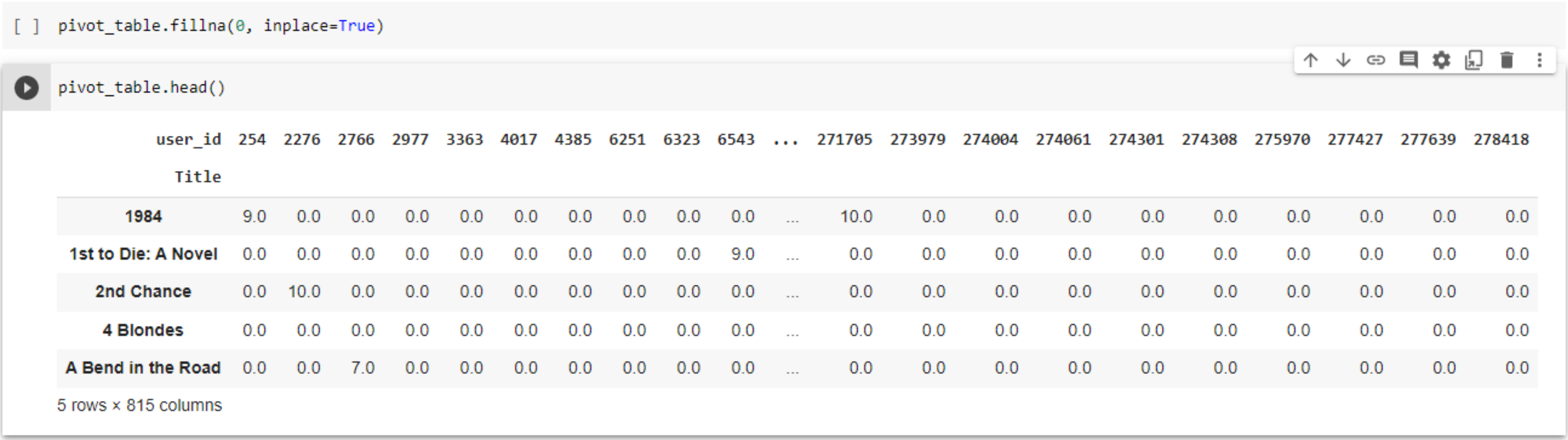
user_id	254	2276	2766	2977	3363	4017	4385	6251	6323	6543	...	271705	273979	274004	274061	274301	274308	275970	277427	277639	278418
Title																					
1984	9.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	10.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN
1st to Die: A Novel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	9.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2nd Chance	NaN	10.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	...	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	0.0	NaN
4 Blondes	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
A Bend in the Road	0.0	NaN	7.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 815 columns

Pivot Table

A pivot table is a similar operation that is commonly seen in spreadsheets and other programs that operate on tabular data. The pivot table takes simple column-wise data as input, and groups the entries into a two-dimensional table that provides a multidimensional summarization of the data.

4. Replaced the Nan values of the pivot table with 0s to create a sparse matrix.



```
[ ]  pivot_table.fillna(0, inplace=True)
```

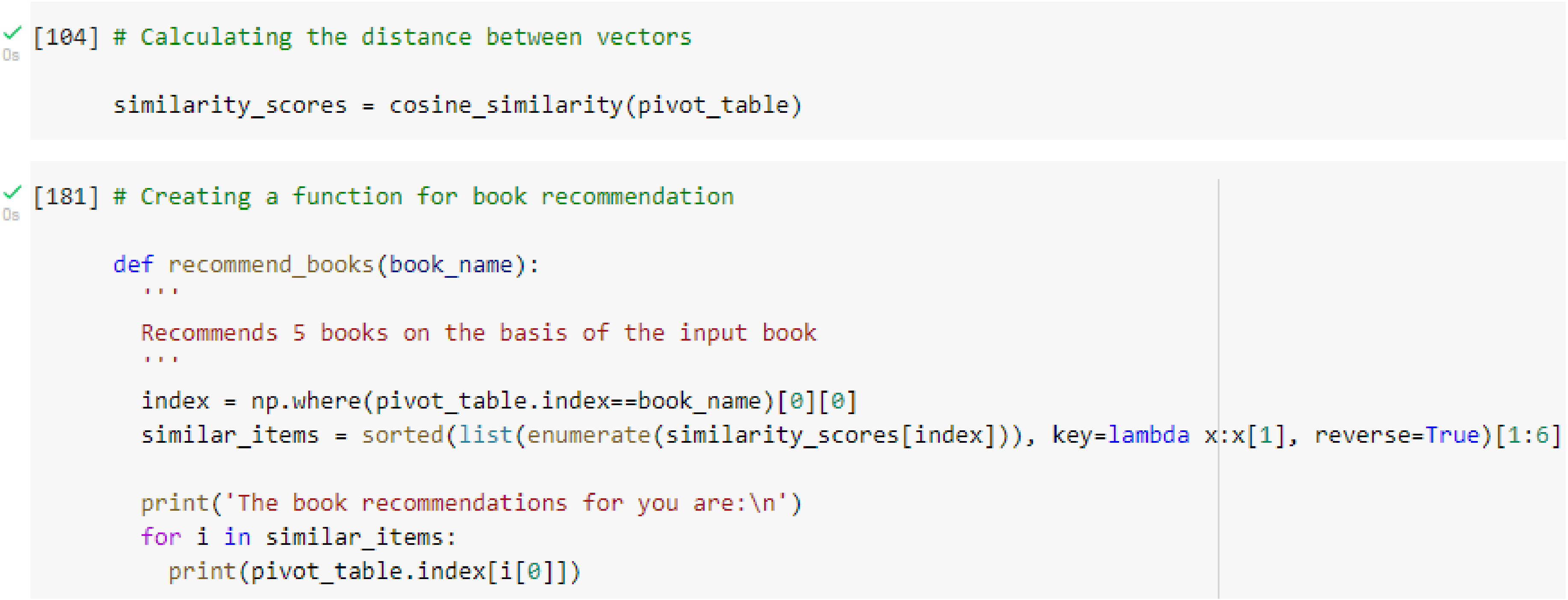
```
pivot_table.head()
```

user_id Title	254	2276	2766	2977	3363	4017	4385	6251	6323	6543	...	271705	273979	274004	274061	274301	274308	275970	277427	277639	278418
1984	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	10.0	0.0	0.0	0.0	0.0	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	0.0	0.0	0.0	9.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2nd Chance	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4 Blondes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A Bend in the Road	0.0	0.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 815 columns

Cosine Similarity

Cosine similarity is a metric used to measure how similar the two items or documents are irrespective of their size. It measures the cosine of an angle between two vectors projected in multi-dimensional space. This allows us to measure the similarity of a document of any type. Due to a multi-dimensional array, any number of variables (which are treated as dimensions) can be used, which in turn supports large sized documents

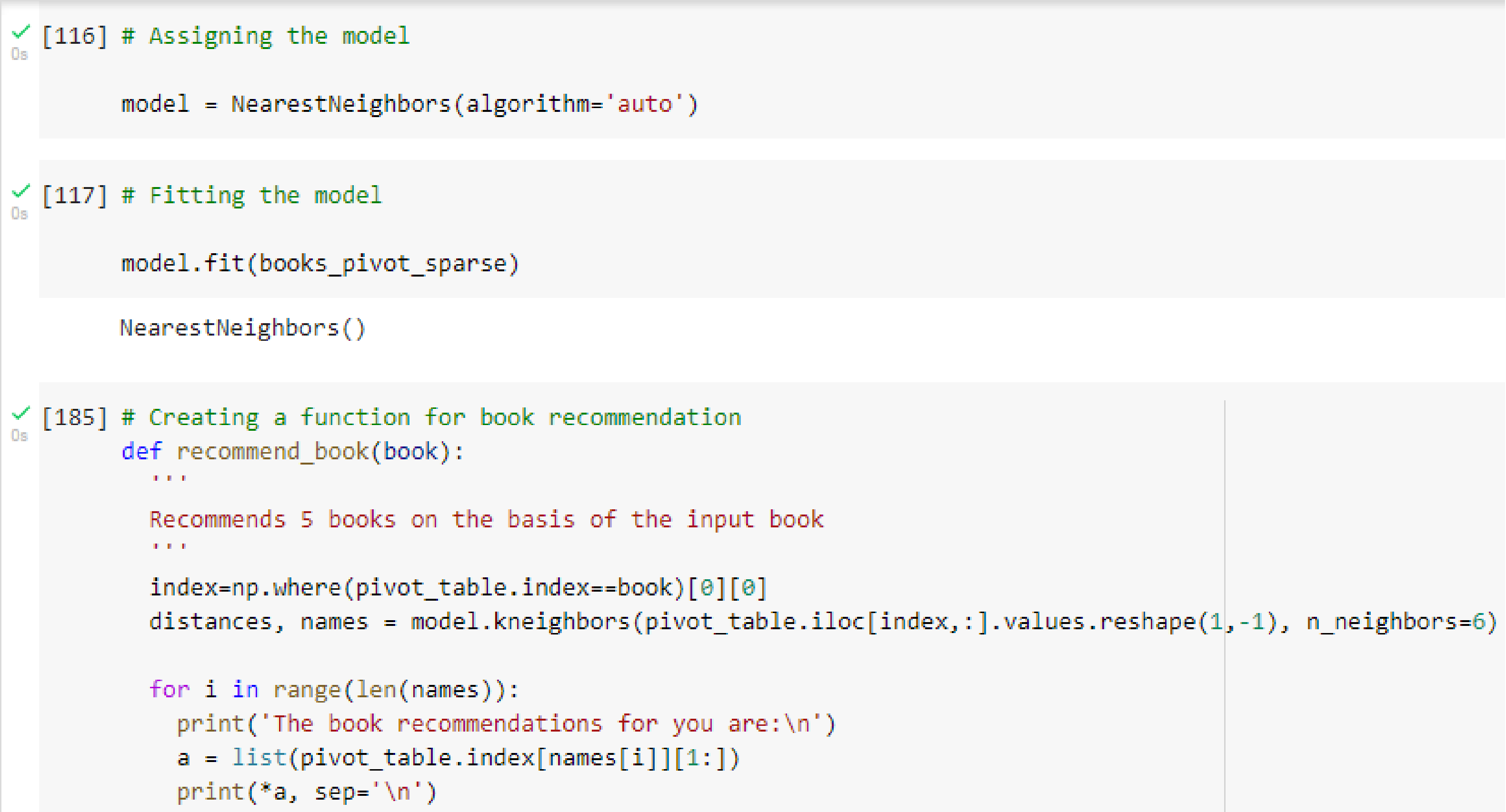


```
[104] # Calculating the distance between vectors

      similarity_scores = cosine_similarity(pivot_table)
```

```
[181] # Creating a function for book recommendation

      def recommend_books(book_name):
        '''
        Recommends 5 books on the basis of the input book
        '''
        index = np.where(pivot_table.index==book_name)[0][0]
        similar_items = sorted(list(enumerate(similarity_scores[index])), key=lambda x:x[1], reverse=True)[1:6]

        print('The book recommendations for you are:\n')
        for i in similar_items:
          print(pivot_table.index[i[0]])
```

Nearest Neighbors

Unsupervised learner for implementing neighbor searches. NearestNeighbors implements unsupervised nearest neighbors learning. It acts as a uniform interface to three different nearest neighbors algorithms: BallTree, KDTree, and a brute-force.



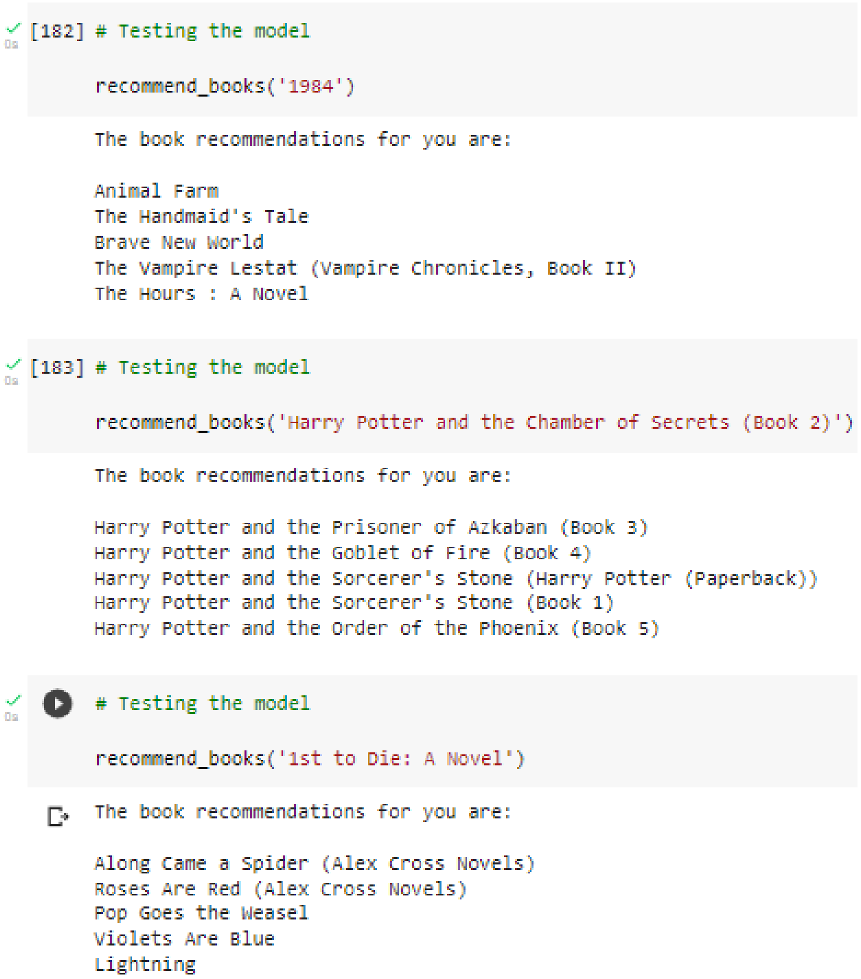
```
[116] # Assigning the model

      model = NearestNeighbors(algorithm='auto')
```

```
[117] # Fitting the model

      model.fit(books_pivot_sparse)
```

```
      NearestNeighbors()
```

```
[185] # Creating a function for book recommendation
      def recommend_book(book):
        '''
        Recommends 5 books on the basis of the input book
        '''
        index=np.where(pivot_table.index==book)[0][0]
        distances, names = model.kneighbors(pivot_table.iloc[index,:].values.reshape(1,-1), n_neighbors=6)

        for i in range(len(names)):
          print('The book recommendations for you are:\n')
          a = list(pivot_table.index[names[i]][1:])
          print(*a, sep='\n')
```

COSINE SIMILARITY RESULTS

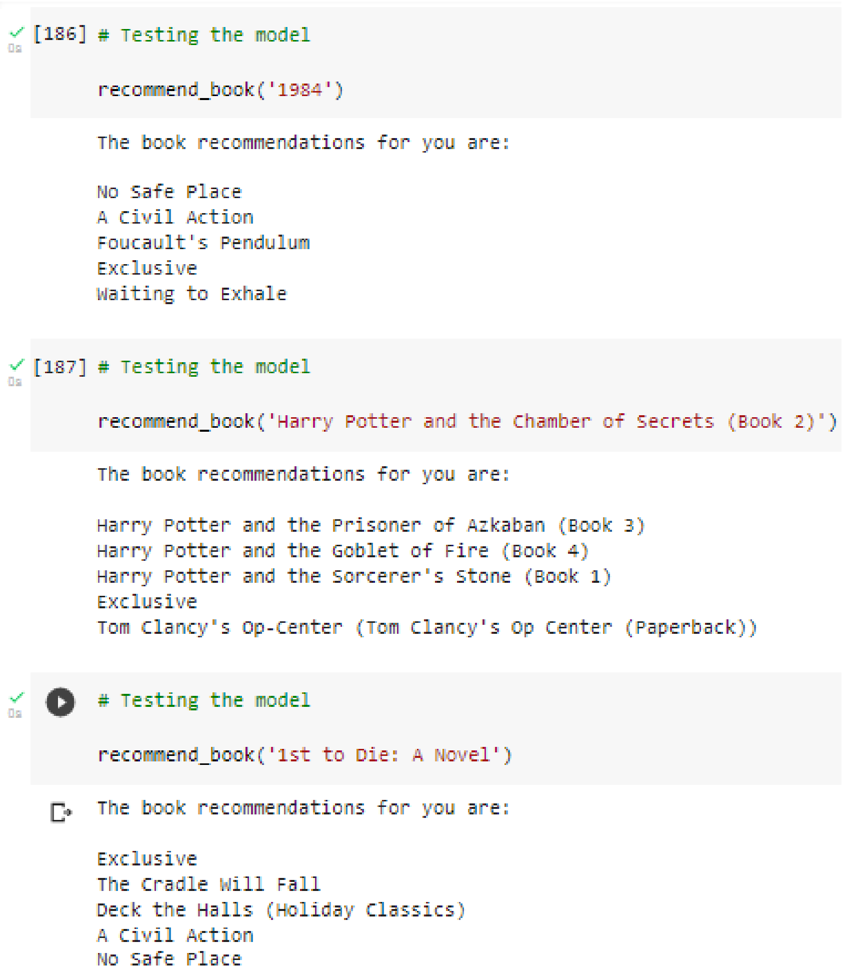


```
[182] # Testing the model

      recommend_books('1984')

      The book recommendations for you are:

      Animal Farm
      The Handmaid's Tale
      Brave New World
      The Vampire Lestat (Vampire Chronicles, Book II)
      The Hours : A Novel
```

```
[183] # Testing the model

      recommend_books('Harry Potter and the Chamber of Secrets (Book 2)')

      The book recommendations for you 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 (Harry Potter (Paperback))
      Harry Potter and the Sorcerer's Stone (Book 1)
      Harry Potter and the Order of the Phoenix (Book 5)
```

```
      # Testing the model

      recommend_books('1st to Die: A Novel')

      The book recommendations for you are:

      Along Came a Spider (Alex Cross Novels)
      Roses Are Red (Alex Cross Novels)
      Pop Goes the Weasel
      Violets Are Blue
      Lightning
```

NEARESTNEIGHBORS RESULTS



```
[186] # Testing the model

      recommend_book('1984')

      The book recommendations for you are:

      No Safe Place
      A Civil Action
      Foucault's Pendulum
      Exclusive
      Waiting to Exhale
```

```
[187] # Testing the model

      recommend_book('Harry Potter and the Chamber of Secrets (Book 2)')

      The book recommendations for you 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)
      Exclusive
      Tom Clancy's Op-Center (Tom Clancy's Op Center (Paperback))
```

```
      # Testing the model

      recommend_book('1st to Die: A Novel')

      The book recommendations for you are:

      Exclusive
      The Cradle Will Fall
      Deck the Halls (Holiday Classics)
      A Civil Action
      No Safe Place
```

CONCLUSION

- The authors ‘Agatha Christie’, ‘William Shakespeare’ and ‘Stephen King’ have written the maximum number of books.
- The year 2002 can be termed as the year in which the maximum number of books were published.
- The publishers ‘Harlequin’, ‘Silhouette’ and ‘Pocket’ have published a maximum number of books.
- The book titles such as ‘Selected Poems’, ‘Little Women’ and ‘Wuthering Heights’ have been published for the maximum number of times.
- Maximum readers of the books and novels can be seen in the United States followed by Canada and the United Kingdom.
- The average reading age of the users can be seen around 20-40. The age at which the users read the maximum is about 28-30 years of age so it is advisable to create the reading content on this basis of the age.
- The maximum ratings received by the books can be seen to be 0 but if 0 ratings are removed then 8, 7 and 10 can be seen as the maximum rating and thus it can be concluded as good content books.
- The books ‘Harry Potter and the Chamber of Secrets (Book 2)’ and ‘Harry Potter and Sorcerer’s Stone’ can be declared as the best and the most popular books as these books have received highest number of ratings, which means a large number of people have read those books and also these books lie in the group of top 10 most popular books on the basis of average ratings.
- The author ‘J.K. Rowling’ (writer of Harry Potter) can be termed as the most popular author as he has received the highest number of ratings, which means the author has reached a large number of audience and also has a good average rating.
- The publisher ‘Five Star (ME)’ has the highest average rating amongst all the publishers but the publisher ‘Andrews McMeel Publishing’ has received the highest number of ratings amongst all the publishers and thus it has maximum reach to the audience also it has good average rating thus it can be termed as the optimum priced publisher as many people can buy and read the books published by it.

Thank you