Capstone Project -4

BOOK RECOMMENDATION SYSTEM

(Unsupervised machine learning algorithm)

By

Nikhil Khushal Nimje

smart sensor systems
Microsystem Technol
Köhler · Fritzsche
Nanotechnology
3G Evolution
HSPA and LTE for Mobile Broadband
Control Techniques for Complex Net
Security and Quality of Service in Ad Hoc Wireless
Bing
Emerging Technologies in Wireless LANs
NEXT GENERATION IPTV SERVICES AND TECH

Introduction

- Today the amount of information in the internet growth very rapidly and people need some instruments to find and access appropriate information. One of such tools is called recommendation system. Recommendation systems help to navigate quickly and receive necessary information. Generally they are used in Internet shops to increase the profit. This paper proposes a quick and intuitive book recommendation system that helps readers to find appropriate book to read next. The overall architecture is presented with it's detailed description. We used a collaborative filtering method based on Pearson correlation coefficient. Finally the experimental results based on the online survey are provided with some discussions.

❖ Points to be Discuss

- Problem Statement
- Data Summary
- Data Pipeline
- Feature Engineering
- Data Cleaning
- EDA
- Preparing Data for Model
- Applying Model
- Conclusion
- Challenges
- Future Scope

Problem Statement

- During the last few decades, recommender systems have taken more and more place in our lives. From e-commerce to online advertisement, 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.
- Recommender systems are algorithms aimed at suggesting relevant items to users. The main objective is to create a book recommendation system for users.

Data Summary

- The dataset is comprised of three csv files:: User_data, Book_data, Rating_data

- Users_dataset.
  - User-ID (unique for each user)
  - Location (contains city, state and country separated by commas)
  - Age

□ Shape of Dataset - (278858, 3)

Data Summary

Books_dataset.

- ISBN (unique for each book)
- Book-Title
- Book-Author
- Year-Of-Publication
- Publisher
- Image-URL-S
- Image-URL-M
- Image-URL-L
- Shape of Dataset - (271360, 8)

Ratings_dataset.

- User-ID
- ISBN
- Book-Rating
- Shape of Dataset - (1149780, 3)

Data Pipeline

- Primary Inspection: Observed irregularities in the data set and unique values for different columns.
- Processing & Feature Engineering: Handled missing values, capped outliers and engineered features for further analysis. Data set was split for building different explicit rank based and implicit rank based recommender systems.
- EDA: Exploratory analysis was performed on columns like Book-Rating, Location, Book-Author to review trends and patterns emerging in the data set.
- Applying Simple Models: Models, based on mean ratings and K-Nearest-Neighbourhood Algorithm, were built to provide simple recommendations .

Data Pipeline

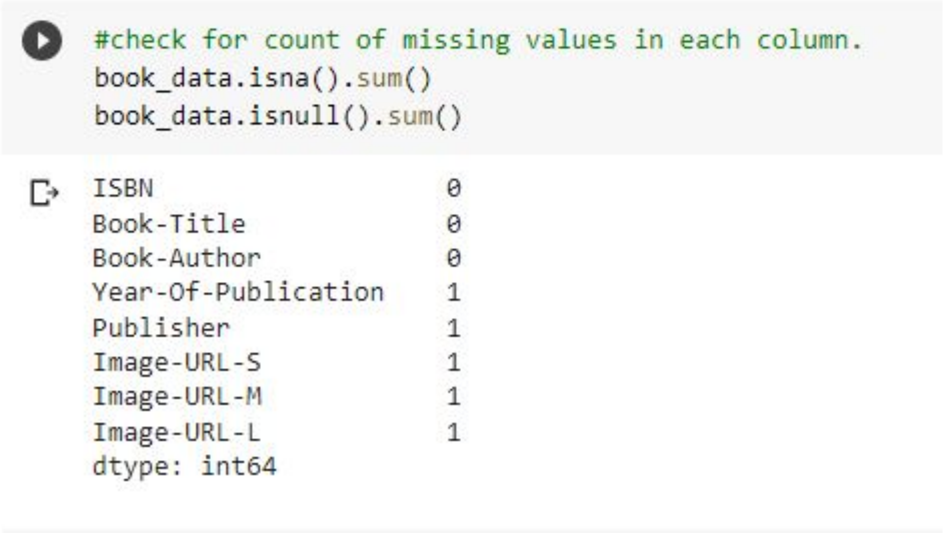
- Applying Collaborative Filtering Model: SVD model based collaborative filtering system was built to provide recommendations based on user-user similarity, for explicitly ranked items.
- Applying Memory based Filtering: K-Nearest-Neighbourhood Algorithm, was used to make recommendations based upon user age, for implicitly rated items.
- Content Based Solution: A model was built to recommend new books, based upon the content description of a user’s past purchase.

Feature Engineering

- Feature Engineering on Location: To analyse user country information, a function get_country() was developed to extract country information from Location column.
- Engineering Book Descriptions: Description for Book-Titles, was fetched from Google Books API, in order to perform implicit, content similarity based recommendations.
- Feature Engineering Age: Age Column was converted into bin, to better reflect preference of a user with respect to their age

Data Cleaning

Null Value Imputation:



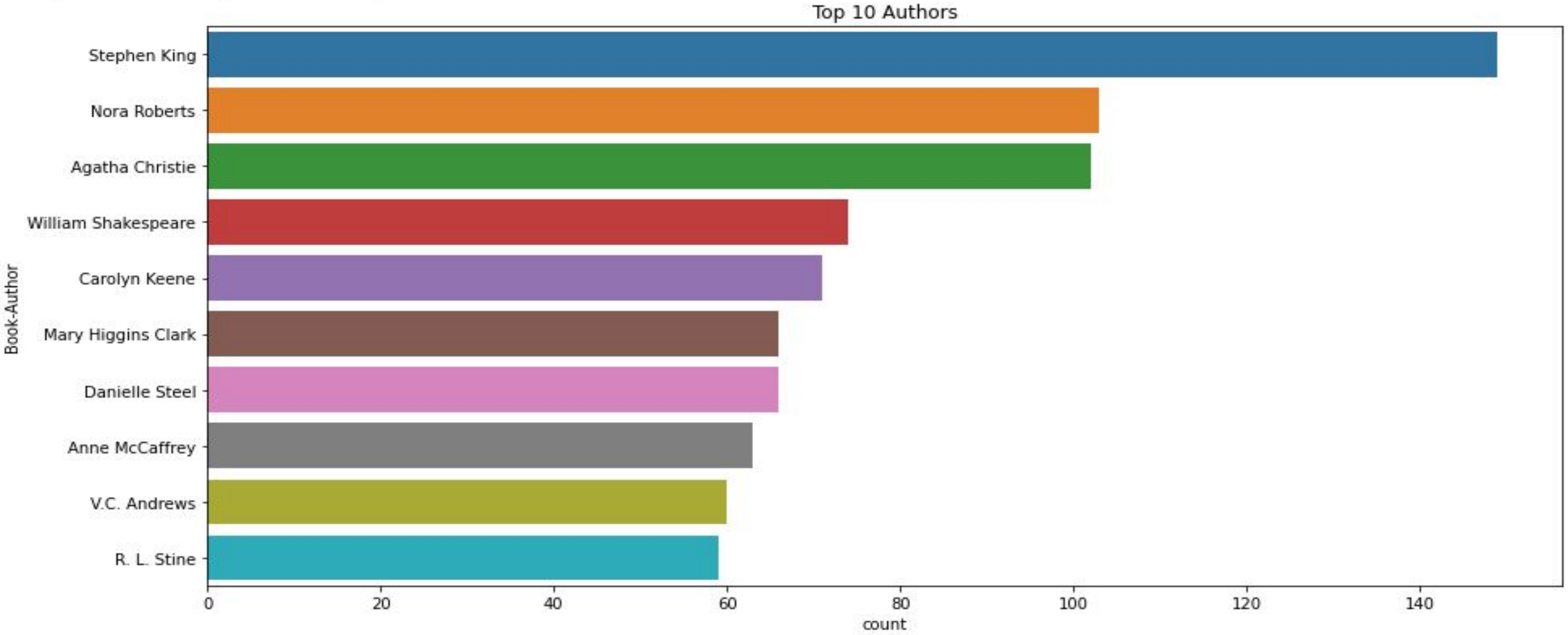
```
#check for count of missing values in each column.
book_data.isna().sum()
book_data.isnull().sum()
```

```
ISBN                   0
Book-Title             0
Book-Author            0
Year-Of-Publication    1
Publisher              1
Image-URL-S            1
Image-URL-M            1
Image-URL-L            1
dtype: int64
```

EDA

Observations from Book_data (Authors)

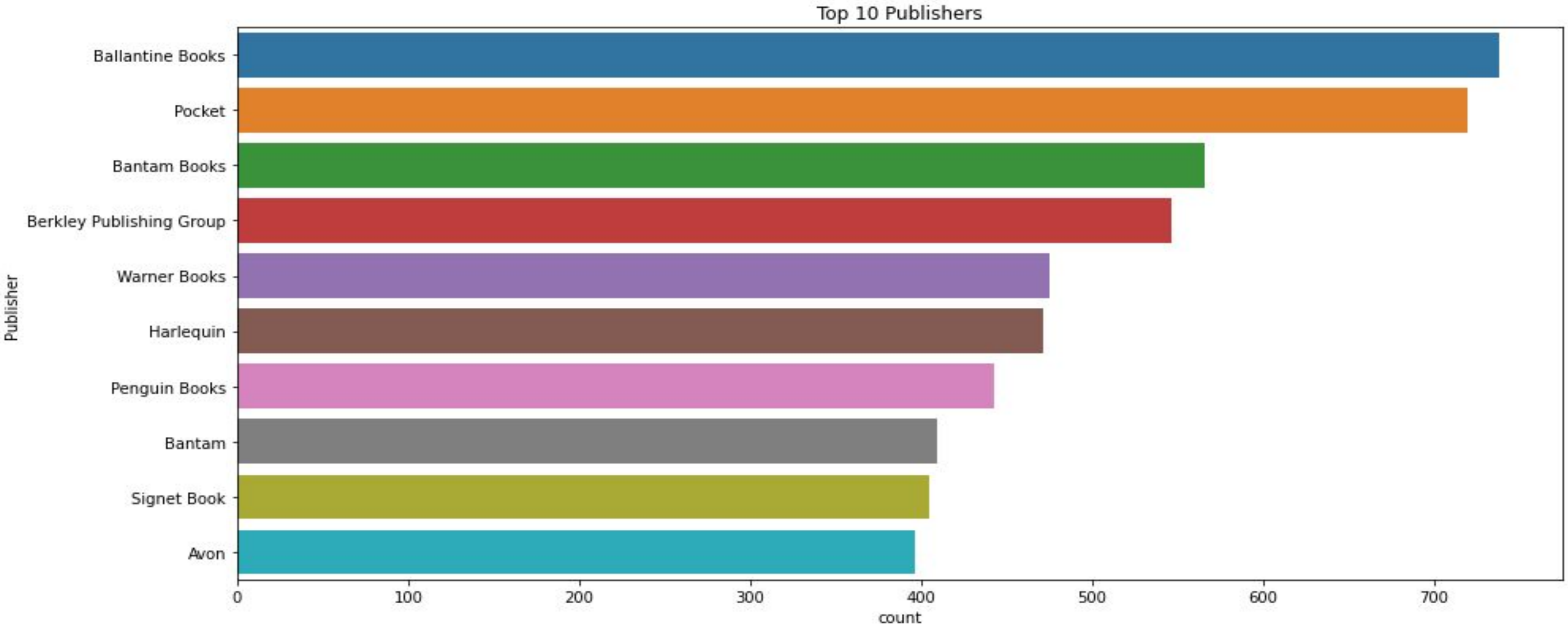
- Stephen king wrote highest number of books in our given dataset



EDA

Observations from Book_df (Publishers)

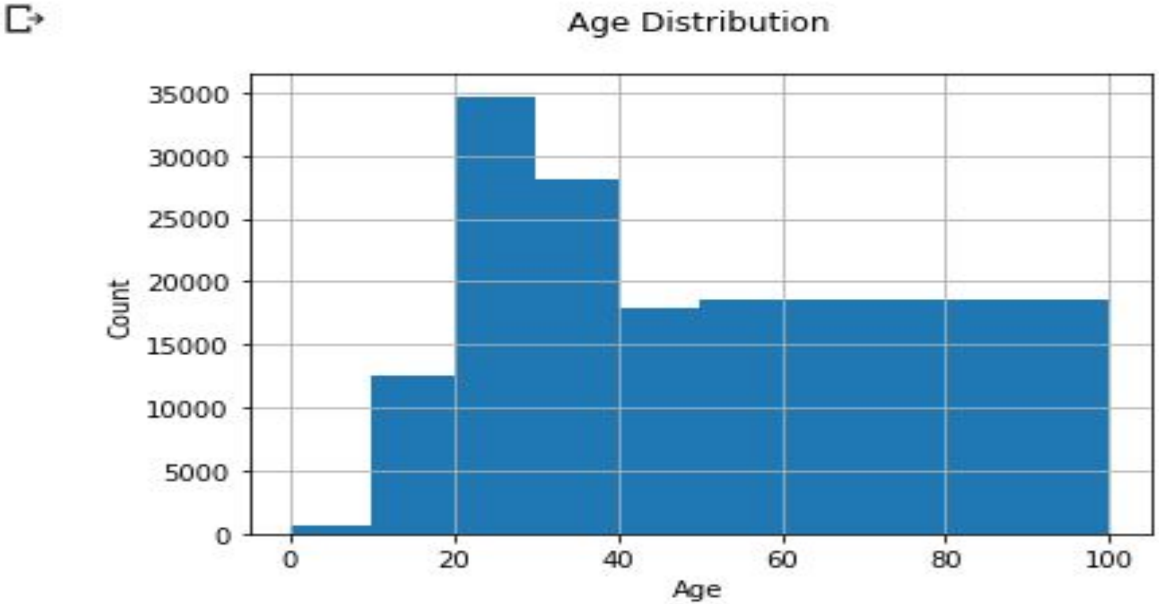
- Ballantine books published highest number of books in our given



EDA

Observations from User_data (Age)

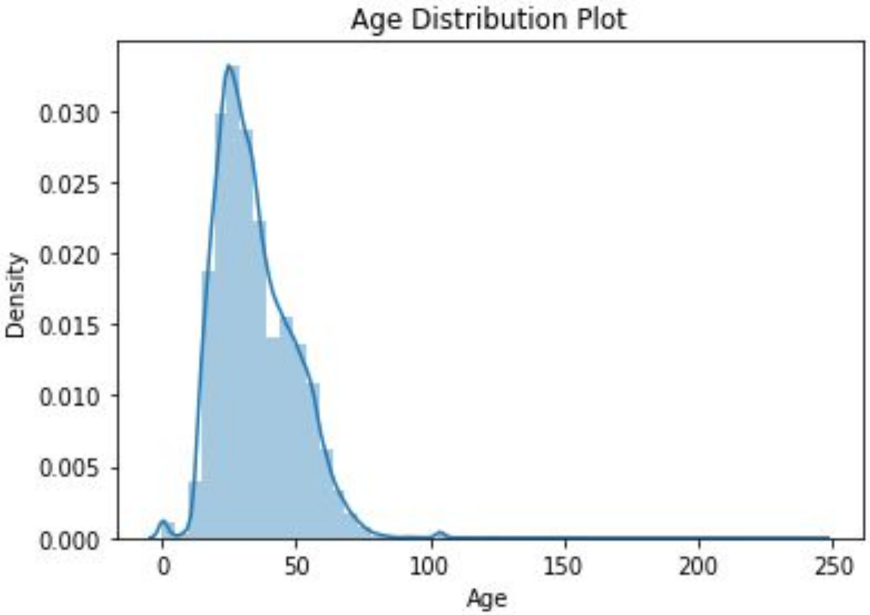
- The Age range distribution is right skewed
- Most active readers lie in age group 20- 40



EDA

Observations from User_data (Age)

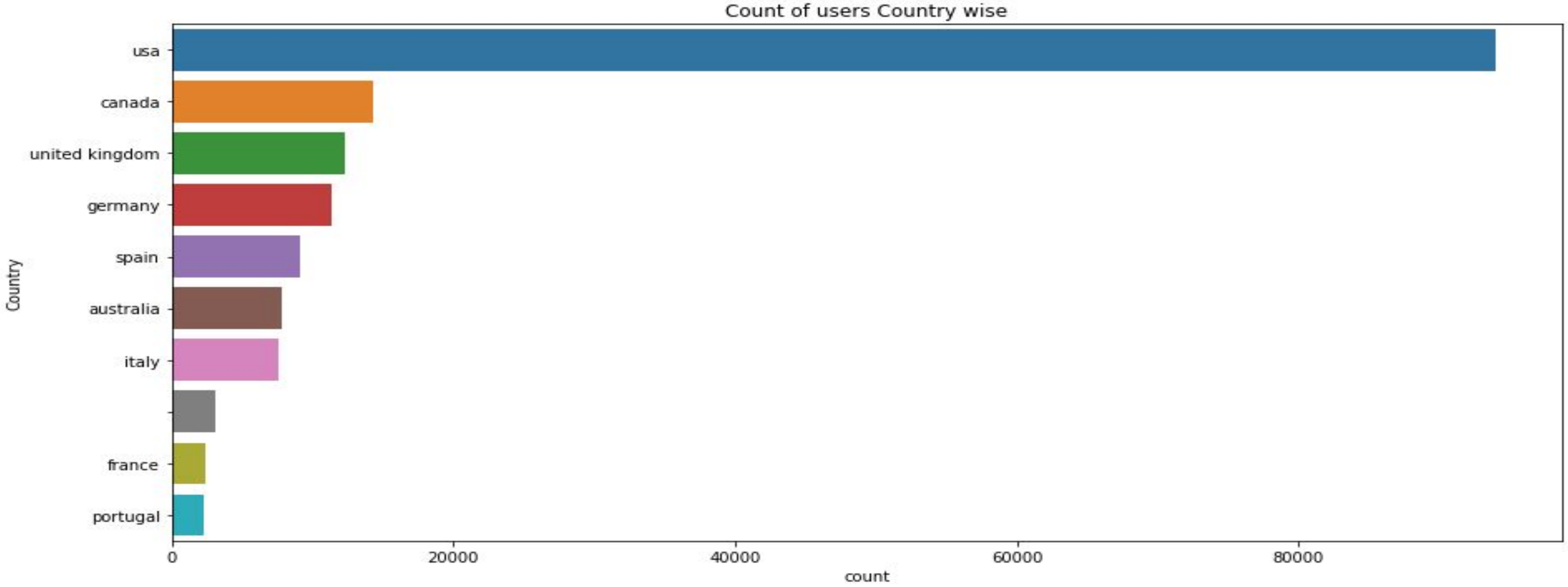
- The Age range given here is from 0 To 250.
- Outliers in the Age column.



EDA

Observations from User_data (Location)

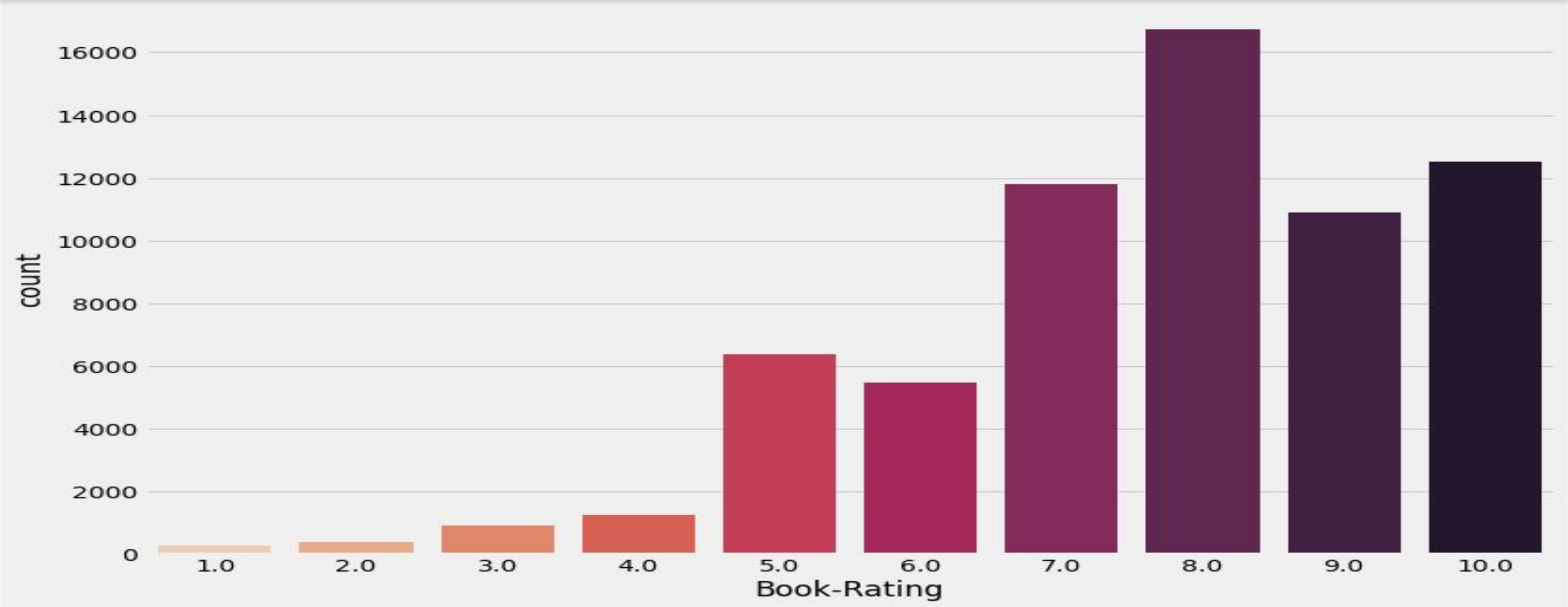
- Splitting Location column and analysing country.
- Most active readers are from USA.



EDA

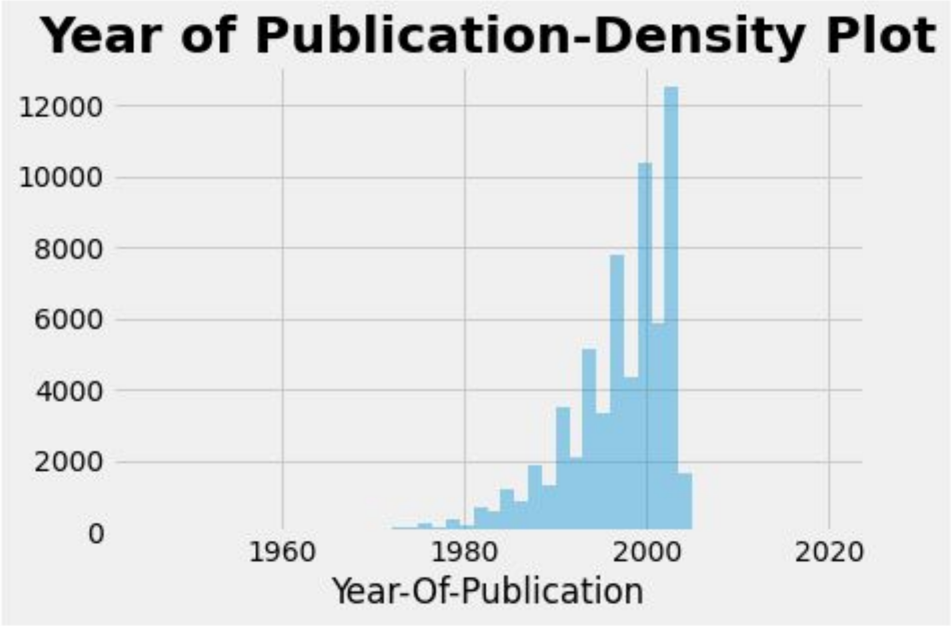
Observations from Rating_data (Book_Rating)

- Higher ratings are more common amongst users
- Rating 8 has been rated the highest number of times



EDA

- Most books got published in early 2000's



Preparing Data for Model

❑ Cleaning Year of Publication:

It was observed that there is noise in the Year of Publication features :-

- String Noise Values - such as 'DK Publishing Inc' and 'Gallimard.
- Integer Noise Values - Since this data was collected in August 2006, so any year value greater than 2006 is a noise value.
- Therefore, after cleaning the dataset based upon Year-Of-Publication Feature, we lost only a miniscule amount of 1.3% data.

❑ Selecting Books with Optimum Number of Ratings:

- Building a recommendation system requires a lot of data.
- Recommendations should be relevant, otherwise they can cause a nuisance to the customers.
- So, we have set a threshold number of ratings per book in order to get optimal recommendations for our users.

Preparing Data for Model

❑ Defining Optimum Reader:

- We can't take every user's rating at face value because if the user is a novice reader with only an experience of reading a couple of books, his/her ratings might not be much relevant for finding similarity among books.
- Therefore, as a general rule of thumb, we're choosing only those Users who have rated at least 10 Books for building the recommendation system.

Applying Model

1.)Popularity Based Recommendation

Book weighted average formula:

Weighted Rating(WR)$=[vR/(v+m)]+[mC/(v+m)]$

Where,

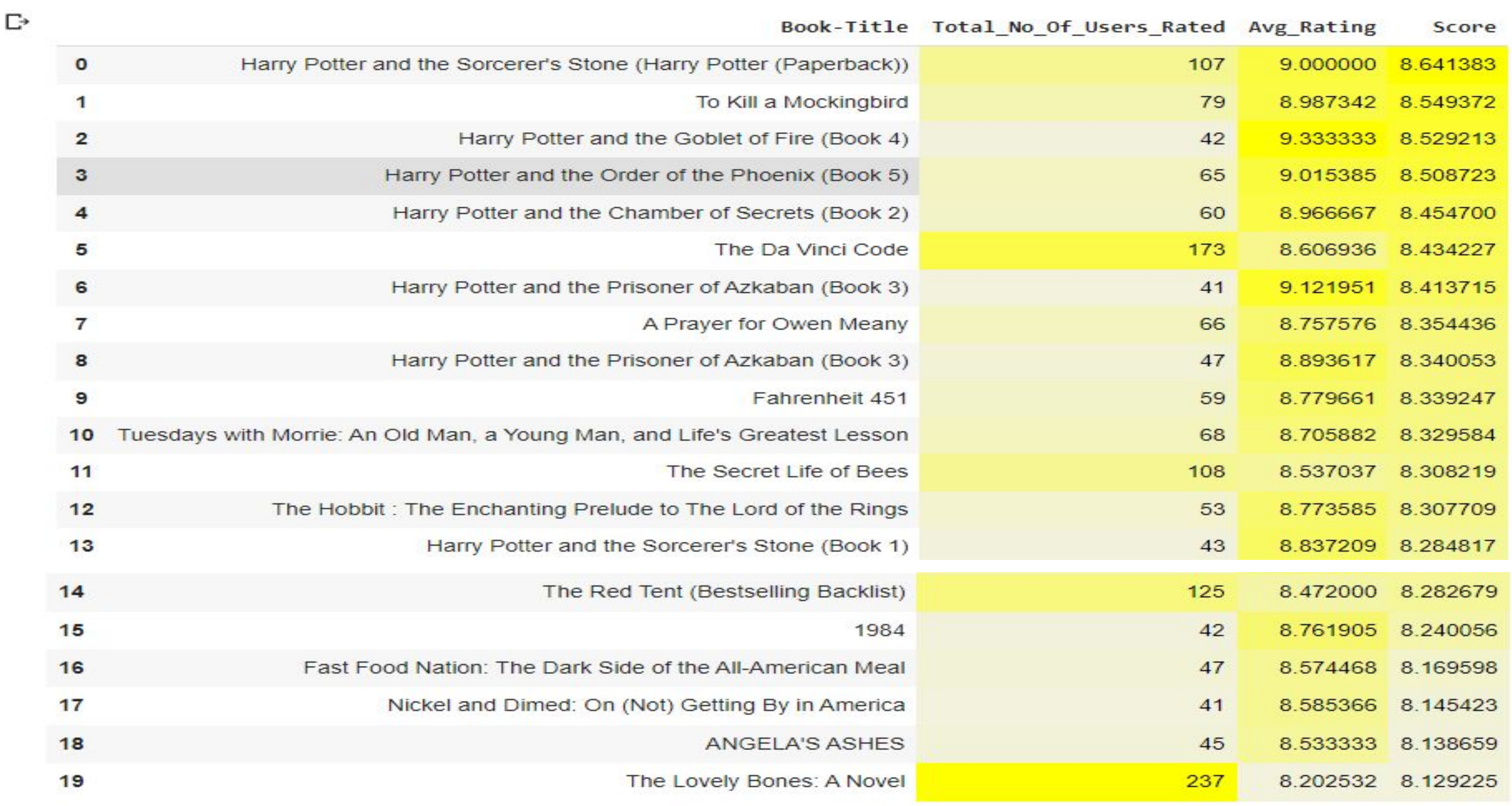
v is the number of votes for the books;

m is the minimum votes required to be listed in the chart;

R is the average rating of the book; and

C is the mean vote across the whole report.

Popularity Based Recommendation

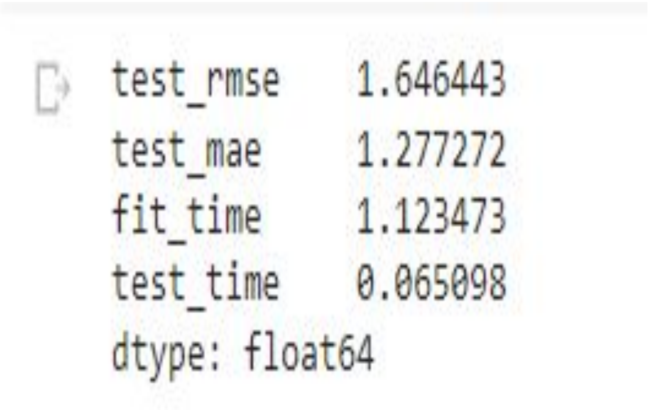


	Book-Title	Total_No_Of_Users_Rated	Avg_Rating	Score
0	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	107	9.000000	8.641383
1	To Kill a Mockingbird	79	8.987342	8.549372
2	Harry Potter and the Goblet of Fire (Book 4)	42	9.333333	8.529213
3	Harry Potter and the Order of the Phoenix (Book 5)	65	9.015385	8.508723
4	Harry Potter and the Chamber of Secrets (Book 2)	60	8.966667	8.454700
5	The Da Vinci Code	173	8.606936	8.434227
6	Harry Potter and the Prisoner of Azkaban (Book 3)	41	9.121951	8.413715
7	A Prayer for Owen Meany	66	8.757576	8.354436
8	Harry Potter and the Prisoner of Azkaban (Book 3)	47	8.893617	8.340053
9	Fahrenheit 451	59	8.779661	8.339247
10	Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	68	8.705882	8.329584
11	The Secret Life of Bees	108	8.537037	8.308219
12	The Hobbit : The Enchanting Prelude to The Lord of the Rings	53	8.773585	8.307709
13	Harry Potter and the Sorcerer's Stone (Book 1)	43	8.837209	8.284817
14	The Red Tent (Bestselling Backlist)	125	8.472000	8.282679
15	1984	42	8.761905	8.240056
16	Fast Food Nation: The Dark Side of the All-American Meal	47	8.574468	8.169598
17	Nickel and Dimed: On (Not) Getting By in America	41	8.585366	8.145423
18	ANGELA'S ASHES	45	8.533333	8.138659
19	The Lovely Bones: A Novel	237	8.202532	8.129225

Applying model

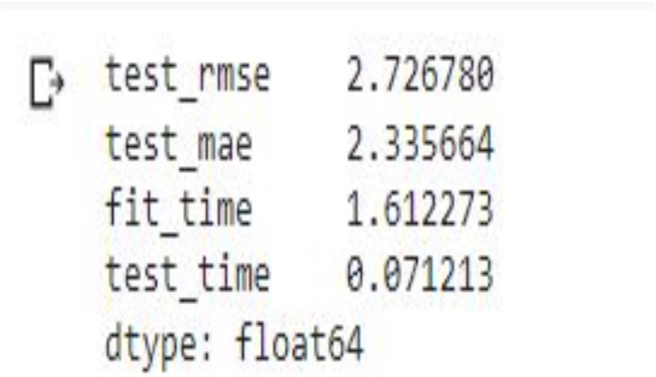
2.)Model based collaborative filtering

SVD



```
test_rmse     1.646443
test_mae      1.277272
fit_time      1.123473
test_time     0.065098
dtype: float64
```

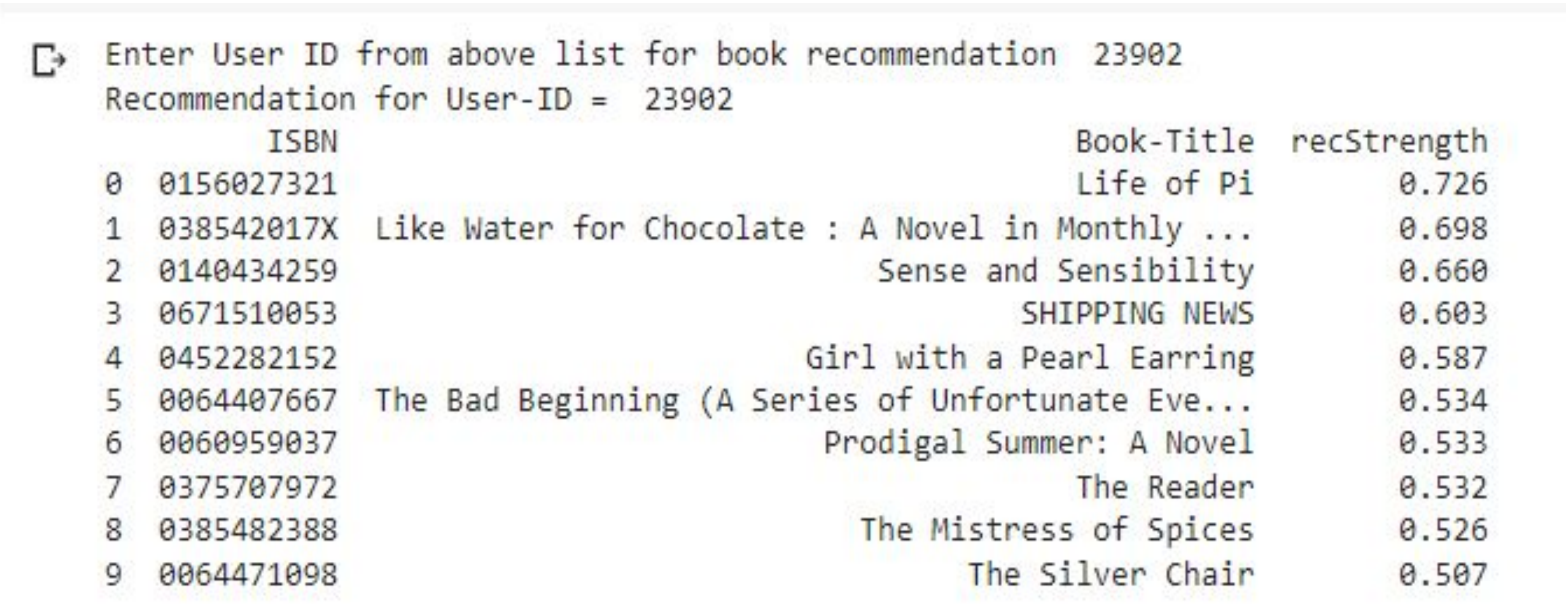
NMF



```
test_rmse     2.726780
test_mae      2.335664
fit_time      1.612273
test_time     0.071213
dtype: float64
```

Applying model

3) Collaborative Filtering based Recommendation System--(User-Item based)



```
Enter User ID from above list for book recommendation  23902
Recommendation for User-ID =  23902
         ISBN                                          Book-Title  recStrength
0  0156027321                                          Life of Pi        0.726
1  038542017X  Like Water for Chocolate : A Novel in Monthly ...        0.698
2  0140434259                               Sense and Sensibility        0.660
3  0671510053                                       SHIPPING NEWS        0.603
4  0452282152                           Girl with a Pearl Earring        0.587
5  0064407667  The Bad Beginning (A Series of Unfortunate Eve...        0.534
6  0060959037                            Prodigal Summer: A Novel        0.533
7  0375707972                                          The Reader        0.532
8  0385482388                              The Mistress of Spices        0.526
9  0064471098                                    The Silver Chair        0.507
```

Conclusion

- In EDA, the Top-10 most rated books were essentially novels. Books like The Lovely Bone, The da vinci code and The Secret Life of Bees were very well perceived.
- Majority of the readers were of the age bracket 20-30 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 Stephen King, Nora Roberts & Agatha Christie.
- For modelling, it was observed that for model based collaborative filtering SVD technique worked way better than NMF with lower Mean Absolute Error (MAE) .
- publisher with the most books was Ballantine books, pocket & Bentam book.
- most books got published in early 2000's.

Challenges

- Handling of sparsity was a major challenge as well since the user interactions were not present for the majority of the books.
- Understanding the metric for evaluation was a challenge as well.
- Since the data consisted of text data, data cleaning was a major challenge in features like Location etc..
- Decision making on missing value imputations and outlier treatment was quite challenging as well.
- Dealing and filtering data, to reach to the most recommendable users was an adventurous task to do .
- Exploring literature and resources to understand the problem and to find the solution was a little exhaustive.
- Deadlines felt a little strained. But it all worked out for the best.

Future Scope

- Given more information regarding the books dataset, namely features like Genre, Description etc, we could implement a content-filtering based recommendation system and compare the results with the existing collaborative-filtering based system.
- We would like to explore various clustering approaches for clustering the users based on Age, Location etc., and then implement voting algorithms to recommend items to the user depending on the cluster into which it belongs.

THANK
YOU!