# Capstone Project-

# BOOK RECOMMENDATION SYSTEM

Presented By :-
Aniket Satpute
Kaiwalya Zankar
(AlmaBetter Trainee)

# Content

- Introduction
- Scope of project
- Problem Statement
- Data Summary
- Process flow
- Library used
- Data preprocessing
- EDA
- Different Recommendation Model
- Challenges
- Conclusion
- Q & A

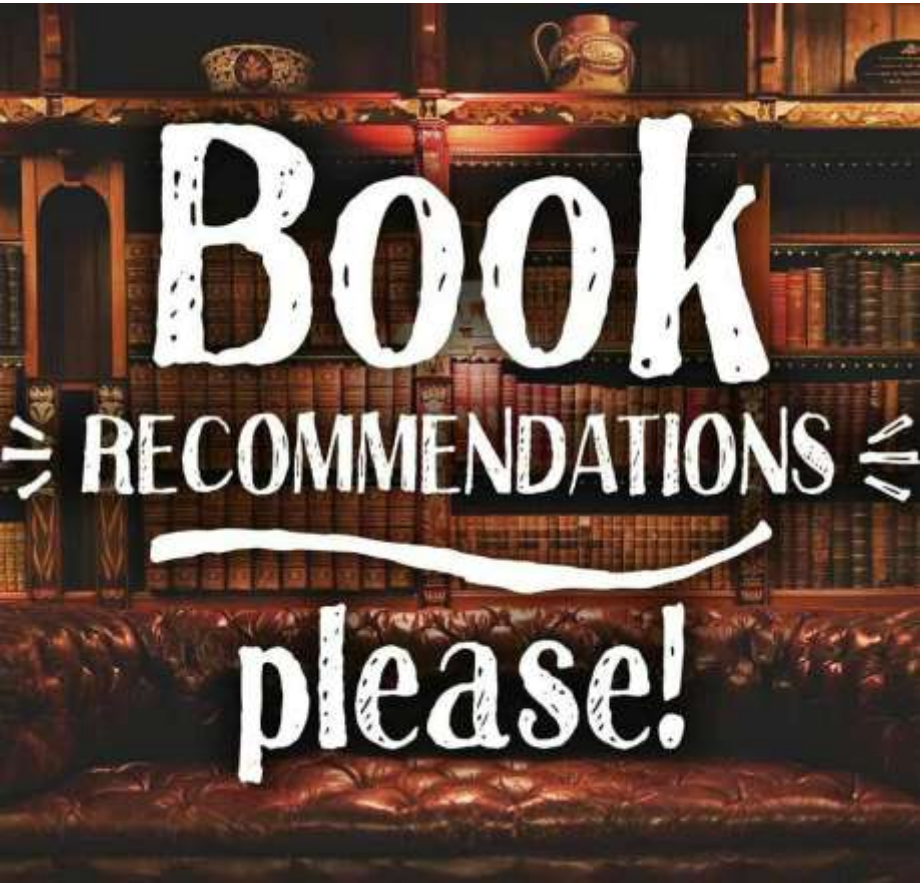
# Introduction

- **Book recommendation is created and deployed in this approach of work, which helps in recommending books. Recommendation achieved by the users feedbacks and rating, this is the online which analyze the ratings, comments and reviews of user, negative positive nature of comments using opinion mining.**
- **Whenever we search a book and we get lots of book having same name at that time recommendation model helps a lot.**
- **This is the trustworthy approach, which is used in this paper where selection is based on the services of cloud using collaborative filtering.**

# scope of project

- The main objective is to create a machine learning model to recommend relevant books to users based on popularity and user interests.
- In addition to the ML Model prediction, we also have taken into account the book recommendation for a totally new user.

# Problem Statement




During the last few decades, with the rise of Youtube, Amazon, Netflix, and many other such web services, recommender systems have become much more important in our lives in terms of providing highly personalized and relevant content.

**The main objective is to create a recommendation system to recommend relevant books to users based on popularity and user interests.**

# Data Summary

The dataset is comprised of three csv files:: User_df, Books_df, Ratings_df

**Users_dataset.**

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

Shape of Dataset - (278858, 3)

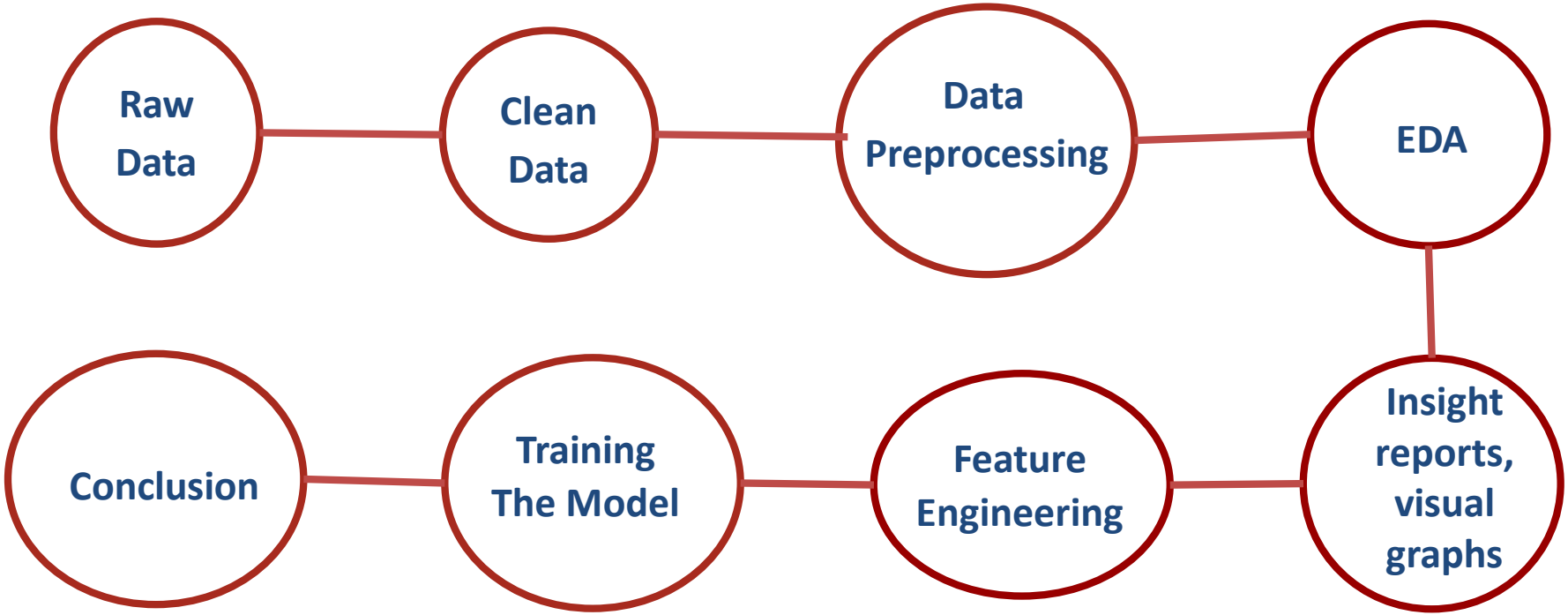
**Books_dataset.**

- 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)

# Process Flow-



# Libraries used-

1. numpy
2. Pandas
3. scipy
4. matplotlib.pyplot
5. seaborn
6. imblearn
7. Sklearn
8. statsmodel
9. Math
10. Xgboost
11. warnings

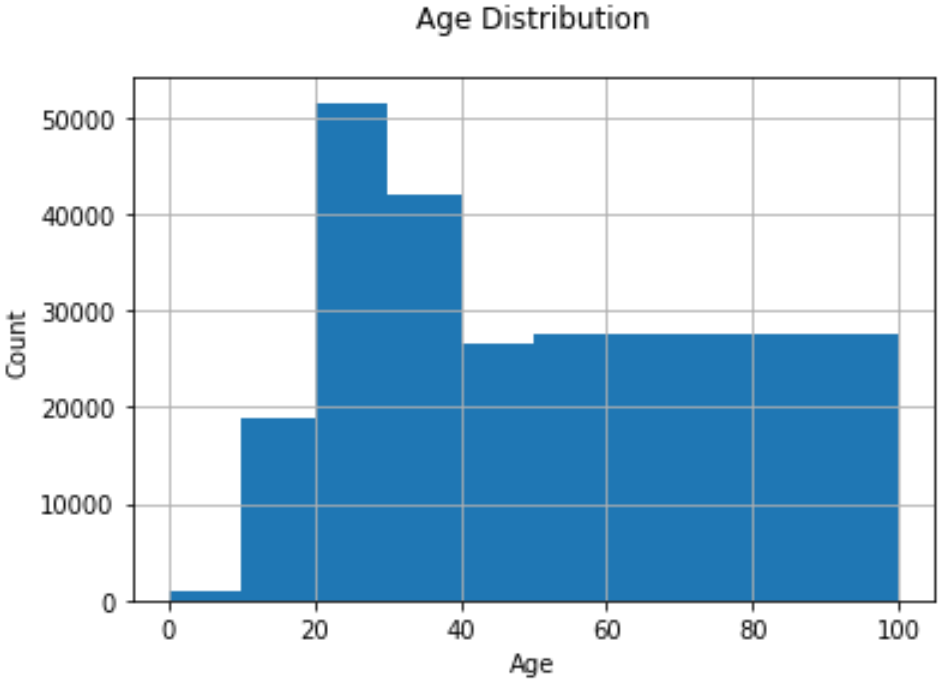
# Data Preprocessing

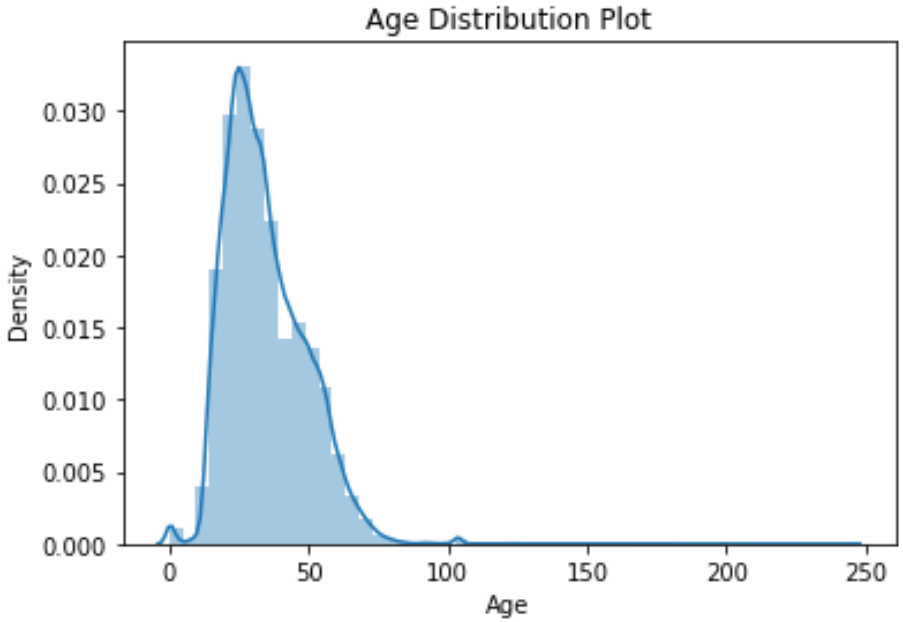
# 1.Users Dataset

| | 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 |

Age have around 39% missing values

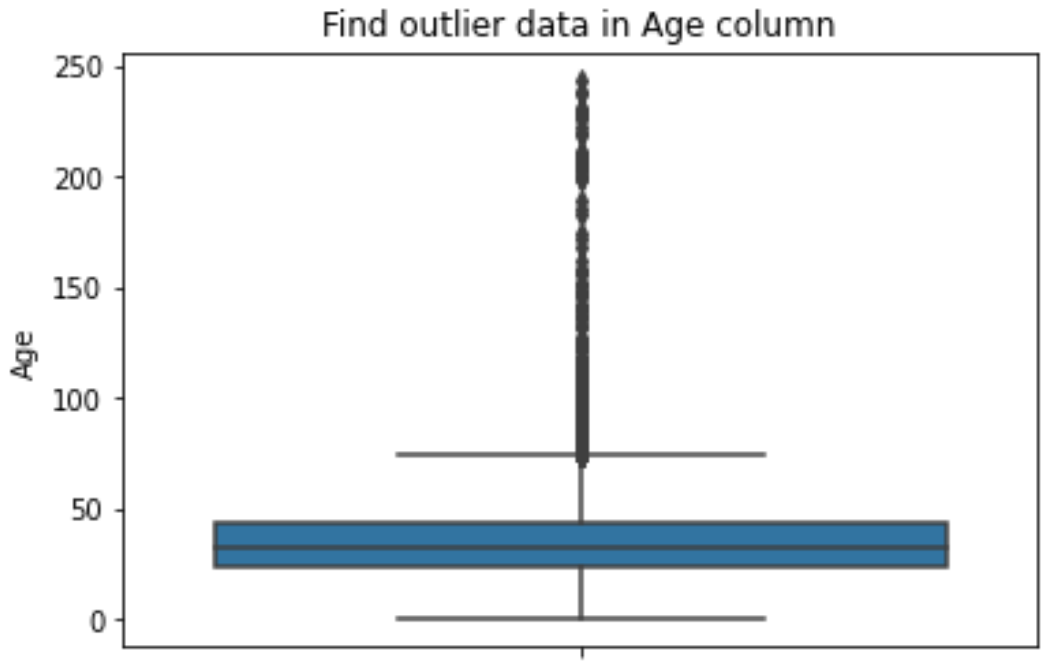
# Age Distribution



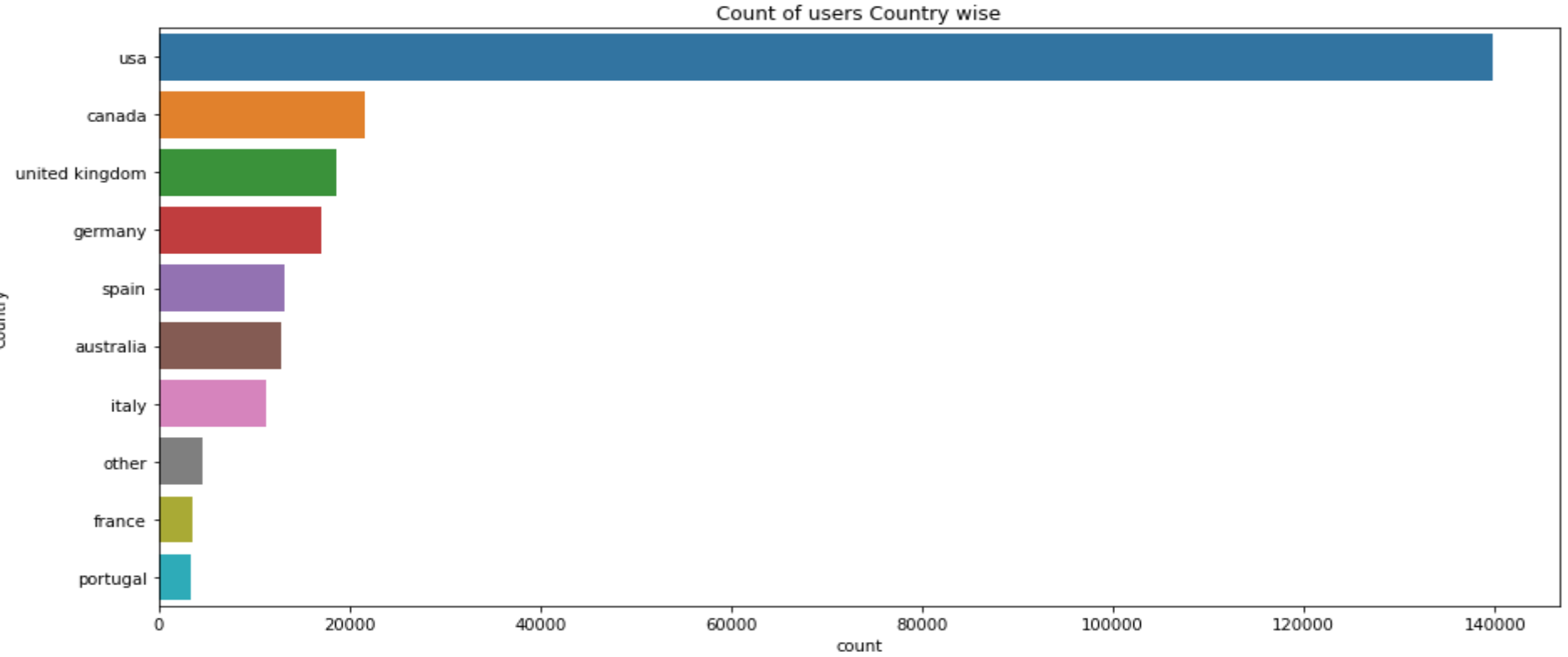

- The Age range distribution is right skewed
- Most active readers lie in age group 20-40

# Outliers in age




- Outliers in Age column
- Age has positive Skewness (right tail) so we can use median to fill Nan values,

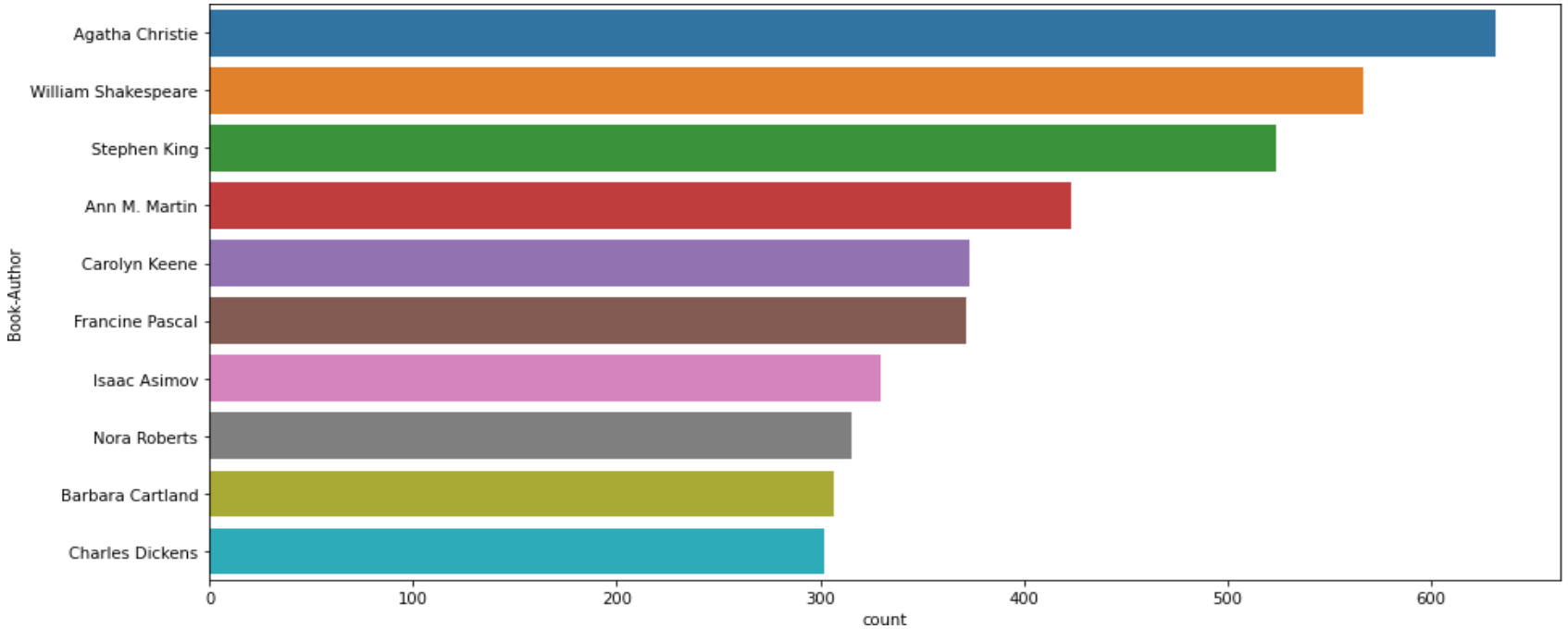
# Users as per location




Most number of users are from USA.

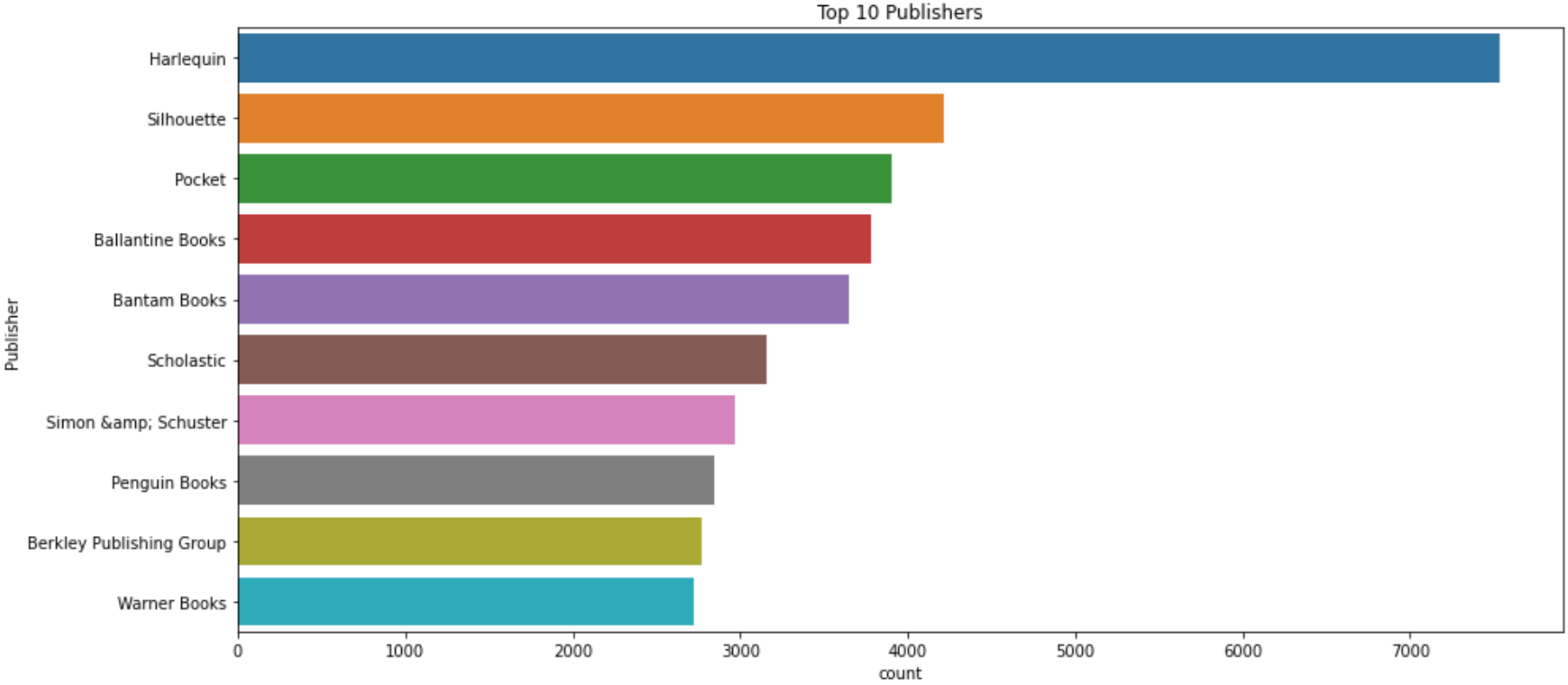
# 2. Books Data

## Top 10 authors with most books written



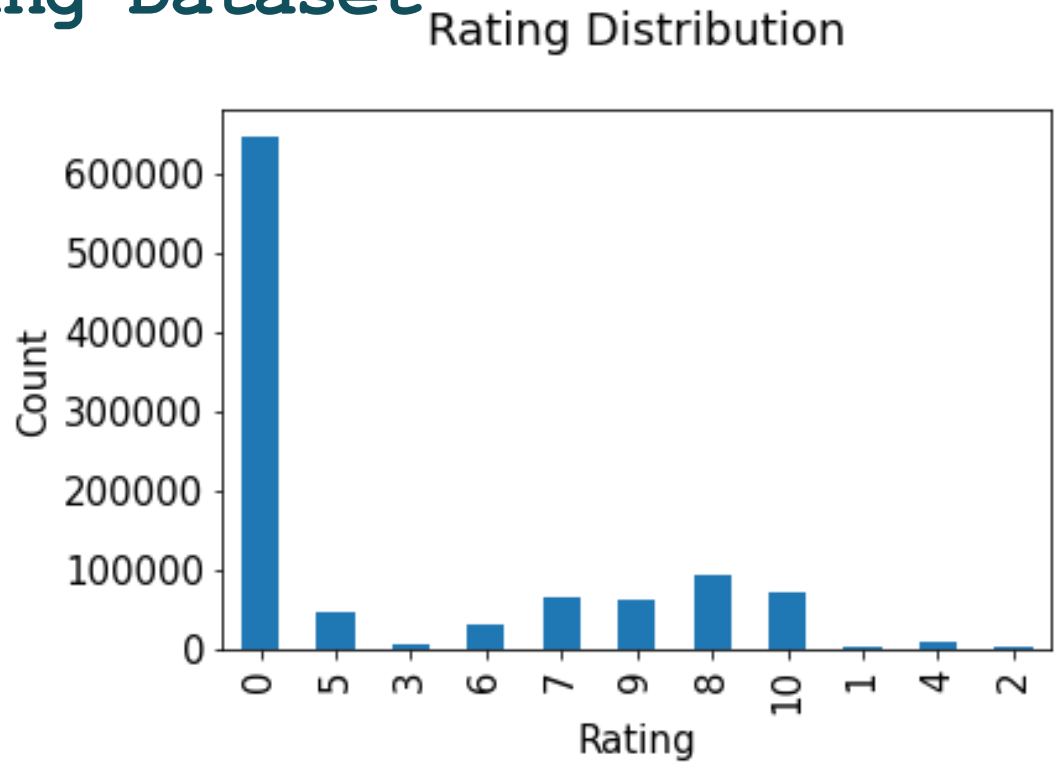

Agatha Christie wrote highest number of books in our given dataset

# Top 10 publisher with most books



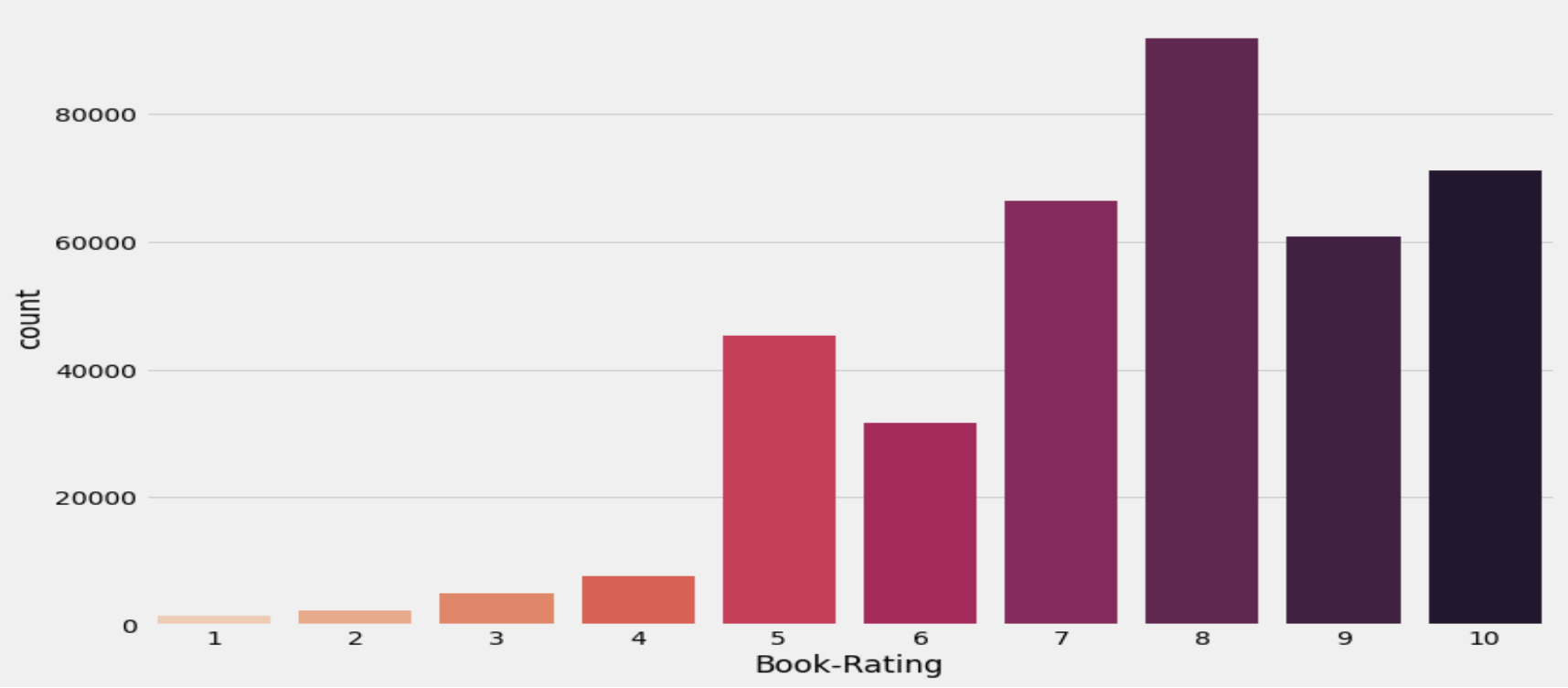

Harlequin published highest number of books in our given dataset

# 3. Rating Dataset




The ratings are very unevenly distributed, and the vast majority of ratings

# Book Rating




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

# Models used

# Different Models

## 1.)Popularity Based Recommendation

Book weighted average formula:

$$\textbf{Weighted Rating(WR)}=[vR/(v+m)]+[mC/(v+m)]$$

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.

| | 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 |

## 2.)Model based collaborative filtering

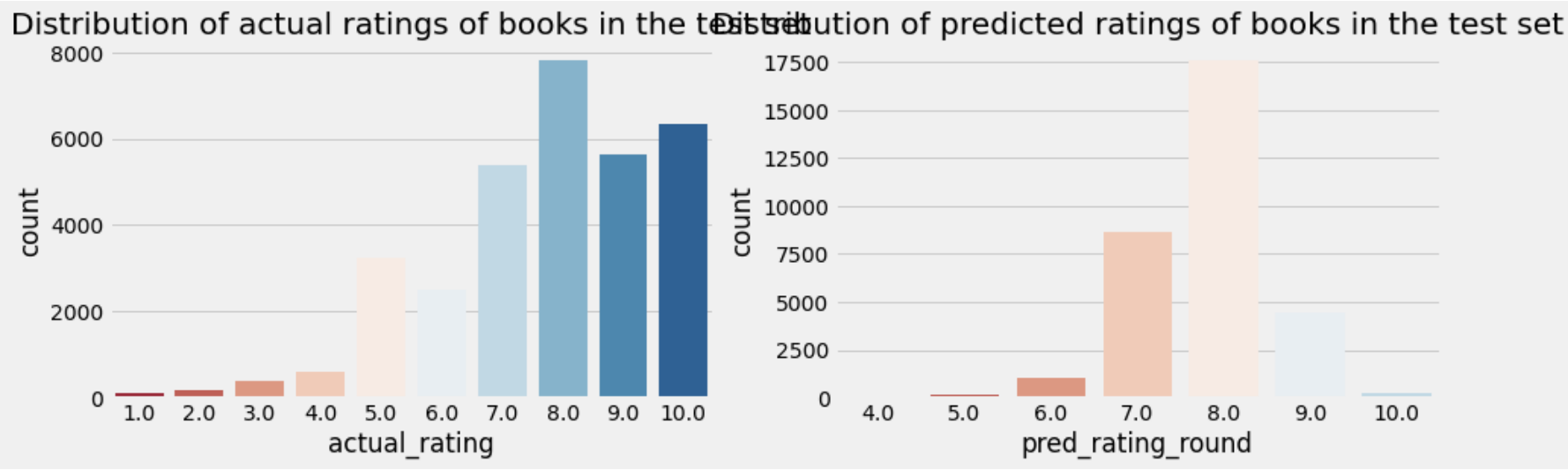
### SVD

| | |
|---|---|
| test_rmse | 1.601489 |
| test_mae | 1.239510 |
| fit_time | 12.735846 |
| test_time | 0.979896 |
| dtype: float64 | |

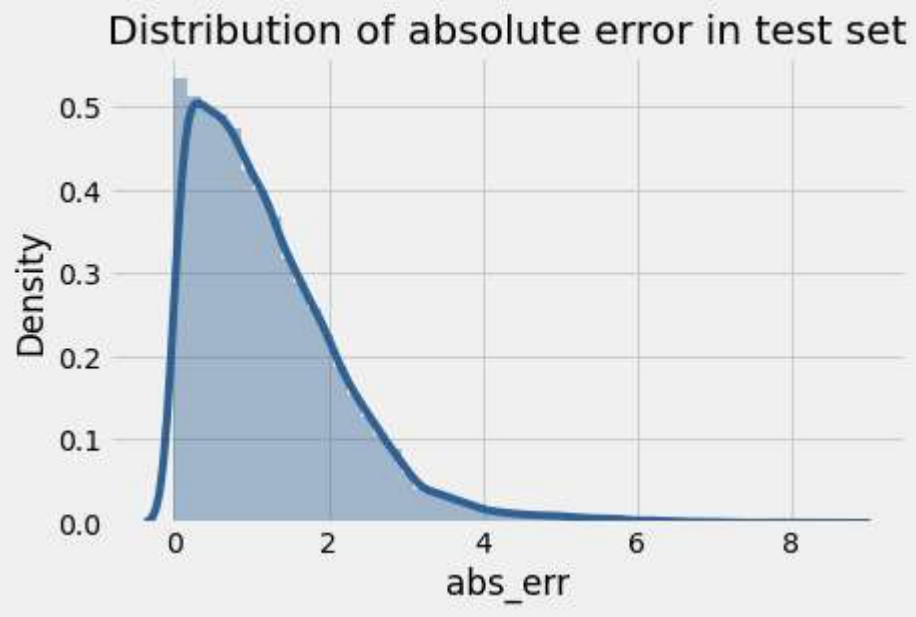
### NMF

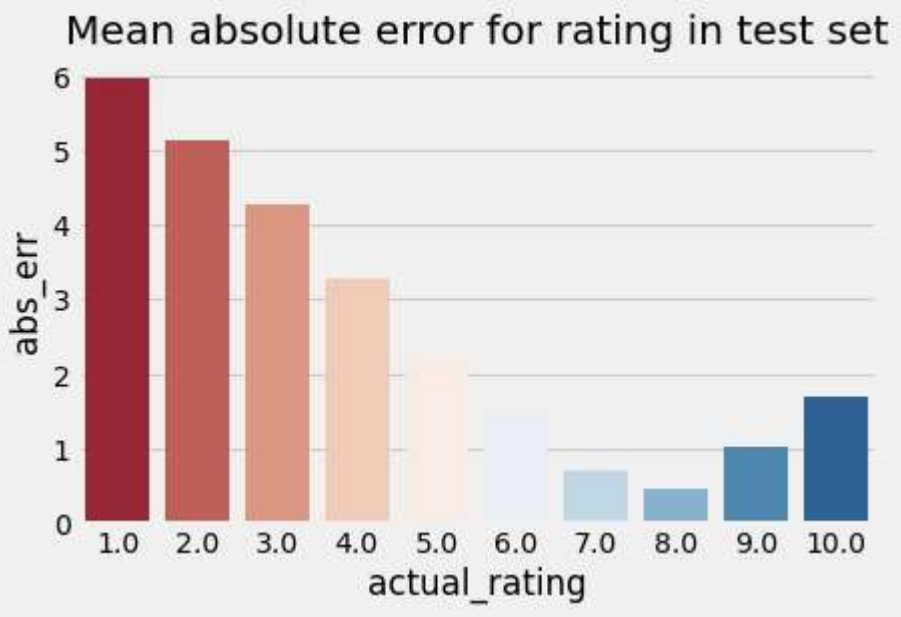
| | |
|---|---|
| test_rmse | 2.618516 |
| test_mae | 2.236896 |
| fit_time | 17.139331 |
| test_time | 0.736626 |
| dtype: float64 | |

# SVD Model Results




The distribution of absolute errors is right skewed, showing that the majority of errors is small.
The long tail indicates the several observation for which the absolute error was close to 10.

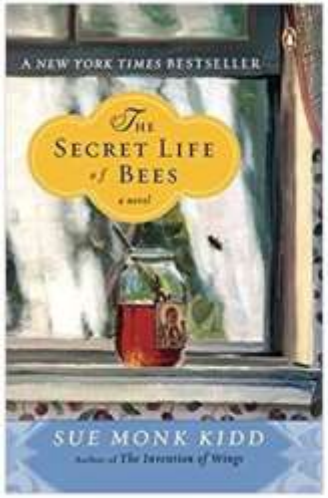

Distribution of absolute error in test set
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
abs_err


Mean absolute error for rating in test set
abs_err
0
1
2
3
4
5
6
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
actual_rating

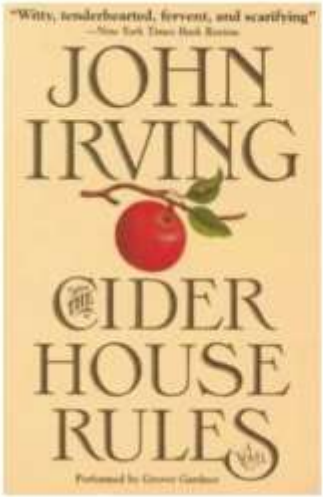
# Analysis of a particular user

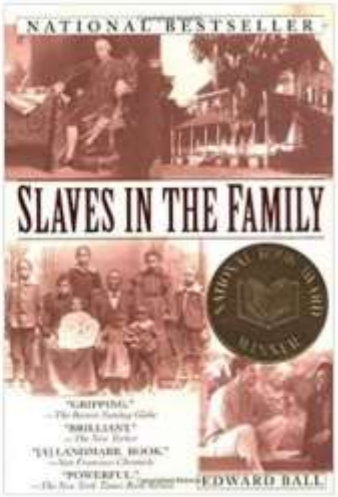
**User-ID - 193458**

**Test set: predicted top rated books**

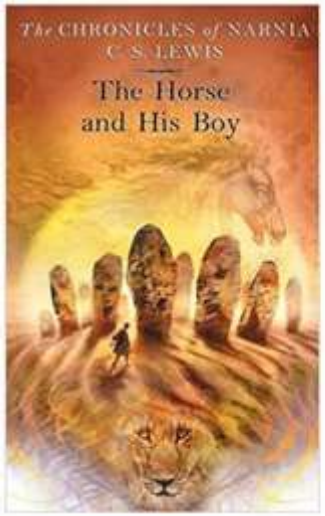




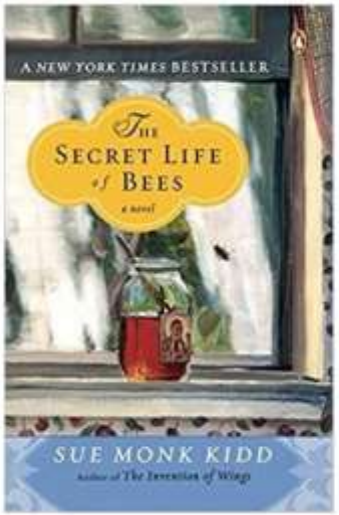



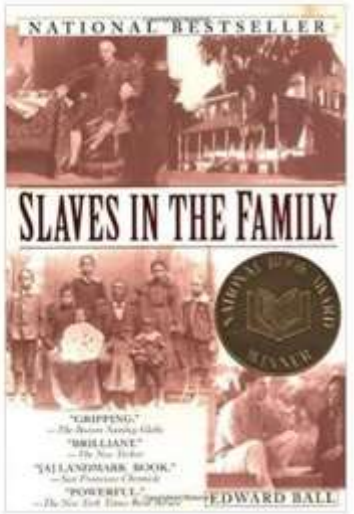



## Test set: actual top rated books





# Collaborative Filtering-(Item-Item based)

## 3.)Collaborative Filtering-(Item-Item based)

| | User_id | Isbn | Book_rating | Avg_rating | Total_No_Of_Users_Rated |
|---|---|---|---|---|---|
| **16** | 276747 | 0060517794 | 9 | 8 | 30 |
| **19** | 276747 | 0671537458 | 9 | 7.176471 | 17 |
| **20** | 276747 | 0679776818 | 8 | 7.476190 | 21 |
| **59** | 276772 | 0553572369 | 7 | 6.625000 | 8 |
| **61** | 276772 | 3499230933 | 10 | 7.166667 | 6 |

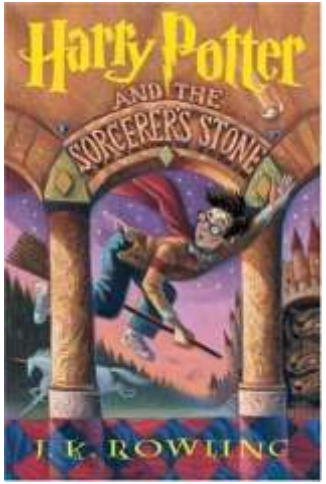
# Different Models

## SVD and Correlation

Recommendations for Harry Potter and the Sorcerer's Stone (Book 1)

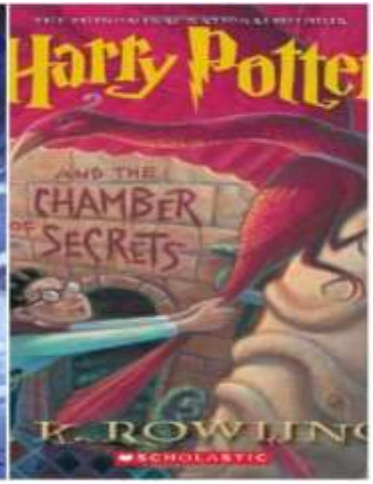
Input

Output









# 4.)Collaborative Filtering-(User-Item based)

| | User-ID | Isbn | Book_rating | Avg_rating | Total_No_Of_Users_Rated |
|---|---|---|---|---|---|
| **1** | 276726 | 0155061224 | 5 | 5.000 | 1 |
| **3** | 276729 | 052165615X | 3 | 3.000 | 1 |
| **4** | 276729 | 0521795028 | 6 | 6.000 | 1 |
| **8** | 276744 | 038550120X | 7 | 7.580 | 81 |
| **16** | 276747 | 0060517794 | 9 | 8.000 | 30 |

# Model Results

| | hits@5_count | hits@10_count | interacted_count | recall@5 | recall@10 | User-ID |
|---|---|---|---|---|---|---|
| **10** | 256 | 331 | 1389 | 0.184 | 0.238 | 11676 |
| **31** | 184 | 243 | 1138 | 0.162 | 0.214 | 98391 |
| **45** | 23 | 29 | 380 | 0.061 | 0.076 | 189835 |
| **30** | 81 | 104 | 369 | 0.220 | 0.282 | 153662 |
| **70** | 27 | 36 | 236 | 0.114 | 0.153 | 23902 |
| **7** | 29 | 47 | 204 | 0.142 | 0.230 | 235105 |
| **47** | 25 | 30 | 203 | 0.123 | 0.148 | 76499 |
| **50** | 22 | 37 | 193 | 0.114 | 0.192 | 171118 |
| **42** | 59 | 71 | 192 | 0.307 | 0.370 | 16795 |
| **43** | 21 | 30 | 188 | 0.112 | 0.160 | 248718 |

# 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**.
•For modelling, it was observed that for **model based** collaborative filtering SVD technique worked way better than NMF with lower Mean Absolute Error (MAE)
.
•Amongst the memory based approach, **item-item CF performed better** than **user-user CF** because of lower computation

# 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.

# Q & A