Capstone Project

BOOK RECOMMENDATION SYSTEM

Sumaya Bai A R

Content

- Problem statement
- Data Summary
- Analysis of different datasets
- Data Cleaning
- Outlier treatment
- Imputing missing values
- Different Recommendation Model
- Challenges
- Conclusion
- Future Scope

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.

Recommendation system really is a value add thing to the E-commerce system, which helps both the buyer and seller. It makes it easier for buyers to buy things as per their taste and needs.

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)

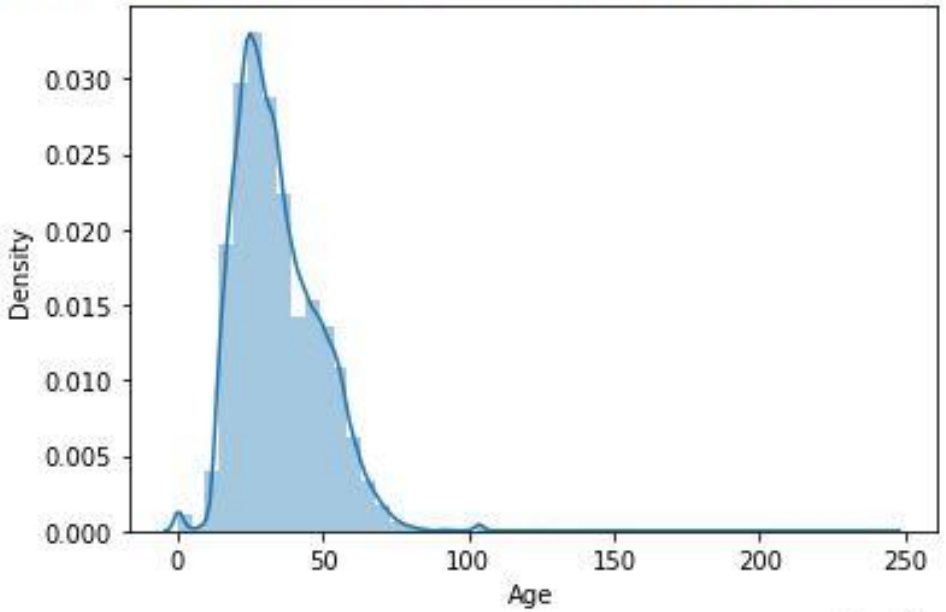

Exploratory
Data Analysis

Analysis on Users_df (Age)

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

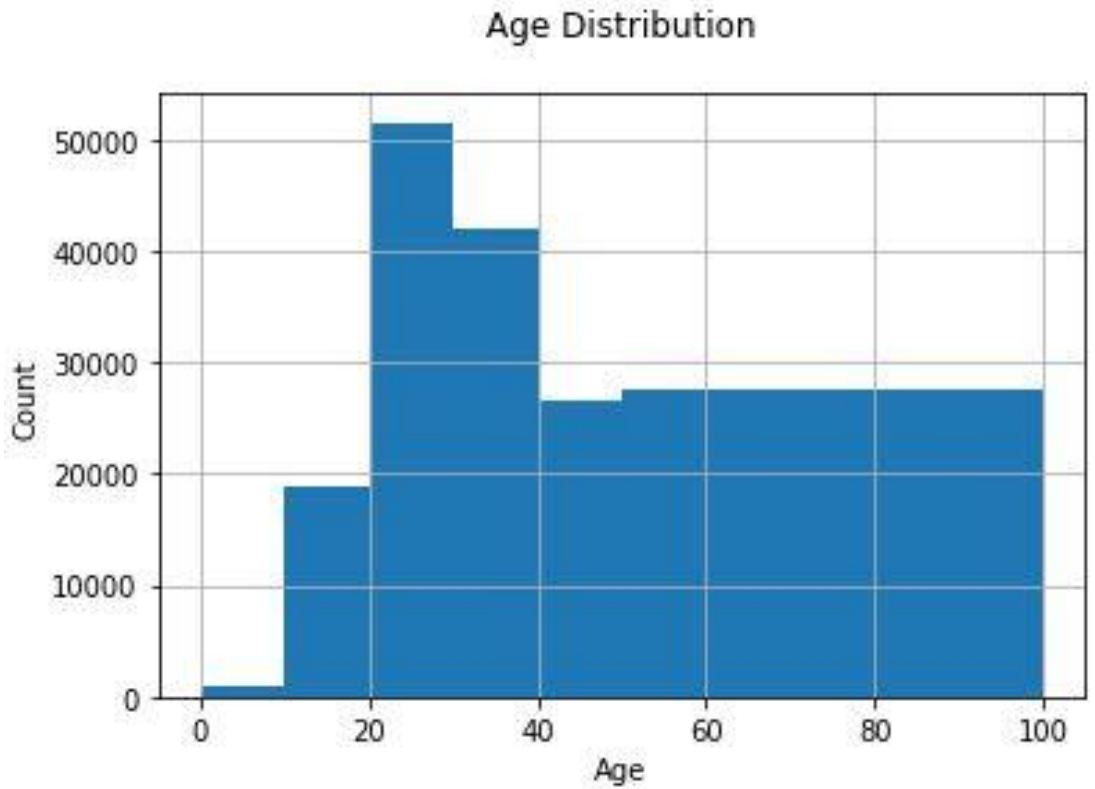
```
1 sns.distplot(users.Age)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5a11ac00d0>
```



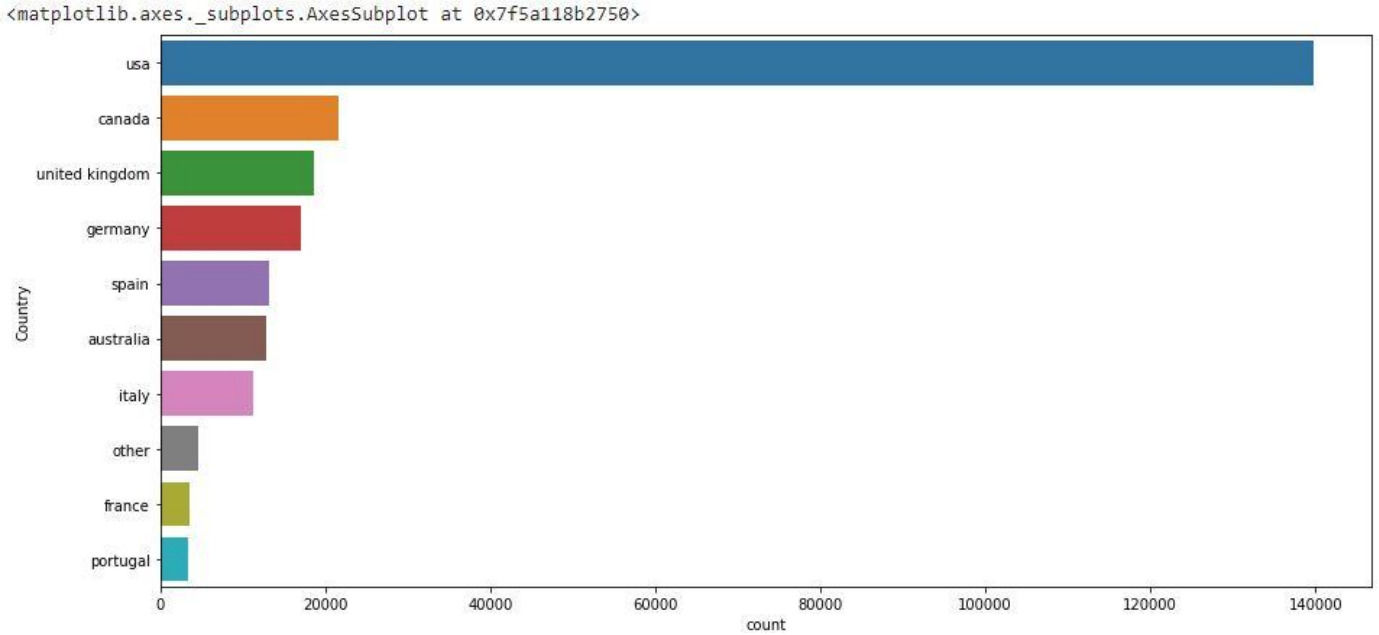
Analysis on Users_df (Age)

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



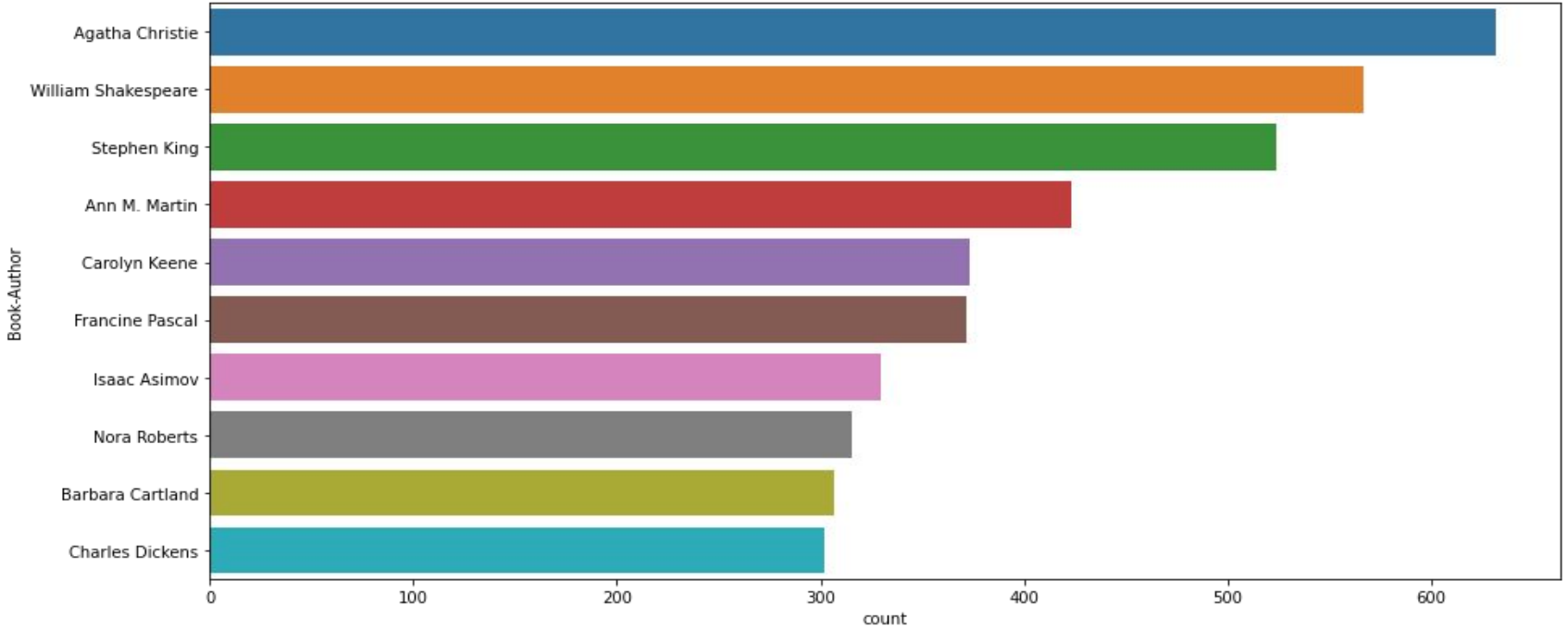
Analysis on Users_df (Location)

- Changing Location based on Countries
- Most active readers are from USA.



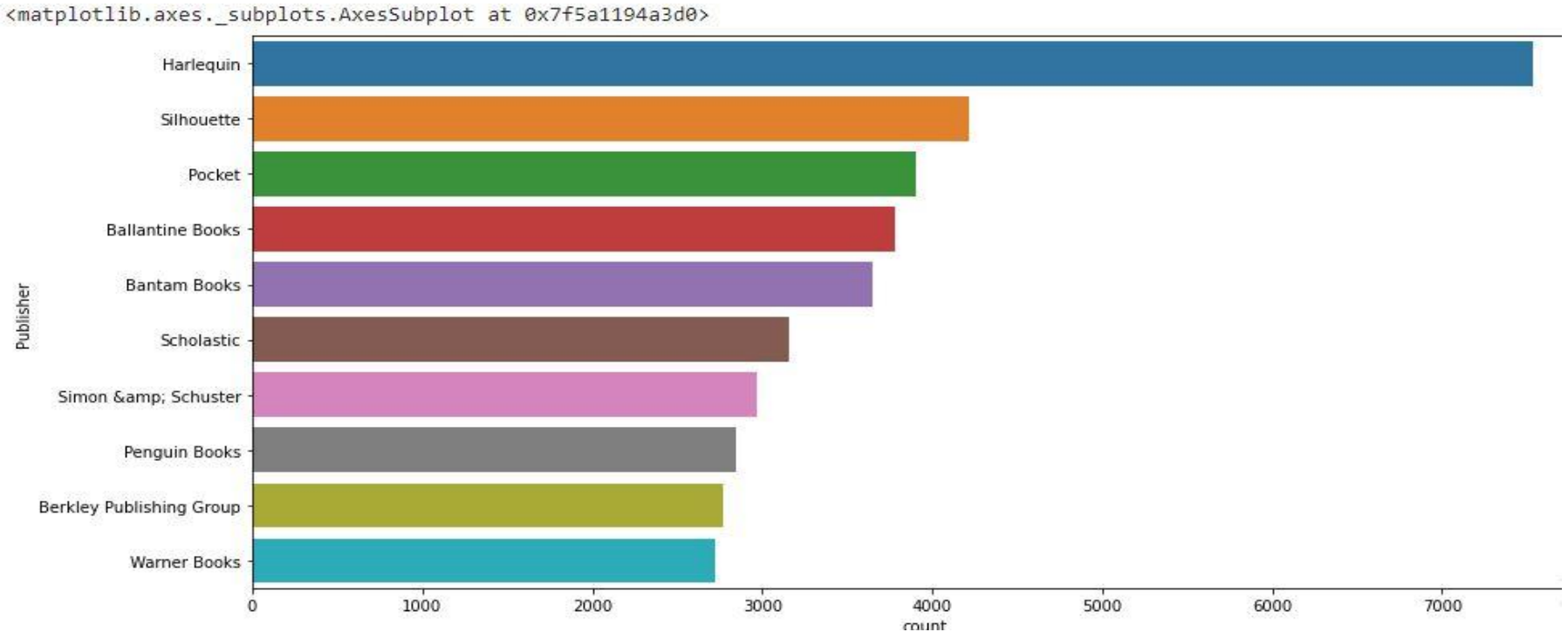
Analysis on Books_df (Authors)

Agatha Christie wrote highest number of books in our given dataset



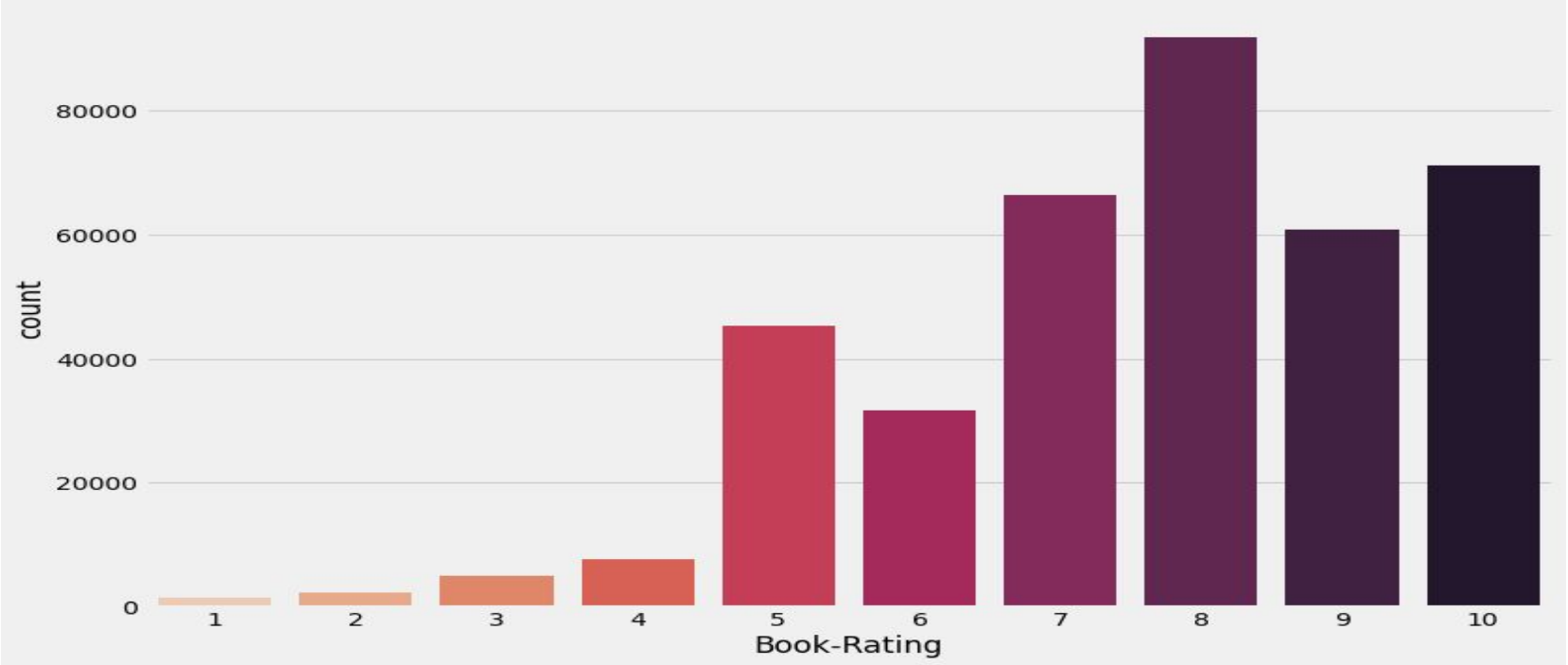
Analysis on Books_df (Publishers)

Harlequin published highest number of books in our given dataset



Analysis on Ratings_df (Book_Rating)

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



Data Preprocessing

1. Treating MIssing Values

Age column has 40% missing values

	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 Cleaning

1. Null Value Imputation:

```
books_df.isnull().sum()
```

```
ISBN                 0
Book-Title           0
Book-Author          1
Year-Of-Publication  0
Publisher            2
Image-URL-S          0
Image-URL-M          0
Image-URL-L          3
dtype: int64
```

Replacing strings by int values

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

Modelling

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 recommended books

Book-Title	Book-Author	mean	count	weighted rating	Year-Of-Publication
Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling	6.541237	194	5.985285	2000.0
Harry Potter and the Chamber of Secrets (Book 2)	J. K. Rowling	6.611765	170	5.978717	1999.0
Free	Paul Vincent	7.962963	54	5.973507	2003.0
Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	6.467005	197	5.929681	1999.0
Harry Potter and the Sorcerer's Stone (Book 1)	J. K. Rowling	6.363095	168	5.767724	1998.0
Harry Potter and the Order of the Phoenix (Boo...	J. K. Rowling	5.571856	334	5.320583	2003.0
The Fellowship of the Ring (The Lord of the Ri...	J. R. R. Tolkien	6.206349	63	5.036522	1999.0
Griffin & Sabine: An Extraordinary Corresp...	Nick Bantock	6.041667	72	5.024219	1991.0
Falling Up	Shel Silverstein	6.921053	38	5.008320	1996.0
The Stand (The Complete and Uncut Edition)	Stephen King	6.175439	57	4.942104	1990.0
Ender's Game (Ender Wiggins Saga (Paperback))	Orson Scott Card	5.302564	195	4.942059	1994.0
The Little Prince	Antoine de Saint-ExupÃ©ry	5.797468	79	4.918397	1968.0
The Secret Life of Bees	Sue Monk Kidd	5.500000	96	4.815270	2002.0
Harry Potter and the Sorcerer's Stone (Harry P...	J. K. Rowling	4.900175	571	4.786846	1999.0
The Hobbit : The Enchanting Prelude to The Lor...	J.R.R. TOLKIEN	5.007117	281	4.777967	1986.0
To Kill a Mockingbird	Harper Lee	4.920308	389	4.756743	1988.0
The Two Towers (The Lord of the Rings, Part 2)	J. R. R. Tolkien	6.230769	39	4.674876	1999.0
The Horse and His Boy	C. S. Lewis	6.216216	37	4.624872	1994.0
The Perks of Being a Wallflower	Stephen Chbosky	5.194175	103	4.623134	1999.0

Collaborative filtering Recommendation System

K- Nearest Neighbour

```
Recommendations for Blood Shot (V.I. Warshawski Novels (Paperback)):

1: Burn Marks (V.I. Warshawski Novels (Paperback)), with distance of 0.6264540101376146:
2: Sanctuary (Peter Decker &amp;amp; Rina Lazarus Novels (Paperback)), with distance of 0.6740584459682026:
3: Praying for Sleep, with distance of 0.6908030539491267:
4: Speaking in Tongues, with distance of 0.6947150289000884:
5: The Stone Monkey (Lincoln Rhyme Novels (Paperback)), with distance of 0.7072004828481997:
```

Singular Value Decomposition

	user_id	isbn	actual_rating	pred_rating	impossible	pred_rating_round	abs_err
15594	62862	0385335482	8.0	7.978811	False	8.0	0.021189
30626	193938	0385497288	8.0	7.882566	False	8.0	0.117434
27451	234401	0812540026	8.0	7.316338	False	7.0	0.683662
14130	89602	0060987529	8.0	6.649098	False	7.0	1.350902
18074	86189	0312186886	10.0	7.303280	False	7.0	2.696720


Evaluation

Evaluation

In Recommender Systems, there are a set metrics commonly used for evaluation. We choose to work with **Top-N accuracy metrics**, which evaluates the accuracy of the top recommendations provided to a user, comparing to the items the user has actually interacted in test set.

This evaluation method works as follows:

- For each user
  - For each item the user has interacted in test set
    - Sample 100 other items the user has never interacted.
    - Ask the recommender model to produce a ranked list of recommended items, from a set composed of one interacted item and the 100 non-interacted items
    - Compute the Top-N accuracy metrics for this user and interacted item from the recommendations ranked list
- Aggregate the global Top-N accuracy metrics

For a random set of users, the recall strength associated with its results is mentioned below.

```
Enter User ID from above list for book recommendation:  166596
Recommendation for User-ID =  166596
         ISBN                                               Book-Title  recStrength
0  0316666343                             The Lovely Bones: A Novel         0.39
1  0312195516                   The Red Tent (Bestselling Backlist)         0.21
2  0385504209                                    The Da Vinci Code         0.20
3  0142001740                                The Secret Life of Bees         0.16
4  0060928336    Divine Secrets of the Ya-Ya Sisterhood: A Novel         0.15
5  0446310786                                 To Kill a Mockingbird         0.15
6  0743418174                                         Good in Bed         0.14
7  0375727345                               House of Sand and Fog         0.14
8  0316769487                              The Catcher in the Rye         0.14
9  0446672211  Where the Heart Is (Oprah's Book Club (Paperba...         0.13
]
```

Conclusion

- 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 better.

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