Capstone Project - 4

Book Recommendation System

1. Subodh Shankar Dooganavar
2. Kasmin Talukdar

Introduction

- During a previous couple of decades, with the upward push of YouTube, Amazon, Netflix, and many different internet services, recommender systems have taken a very important place in our lives.

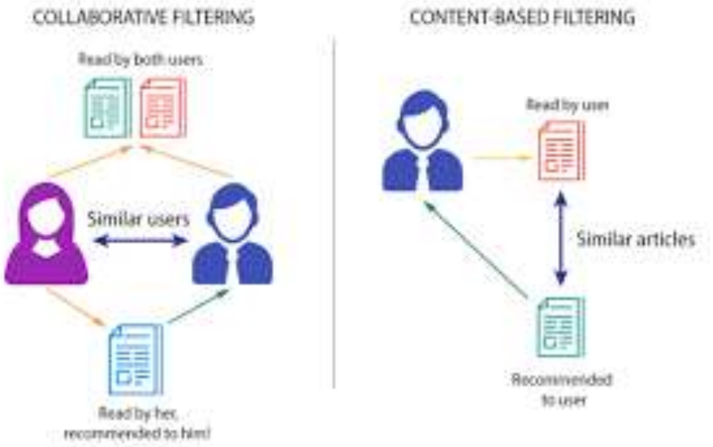
- From e-commerce (propose to consumers articles that might hobby them) to online advertisement (propose to customers the proper contents, matching their preferences), recommender systems are these days unavoidable in our everyday online journeys.

- In a totally trendy way, recommender systems are algorithms aimed toward suggesting applicable objects to users (items being films to watch, textual content to read, merchandise to buy, or whatever else relying on industries).

Problem Statement

- 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.

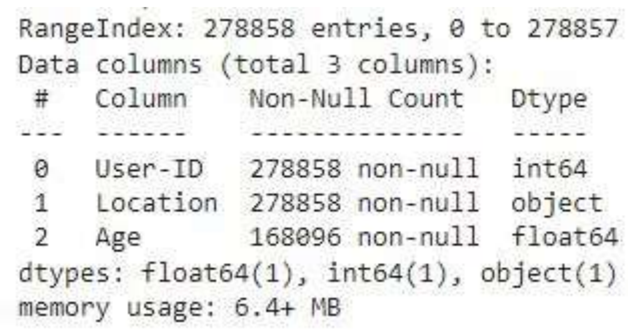
- The main objective is to create a machine learning model to recommend relevant books to users based on popularity and user interests.



Data Summary

Details of dataset – Users.csv

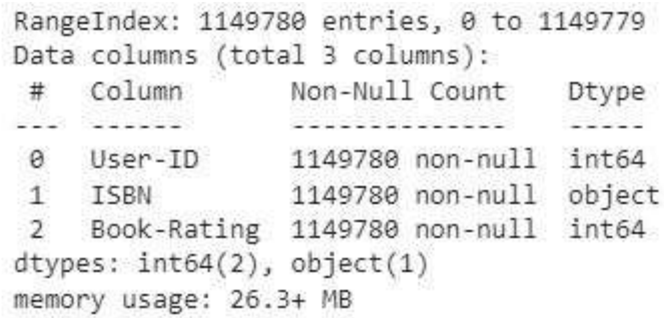
- Number of rows – 278857
- Number of columns – 3
- Datatypes - int64,object,float64



```
RangeIndex: 278858 entries, 0 to 278857
Data columns (total 3 columns):
 #   Column    Non-Null Count   Dtype  
---  ------    --------------   -----  
 0   User-ID   278858 non-null  int64  
 1   Location  278858 non-null  object 
 2   Age       168096 non-null  float64
dtypes: float64(1), int64(1), object(1)
memory usage: 6.4+ MB
```

Details of dataset – Ratings.csv

- Number of rows – 1140779
- Number of columns – 3
- Datatypes - int64 and object

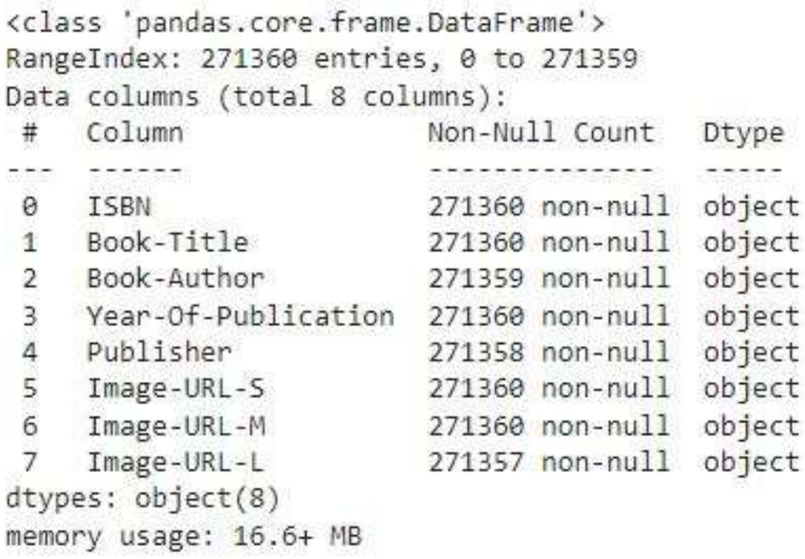


```
RangeIndex: 1149780 entries, 0 to 1149779
Data columns (total 3 columns):
 #   Column       Non-Null Count    Dtype 
---  ------       --------------    ----- 
 0   User-ID      1149780 non-null  int64 
 1   ISBN         1149780 non-null  object
 2   Book-Rating  1149780 non-null  int64 
dtypes: int64(2), object(1)
memory usage: 26.3+ MB
```

Data Summary(Cont.)

Details of dataset – Books.csv

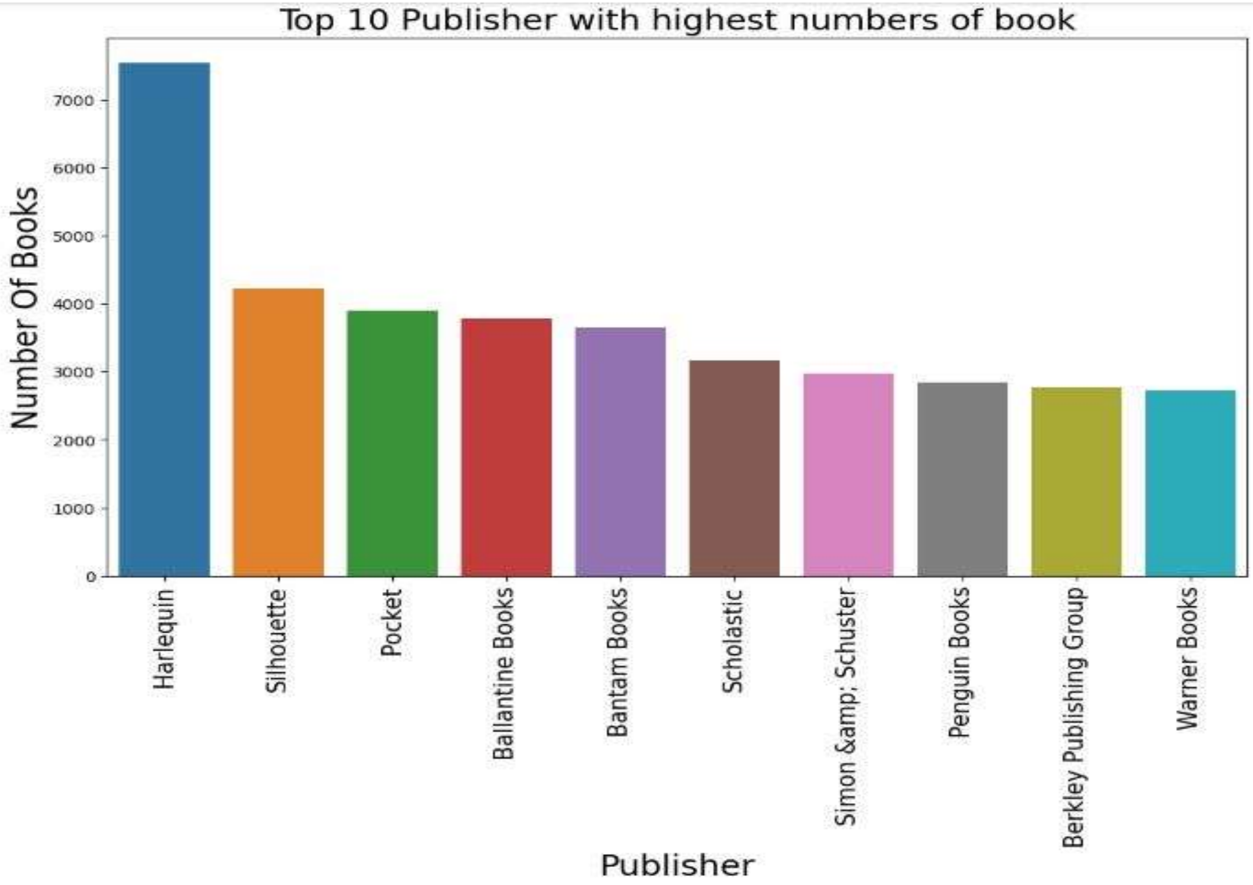
- Number of rows – 271360
- Number of columns – 8
- Datatypes - object



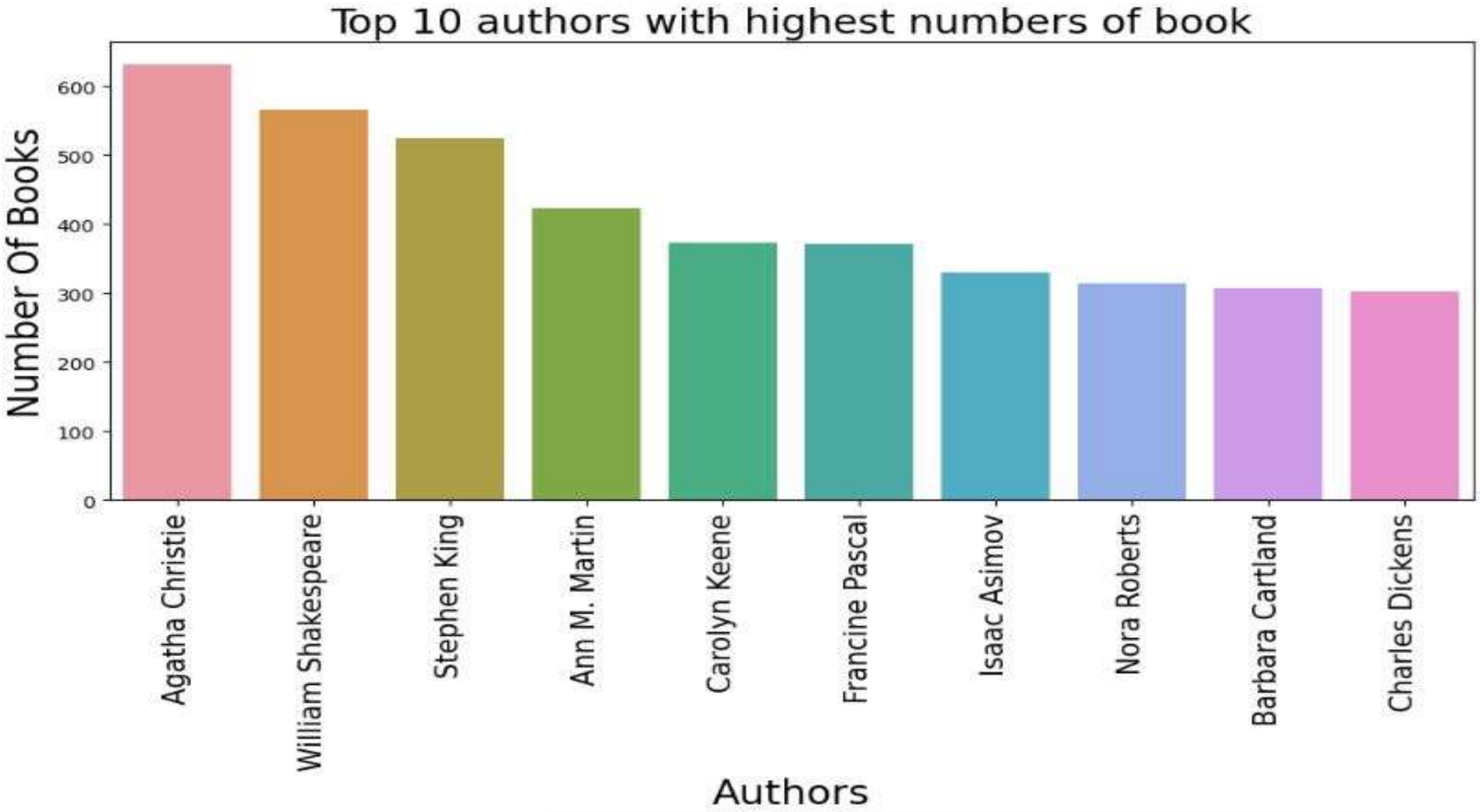
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271360 entries, 0 to 271359
Data columns (total 8 columns):
 #   Column               Non-Null Count   Dtype 
---  ------               --------------   ----- 
 0   ISBN                 271360 non-null  object
 1   Book-Title           271360 non-null  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
dtypes: object(8)
memory usage: 16.6+ MB
```

Exploratory Data Analysis

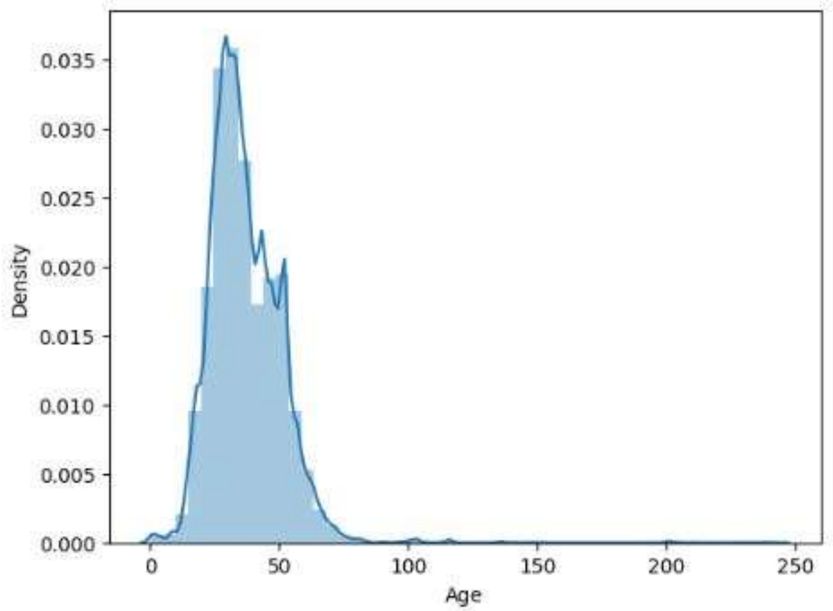
Finding Top 10 Publishers

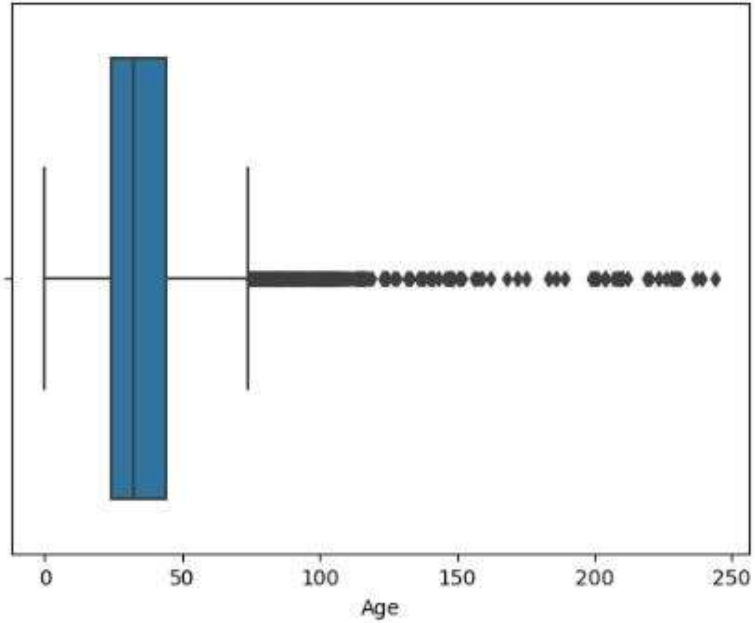


Finding Top 10 Authors



Distribution of Age of Users

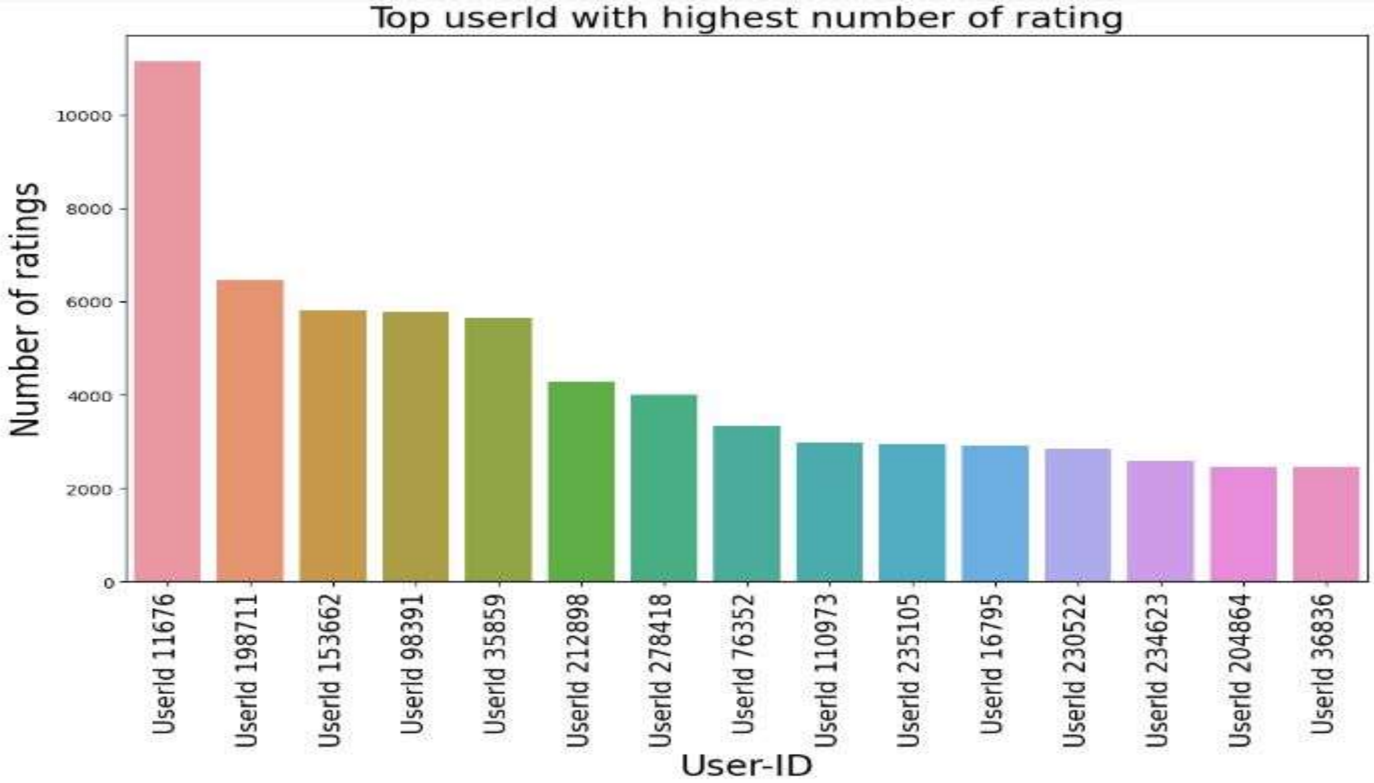




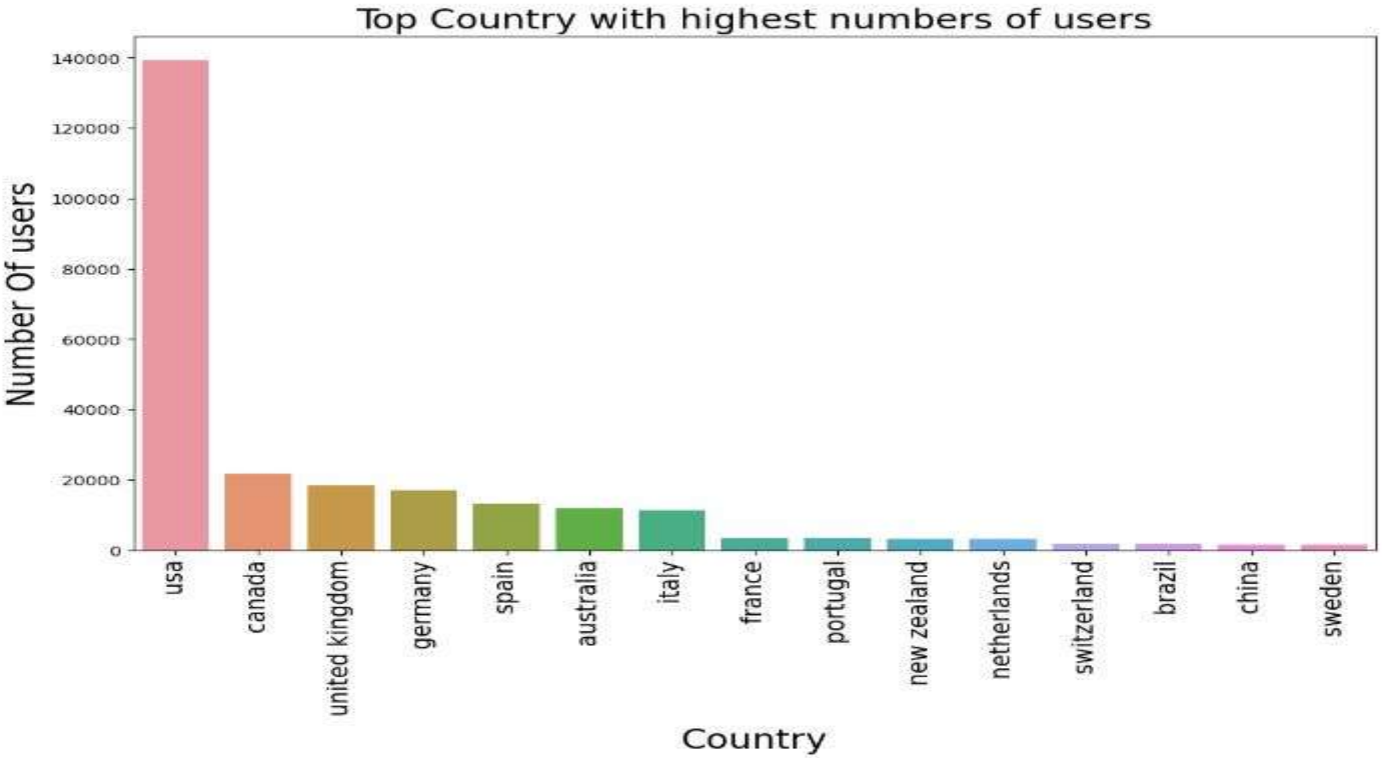
- We can see that majority of the readers are of the age of 20–35

- Some outliers present in the age column

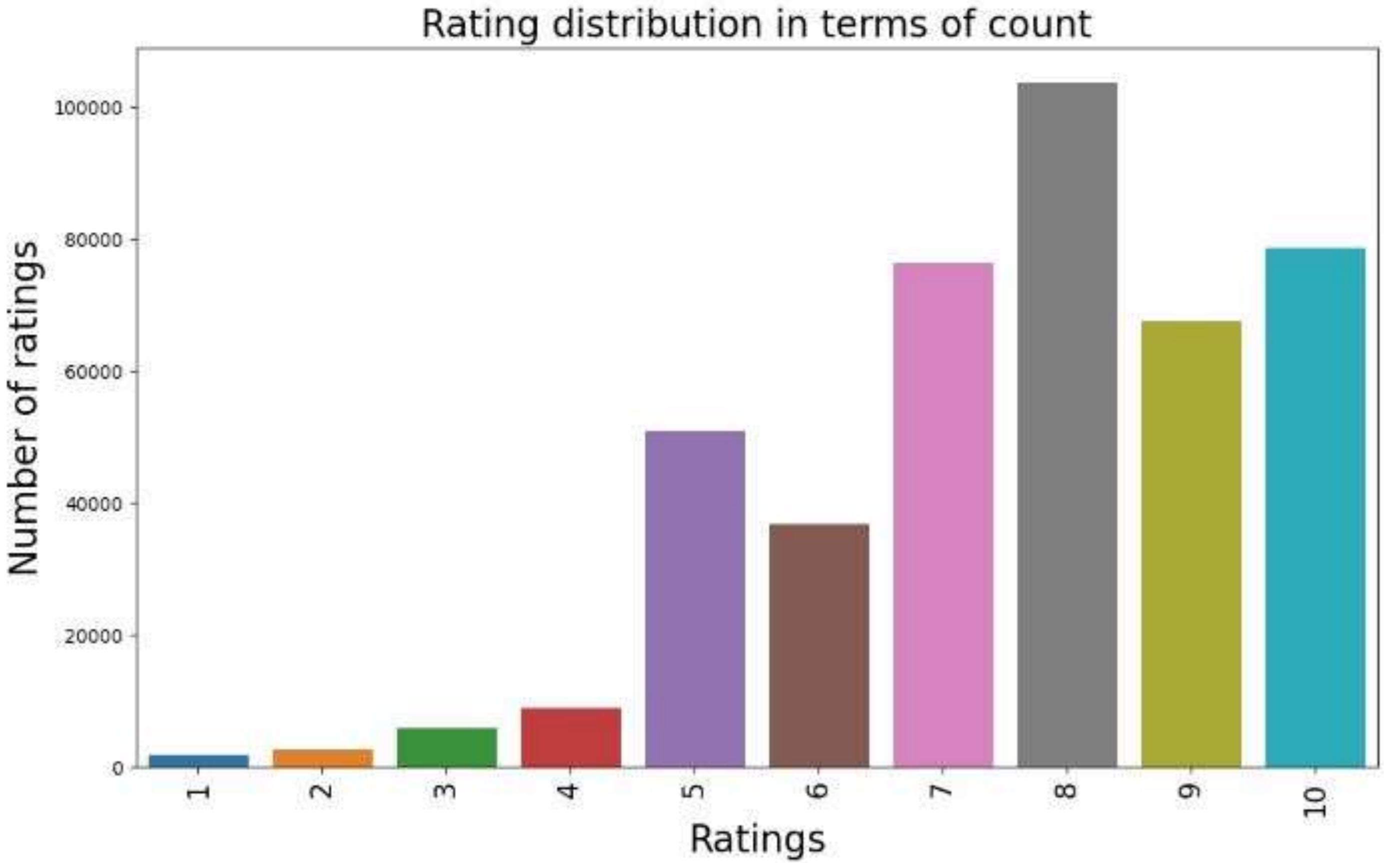
Users with highest ratings



Finding Countries with most number of Users



Rating Distribution



Insights from EDA

- Agatha Christie is the Top authors with highest numbers of book

- Harlequin is the Top Publisher with highest numbers of book

- USA is the Top Country with highest numbers of users

- There are some outlier in the age column. majority of the users are of age 20-35

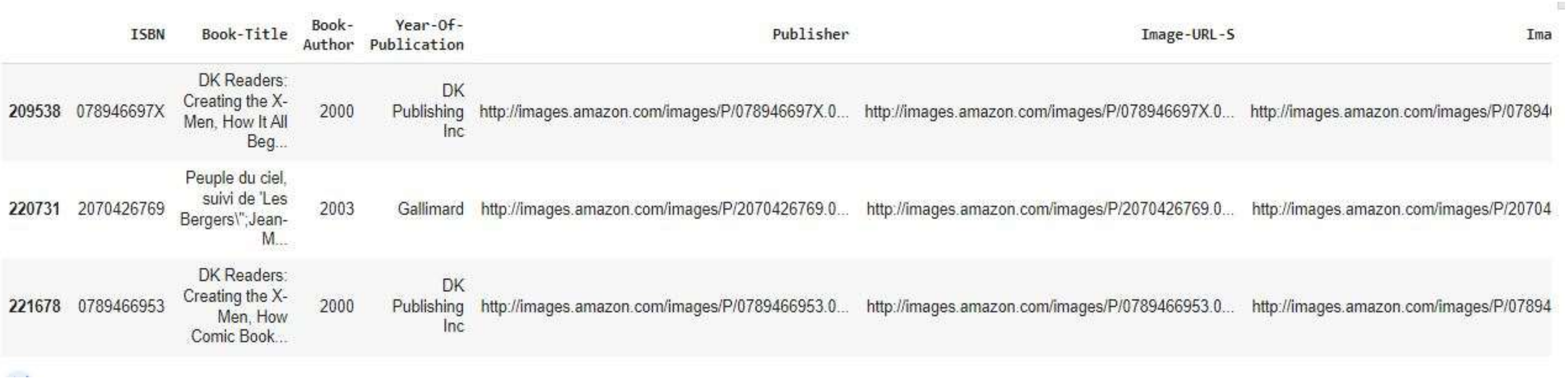
- Maximum of book have good rating.8 is the most common rating for most number of book

Data Preprocessing

Data preprocessing on Year-Of-Publication

- From the analysis part we get that the Year-Of-Publication was wrongly mentioned for some of the rows.

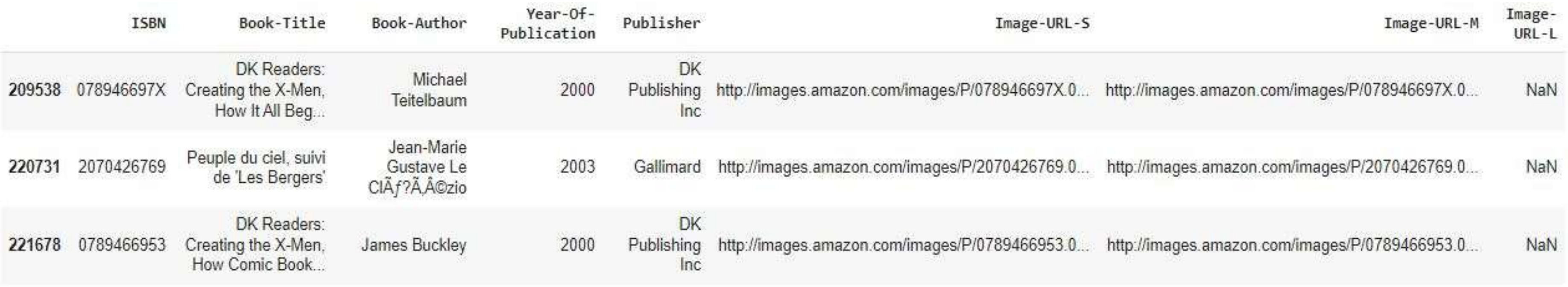
- Diving deep into the Books dataframe we got to know that for these rows there was actually a column mismatch.



	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Ima
209538	078946697X	DK Readers: Creating the X-Men, How It All Beg...	2000	DK Publishing Inc	http://images.amazon.com/images/P/078946697X.0...	http://images.amazon.com/images/P/078946697X.0...	http://images.amazon.com/images/P/07894
220731	2070426769	Peuple du ciel, suivi de 'Les Bergers\";Jean-M...	2003	Gallimard	http://images.amazon.com/images/P/2070426769.0...	http://images.amazon.com/images/P/2070426769.0...	http://images.amazon.com/images/P/20704
221678	0789466953	DK Readers: Creating the X-Men, How Comic Book...	2000	DK Publishing Inc	http://images.amazon.com/images/P/0789466953.0...	http://images.amazon.com/images/P/0789466953.0...	http://images.amazon.com/images/P/07894

Data Preprocessing on Year-Of-Publication

- Making required corrections to 'Year of Publication' column



	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Image-URL-M	Image-URL-L
209538	078946697X	DK Readers: Creating the X-Men, How It All Beg...	Michael Teitelbaum	2000	DK Publishing Inc	http://images.amazon.com/images/P/078946697X.0...	http://images.amazon.com/images/P/078946697X.0...	NaN
220731	2070426769	Peuple du ciel, suivi de 'Les Bergers'	Jean-Marie Gustave Le ClÃƒ?Ã‚Â©zio	2003	Gallimard	http://images.amazon.com/images/P/2070426769.0...	http://images.amazon.com/images/P/2070426769.0...	NaN
221678	0789466953	DK Readers: Creating the X-Men, How Comic Book...	James Buckley	2000	DK Publishing Inc	http://images.amazon.com/images/P/0789466953.0...	http://images.amazon.com/images/P/0789466953.0...	NaN

Data Preprocessing (Continued)

1) Converting Year-Of Publication column from string to integer.

2) For the anomalous entries we first fill them with Nan values.

3) Replacing Age NaN values with median.

Building Recommender System

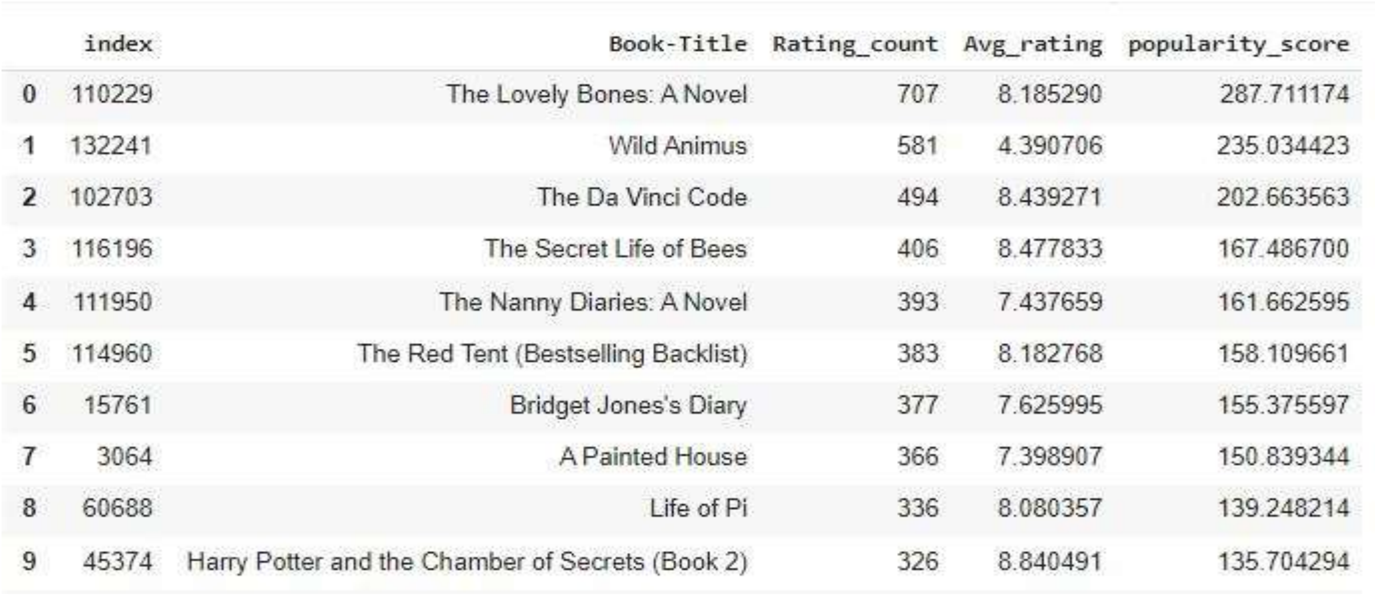
Models used to build Recommendation System

- Popularity Based Approach
- Collaborative Filtering with similarity score
- Collaborative Filtering with SVD method

Popularity Based Approach

Considering book with more than 200 review
popularity_score=0.7*Avg_rating+0.3*Rating_count

- Avg_rating=Average rating of the book
- Rating_count=Number of rating of the book

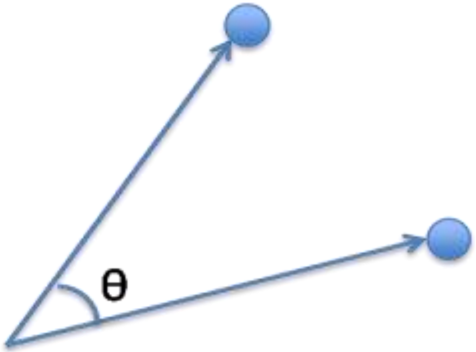


	index	Book-Title	Rating_count	Avg_rating	popularity_score
0	110229	The Lovely Bones: A Novel	707	8.185290	287.711174
1	132241	Wild Animus	581	4.390706	235.034423
2	102703	The Da Vinci Code	494	8.439271	202.663563
3	116196	The Secret Life of Bees	406	8.477833	167.486700
4	111950	The Nanny Diaries: A Novel	393	7.437659	161.662595
5	114960	The Red Tent (Bestselling Backlist)	383	8.182768	158.109661
6	15761	Bridget Jones's Diary	377	7.625995	155.375597
7	3064	A Painted House	366	7.398907	150.839344
8	60688	Life of Pi	336	8.080357	139.248214
9	45374	Harry Potter and the Chamber of Secrets (Book 2)	326	8.840491	135.704294

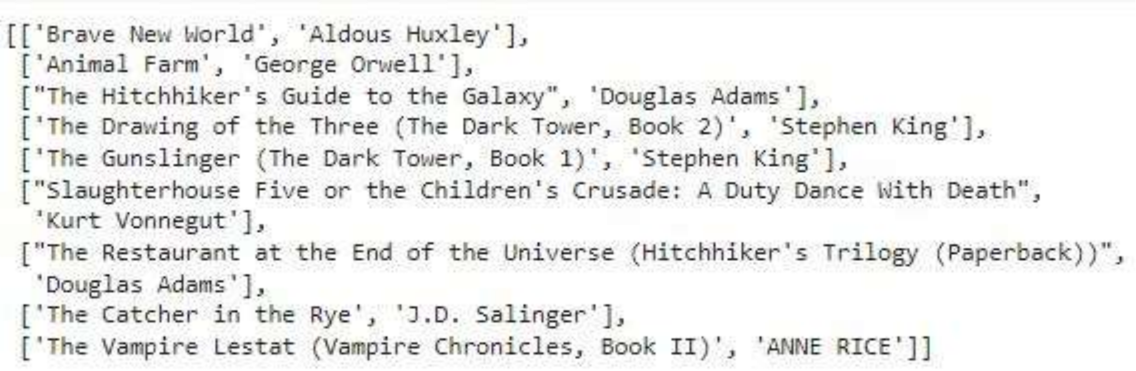
Collaborative Filtering Based Recommender System using cosine similarity

- cosine similarity

$$sim(A,B)=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|}$$



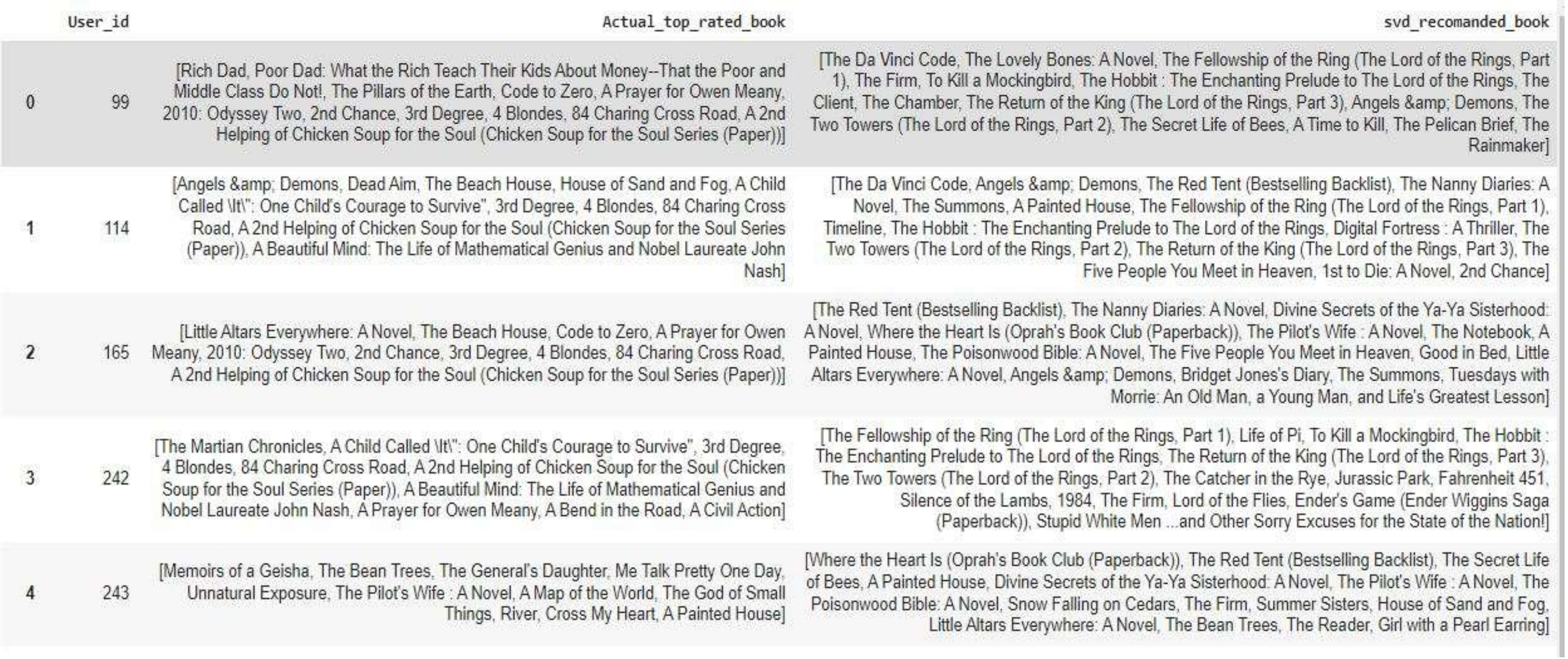
- Recommendation for ‘1984’ book



```
[['Brave New World', 'Aldous Huxley'],
 ['Animal Farm', 'George Orwell'],
 ["The Hitchhiker's Guide to the Galaxy", 'Douglas Adams'],
 ['The Drawing of the Three (The Dark Tower, Book 2)', 'Stephen King'],
 ['The Gunslinger (The Dark Tower, Book 1)', 'Stephen King'],
 ["Slaughterhouse Five or the Children's Crusade: A Duty Dance With Death",
  'Kurt Vonnegut'],
 ["The Restaurant at the End of the Universe (Hitchhiker's Trilogy (Paperback))",
  'Douglas Adams'],
 ['The Catcher in the Rye', 'J.D. Salinger'],
 ['The Vampire Lestat (Vampire Chronicles, Book II)', 'ANNE RICE']]
```

Collaborative Filtering with SVD

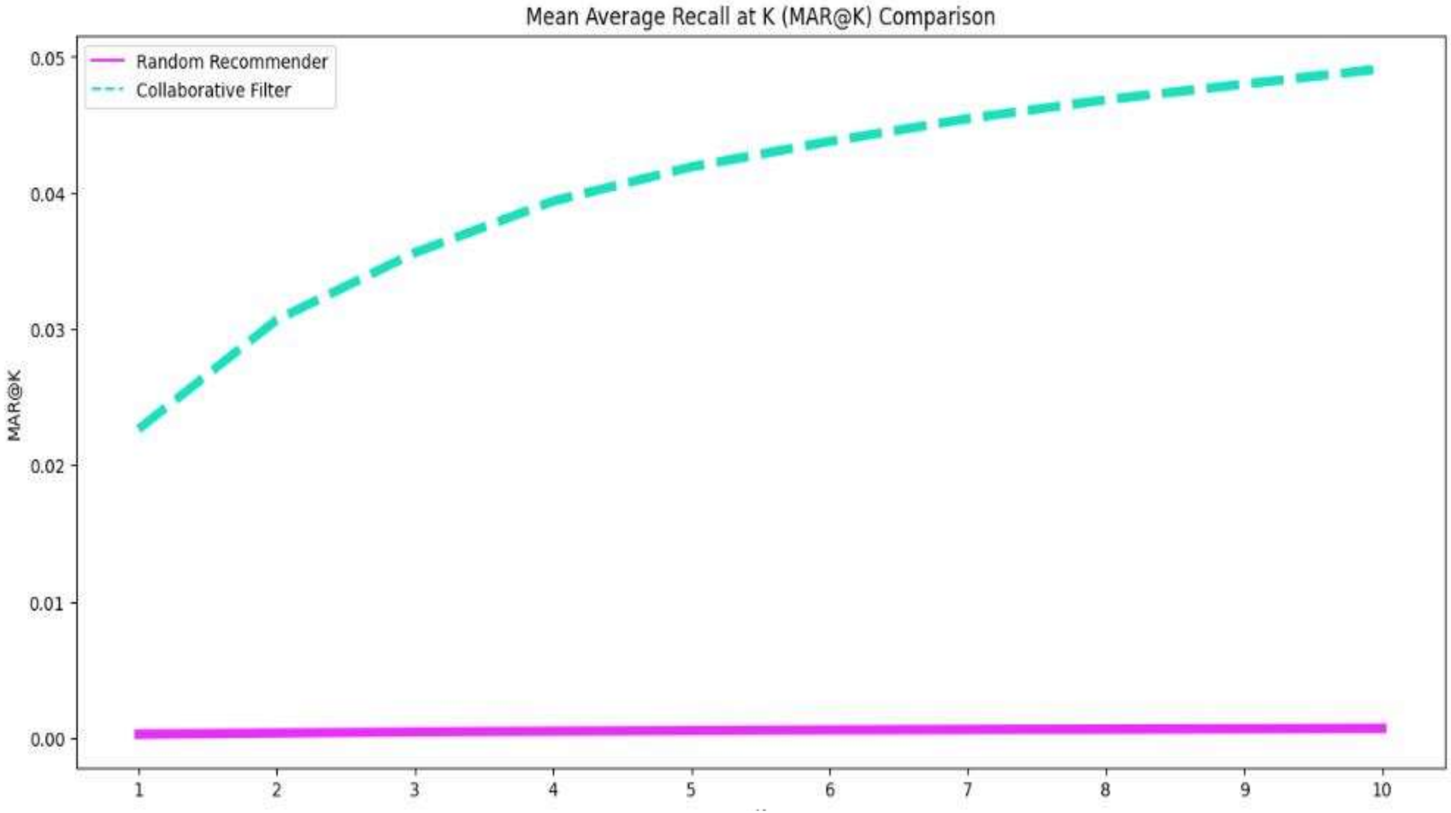
- Top 10 recommended book by SVD model for each user ID



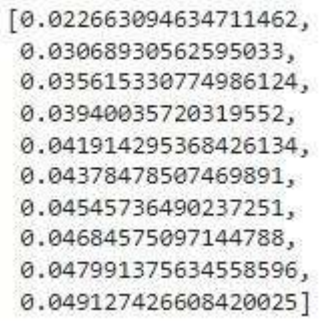
	User_id	Actual_top_rated_book	svd_recomanded_book
0	99	[Rich Dad, Poor Dad: What the Rich Teach Their Kids About Money--That the Poor and Middle Class Do Not!, The Pillars of the Earth, Code to Zero, A Prayer for Owen Meany, 2010: Odyssey Two, 2nd Chance, 3rd Degree, 4 Blondes, 84 Charing Cross Road, A 2nd Helping of Chicken Soup for the Soul (Chicken Soup for the Soul Series (Paper))]	[The Da Vinci Code, The Lovely Bones: A Novel, The Fellowship of the Ring (The Lord of the Rings, Part 1), The Firm, To Kill a Mockingbird, The Hobbit : The Enchanting Prelude to The Lord of the Rings, The Client, The Chamber, The Return of the King (The Lord of the Rings, Part 3), Angels &amp; Demons, The Two Towers (The Lord of the Rings, Part 2), The Secret Life of Bees, A Time to Kill, The Pelican Brief, The Rainmaker]
1	114	[Angels &amp; Demons, Dead Aim, The Beach House, House of Sand and Fog, A Child Called \It\": One Child's Courage to Survive", 3rd Degree, 4 Blondes, 84 Charing Cross Road, A 2nd Helping of Chicken Soup for the Soul (Chicken Soup for the Soul Series (Paper)), A Beautiful Mind: The Life of Mathematical Genius and Nobel Laureate John Nash]	[The Da Vinci Code, Angels &amp; Demons, The Red Tent (Bestselling Backlist), The Nanny Diaries: A Novel, The Summons, A Painted House, The Fellowship of the Ring (The Lord of the Rings, Part 1), Timeline, The Hobbit : The Enchanting Prelude to The Lord of the Rings, Digital Fortress : A Thriller, The Two Towers (The Lord of the Rings, Part 2), The Return of the King (The Lord of the Rings, Part 3), The Five People You Meet in Heaven, 1st to Die: A Novel, 2nd Chance]
2	165	[Little Altars Everywhere: A Novel, The Beach House, Code to Zero, A Prayer for Owen Meany, 2010: Odyssey Two, 2nd Chance, 3rd Degree, 4 Blondes, 84 Charing Cross Road, A 2nd Helping of Chicken Soup for the Soul (Chicken Soup for the Soul Series (Paper))]	[The Red Tent (Bestselling Backlist), The Nanny Diaries: A Novel, Divine Secrets of the Ya-Ya Sisterhood: A Novel, Where the Heart Is (Oprah's Book Club (Paperback)), The Pilot's Wife : A Novel, The Notebook, A Painted House, The Poisonwood Bible: A Novel, The Five People You Meet in Heaven, Good in Bed, Little Altars Everywhere: A Novel, Angels &amp; Demons, Bridget Jones's Diary, The Summons, Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson]
3	242	[The Martian Chronicles, A Child Called \It\": One Child's Courage to Survive", 3rd Degree, 4 Blondes, 84 Charing Cross Road, A 2nd Helping of Chicken Soup for the Soul (Chicken Soup for the Soul Series (Paper)), A Beautiful Mind: The Life of Mathematical Genius and Nobel Laureate John Nash, A Prayer for Owen Meany, A Bend in the Road, A Civil Action]	[The Fellowship of the Ring (The Lord of the Rings, Part 1), Life of Pi, To Kill a Mockingbird, The Hobbit : The Enchanting Prelude to The Lord of the Rings, The Return of the King (The Lord of the Rings, Part 3), The Two Towers (The Lord of the Rings, Part 2), The Catcher in the Rye, Jurassic Park, Fahrenheit 451, Silence of the Lambs, 1984, The Firm, Lord of the Flies, Ender's Game (Ender Wiggins Saga (Paperback)), Stupid White Men ...and Other Sorry Excuses for the State of the Nation!]
4	243	[Memoirs of a Geisha, The Bean Trees, The General's Daughter, Me Talk Pretty One Day, Unnatural Exposure, The Pilot's Wife : A Novel, A Map of the World, The God of Small Things, River, Cross My Heart, A Painted House]	[Where the Heart Is (Oprah's Book Club (Paperback)), The Red Tent (Bestselling Backlist), The Secret Life of Bees, A Painted House, Divine Secrets of the Ya-Ya Sisterhood: A Novel, The Pilot's Wife : A Novel, The Poisonwood Bible: A Novel, Snow Falling on Cedars, The Firm, Summer Sisters, House of Sand and Fog, Little Altars Everywhere: A Novel, The Bean Trees, The Reader, Girl with a Pearl Earring]

Evaluation of SVD model using Recall @k

- Mean Average Recall at k for SVD



- Recall for SVD



```
[0.02266309463471146‍2,
 0.030689305625950‍33,
 0.0356153307749861‍24,
 0.039400357203195‍52,
 0.0419142953684261‍34,
 0.043784785074698‍91,
 0.045457364902372‍51,
 0.046845750971447‍88,
 0.0479913756345585‍96,
 0.0491274266084200‍25]
```

Challenges Faced

- Handling of sparsity became a primary mission as properly because the user interactions had been now no longer present for the majority of the books.

- Understanding the metric for evaluation was a challenge as well.

- Since the information consisted of textual content information, data cleaning became a main task

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.

- Author of most of the books was Agatha Christie, William Shakespeare, Stephen King

- 8 is the most common rating for most number of book. Rating below 5 are in very few in number.

- MAR@K gives that our SVD recommender is able to recall much more then random Recommender.

Thank You