

Capstone Project (Unsupervised Learning)

Book Recommender System

By: Kumari Rashmi

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Home Page Recommendations

D. Recommendations after selecting a Book

E. Evaluation Metrics for Recommender System

F. Conclusion: Key Takeaways

What this presentation is about?

Recommender System:

- It is a subclass of information filtering, that seeks to predict the "ratings" a user will give to new items, and basis these predicted ratings, recommends Top-N rated items to that user
- Helps users by suggesting them most relevant items from a large corpora
- In other words, it is a useful alternative to search algorithms, as they help users discover items they might not have found otherwise

Goals:

- To identify underlying trends and patterns in datasets
- To spot and handle irregularities (if any)
- Draw meaningful insights
- Build a Book Recommender System

Why Recommender Systems are important?

- **Increase Revenue** by directing the items and sales offers to specific users, and thus increasing the likelihood of selling items
- **Increase Customer retention:** It helps business in understanding the needs of customers & eventually, in devising the right strategies to foster and maintain relationship with them
- **Make User's life easier** by handling a large amount of information and by providing them with personalized content recommendations, which further leads to customer satisfaction



Basic Information about Datasets

1. Books Dataset

- It contains information on books, such as their unique ISBN, title, author, publisher, year of publication and image-URLs
- There are 2,71,046 unique books, 98,089 unique authors, 2,38,963 unique titles and 16,287 unique publishers

2. Ratings Dataset

- It contains ratings given by users for different books. Rating ranges from 0 to 10; zero being implicit rating
- ~38% observations are explicit ratings, rest are implicit ratings

3. Users Dataset

- It contains information on 2,78,858 unique users such as their unique user Id, age and location
- Basis location feature, extracted the country of users

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Home Page Recommendations

D. Recommendations after selecting a Book

E. Evaluation Metrics for Recommender System

F. Conclusion: Key Takeaways

Data Preparation & Cleaning

1. Formatting Inconsistent data types of columns:

- Few values in Year of Publication were string, and accordingly, were corrected to int data type

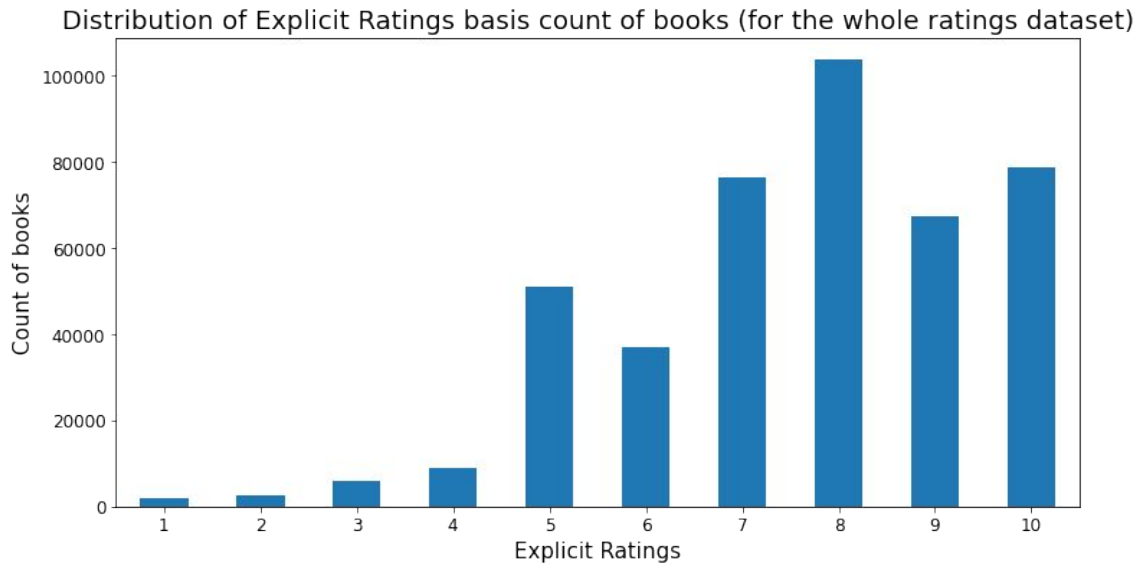
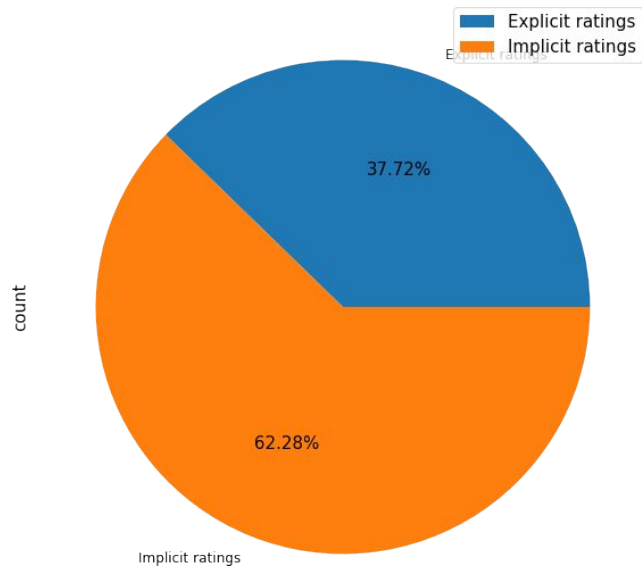
2. Handling Missing Values & Data Irregularities:

- Very few values were missing in Image-URL-L, author and publisher columns, and were ignored
- Age of ~40% users were missing, ~0.15% users were of age >90 years & ~0.33% users were of age <6 years. These ambiguous age values were replaced by Median age
- Since this data was collected in 2004, Publication years such as 0 or after 2004 were replaced by Median year value
- To avoid data ambiguity, columns with textual data were converted to either lower-case or upper-case and punctuations were also removed
- Duplicate values of ISBN from books dataset were dropped

Underlying Patterns and Trends

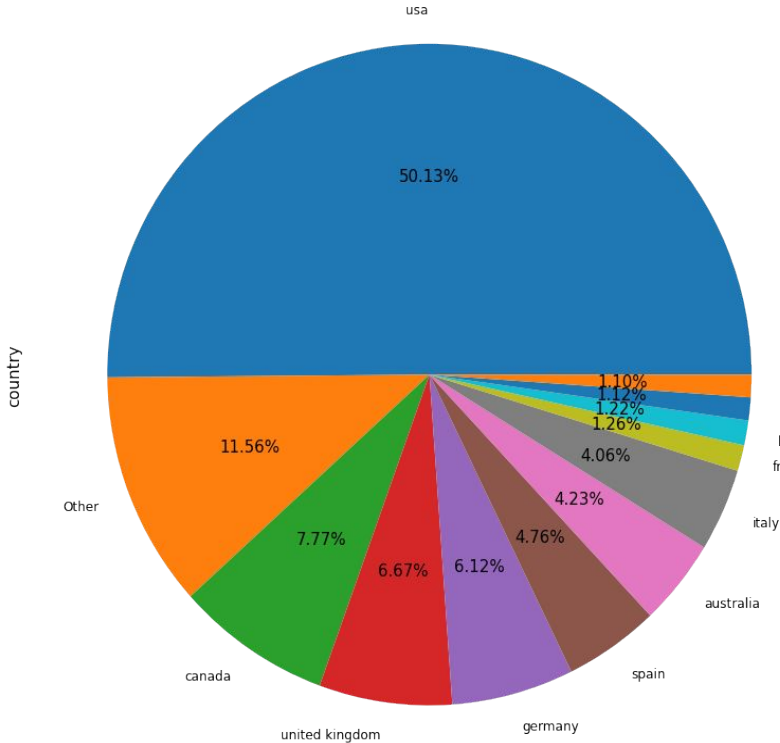
1. Explicit Vs Implicit Ratings

- Majority of ratings provided are of implicit type (~62%)
- Among explicit ratings, rating of 8 has the highest count, followed by 10

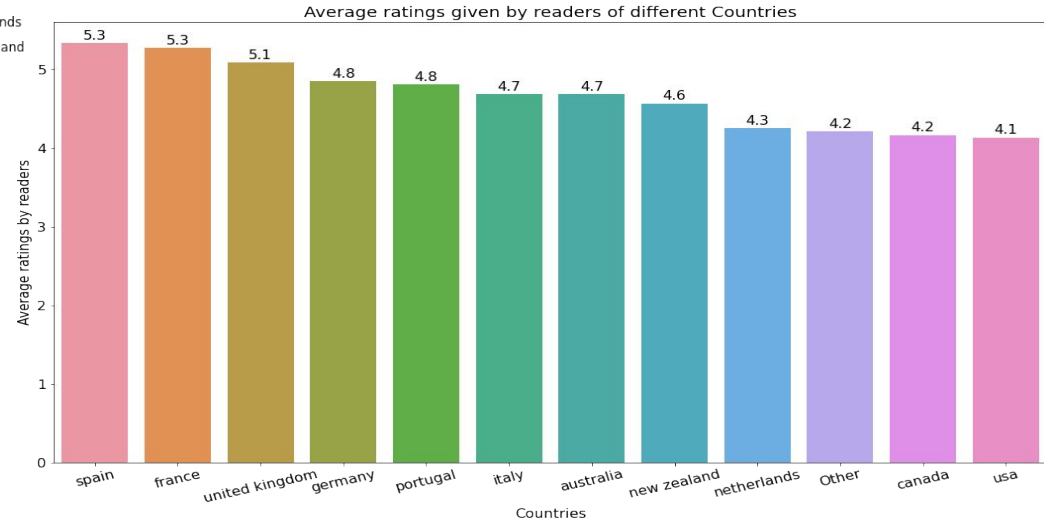


2. Readers and their countries

Distribution of users basis their countries

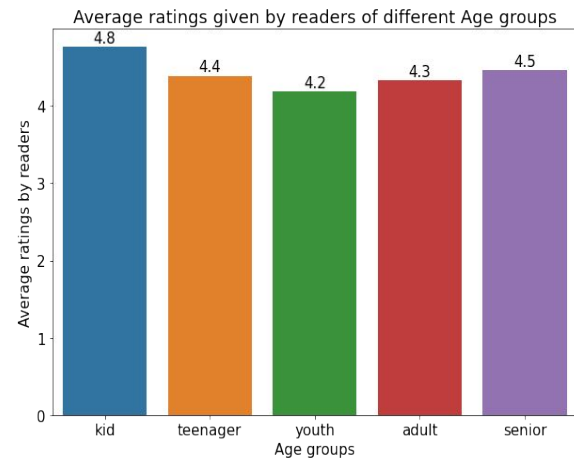
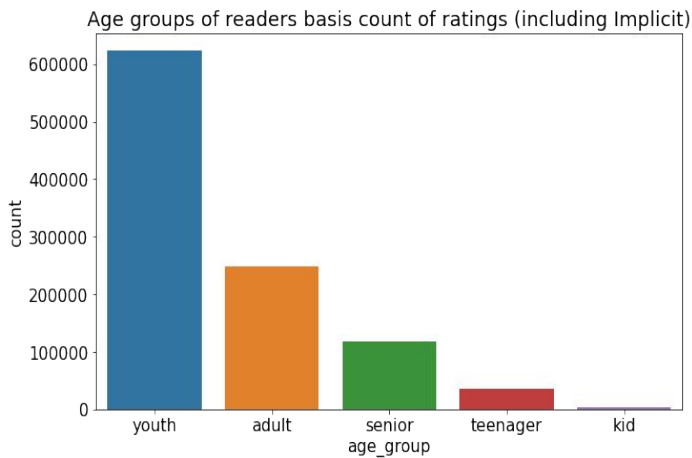
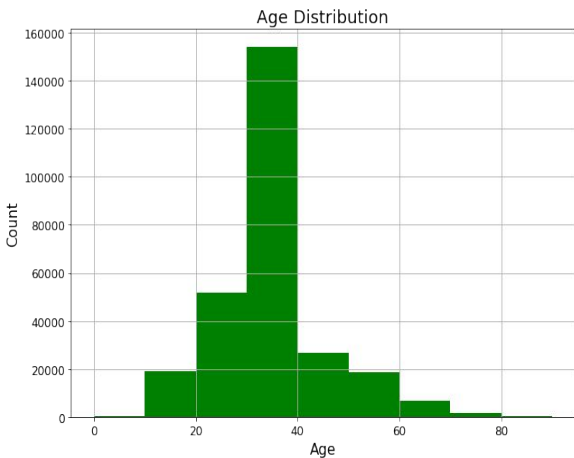


- Majority of readers are from USA (~50%), then followed by Canada and UK
- Average rating by readers of Spain & France is the highest (5.3), while that of USA is lowest (4.1)

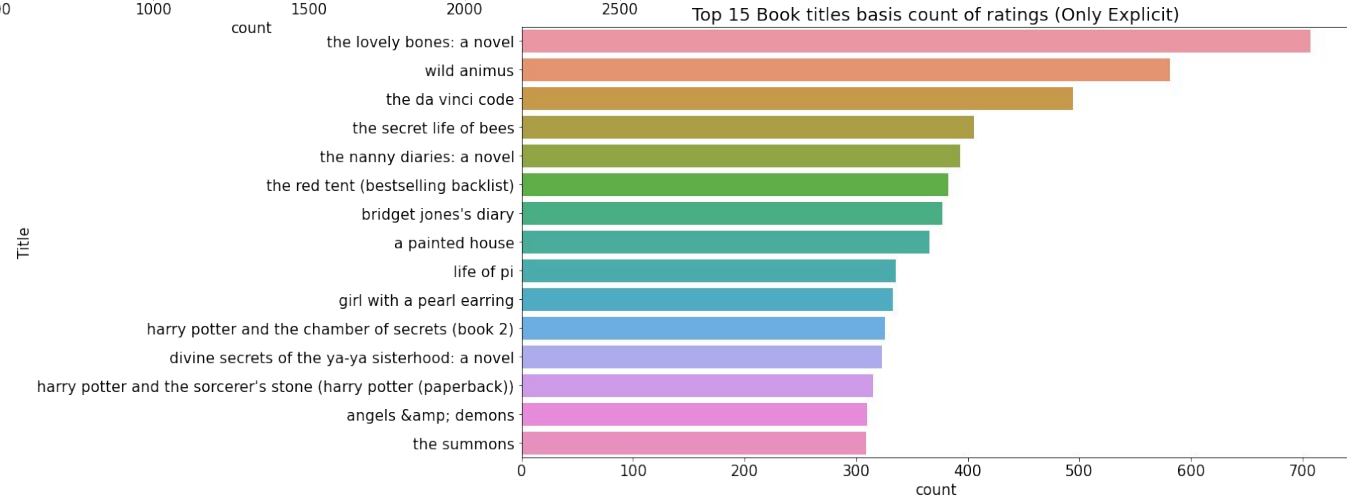
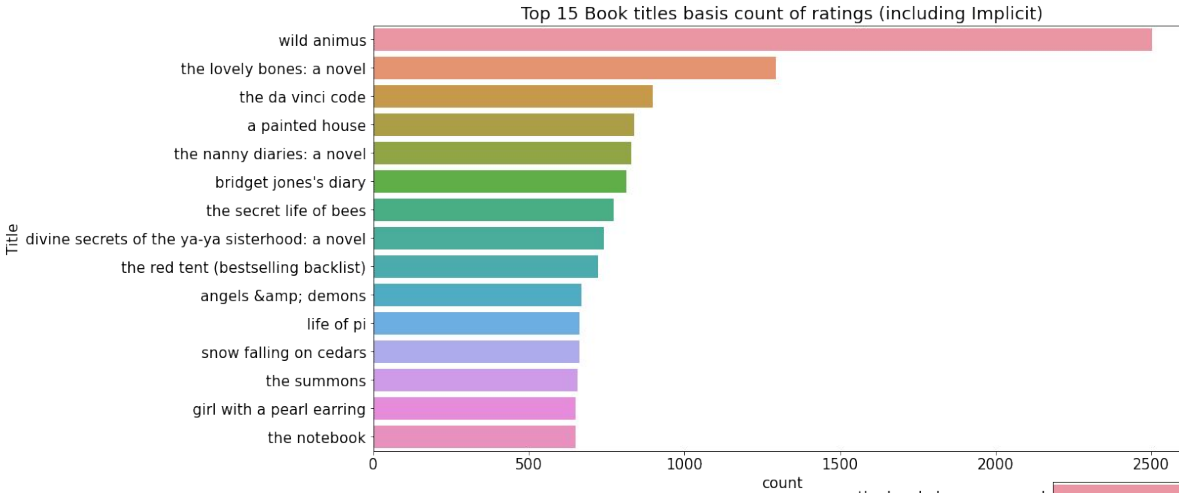


3. Readers and their Ages

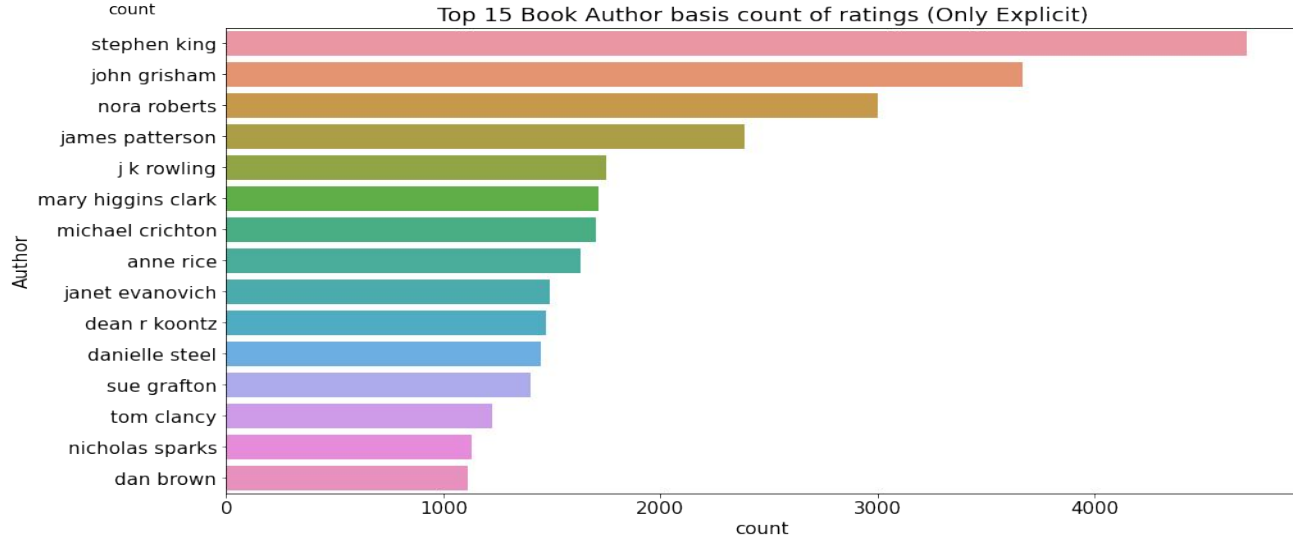
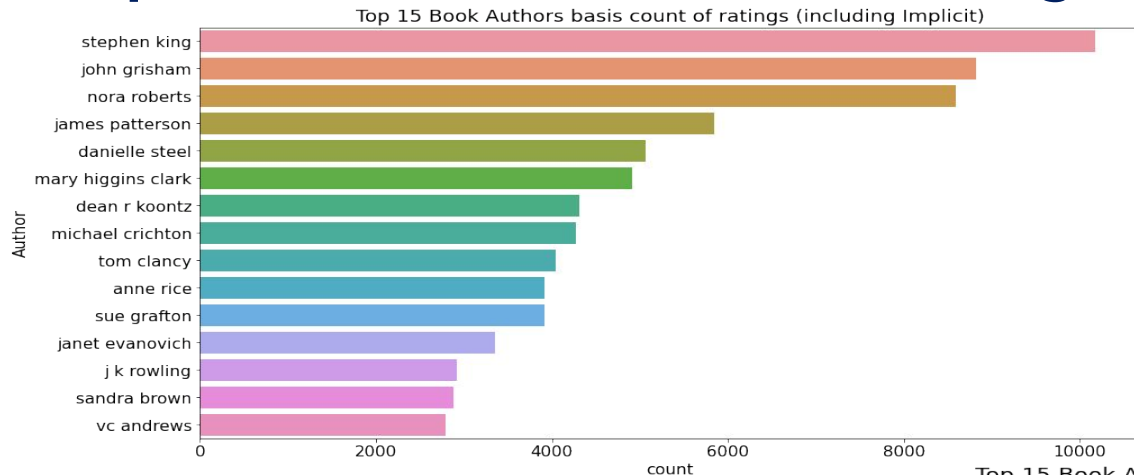
- Basis age values, readers were categorized as Kid(<13 years), Teenager(<20 years), Youth(<36 years), Adult (<51 years) and Senior(>=51 years)
- Readers of youth age group have the highest count, followed by Adult readers
- Average of ratings given by Kids is the highest (4.8), followed by Seniors (4.5)



4. Top 10 Book Titles basis count of Ratings



5. Top 15 Authors basis count of Ratings



Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

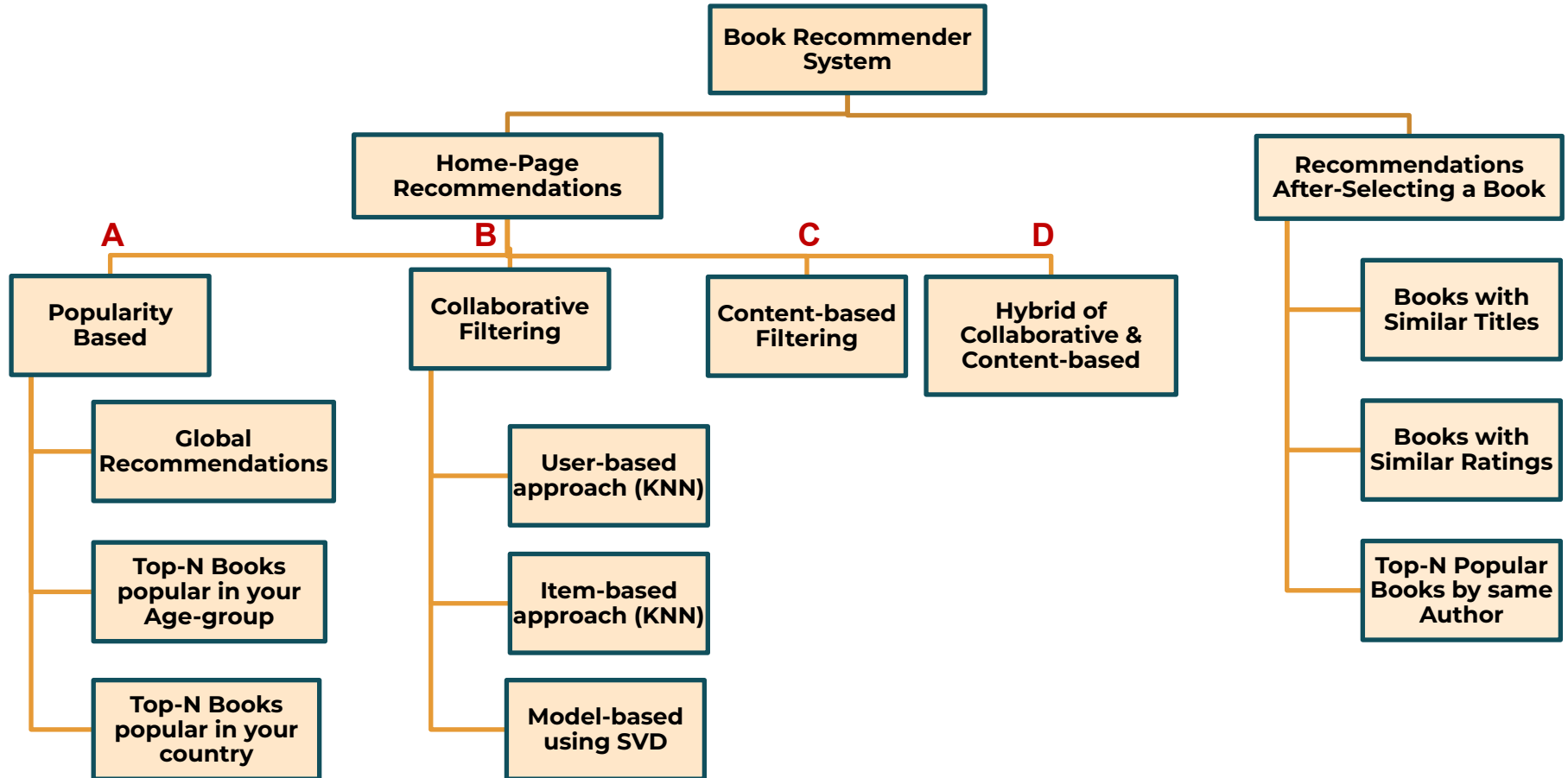
C. Home Page Recommendations

D. Recommendations after selecting a Book

E. Evaluation Metrics for Recommender System

F. Conclusion: Key Takeaways

Overview of Book Recommender System



A. Popularity Based Recommendations

- **Getting a smaller subset of potential candidates (Books):**
 - Basis minimum number of votes per books (i.e., 8 votes per book)
 - Basis minimum number of votes per user (i.e., 15 votes given by a user)
- **Bayesian Weighted Average Rating (IMDB's formula):**
 - For a book with a smaller number of votes, we may not trust its individual average rating and would prefer to go with the average rating for entire available dataset
 - To include the influence of number of votes in book's ratings, we need to arrive at some "Bayesian Average rating" formula. Following IMDB's weighted average rating formula was used:

$$\bullet \quad WR = (v \cdot R / (v + m)) + (m \cdot C / (v + m))$$

Where, R = average rating for the movie, v = number of votes for the movie, m = minimum votes required to be listed in Top-N (I have taken 40 votes for books), C = the mean rating across the whole dataset

Top N Popular Books (Global)

S. No.	Title of Book -- By Author	Average weighted score
1	Harry Potter and the goblet of fire (book 4) by JK Rowling	5.91
2	Free by Paul Vincent	5.78
3	Harry Potter and the prisoner of Azkaban (book 3) by JK Rowling	5.61
4	Harry Potter and the sorcerer's stone (book 1) by JK Rowling	5.38
5	Harry Potter and the order of the phoenix (book 5) by JK Rowling	5.30
6	The Fellowship of the ring (the lord of the rings, part 1) by JRR Tolkien	5.05
7	Harry potter and the chamber of secrets (book 2) by JK Rowling	5.02
8	Ender's game (Ender Wiggins saga (paperback)) by Orson Scott card	4.96
9	Griffin & Sabine: an extraordinary correspondence by Nick Bantock	4.90
10	The two towers (The lord of the rings, part 2) by JRR Tolkien	4.89

Top N Books Popular in your age group


S. No.	Title of Book--By Author	Score
1	Harry potter & the goblet of fire (book 4) by JK Rowling	5.91
2	Free by Paul Vincent	5.78
3	Harry potter & the prisoner of Azkaban (book 3) by JK Rowling	5.60

**Youths**

S. No.	Title of Book--By Author	Score
1	The two towers (the lord of the rings, part 2) by JRR Tolkien	4.89
2	The hobbit: the enchanting prelude to the lord of the rings by JRR Tolkien	4.74
3	The horse and his boy by CS Lewis	4.46

Seniors

S. No.	Title of Book--By Author	Score
1	Falling up by Shel Silverstein	4.83
2	Love you forever by Robert N Munsch	4.55
3	The little prince by Antoine de Saintexup�ry	4.54

**Adults**

Top N Books Popular in your Country

S.N.	Title of Book--By Author	Score
1	The horse and his boy by CS Lewis	4.46
2	A clockwork orange (Norton paperback fiction) by Anthony Burgess	4.07
3	The last battle by CS Lewis	4.06

United Kingdom

S. N.	Title of Book--By Author	Score
1	Harry Potter and the order of the phoenix (book 5) by JK Rowling	5.30
2	The two towers (the lord of the rings, part 2) by JRR Tolkien	4.89
3	Falling up by Shel Silverstein	4.83

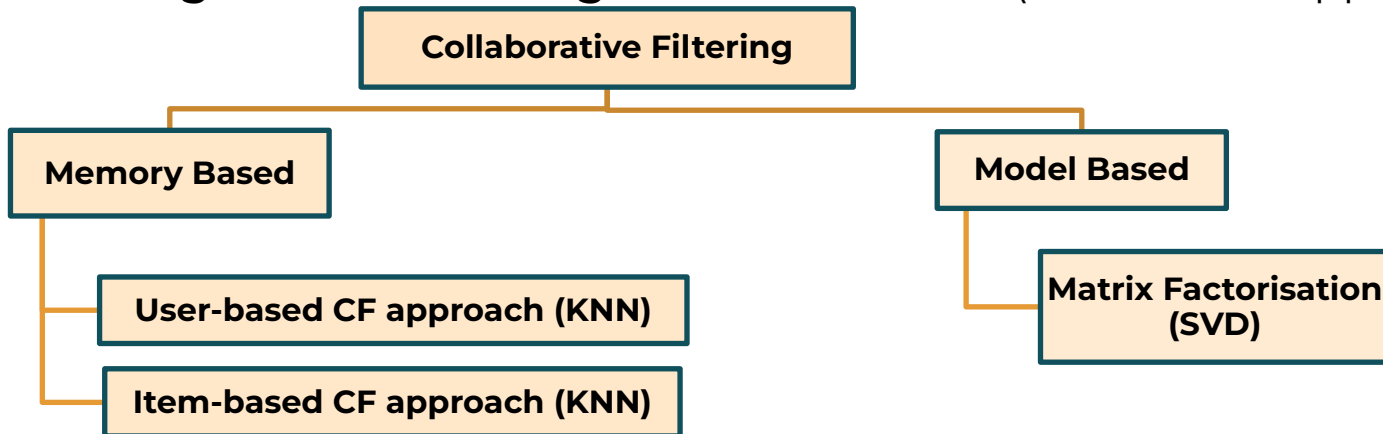
USA

S. N.	Title of Book--By Author	Score
1	El Ocho by Katherine Neville	3.48
2	El Ultimo Caton by Matilde Asensi	3.41
3	Historia de una gaviota y del gato (andanzas) by Luis sepulveda	3.29

Spain

B. Collaborative Filtering

- It approaches to build a model from a user's past behavior and similar decisions made by other users
- It generates recommendations by locating peer users/items with a rating history, similar to the current user or item
- Underlying assumptions:
 - **“Similar people will have similar taste”** (User-Based approach)
 - **“A User gives similar ratings to similar items”** (Item-Based approach)




User-Based CF Approach (Using KNN)

Steps taken:

1. Created a user-item interaction matrix taking users' ratings as values
2. Built a KNN-model using cosine distances & above user-item matrix
3. For a given user-A, got 5-most similar users basis cosine distances calculated from above KNN-model
4. Calculated weighted average (taking similarities as weights) of ratings by similar users for books unread by user-A
5. Basis above scores, rank ordered the list of unread books by user-A & recommended Top-N books

For User ID: 243




S. No.	Titles of Book -- By Author	Weighted average score
1	The bean trees by Barbara Kingsolver	5.60
2	The pilot's wife : a novel by Anita Shreve	5.20
3	Good in bed by Jennifer Weiner	5.19
4	Atonement : a novel by Ian McEwan	5.00
5	To kill a mockingbird by Harper lee	3.80

Item-Based CF Approach (Using KNN)

Steps taken:

1. Built a KNN-model using cosine as metric
2. For a given user-A, got the list of all unread books, and then found 5-most similar books against each unread book
3. Predicted rating by user-A for each unread book, by calculating the weighted average of ratings (taking similarity as weights) for similar books by user-A
4. After predicting ratings for all unread books by user-A, rank ordered it and recommended Top-N books

For User ID: 243



S. No.	Titles of Book -- By Author	Weighted average score
1	Animal Dreams by Barbara Kingsolver	10.00
2	Cruel & Usual by Patricia D Cornwell	9.00
3	From potter's field by Patricia D Cornwell	9.00
4	Point of origin by Patricia D Cornwell	9.00
5	The Body Farm by Patricia Daniels Cornwell	9.00
6	Drowning Ruth (Oprah's book club) by Christina Schwarz	7.00
7	Promises by Belva Plain	6.00

SVD: Model-Based CF Approach

- It is a matrix factorization technique, which reduces the number of features from N-dimension to K-dimension
- SVD decomposes a given user-item interaction matrix into 3 low-rank orthogonal matrices as:

$$\mathbf{A} = \mathbf{U} * \mathbf{S} * \mathbf{V}^T$$

U: #users x #latent factors (m x r)

S: diagonal matrix (r x r)

V: #items x #latent factors (n x r)

- A: reconstructed matrix with predicted ratings, used for recommendations
- Chosen number of latent factors = 15

For User ID:
243

S. No.	Titles of Book
1	The Secret life of bees
2	The Lovely bones: a novel
3	The Da Vinci code
4	Bridget Jones's diary
5	The Red Tent (bestselling backlist)

For User ID:
16916

S. No.	Titles of Book
1	The Partner
2	A Time to kill
3	The Testament
4	The Runaway Jury
5	A Painted House

C. Content-Based Filtering

- It approaches to build a model, from information about the description and attributes of the items, that a given user has previously consumed to model user's preferences (User profile)
- Various candidate items are compared with items previously rated by the given user (User profile), and the best-matching items are recommended

**For User ID:
243**

S. No.	Titles of Book
1	The Club
2	Secrets of the heart
3	Golden Lies
4	Miles from Nowhere: A round the world bicycle adventure
5	Mother's Day

**For User ID:
16916**

S. No.	Titles of Book
1	Nicolae: The Rise of Antichrist (left behind no. 3)
2	Expiration Date
3	Assassins: Assignment: Jerusalem, target: Antichrist (left behind no. 6)
4	The Cave
5	The Anubis Gates

D. Hybrid Recommender

- Hybrid recommenders combine two or more recommendation strategies in different ways to benefit from their complementary advantages
- I have used weighted approach to combine Model-based CF (SVD) system and content-based recommender system, with more weightage to SVD based CF system (90:10)

For User ID:
243

S. No.	Titles of Book
1	The Bean Trees
2	Me talk pretty one day
3	The Poisonwood Bible
4	Animal dreams
5	Falling Angels

For User ID:
16916

S. No.	Titles of Book
1	Desecration: Antichrist takes the throne (left behind #9)
2	The Indwelling: The Beast takes possession (left behind #7)
3	Nicolae: The rise of antichrist (left behind #3)
4	Tribulation Force: The continuing drama of those left behind (left behind #2)
5	The Remnant: On the brink of Armageddon (left behind #10)

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Home Page Recommendations

D. Recommendations after selecting a Book

E. Evaluation Metrics for Recommender System

F. Conclusion: Key Takeaways

Books with Similar Titles

Steps taken:

1. For entire books set, created vectorized matrix using Titles (text data)
2. For a selected Book, found most N-similar books basis cosine similarity of vectorized matrix

The return of the king (The Lord of the rings, part 3)



S. N.	Titles of Book
1	The Two towers (The Lord of the rings, part 2)
2	The fellowship of the ring (the lord of the rings, part 1)
3	The Lord of the rings (movie art cover)
4	The Hobbit: the enchanting prelude to the lord of the rings
5	The Return Journey

Falling Angels




S. No.	Titles of Book
1	Angels
2	The Sky is falling
3	The Killer angels
4	Snow falling on cedars
5	Angels & Demons

Books with Similar Ratings

Steps taken:


1. Created a user-item interaction matrix taking users' ratings as values
2. For a selected Book, found most N-similar books basis correlation coefficients

1812 (The American story)



S. No.	Titles of Book
1	West of dodge
2	Deep water
3	The day after tomorrow
4	The Hancock boys
5	Road to perdition

Harry potter and the chamber of secrets (book 2)



S. No.	Titles of Book
1	Harry Potter and the prisoner of Azkaban (book 3)
2	Harry Potter and the goblet of fire (book 4)
3	Harry Potter and the sorcerer's stone (book 1)
4	Harry Potter and the order of the phoenix (book 5)
5	Harry Potter and the sorcerer's stone (harry potter (paperback))

Top-N Popular Books by same Author

Steps taken:

1. Get the name of author of selected book
2. Filter out all books by above author and rank order these basis weighted average rating (IMDB's popularity formula)

**The Girl who loved Tom
Gordon by Stephen king**



S. No.	Titles of Book
1	Dolores Claiborne
2	The regulators
3	It (r)
4	Needful things
5	Blood and Smoke

**The Sittaford mystery
by Agatha christie**



S. N.	Titles of Book
1	Halloween party
2	Peril at end house (hercule Poirot mysteries (paperback))
3	Death on the Nile
4	Murder on the orient express
5	The mysterious affair at styles

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Home Page Recommendations

D. Recommendations after selecting a Book

E. Evaluation Metrics for Recommender System

F. Conclusion: Key Takeaways

Accuracy Based Evaluation Metric

RMSE:

- It is square-root of the mean of the squared errors (i.e., difference between the actual value/rating & the predicted value/rating)
- Using Surprise library, implemented different recommender systems and compared all basis RMSE
- SVD based recommender was the best performing model with lowest RMSE value

$$\text{RMSE} = \sqrt{\frac{1}{|\hat{R}|} \sum_{\hat{r}_{ui} \in \hat{R}} (r_{ui} - \hat{r}_{ui})^2}$$

S. No.	Recommender System	RMSE
1	SVD (Model Based CF)	1.545
2	KNNWithMeans (Memory based UB-CF)	1.695
3	KNNBaseline (Memory based UB-CF)	1.696
4	KNNBasic (Memory based UB-CF)	1.833
5	NMF (Model Based CF)	2.508

Ranking Based Evaluation Metric

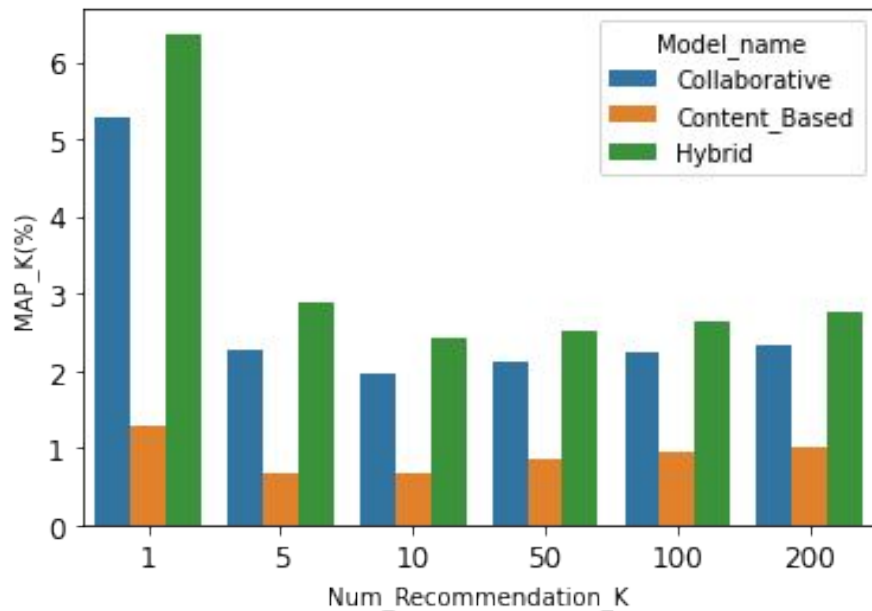
MAP@N:

- **Precision@N** is the fraction of total recommendations (N) that are relevant to the user
- **AP@N**: To assess the performance of a recommender basis positions/ranking of relevant recommendations, we take average of precisions at every position ($k=1$ to N , m = total relevant items)
- **MAP@N** is the mean of “average precisions at each position” for all users

$$P = \frac{\# \text{ of our recommendations that are relevant}}{\# \text{ of items we recommended}}$$

$$AP@N = \frac{1}{m} \sum_{k=1}^N (P(k) \text{ if } k^{th} \text{ item was relevant})$$

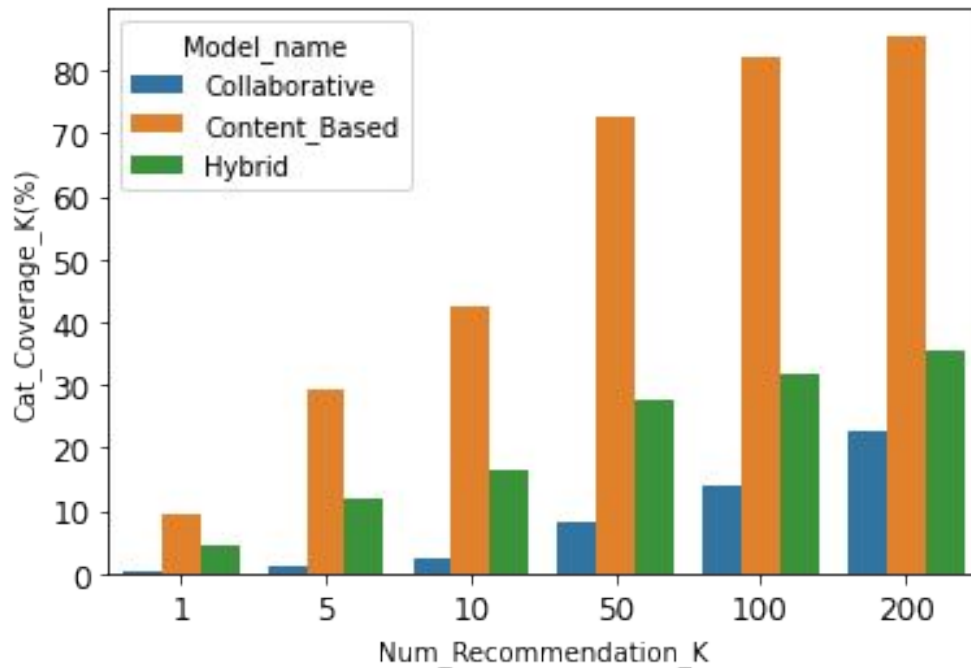
$$MAP = \frac{\sum_{q=1}^Q AveP(q)}{Q}$$



Serendipity Based Evaluation Metric

Catalogue Coverage:

- Coverage is the percent of items in the training data, that model is able to recommend to users
- Coverage does not evaluate if the user enjoys the recommendation or not, instead, it ***assess the system in terms of its ability to bring unexpectedness to the user***
- Low coverage may lead to users' dissatisfaction



Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Home Page Recommendations

D. Recommendations after selecting a Book

E. Evaluation Metrics for Recommender System

F. Conclusion: Key Takeaways

Key Takeaways:

- Majority of readers are from USA, followed by Canada
- The average of ratings given by Kids was the highest, followed by seniors; On average, rating given by readers from Spain & France were the highest
- For new users, built global popularity based recommender systems
- For registered users with no rating history, built demographics based popularity recommender systems
- For users with ratings history, built collaborative, content based and hybrid recommender systems
- For this dataset, Content-based recommender has the highest catalogue coverage for all N (as ~43% for N=10), while Hybrid recommender system has the highest MAP@N for all N (as, MAP@10 = 2.43%)

Challenges faced:

- **Cold Start Problem:** It happens whenever either new item is added to the catalogue or a new user is added, that have either zero or very little interactions. This was overcome by using content-based and popularity based recommendations respectively
- **Sparsity:** Majority of interactions data are sparse, as most customers do not give explicit feedbacks; SVD technique handles sparsity well
- **Lack of data:** Building a good recommender system requires huge amount of data. As the size of dataset was small, the results could be improved by increasing the size of dataset

**END OF PRESENTATION
THANKS**