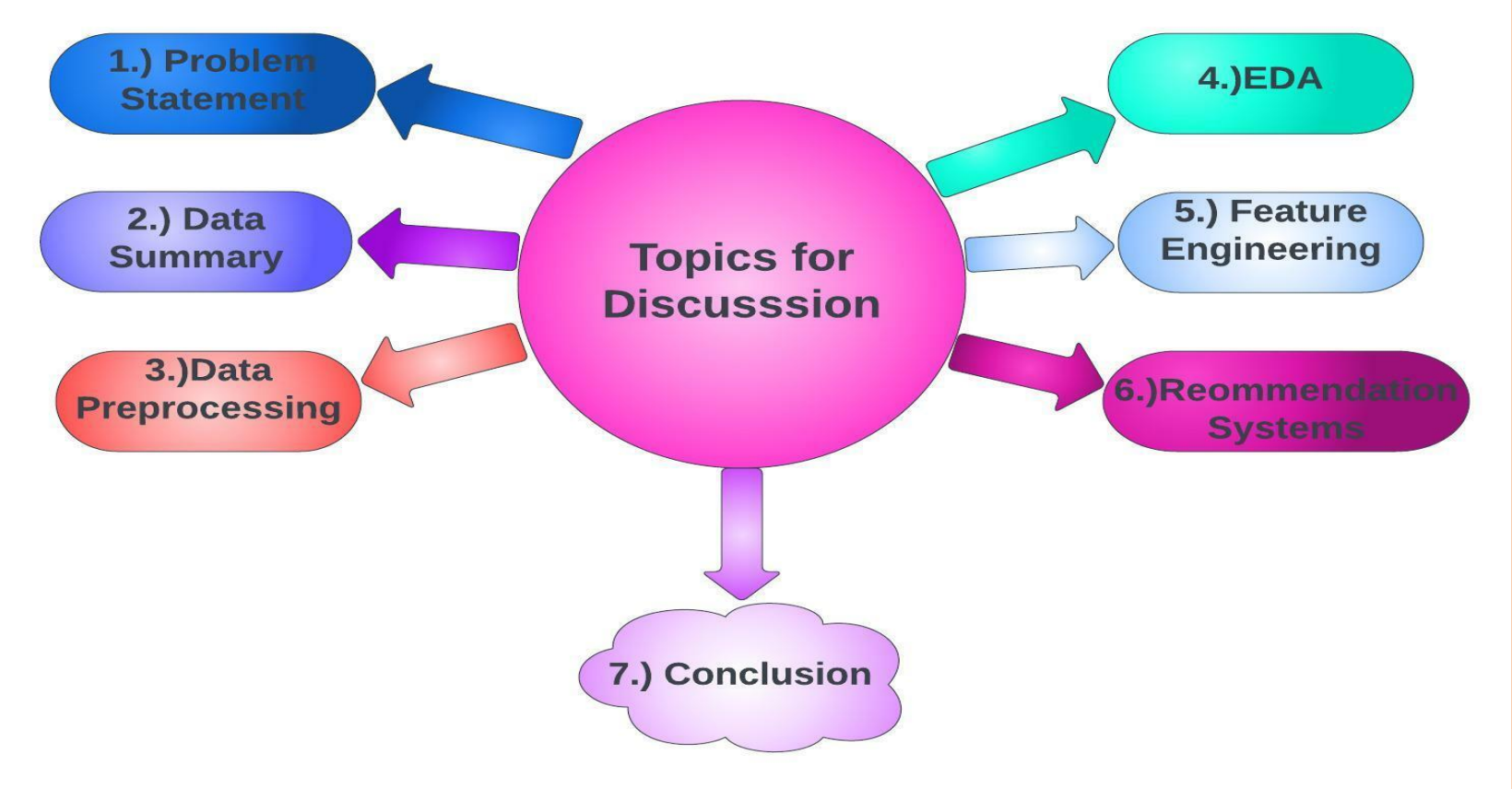
Capstone Project

Book Recommendation System

By-Shailendra, Umesh, Akram, Ifraz

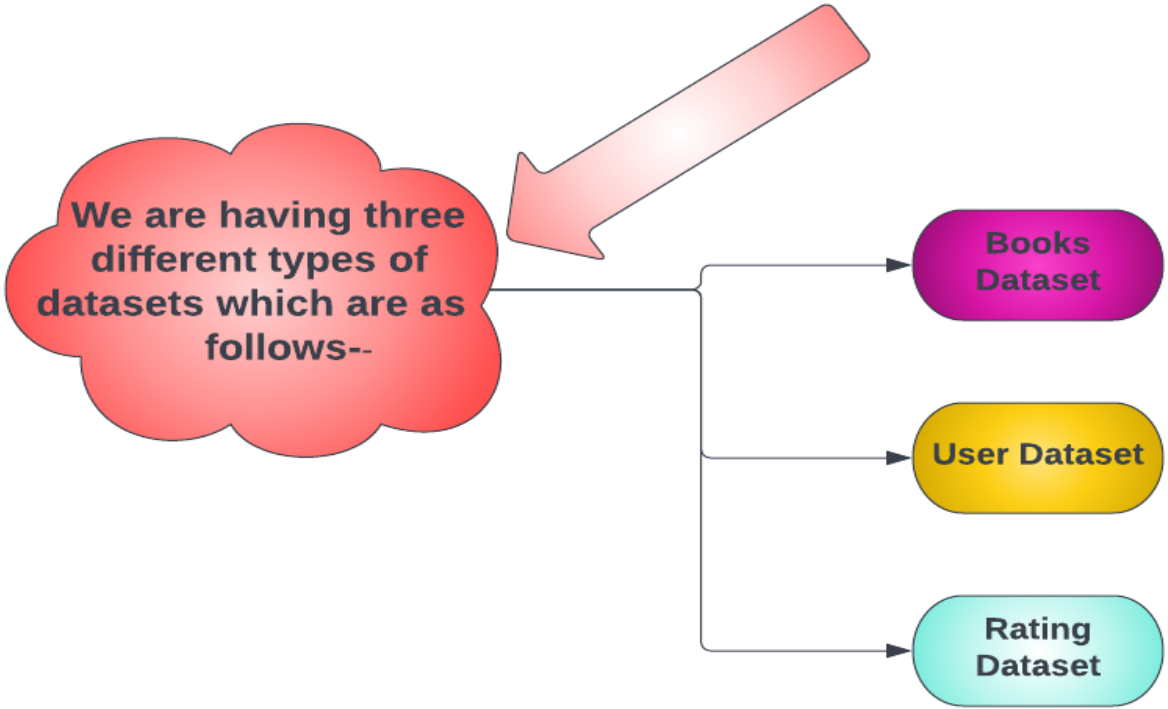
Content

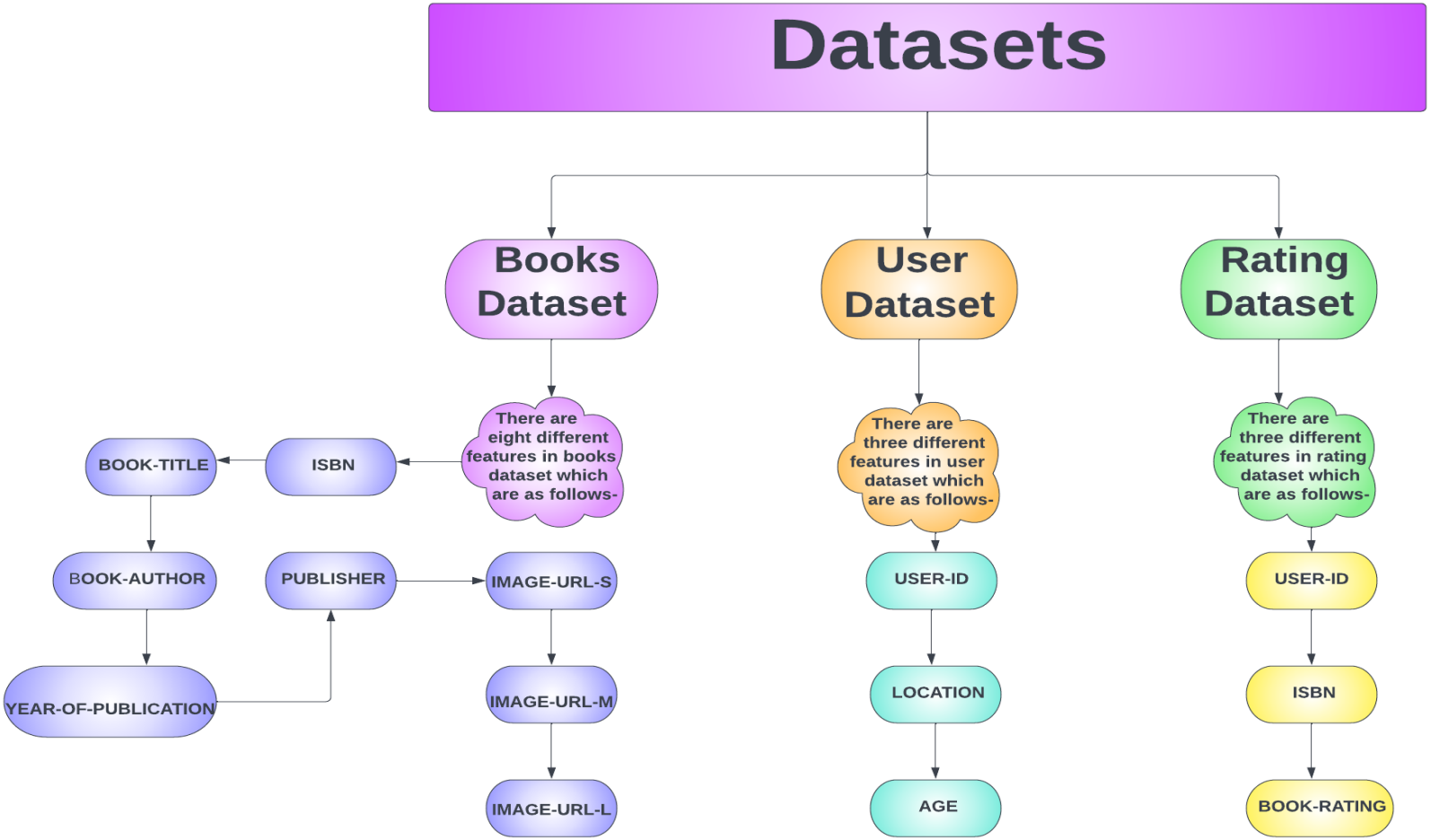

1.) Problem Statement
2.) Data Summary
3.)Data Preprocessing
Topics for Discusssion
4.)EDA
5.) Feature Engineering
6.)Reommendation Systems
7.) Conclusion

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




Datasets
Books Dataset
User Dataset
Rating Dataset
There are eight different features in books dataset which are as follows-
There are three different features in user dataset which are as follows-
There are three different features in rating dataset which are as follows-
BOOK-TITLE
ISBN
BOOK-AUTHOR
PUBLISHER
IMAGE-URL-S
YEAR-OF-PUBLICATION
IMAGE-URL-M
IMAGE-URL-L
USER-ID
LOCATION
AGE
USER-ID
ISBN
BOOK-RATING

Data Preprocessing

We preprocessed all the three datasets.

Books-Data Preprocessing

1.) In Books-data preprocessing,we handled missing values and duplicated values.

2.) Handled outliers in Year-of-Publication column.

3.) Corrected values in Book-Author, Book-Title, and Publisher column.

4.) Dropped image-url-s,image-url-m, and image-url-l columns as they were not of much importance.

User-Data Preprocessing

1.) In User-data preprocessing,we took the age group of 6 to 90.

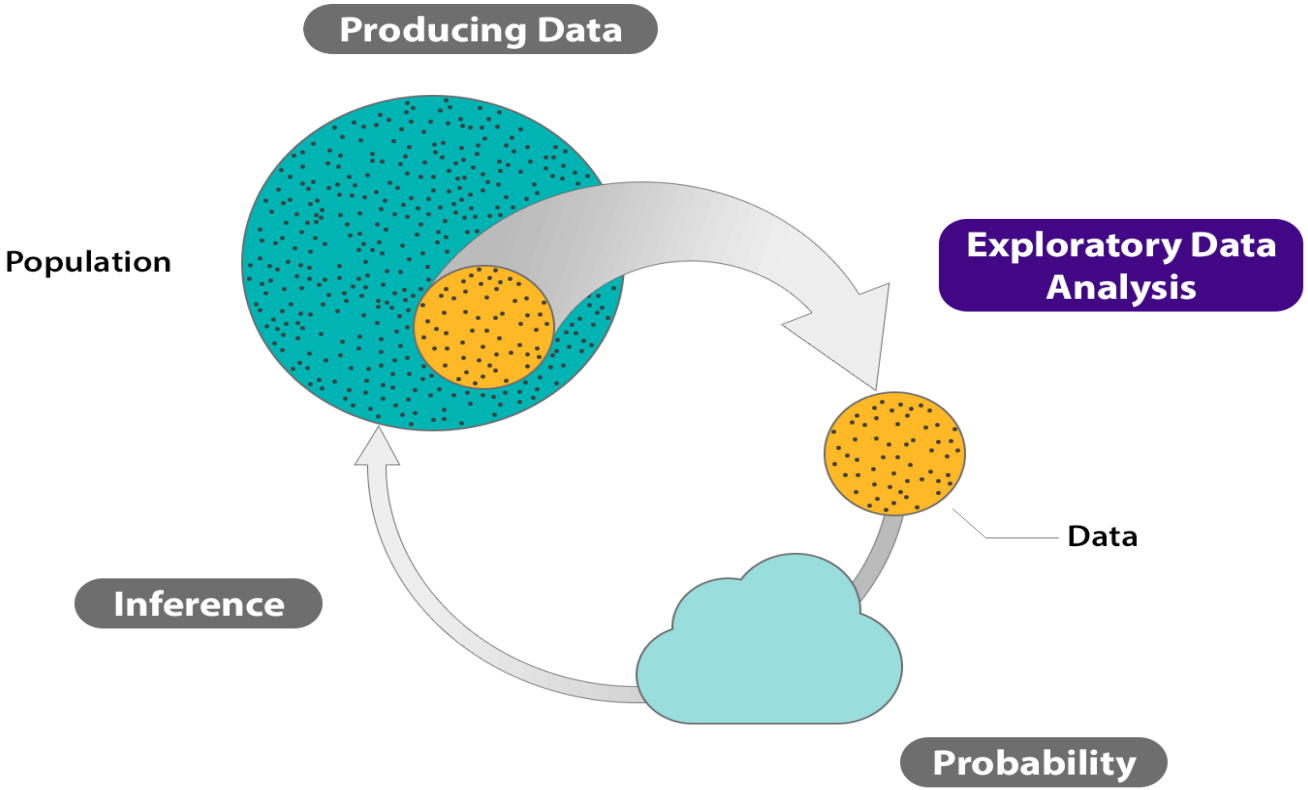
2.) Extracted City,State,and Country features from Location Column.

3.) Handled outliers,missing values, and duplicated values in age column.

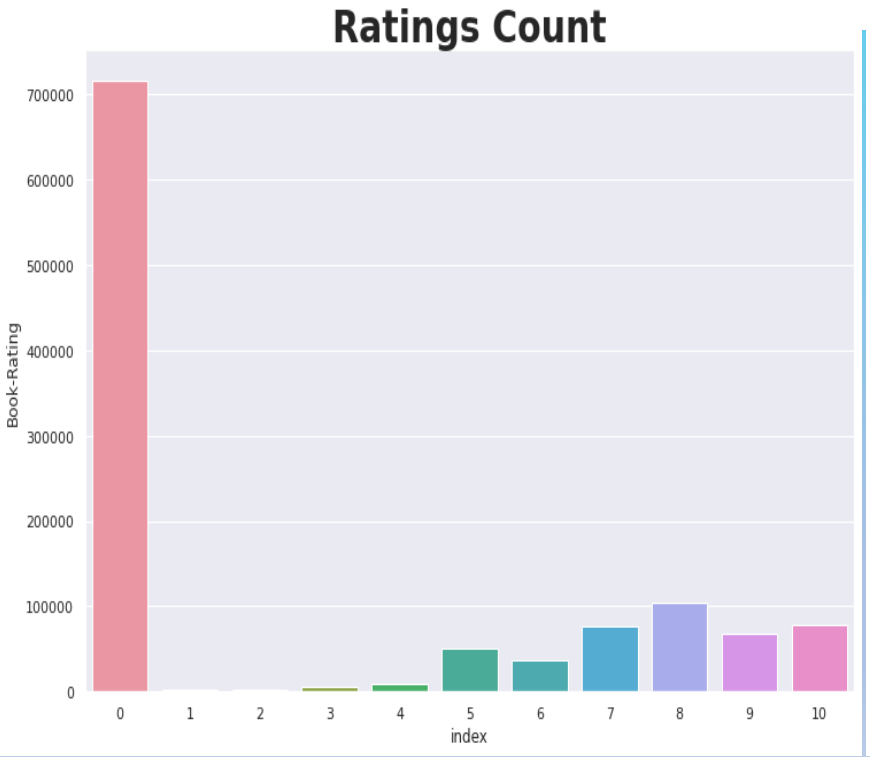
Rating-Data Preprocessing

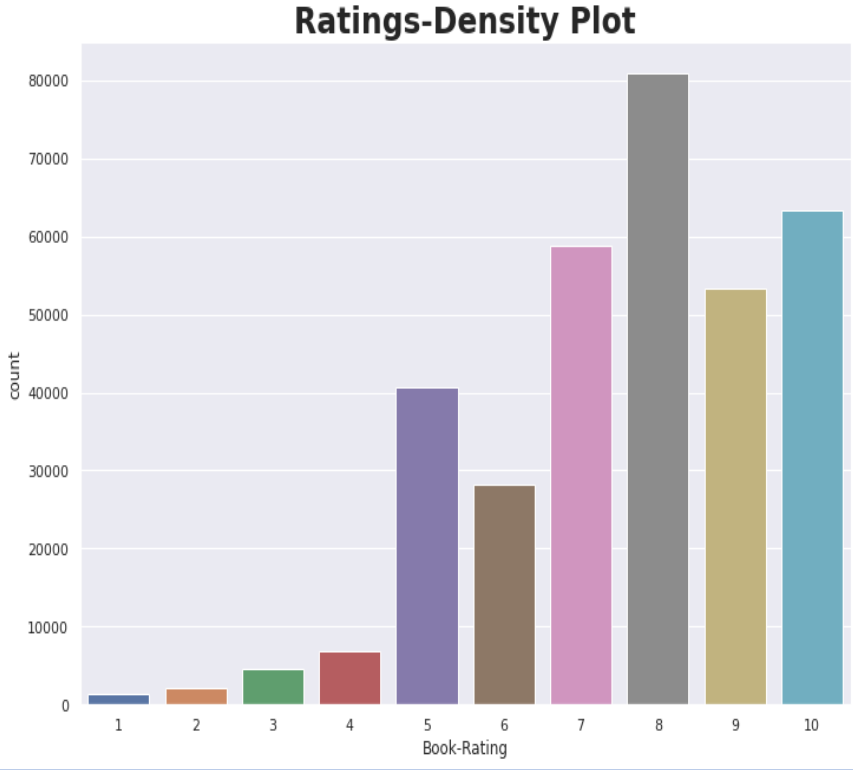
1.) In Rating-data preprocessing, we seprated book-rating into rating implicit and rating explicit.

2.) Handled outliers,missing values, and duplicated values in book-rating.


Producing Data
Population
Exploratory Data Analysis
Data
Inference
Probability

Primary EDA: Ratings, (Explicit + Implicit) vs Explicit





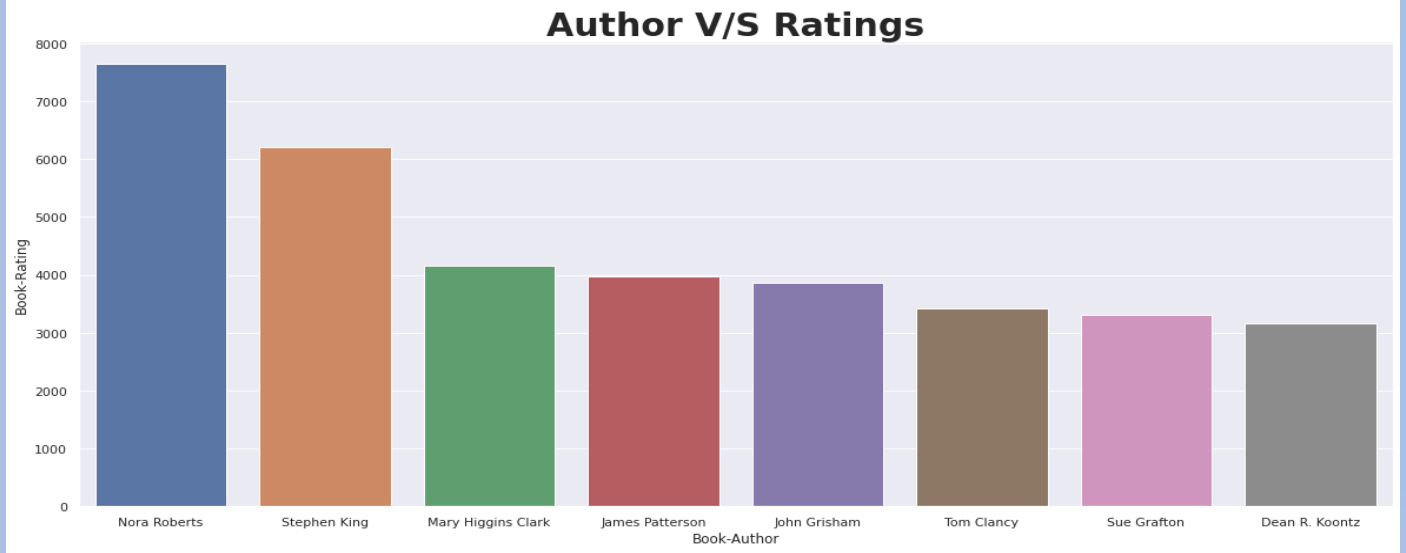
Insights

From the above bar graphs,we can see the following insights--

1.) We can see that most of the Rating is zero that is approximately 700000.

2.) In the Rating density plot, we can see that most of the books have been rated as 8, then 10,then 7 and so on.

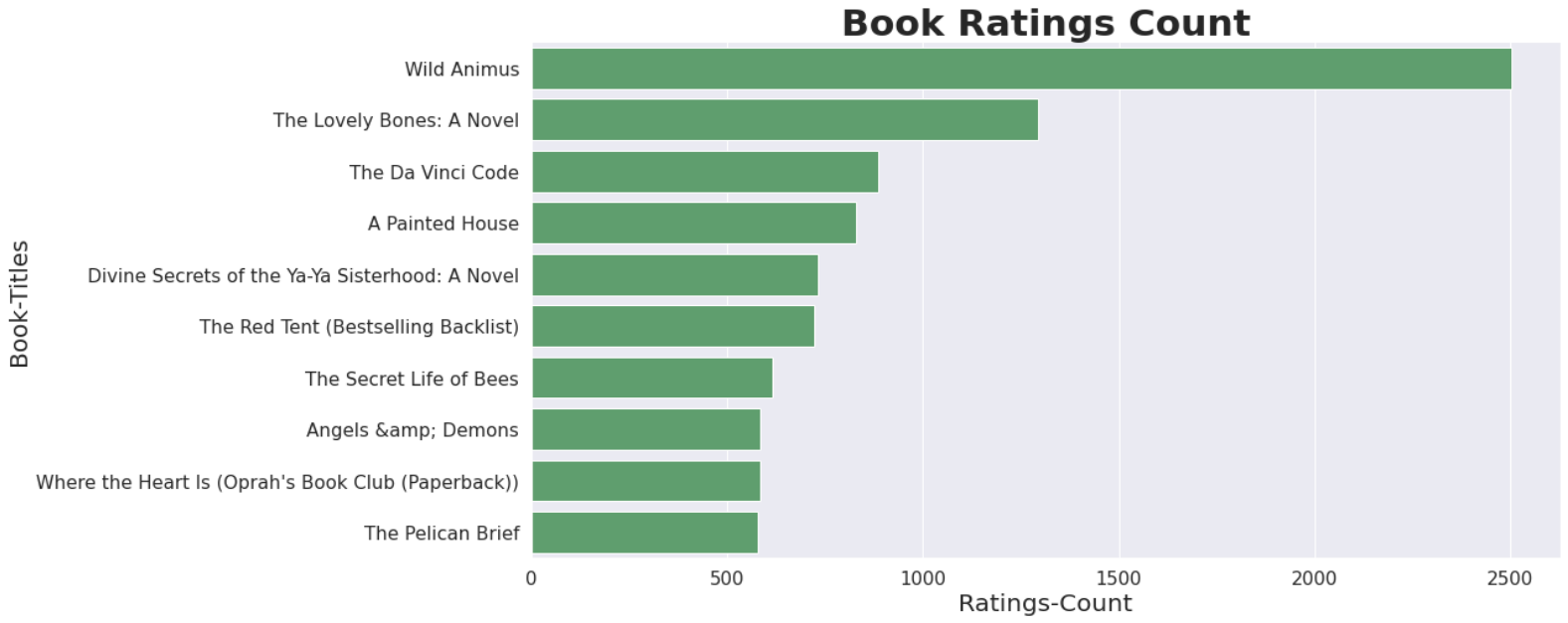
Primary EDA: Author vs Ratings



❑ Insights:-

- ❖ Here, we can observe, most frequently rated Authors.
- ❖ Most frequently rated author is Nora Roberts, followed by Stephen King

Primary EDA: Most Frequently Rated Books

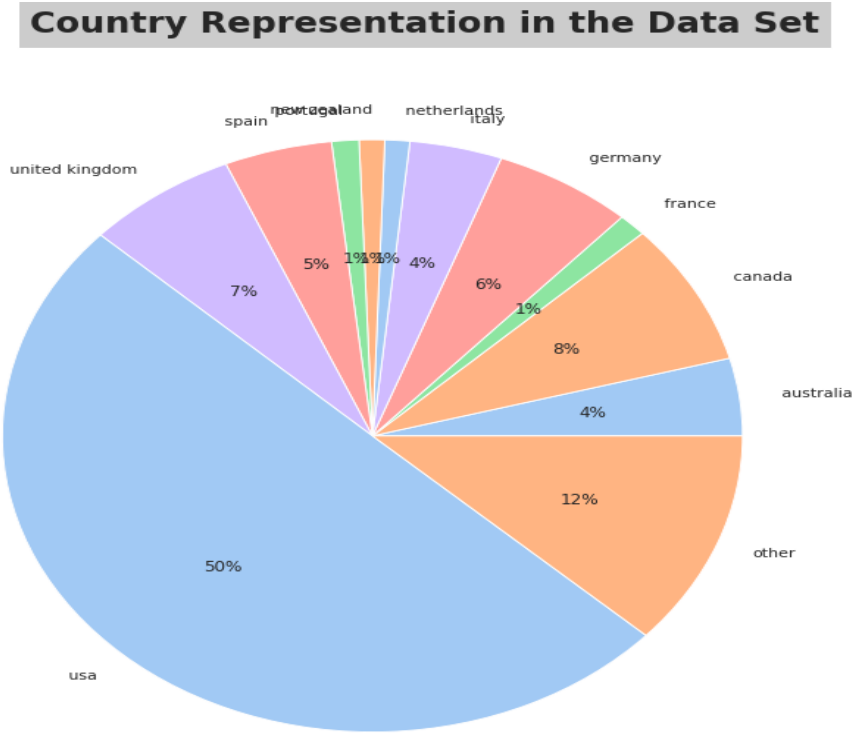


- **Insights:-**
  - Here, we are able to observe, most frequently rated books by the users.
  - Most frequently rated book, happens to be “Wild Animus”.

Primary EDA: Country Representation in the Dataset

❑ Insights:-

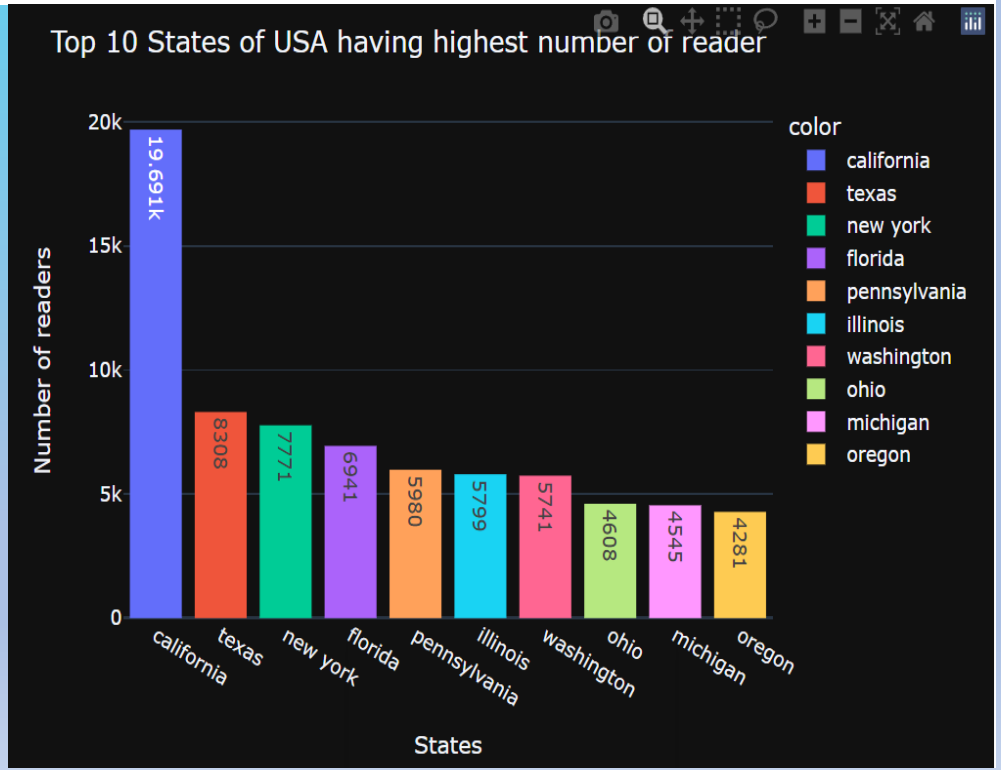
- ❖ Most customers are from the United States of America, followed by Canada, United Kingdom and Germany.
- ❖ Countries with less than 1% customers are labelled as other.



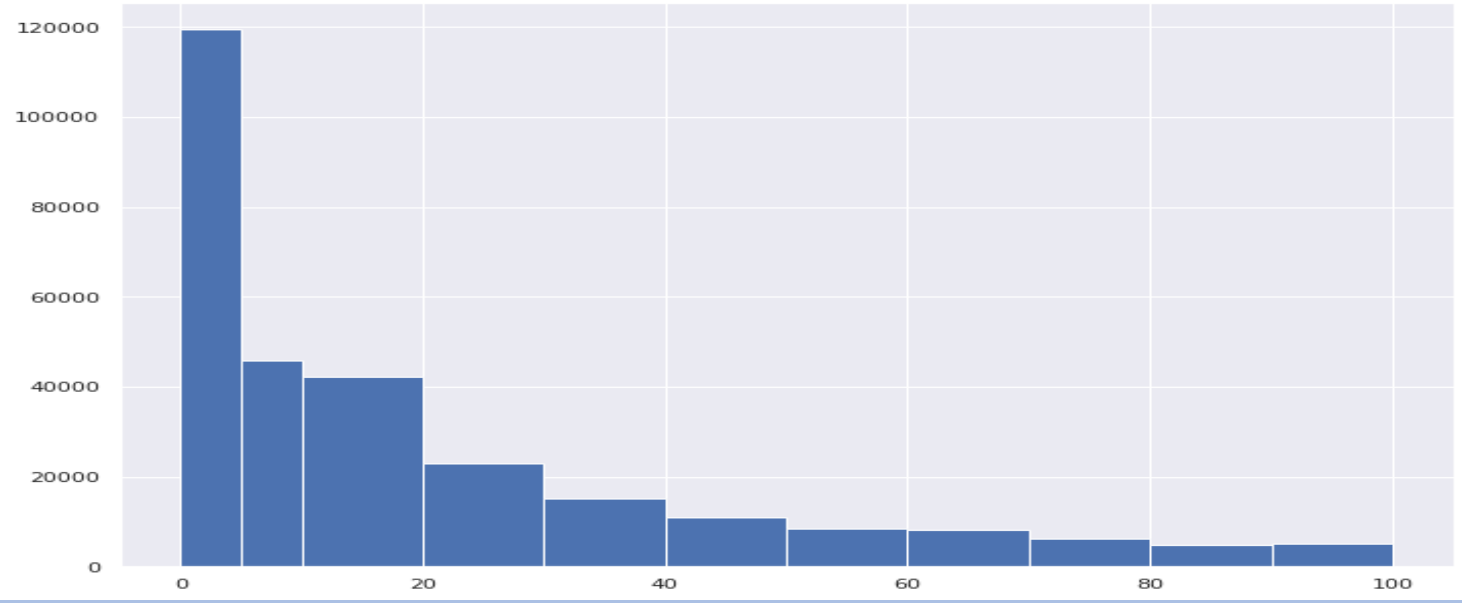
Primary EDA: Top 10 States of USA

❑ Insights:-

- ❖ We can see here that “California” State has highest number of readers of 19.691k.
- ❖ “Texas” is the second state with highest number of user of 8,308.
- ❖ “Oregon” state has least number of users.



Primary EDA:Age vs Rating Density



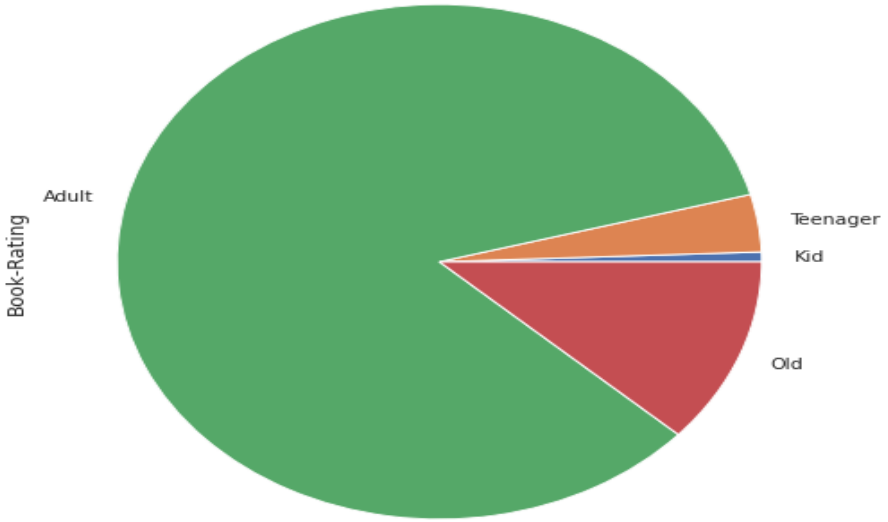
❑ Insights:-

- ❖ Here, we are able to observe, which age bin has contributed most to the Book-Ratings.

Primary EDA: Age Bin Representation in the Dataset

❑ Insights:-

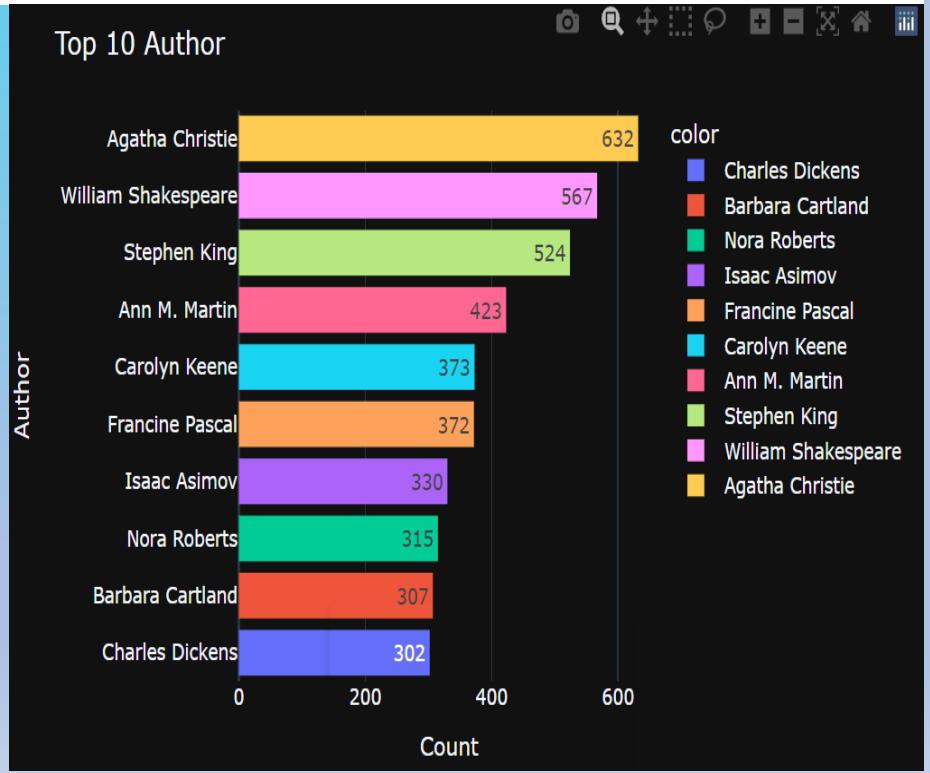
- ❖ Most customers are Adults (20-50yrs).
- ❖ 2nd most represented age group is for old (>50yrs).



Primary EDA: Top 10 Authors

❑ Insights:-

- ❖ We can see that “Agatha Christie” is the best author who has written 632 books.
- ❖ “William Shakespeare” is the second best author who has written 567 books.

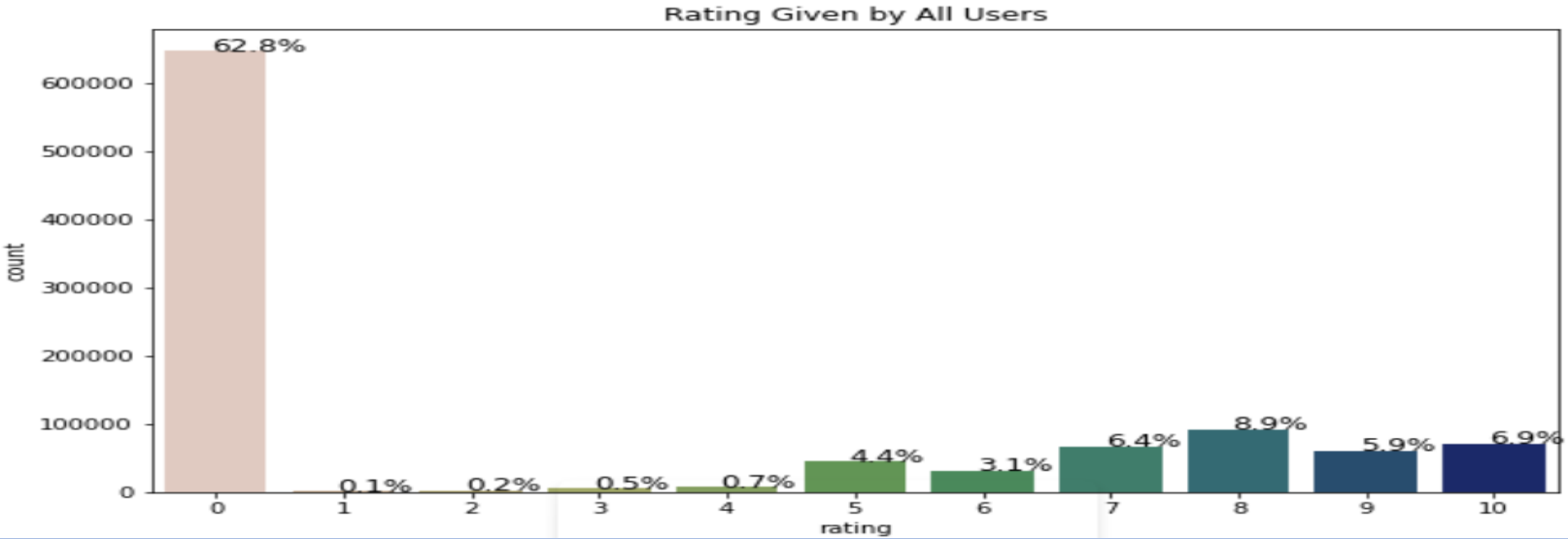


Primary EDA: Top 10 Publishers

❑ Insights:-

- ❖ We can see that “Harlequin” is the best publisher which has published 7,535 books.
- ❖ “Silhouette” is the second best publisher which has published 4,220 books.

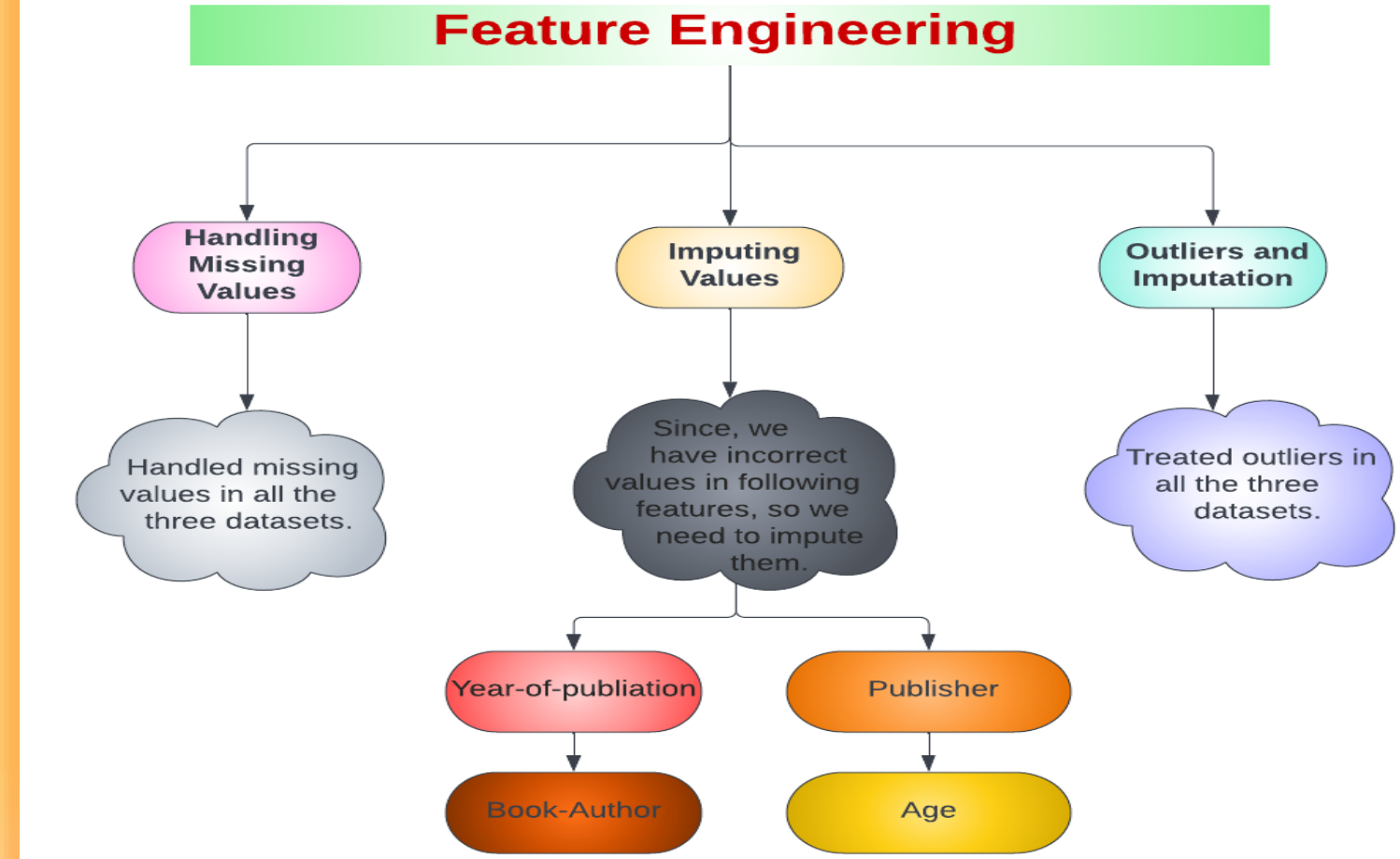
Primary EDA: Rating Given by All Users



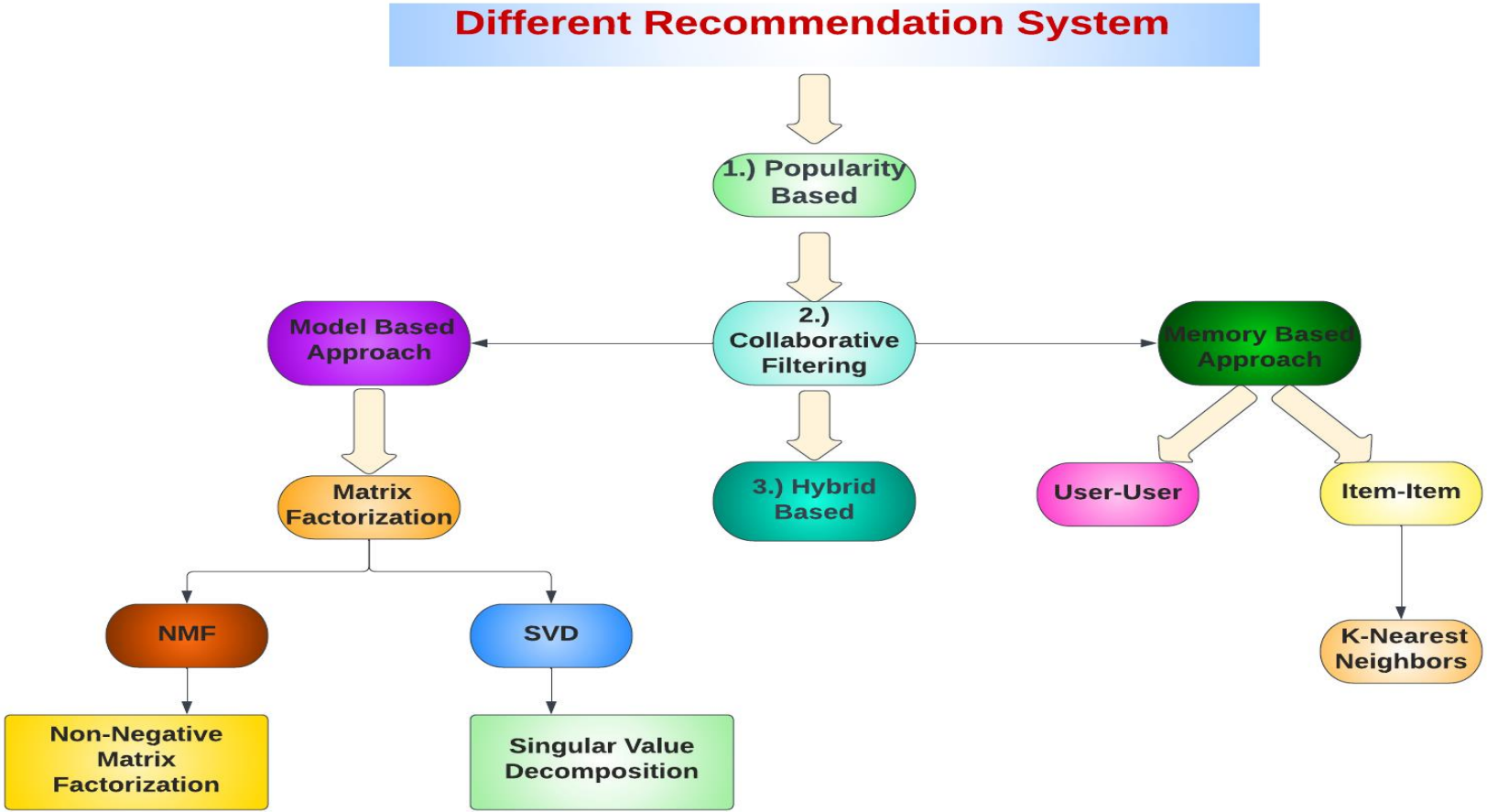
❑ Insights:-

- ❖ We can see that most of the user has given rating as 0 which is approximately 62.8%.

Feature Engineering


Feature Engineering
Handling Missing Values
Imputing Values
Outliers and Imputation
Handled missing values in all the three datasets.
Since, we have incorrect values in following features, so we need to impute them.
Treated outliers in all the three datasets.
Year-of-publiation
Publisher
Book-Author
Age

Recommendation System


Different Recommendation System
1.) Popularity Based
2.) Collaborative Filtering
3.) Hybrid Based
Model Based Approach
Memory Based Approach
Matrix Factorization
User-User
Item-Item
NMF
SVD
K-Nearest Neighbors
Non-Negative Matrix Factorization
Singular Value Decomposition

Popularity Based Recommender System

- So, we can see that popularity metric we can calculate the top books that could be recommended to a user.

	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

Collaborative Filtering (Memory Based Approch)

Memory Based KNN Model

- A KNN model, with cosine similarity as a metric for measuring the distance between different ratings, was used to provide recommendations.

```
recommend('9-11 by Noam Chomsky', n_values=10)
```

```
The Top 9 Recommendations for Users who have read book 9-11 by Noam Chomsky are shown below:-
1: Die Weiss Lowin / Contemporary German Lit by Henning Mankell, with distance of 0.6220355269907728.
2: The First Counsel by Brad Meltzer, with distance of 0.6220355269907728.
3: Schlafes Bruder by Robert Schneider, with distance of 0.6220355269907728.
4: Herzsprung by Ildiko Kurthy, with distance of 0.6220355269907728.
5: Due di due (Bestsellers) by Andrea De Carlo, with distance of 0.6220355269907728.
6: MÃ?Â¶rder ohne Gesicht. by Henning Mankell, with distance of 0.6220355269907728.
7: UN Viejo Que Leia Novelas De Amor/the Old Men Who Read Love Stories (ColecciÃ³n Andanzas) by Luis Sepulveda, with distance of 0.6220355269907728.
8: Vernon God Little: A 21st Century Comedy in the Presence of Death by D. B. C. Pierre, with distance of 0.6220355269907728.
9: Lauf, Jane, lauf. Roman. by Joy Fielding, with distance of 0.6220355269907728.
```

Insight:-

- We can see, that the recommended books, are quite similar in genre to the selected item

Collaborative Filtering (Model Based Approch)

- It makes predictions about the interests of a user by collecting preferences from many users. The underlying assumption is, if a person A has the same opinion as a person B on a set of items, A is more likely to have B's opinion for a given item than that of a randomly chosen person.

```
Global metrics:
{'modelName': 'Collaborative Filtering', 'recall@5': 0.22556281771968045, 'recall@10': 0.29784555797627693}
```

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	User-ID
10	293	356	1389	0.210943	0.256299	11676
31	182	236	1138	0.159930	0.207381	98391
45	27	36	380	0.071053	0.094737	189835
30	90	110	369	0.243902	0.298103	153662
70	28	36	236	0.118644	0.152542	23902
7	29	36	204	0.142157	0.176471	235105
47	21	29	203	0.103448	0.142857	76499
50	27	38	193	0.139896	0.196891	171118
42	65	78	192	0.338542	0.406250	16795
43	22	30	188	0.117021	0.159574	248718

Collaborative Filtering (Model Based Approch)

- We can see, the user: 40943, has rated Harry Potter and the Sorcerer's Stone (Book 1), very highly. Our model, is recommending other parts of the same series. This seems to be consistent with high precision and high recall values that we have obtained thus far.

	User-ID	ISBN	Book-Rating	Book-Title
367478	40943	0671003755	5	She's Come Undone (Oprah's Book Club (Paperback))
367497	40943	0679746048	8	Girl, Interrupted
367499	40943	039480967X	5	Bears on Wheels (Bright & Early Books)
367514	40943	043936213X	10	Harry Potter and the Sorcerer's Stone (Book 1)
367518	40943	0553274295	10	Where the Red Fern Grows

```
recc
```

```
array(['Harry Potter and the Chamber of Secrets (Book 2)',
       'Harry Potter and the Prisoner of Azkaban (Book 3)',
       'Harry Potter and the Goblet of Fire (Book 4)',
       "Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))",
       'The Secret Life of Bees',
       'Harry Potter and the Order of the Phoenix (Book 5)',
       "Harry Potter and the Sorcerer's Stone (Book 1)",
       "Bridget Jones's Diary",
       'The Fellowship of the Ring (The Lord of the Rings, Part 1)',
       'The Nanny Diaries: A Novel'], dtype=object)
```

❑ 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 requirements .
- It is very important to deal with implicit and explicit user ratings, separately.
- For dealing with explicit ratings, we can build simple models based on ratings, and we can also use certain comprehensive models based on Collaborative Filtering approach.
- For dealing with implicit ratings, we can build KNN based models , and we can also use content based models, which utilize the similarity of different contents, to make recommendations.
- It is crucial to be precise about user preferences, otherwise repetitive recommendations can cause nuisance to the user.

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.


Thank
You