Capstone Project

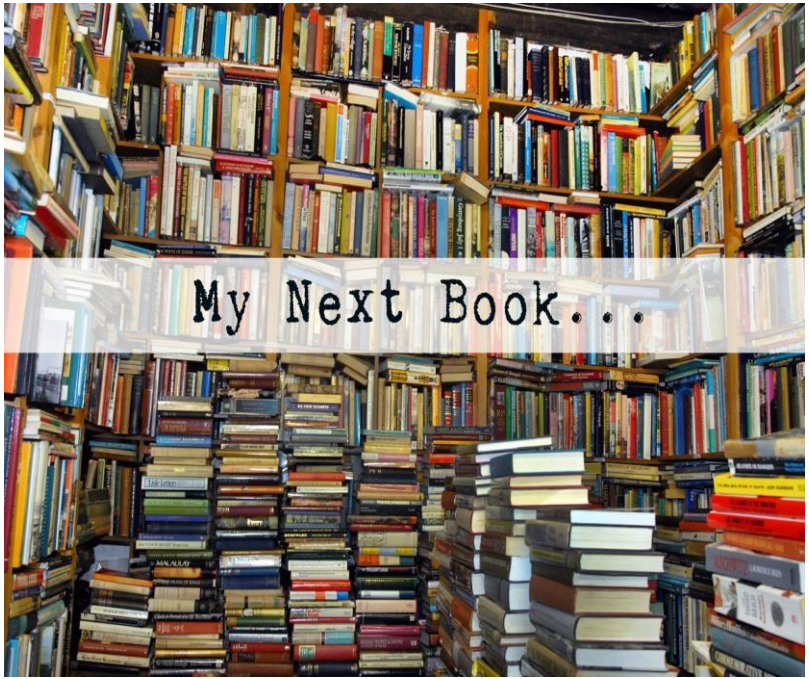
Book Recommendation System

By-Pooja Potdar

Content

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

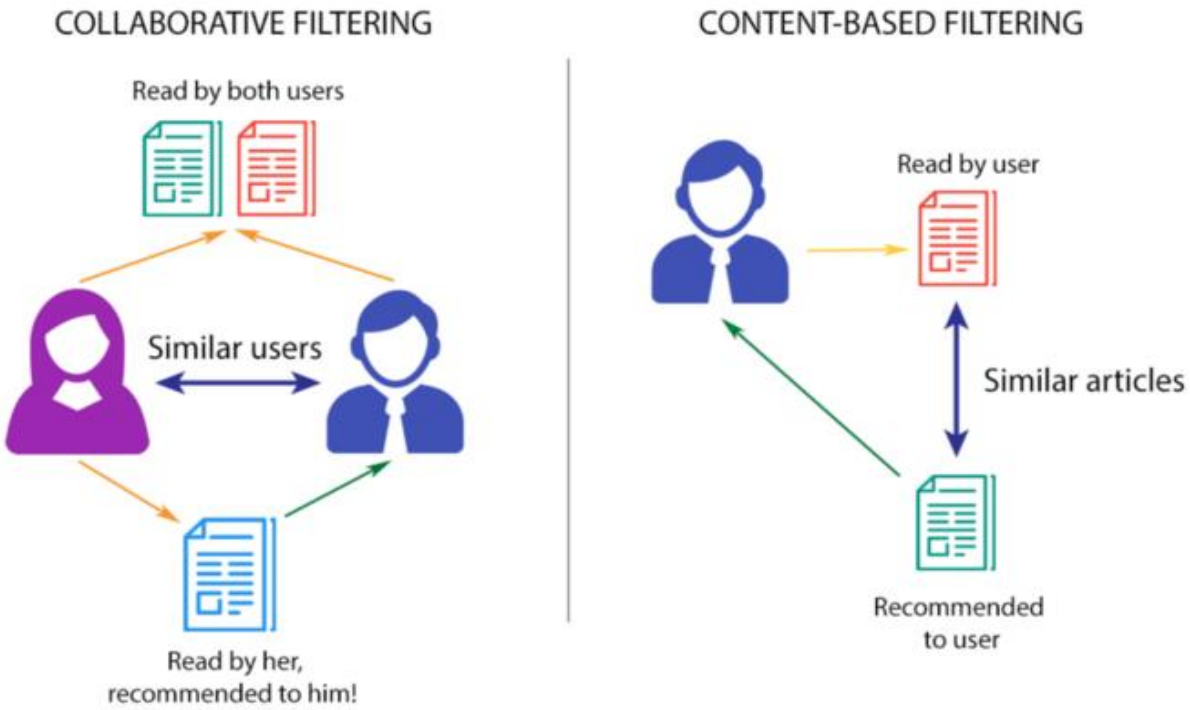
Problem Statement



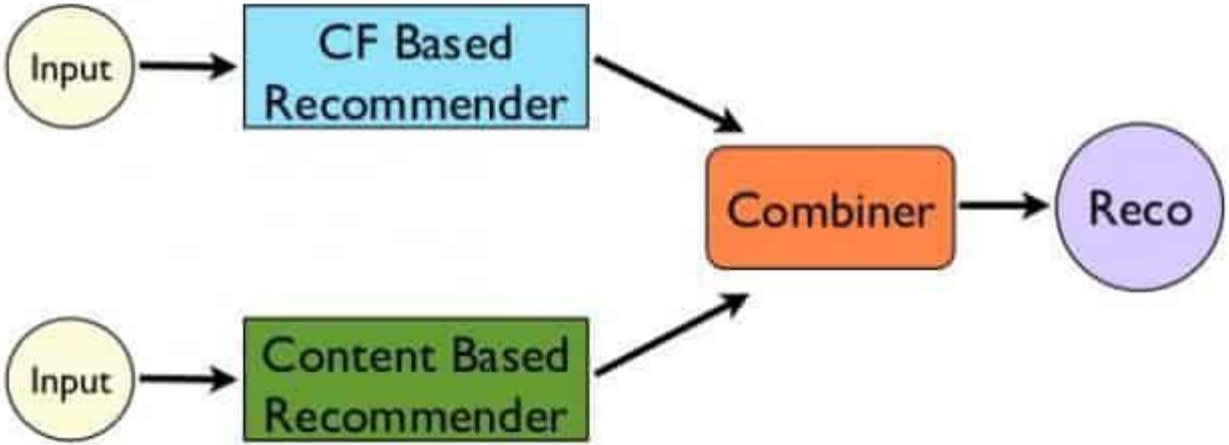
During the last few decades, with the rise of Youtube, Amazon, Netflix, and many other such web services, recommender systems have become much more important in our lives in terms of providing highly personalized and relevant content.

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

Types of filtering are:



Hybrid Recommendations



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

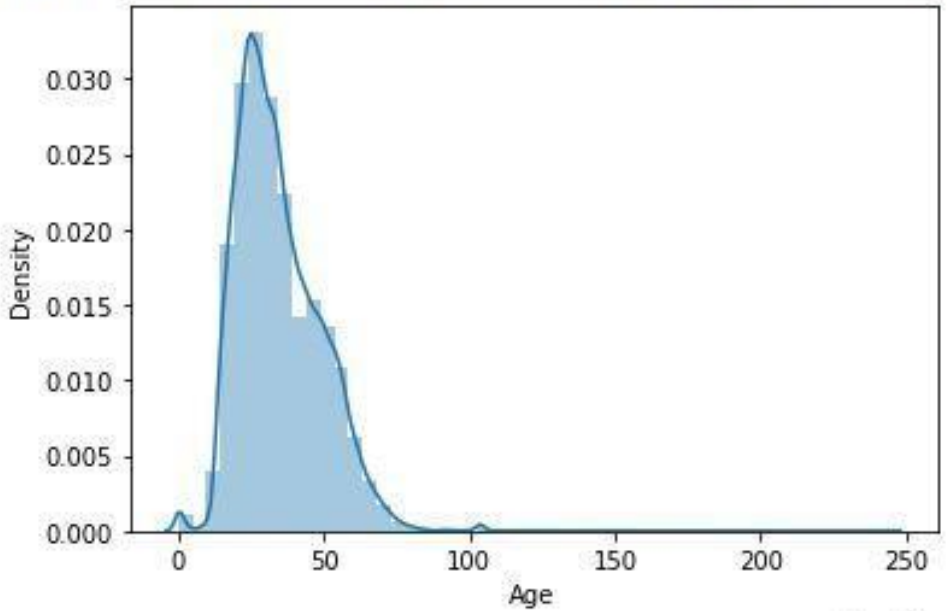
- Book-Rating
- Shape of Dataset - (1149780, 3)

Observations from Users_df (Age)

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

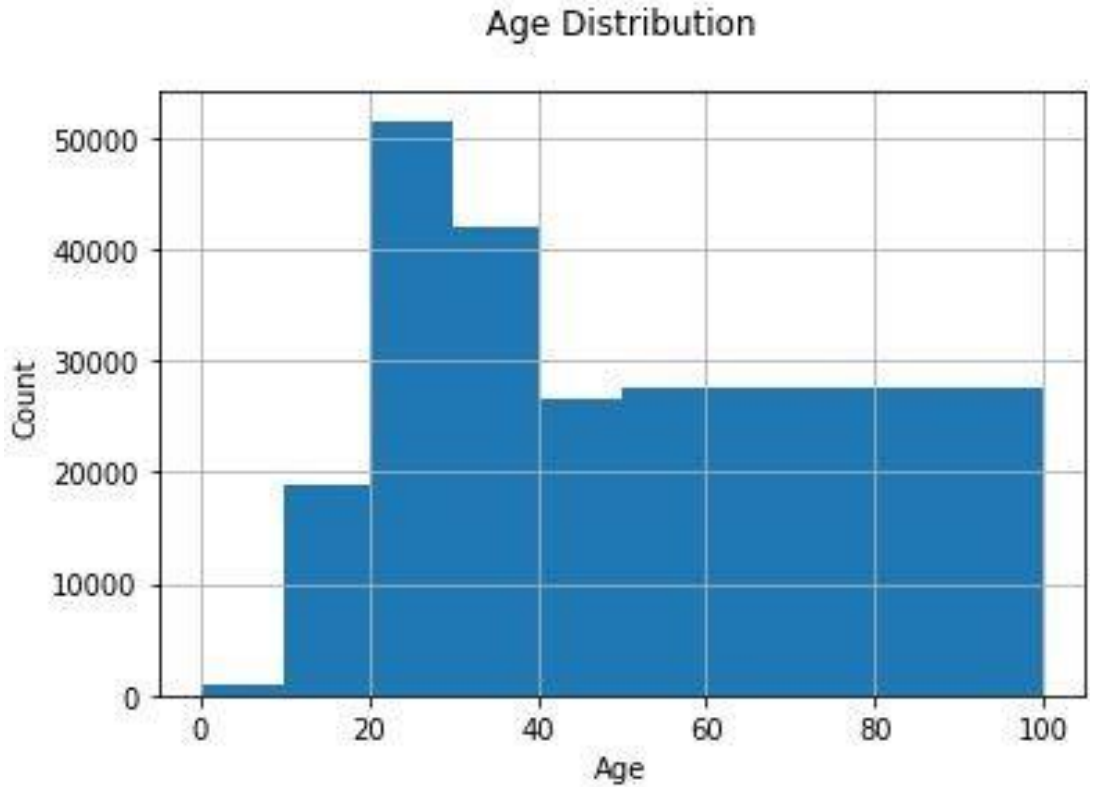
```
1 sns.distplot(users.Age)
```

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



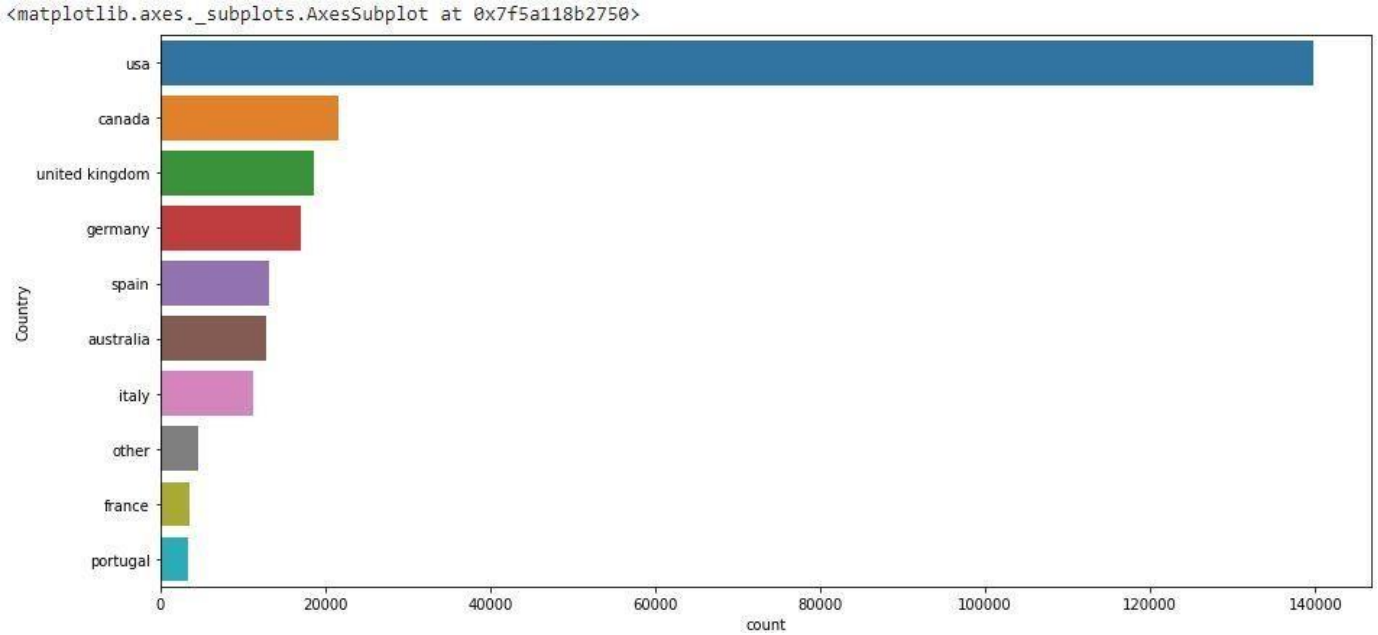
Observations from Users_df (Age)

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



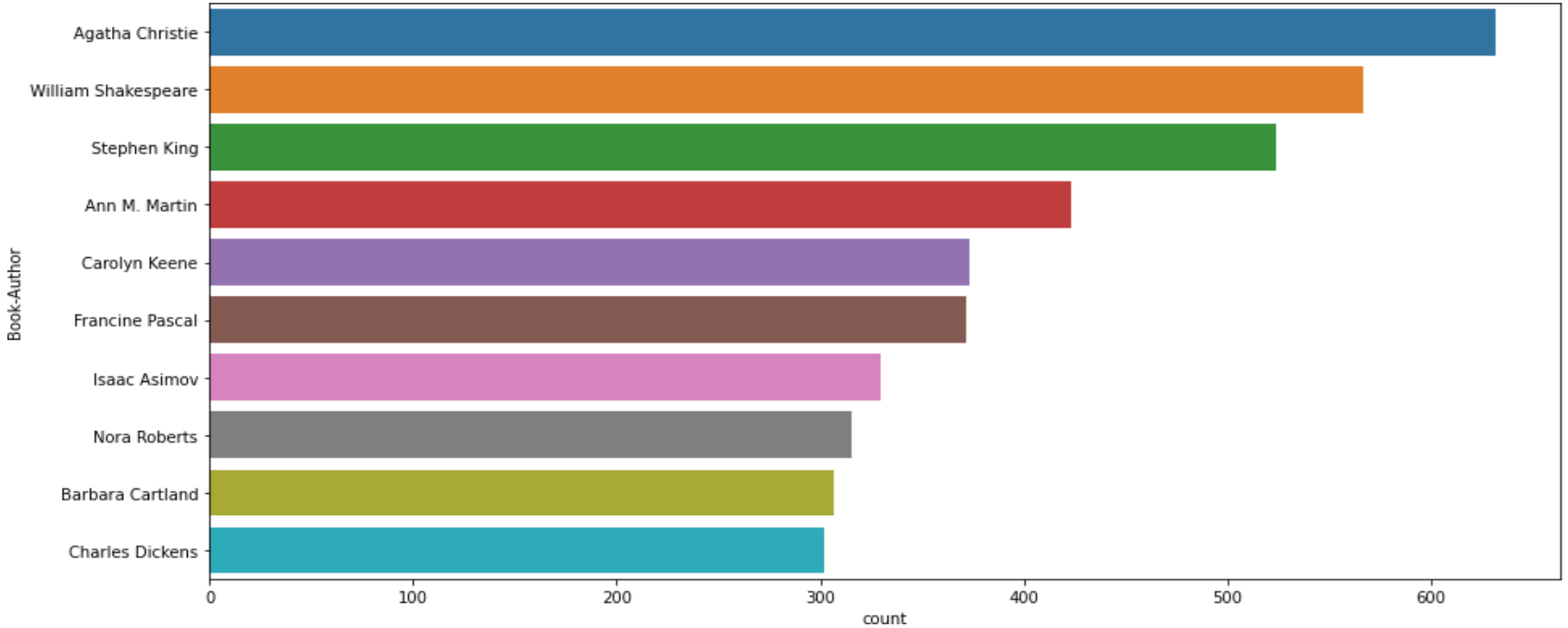
Observations from Users_df (Location)

- Splitting Location column and analysing country.
- Most active readers are from USA.



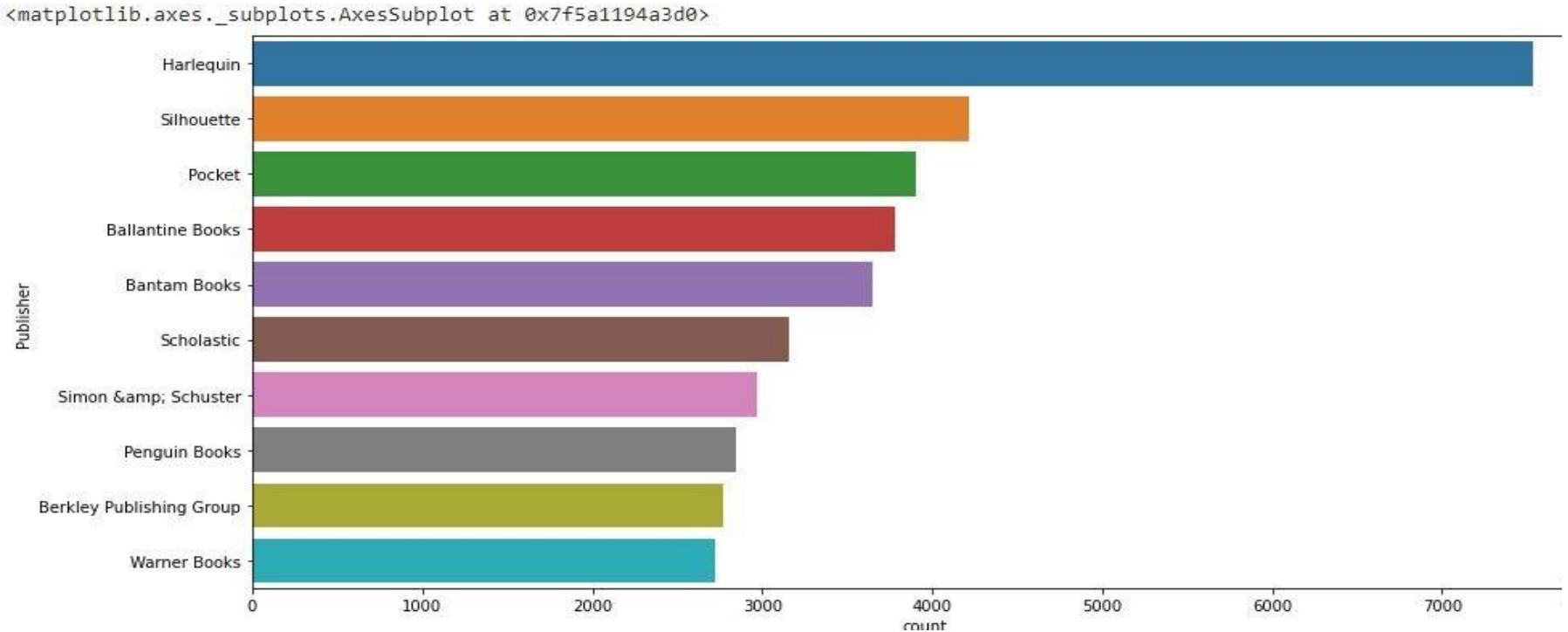
Observations from Book_df (Authors)

Agatha Christie wrote highest number of books in our given dataset



Observations from Book_df (Publishers)

Harlequin published highest number of books in our given dataset



Data Cleaning

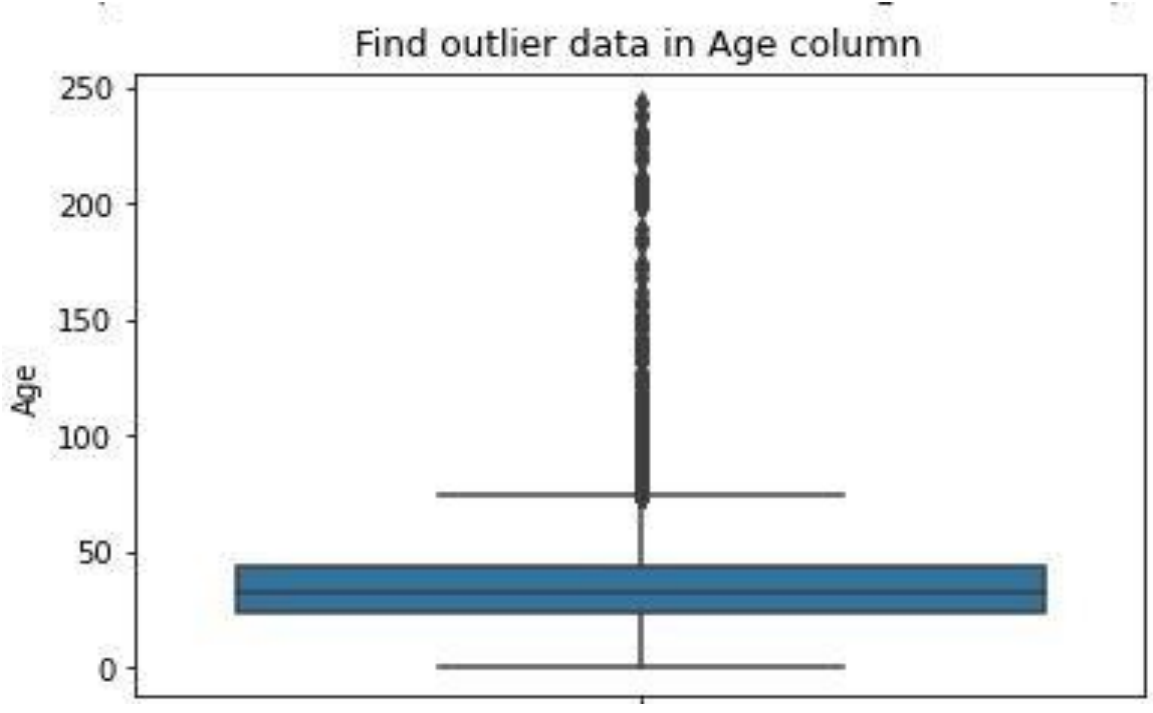
1. **Null Value Imputation:**

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

Imputing missing values

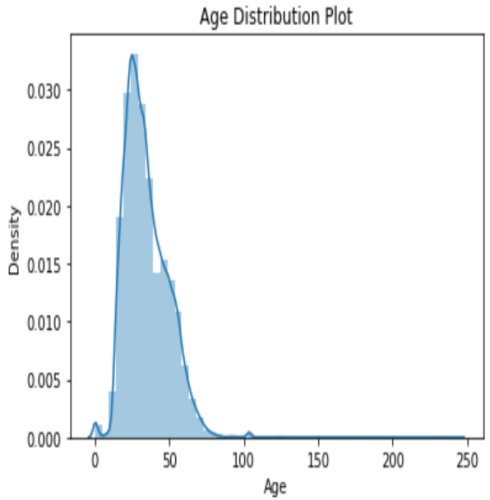
- Outliers in Age column
- Age has positive Skewness (right tail) so we can use median to fill Nan values,



Age Displot

```
sns.distplot(users.Age)
plt.title('Age Distribution Plot')
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be
  warnings.warn(msg, FutureWarning)
Text(0.5, 1.0, 'Age Distribution Plot')
```



✓ 8s completed at 3:13 PM

age value's below 5

and above 100 do not make much sense for our book rating sense so considering it as outliers and removing it

Data Cleaning

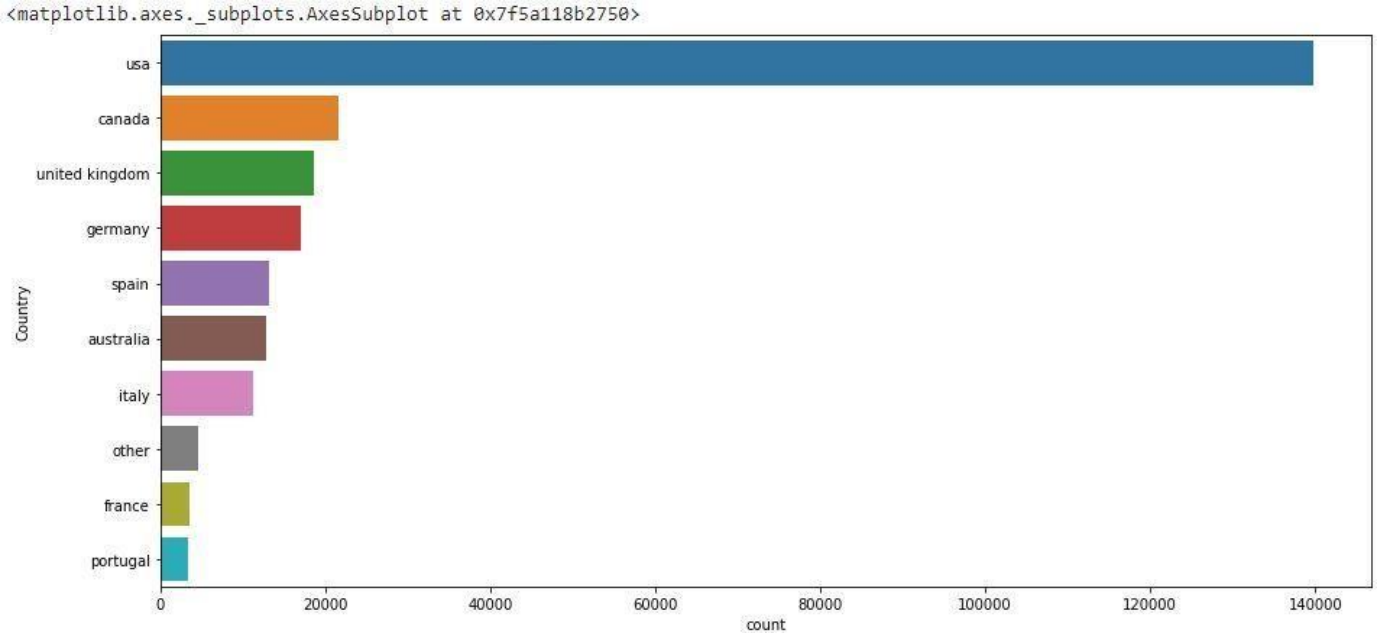
1. Null Value Imputation:

```
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
```

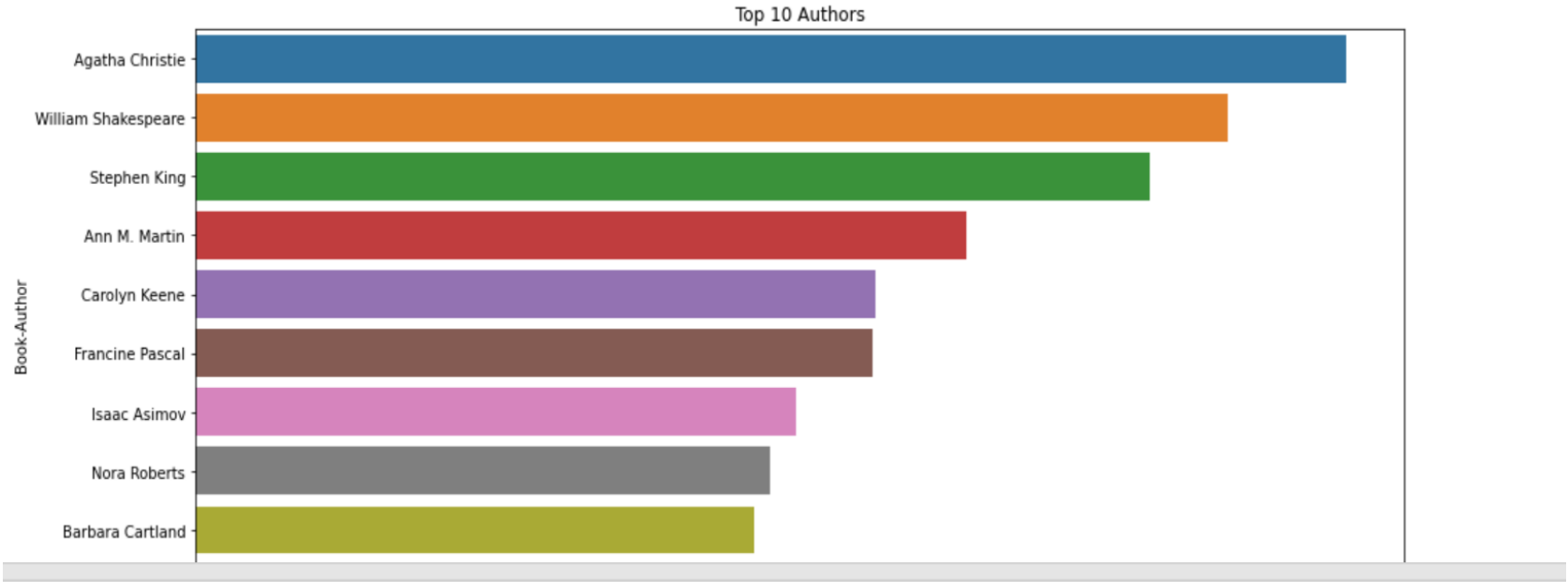
Observations from Users_df (Location)

- Splitting Location column and analysing country.
- Most active readers are from USA.



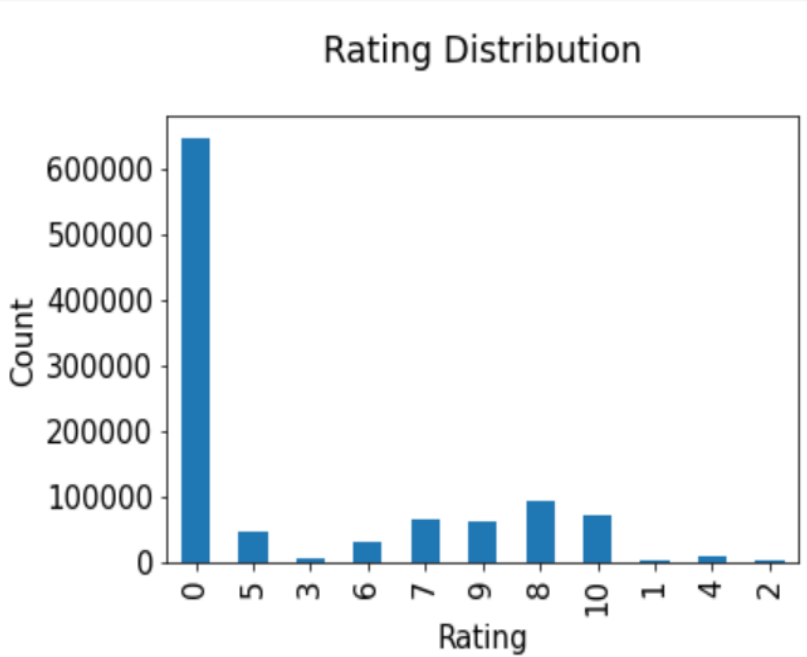
Top 10 authors:

```
Text(0.5, 1.0, 'Top 10 Authors')
```



✓ 8s completed at 3:13 PM

Rating Distribution



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) .

Conclusion

A recommendation system helps an organization to create loyal customers. The recommendation system today are very powerful that they can handle the new customer too who has visited the site for the first time. They recommend the products which are currently trending or highly rated and they can also recommend the products which bring maximum profitto the company.

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.

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