



**Capstone Project**

**Book Recommendation System**

**Rohit Madame**

- ❖ Introduction
- ❖ Problem Statement
- ❖ Data Summary
- ❖ Data Wrangling
- ❖ EDA
- ❖ Model used to build recommendation system
- ❖ Challenges
- ❖ Conclusion

- ❖ During the last few decades, with the rise of YouTube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys.
- ❖ In a very general way, recommender systems are algorithms aimed at suggesting relevant
  - items to users (items being movies to watch, text to read, products to buy, or anything else depending 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 book recommendation system for users.
- ❖ The main objective is to predict the same genre books.

# Data Summary

- The Book-Crossing dataset comprises 3 files.

## ❖ Users

- Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these
- fields contain NULL values.

## ❖ Books

- Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title,
- Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web
- Services. Note that in the case of several authors, only the first is provided. URLs linking

to cover images are also given, appearing in three different flavors (Image-URL-S,

- Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the
- Amazon website.

## ❖ Ratings

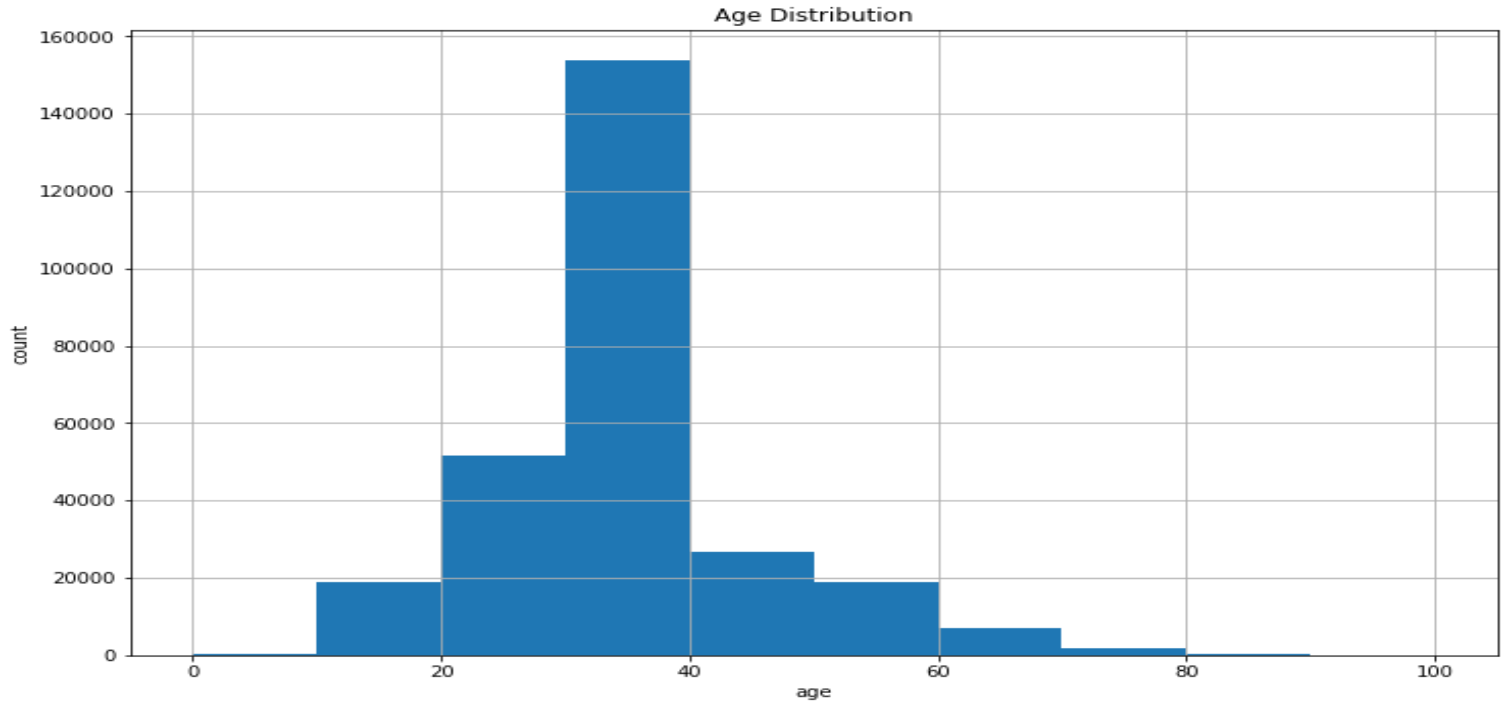
- Contains the book rating information. Ratings (Book-Rating) are either explicit,
- expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit,
- expressed by 0

# Data Wrangling



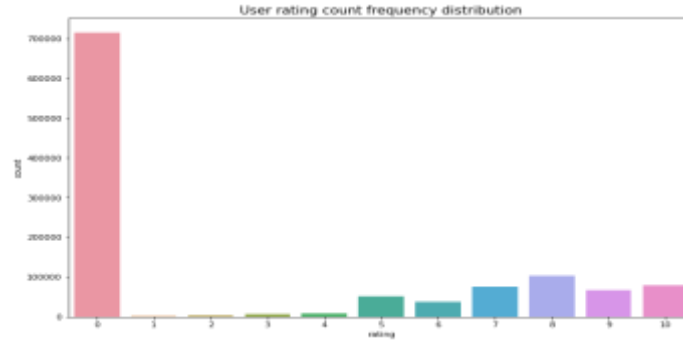
- ❖ First We are changing columns name as it contains spaces and hyphen.
- ❖ Then dropping image column as we don't need to show image.
- ❖ Book data having year as data type object so we need to convert it to proper type.
- ❖ Year column contains missing values so we have filled it with mean.

- ❖ Most of the book readers are between the age group of 30 to 40.





- ❖ In the below figure most reader's did not rated , it will impact our prediction.



- ❖ So we again plotting the count plot dropping zero rating



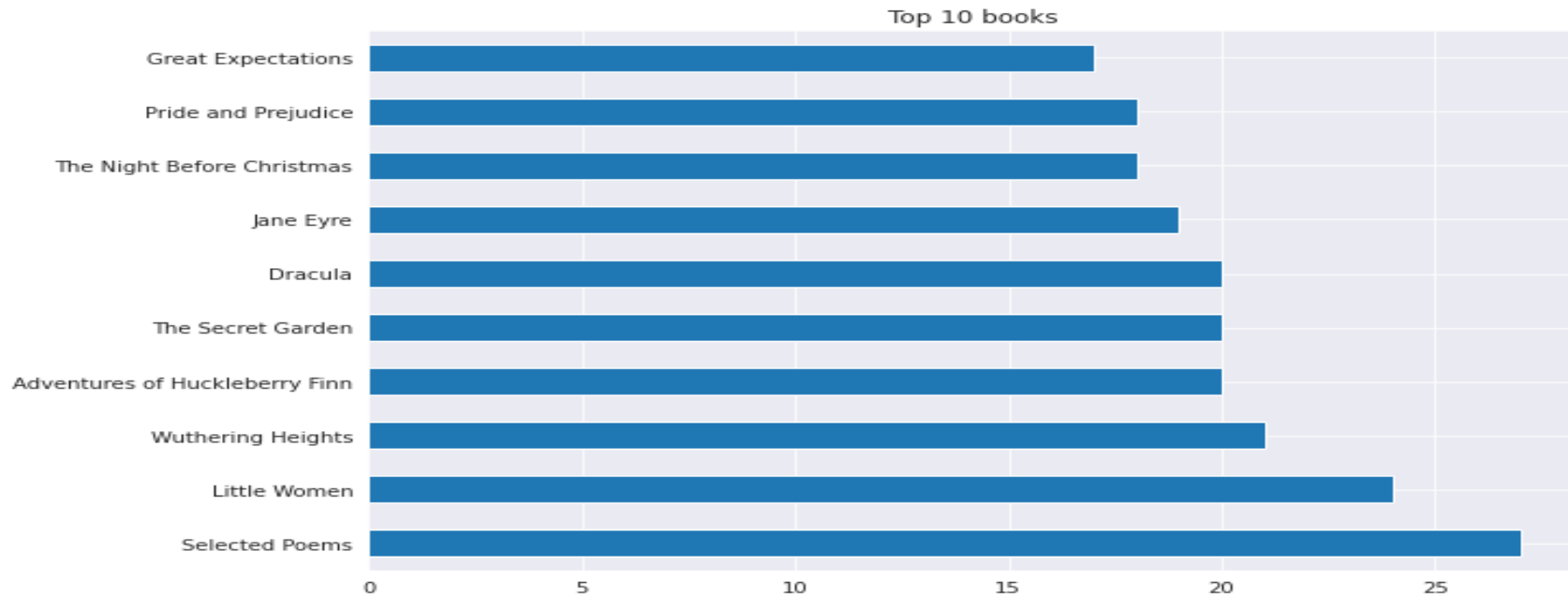
❖ Maximum number of books published in year 1996 followed by 2002 and year 1999.



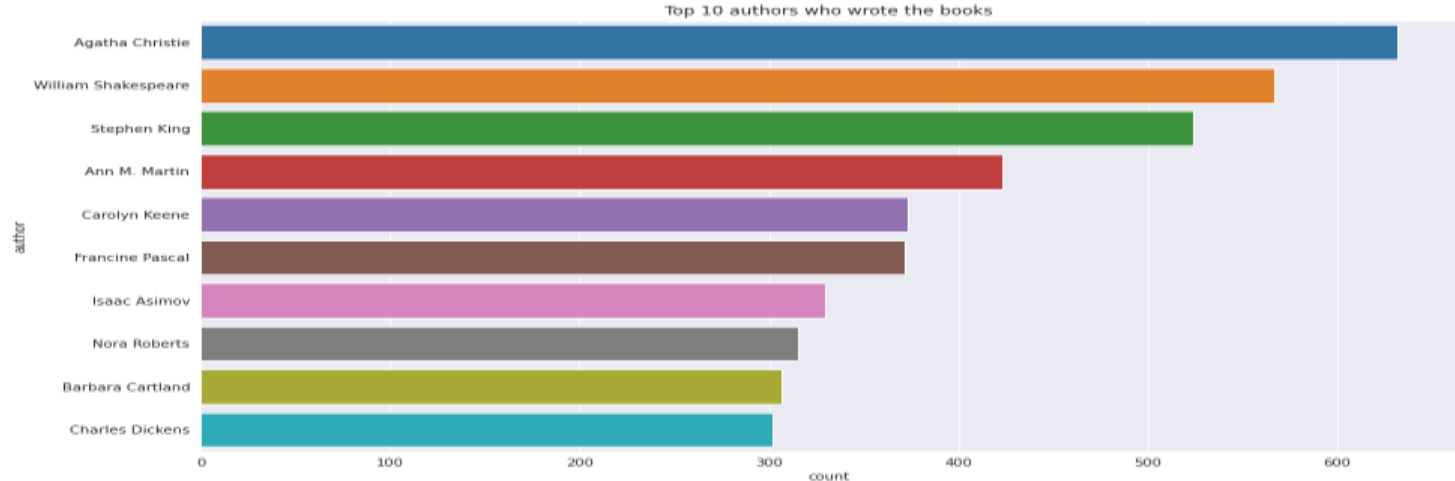
❖ Minimum books published in 1985.



- ❖ We have plotted bar plot for top ten books.
- ❖ Selected Poems is on number one and Little Women is on second number.



- ❖ Agatha Christie is the best author among all and William Shakespeare is second most liked author.



# Models Used to build recommendation system



- ❖ **Popularity Based Recommendation(Base model)**
- ❖ **Collaborative Filtering**

# Popularity Based Recommendation

- ❖ It is a type of recommendation system which works on the principle of popularity
  - and or anything which is in trend. These systems check about the books which are
  - in trend or are most popular among the users and directly recommend them.
  
- ❖ I have sorted top 10 books from the dataset on the basis of number of ratings
  - Received.

# Collaborative Filtering

- ❖ Collaborative Filtering is considered to be one of the very smart recommender systems that work on the similarity between different users and also items that are widely used as an e-commerce website and also online movie websites. It checks about the taste of similar users and makes recommendations.
- ❖ The similarity is not restricted to the taste of the user, moreover there can be consideration of similarity between different items also. The system will give more efficient recommendations if we have a large volume of information about users and items.

# Using KNN

- ❖ For this model we used KNN to find clusters of similar users based on common book ratings, and make predictions using the average rating of top-k nearest neighbors.
- ❖ In this model we assume that users who have given ratings more than 200 are users who read at least 20 books(suppose on user given rating 10/10 so minimum he read books(200 ratings/10 ratings per book=20).For statistical significance we should consider only the data of user who given more than 200 ratings.
- ❖ To ensure statistical significance, users with less than 200 ratings, and books with less than 100 ratings are excluded.



# Using SVD

- ❖ The SVD(Singular Value Decomposition) is used as a collaborative filtering technique. It uses a matrix structure where each row represents a user, and each column represents an item. The elements of this matrix are the ratings that are given to items by users.

# Challenges

- ❖ Since the data consisted of text data, data cleaning was a major challenge in feature engineering.
- ❖ Understanding the metric for evaluation was a challenge as well.

# Conclusion



- ❖ Recommendation system is unturned to exist in the e-commerce businesses with the help of collaborative or content-based filtering to predict different items and yes, users are most satisfied with the products recommended to them.
- ❖ As we have seen the age group between 30 to 40 reads the most books.
- ❖ We have removed the readers who did not give rating and then we found that most of the users gave 8 stars.
- ❖ KNN model gives good recommendation for books.

- **Thank You !**