INTRODUCTION TO DATA SCIENCE (CAP 5771)

PROJECT REPORT

MILESTONE 1: Data Collection, Preprocessing and Exploratory Data Analysis

(Feb 5, 2025 – Feb 21, 2025)

Name: Pranay Reddy Pullaiahgari

UF-ID: 6238-1134

Project Title: BOOK RECOMMENDATION SYSTEM USING CONTENT-BASED AND COLLABORATIVE FILTERING TECHINQUES

Project Objective

The objective of the project is to develop a recommendation system that suggests books to the users based on the reading preferences and historical ratings. The goal of the project is to implement content-based filtering and collaborative filtering techniques to provide personalized book recommendations. The project aims to analyze the user behavior and books metadata to improve the recommendation system's accuracy.

This project will be implemented as a web application developed using Streamlit. This enables users to input their preferences, ratings, and then the system will generate recommendations based on user's profile of historical ratings and book preferences.

Datasets used

Three datasets taken from Kaggle are used to build this recommendation engine.

1. Books.csv:

It contains book metadata. Attributes are ISBN(Book-ID), Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL-S, Image-URL-M, Image-URL-L. It has 271360 records in total.

2. Ratings.csv:

It includes book ratings given by the users on the scale of 0 to 10. Attributes are User-ID, ISBN, Book-Rating. It has 1149780 records in total.

3. Users.csv:

It has the demographic information of the users, including User-ID, Location and Age. It has 278858 records.

Tech Stack

The following are the tools, technologies and libraries required for this project:

Python – the core programming language for building this project.

Pandas & Numpy – Python libraries used for data manipulation and numerical computing.

Matplotlib & Seaborn – Python Library used for data visualization.

Surprise – a python scikit for building and analyzing recommender systems.

Scikit-Learn – used for feature engineering and model evaluation.

SqLite – Used for storing book and user data.

Streamlit – Python based web application development framework.

Project Timeline

Milestone 1: Data collection, Preprocessing and Exploratory Data Analysis. (Feb 5 – Feb 21)

Milestone 2: **Feature Engineering, Feature Selection and Data Modeling**

Feature Engineering (Feb 21 – Feb 26)

Feature Selection (Mar 1 – Mar 4)

Data Modeling (Mar 6 – Mar 19)

Milestone 3: **Evaluation, Interpretation and Tool Development**

Evaluation (Mar 24 – Mar 30)

Intepretation (April 1 – April 3)

Tool Development (April 4 – April 22)

EXPLORATORY DATA ANALYSIS – KEY INSIGHTS

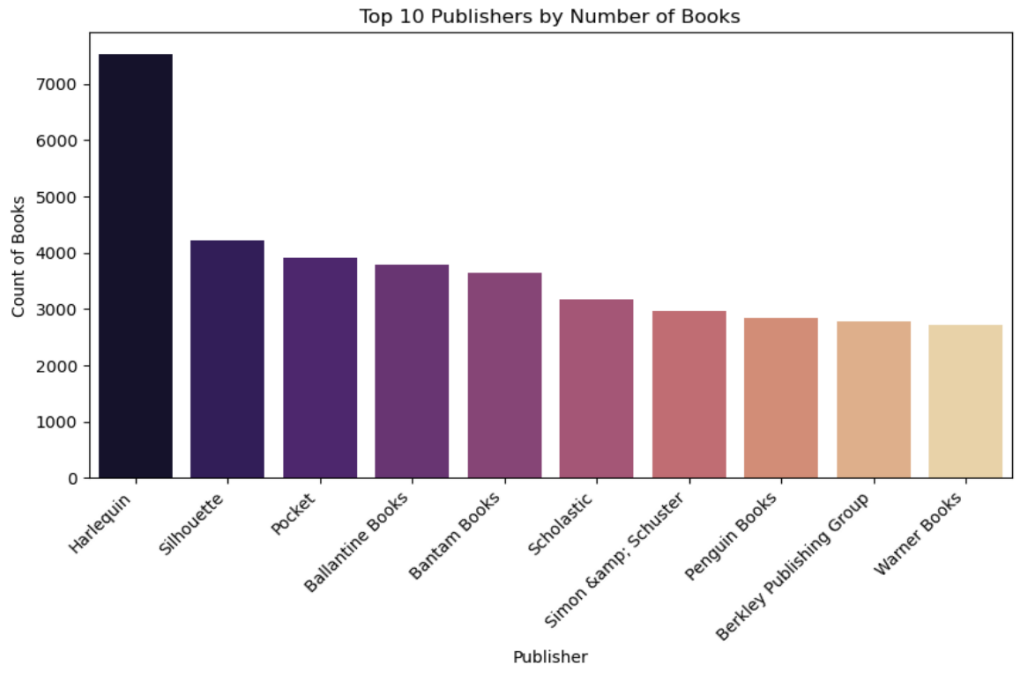
EDA helps us in understanding the dataset structure, detect anomalies, and extracting valuable insights for predictive model development.

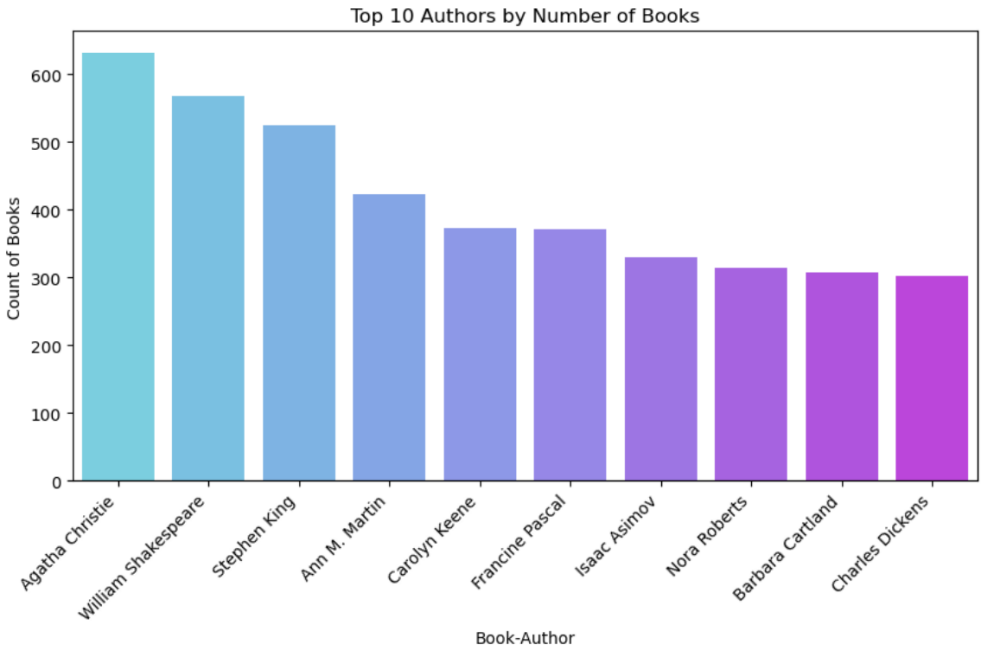
Books.csv Dataset

Boxplot for Year of Publication Distribution

We observe that there are some years which are unrealistic, example 0, and others <1800, which are not significant. These incorrect years can distort recommendations based on books age. So we can filter out the books that are published before 1800.

Bar plots for Top Publishers and Authors

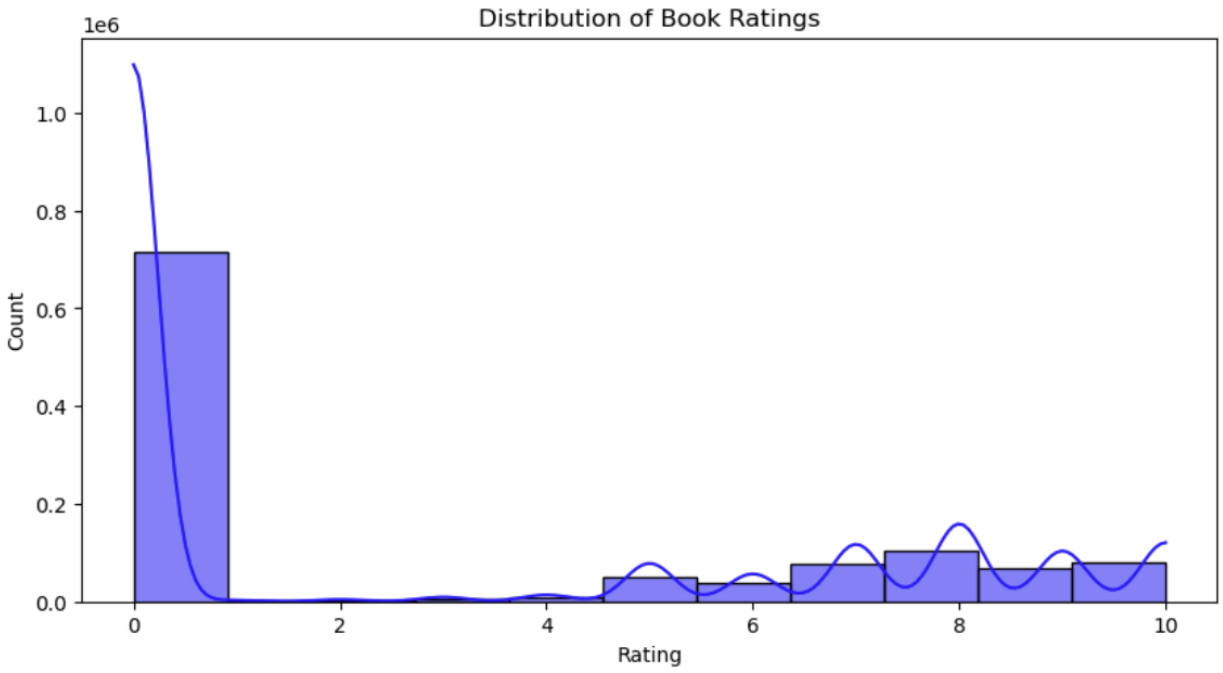




There are small number of publishers and authors responsible for a large portion of books. This visualization helps in identifying key contributors in the dataset, and understand the diversity of book sources available.

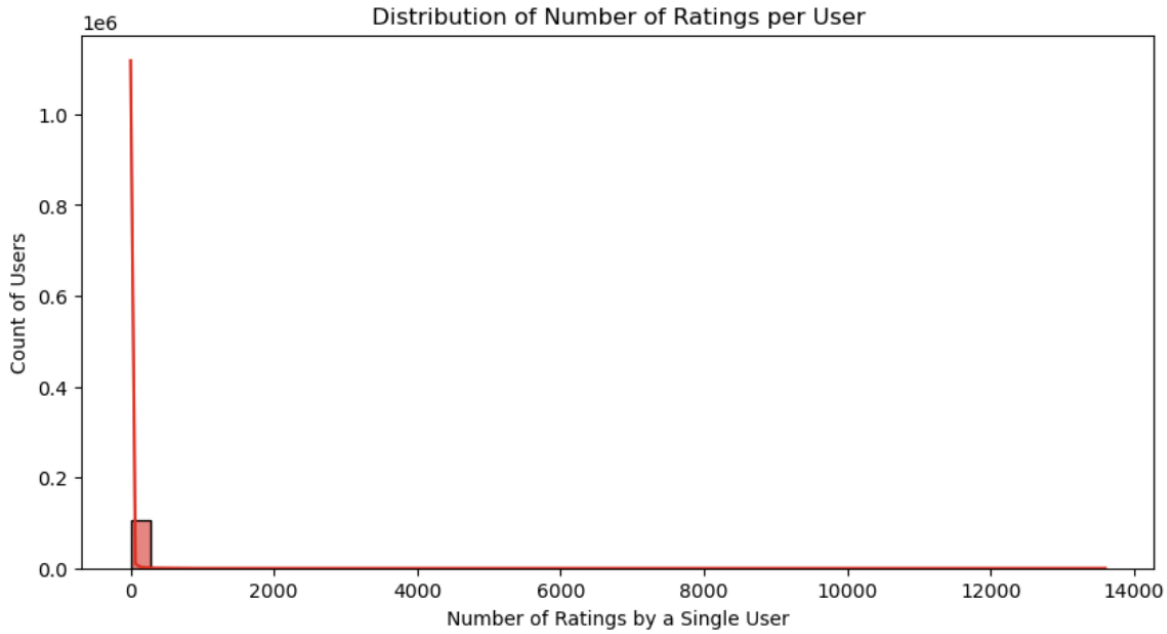
Ratings.csv Dataset

Histogram for ratings distribution



Ratings are highly skewed toward 0 or 10. This type of binary rating tendency might impact model performance. We can consider applying normalization or adjusting weighting to handle skewness.

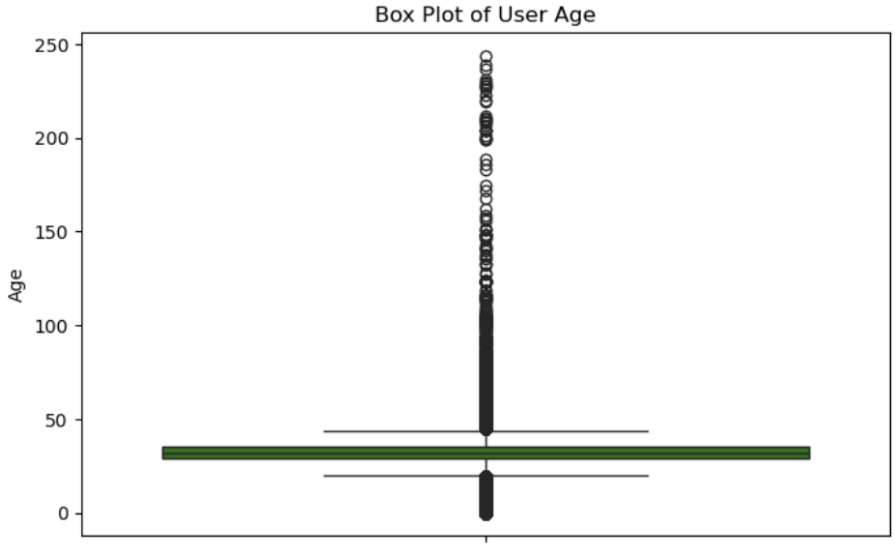
Histogram for Number of Ratings per user

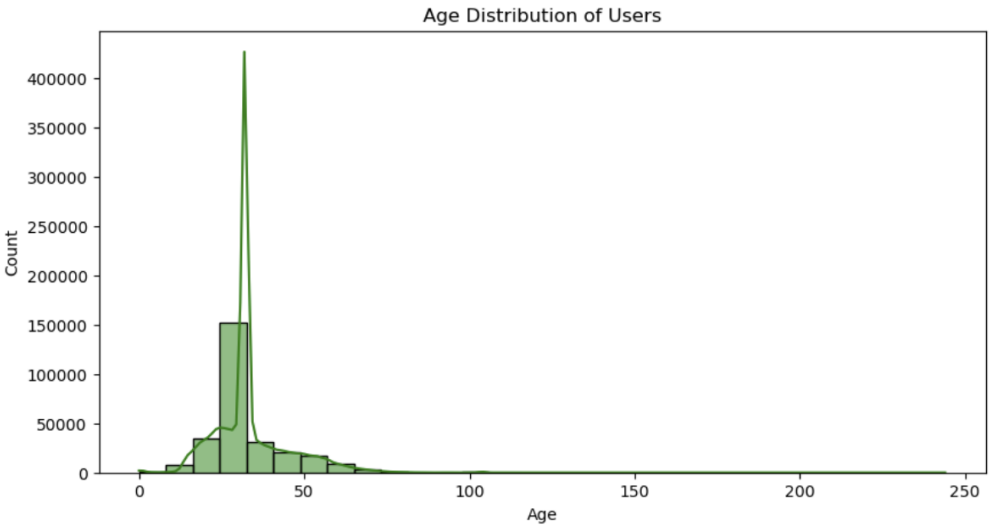


Most of the users have rated very few books, whereas a few users have rated thousands. This sparse data makes collaborative filtering less effective for users with minimal interactions. We consider filtering out the inactive users who rated <5 books, so that model performance can be improved.

Users.csv Dataset

Box plot and Histogram for Age Distribution

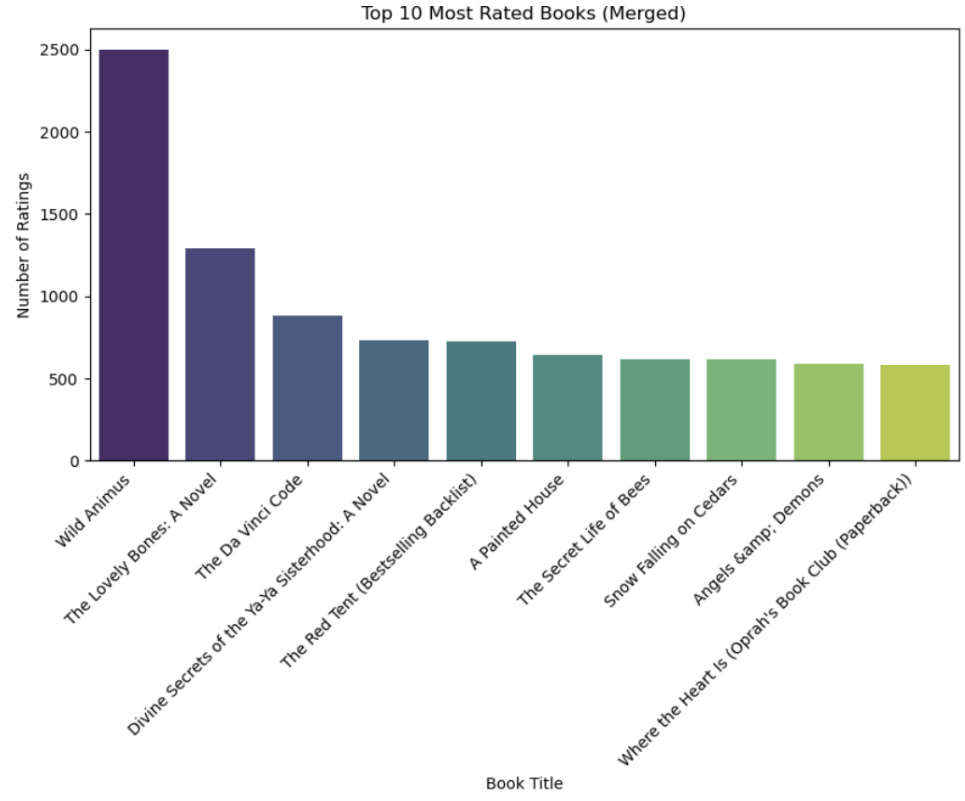




Few users have unrealistic ages like 0 and 250. Such outliers can distort demographic based recommendations, and so we clip the age values to the range 5 to 100.

Final Dataset after merging "df"

Top10 Most rated books



This highlights the most popular books based on rating, and can be useful in popularity-based recommendations.