BOOK RECOMMENDATIONS SYSTEM

People who love reading
by Mili Thakrar

INTRODUCTION

This initiative leverages machine learning to deliver personalized book recommendations, helping readers discover titles that truly resonate with their unique tastes in a sea of literature.

PROBLEM AREA

- Too many books, hard to choose.
- Reduced motivation to read.
- Overwhelmed by recommendations.
- Hard to find niche interests.
- Generic suggestions don’t fit.

DATA SCIENCE SOLUTION

BUILDING A SMART BOOK MATCH MAKER

Hybrid Recommendation Engine:

- Collaborative + Content-based filtering
- Considers: reading history, ratings, similar readers, book "DNA"

Process:

- Analyze reader data
- Decode book essence
- Develop reader-book matching algorithms
- Fine-tune recommendations

DATA OVERVIEW

BOOKS

Field Name	Type	Description
ISBN	string	International Standard Book Number, unique identifier for books
Title	string	The title of the book
Author	string	The name of the book's author
Publisher	string	The name of the book's publisher
Publication_year	int	The year the book was published
Image_URL	string	URL link to the book's cover image

USERS

Field Name	Type	Description
User_id	float	Unique identifier for each user
Age	float	Age of the user
Location	string	Location where the user is located

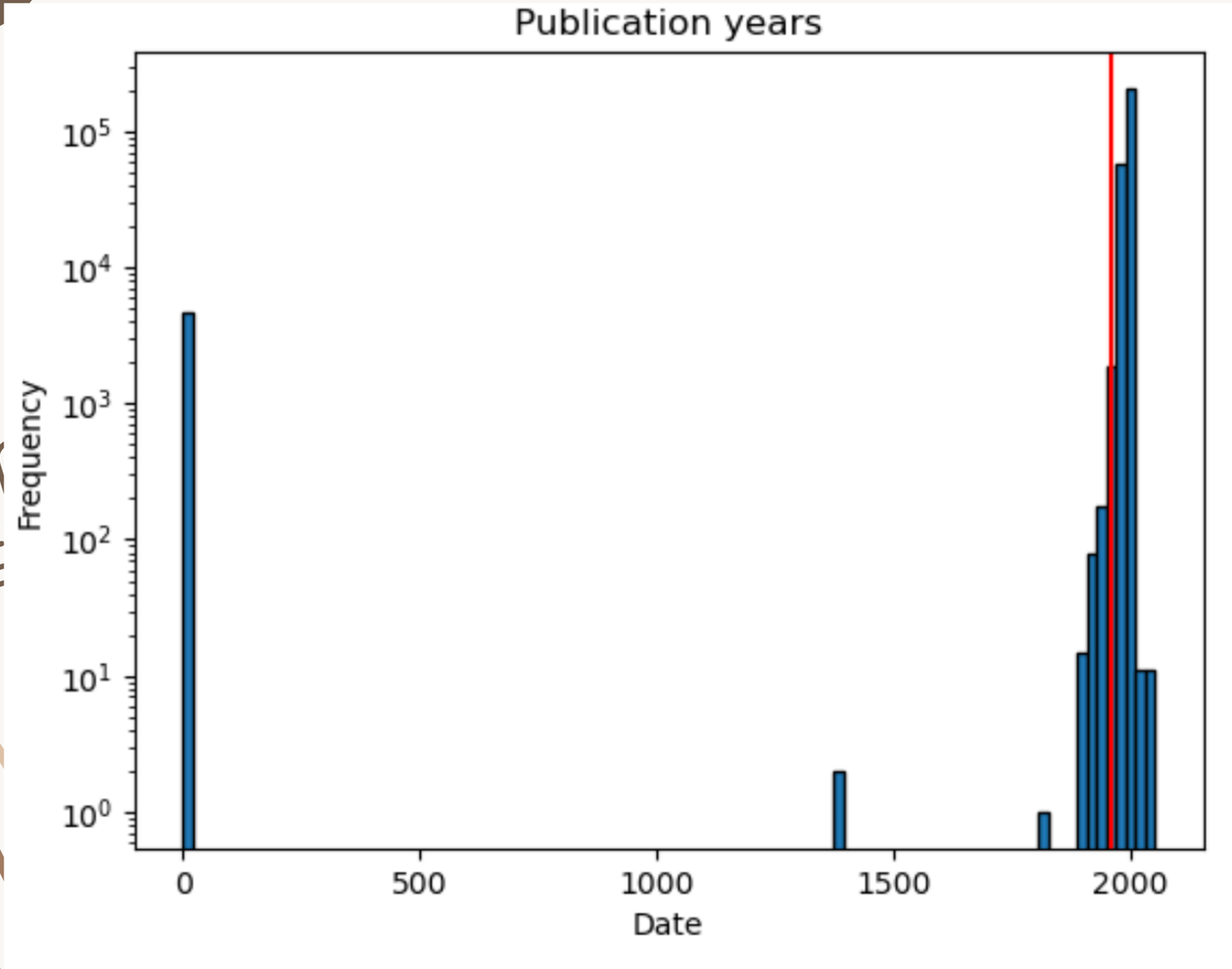
RATINGS

Field Name	Type	Description
User_id	float	Unique identifier for each user
ISBN	string	International Standard Book Number of the rated book
Ratings	float	User's rating of the book, scale of 1-10

DATA PREPROCESSING

BOOKS

Initial distribution of publication years before cleaning



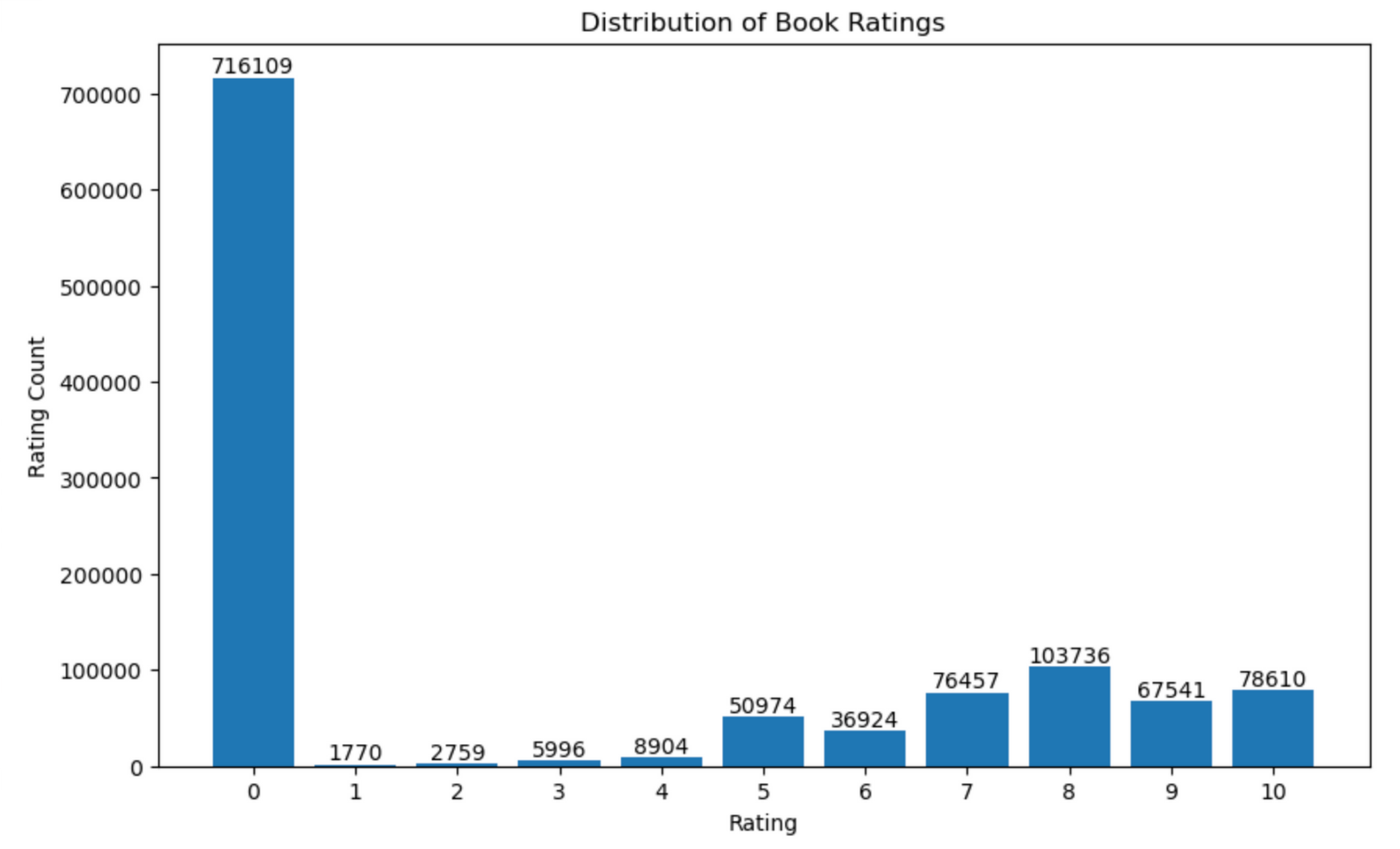
Percentage of null values in the dataset

```
Image_URL           0.001106
Author              0.000737
Publisher           0.000737
ISBN                0.000000
Title               0.000000
Publication_year    0.000000
dtype: float64
```

- Classified extreme years as "Unknown"
- Manually corrected future dates
- Dropped minimal null entries (<1% of dataset)
- Created year categories: Unknown, Pre-1950, 1950-1979, 1980-1999, 2000-present

DATA PREPROCESSING

RATINGS

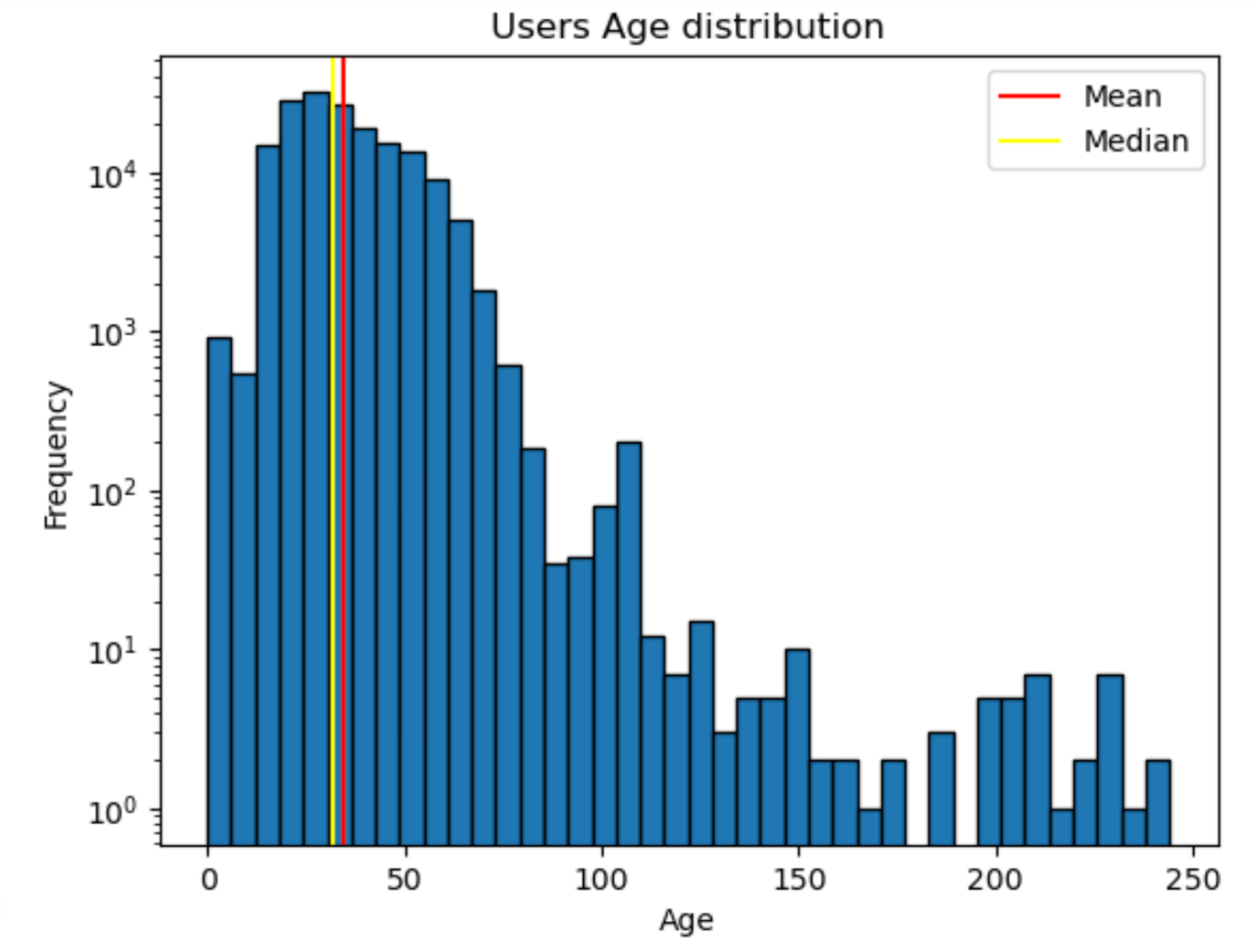


- **Added 'Total_Ratings' column:** Count of ratings per book
- **Created 'Avg_Rating' column:** Mean rating for each book
- Enables analysis of book popularity and reception
- Facilitates identification of highly-rated vs. frequently-rated books
- Supports more nuanced recommendationS

DATA PREPROCESSING

USERS

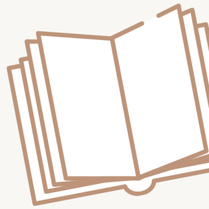
Initial distribution of age column before cleaning



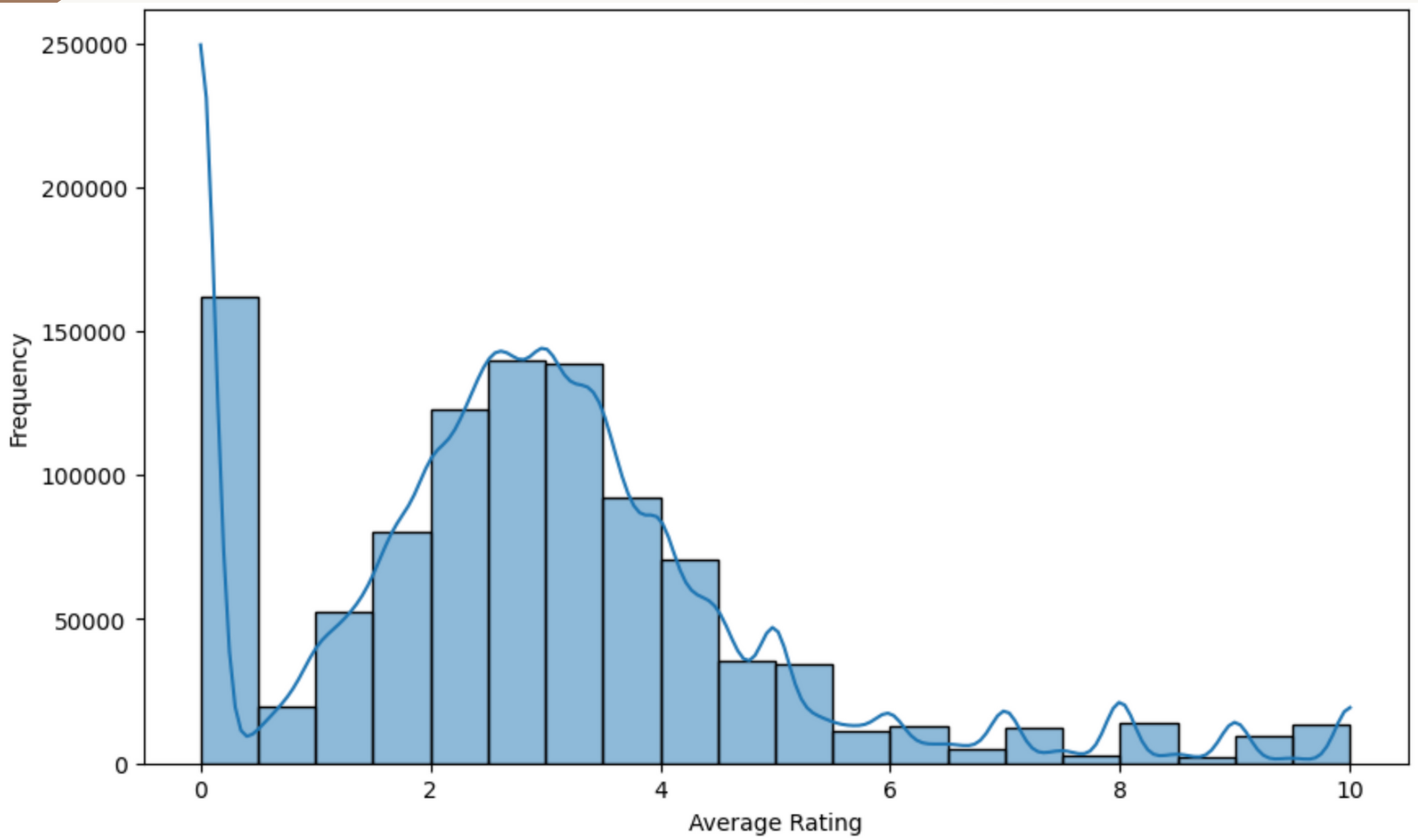
- Handled extreme age values: Removed ages <5 and >100
- Created age categories
- Split location into city, state, and country
- Enables demographic analysis and location-based insights
- Supports age-specific and geographical recommendations

FINAL DATASET

Field Name	Type	Description
ISBN	string	International Standard Book Number, unique identifier for books
Title	string	The title of the book
Author	string	The name of the book's author
Ratings	float	User's rating of the book, scale of 1-10
Total_num_of_ratings	float	Total number of ratings for the book
Avg_ratings	float	Average rating score for the book
Publisher	string	The name of the book's publisher
Publication_year	int	The year the book was published
Year_Category	string	Categorized time period of publication
User_id	float	Unique identifier for each user
Age	float	Age of the user
Age_Category	string	Categorized Age into age ranges
City	string	City where the user is located
State	string	State or region where the user is located
Country	string	Country where the user is located
Image_URL	string	URL link to the book's cover image

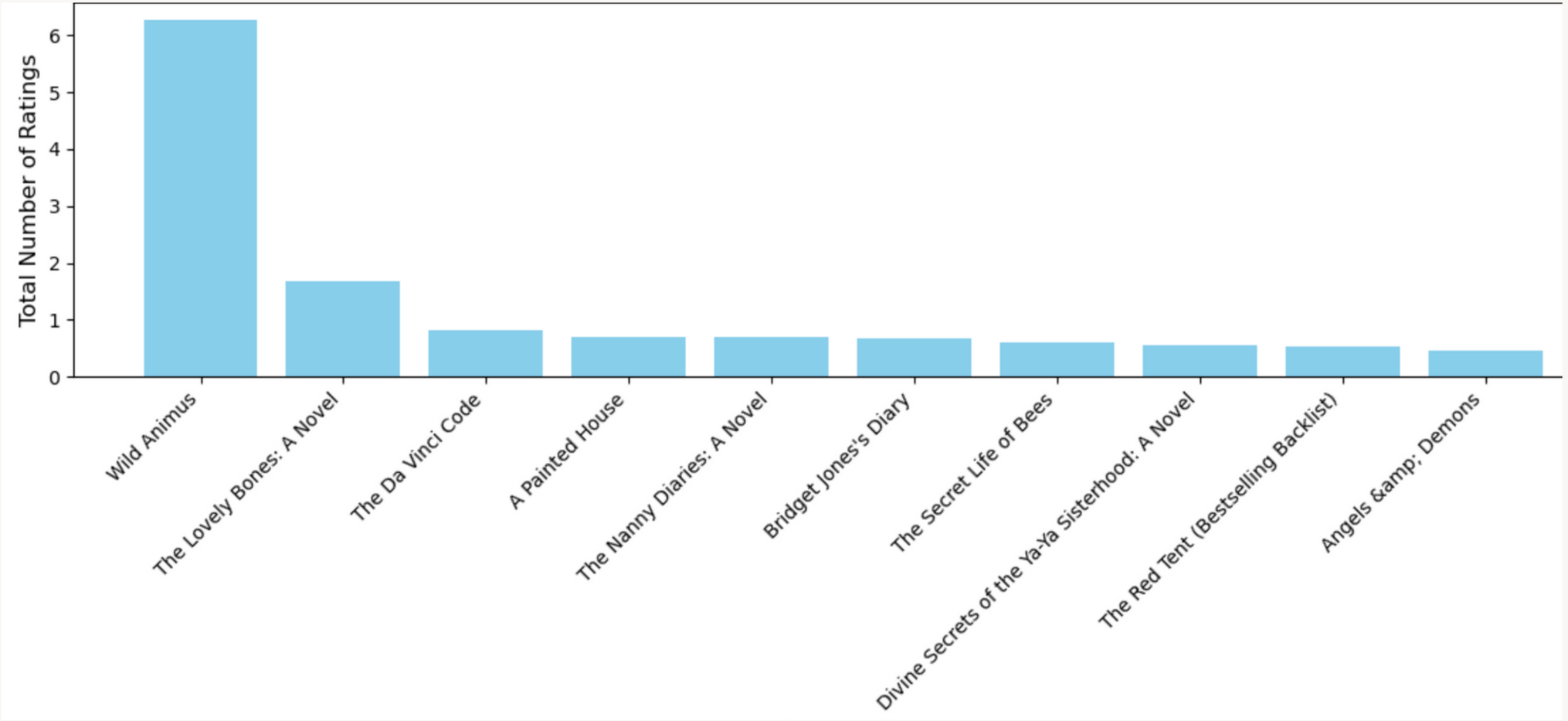


Distribution of Average Book Rating



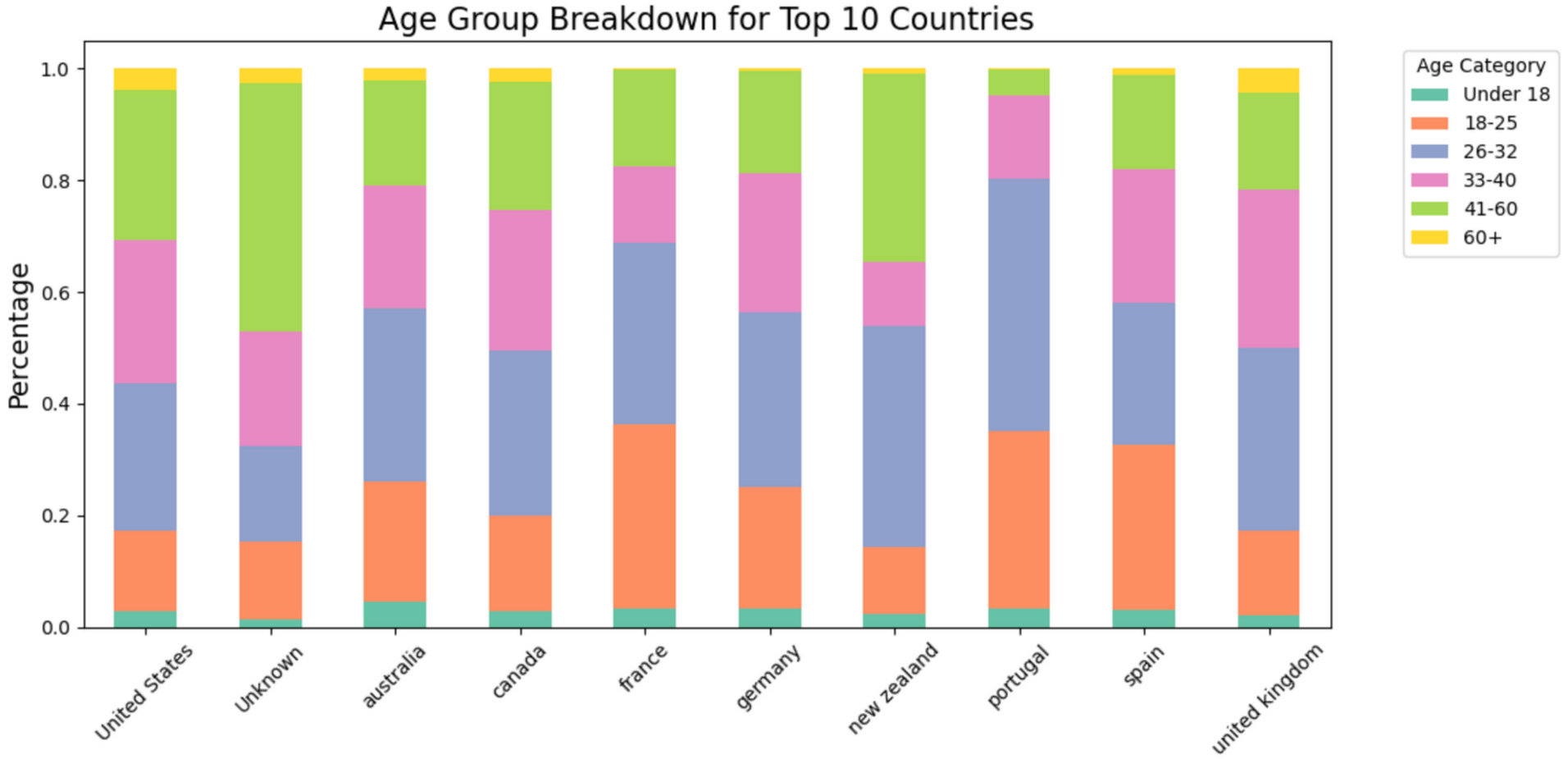
- Many users don't rate books
- Average ratings near-normal, slightly right-skewed
- Ratings peak at 2.5
- Active users give moderate ratings
- Some users show preference for higher ratings
- Suggests positive bias in engaged users

Top 10 Books



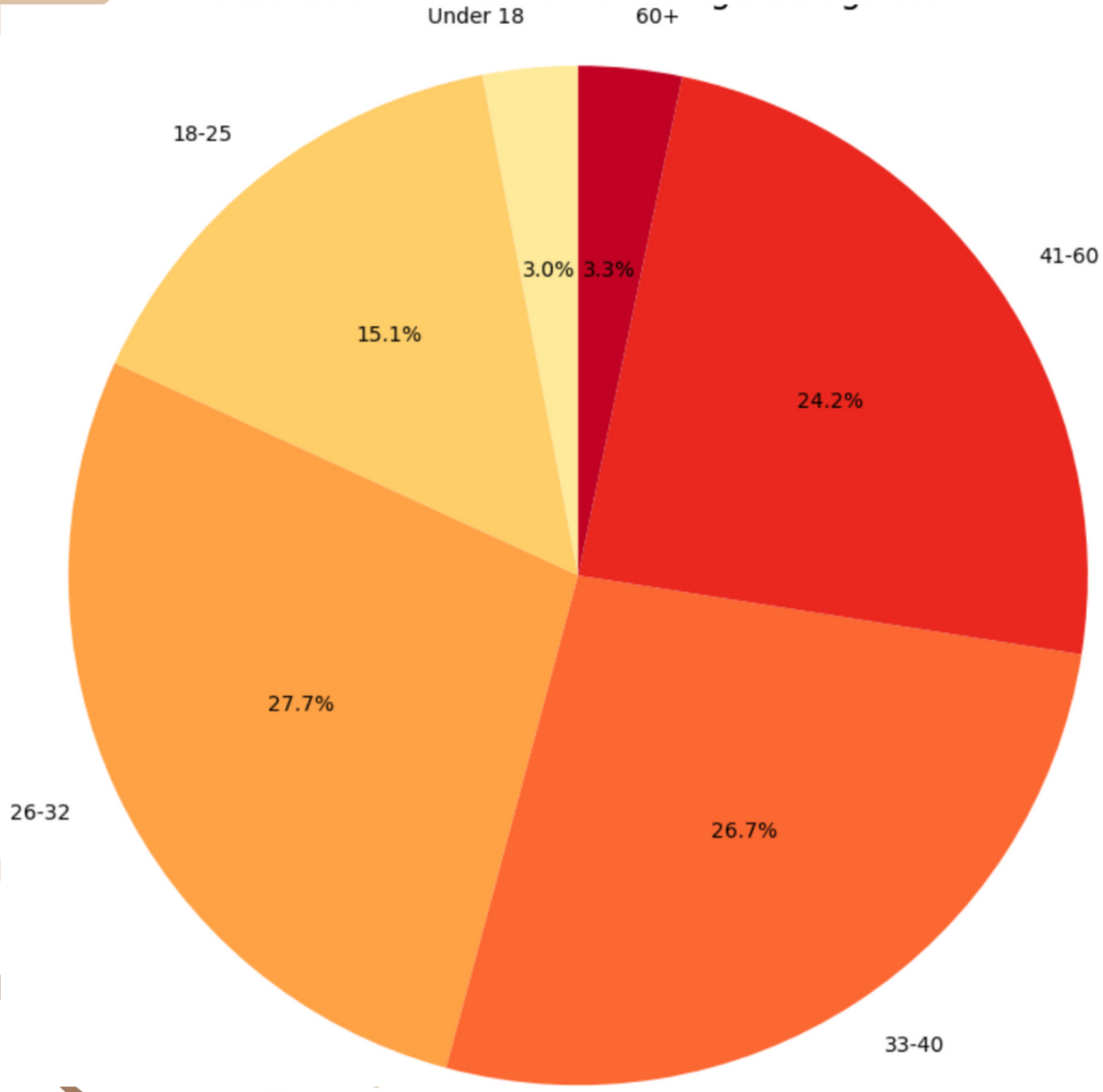
- Top Books: Wild Animus leads, followed by The Da Vinci Code.
- Issues: Total ratings misrepresent quality due to no weighting, popularity bias, and outliers.
- Solution: Introduce a weighted rating system with volume and smoothing factors.

Top 10 User Countries with Age breakdown



- The US has the highest user count.
- Significant representation from Germany, France and Portugal.
- US and Canada have even distributions; Portugal and New Zealand skew younger.
- Under-18 Representation: Very low percentage of users under 18.

Distribution of Users Across Age Categories



- Balanced representation among users aged 26-60
- 15% of users in 16-25 age range
- Smaller percentages in extreme age categories
- Data cleaning removed unrealistic age entries

NEXT STEPS

1. **Content Based Recommendations**:
   - extract meaningful features from book titles using NLP
2. **Hybrid Model** (Weighted/Mixed):
   - Adjusts component importance.
   - Presents diverse recommendations

IMPACT

- Increases user satisfaction and engagement
- Reduces consumer search and uncertainty, improving the decision-making process for users
- Boosts motivation to read, leading to more time spent reading

THANK YOU