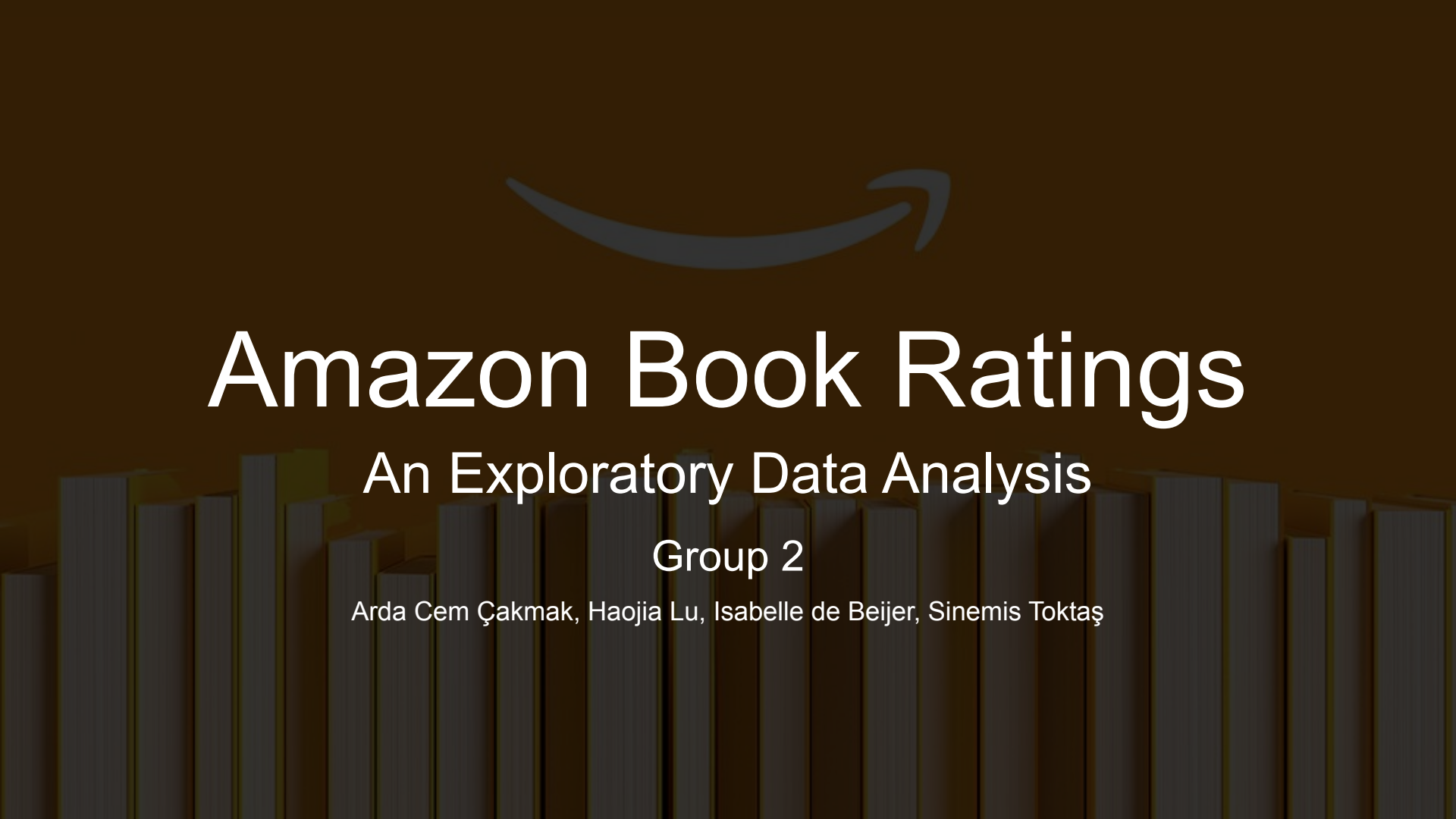


Amazon Book Ratings

An Exploratory Data Analysis

Group 2

Arda Cem Çakmak, Haojia Lu, Isabelle de Beijer, Sinemis Toktaş

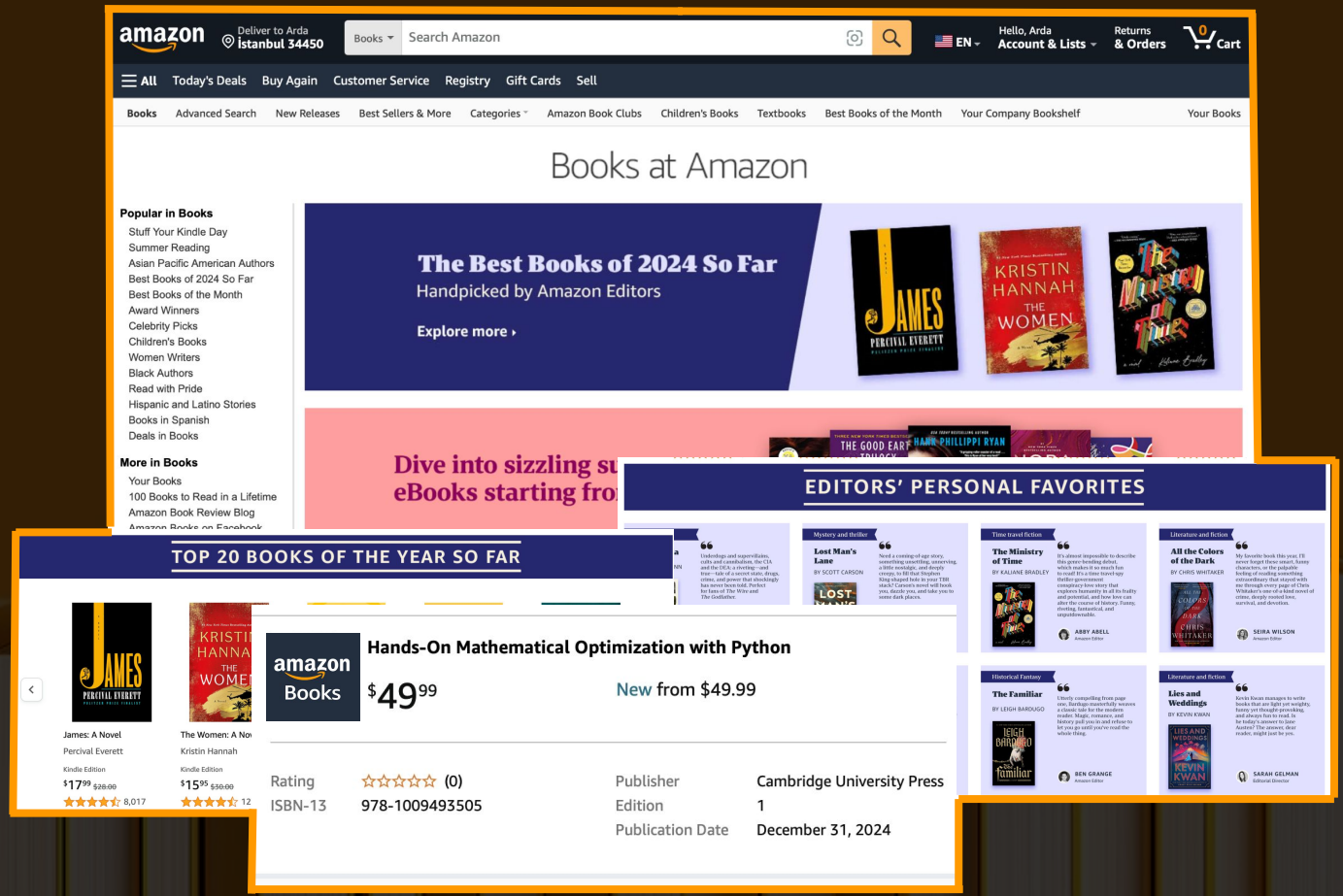
What You'll be Hearing from Us

- Introduction
- Our Goal
- Dataset Description
- Data Cleaning Steps
- Results
- Conclusion
- Further Work
- Q&A

Introduction

Amazon Books is an online bookstore platform that also allows its users to rate the books.

Our data is a collection of books, ratings, and users found on Amazon Books platform.



Our Goal: Explore, Uncover, Understand

Trends in books, publishers and authors

Behaviour of users based on age and location

Dataset Description: Books Dataset obtained from Kaggle [1]

ratings.csv (30.68 MB)

Detail | Compact | Column

User-ID	ISBN	Book-Rati...
276725	034545104X	0
276726	0155061224	5
276727	0446520802	0
276729	052165615X	3
276729	0521795028	6
276733	2080674722	0
276736	3257224281	8
276737	0600570967	6
276744	038550120X	7
276745	342310538	10

User IDs, ISBN Number of Rated Book, Rating Score (0 to 10).

books.csv (77.79 MB)

Detail | Compact | Column

ISBN	Book-Title	Book-Aut...	Year-Of-P...	Publisher	Image-UR...	Image-UR...	Image-UR...
0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.01.THUMBZZZ.jpg	http://images.amazon.com/images/P/0195153448.01.MZZZZZZZ.jpg	http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg
0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.01.THUMBZZZ.jpg	http://images.amazon.com/images/P/0002005018.01.MZZZZZZZ.jpg	http://images.amazon.com/images/P/0002005018.01.LZZZZZZZ.jpg
0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.01.THUMBZZZ.jpg	http://images.amazon.com/images/P/0060973129.01.MZZZZZZZ.jpg	http://images.amazon.com/images/P/0060973129.01.LZZZZZZZ.jpg
0374157065	Flu: The Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.01.THUMBZZZ.jpg	http://images.amazon.com/images/P/0374157065.01.MZZZZZZZ.jpg	http://images.amazon.com/images/P/0374157065.01.LZZZZZZZ.jpg
0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company	http://images.amazon.com/images/P/0393045218.01.THUMBZZZ.jpg	http://images.amazon.com/images/P/0393045218.01.MZZZZZZZ.jpg	http://images.amazon.com/images/P/0393045218.01.LZZZZZZZ.jpg

ISBN Number, Book Title, Author, Publication Tear, Publisher, Book cover image URLs of different sizes for each book.

users.csv (12.28 MB)

Detail | Compact | Column

User-ID	Location	Age
1	nyc, new york, usa	NULL
2	stockton, california, usa	18
3	moscow, yukon territory, russia	NULL
4	porto, v.n.gaia, portugal	17
5	farnborough, hants, united kingdom	NULL
6	santa monica, california, usa	61
7	washington, dc, usa	NULL
8	timmins, ontario, canada	NULL

User IDs, Locations, Ages of Users.

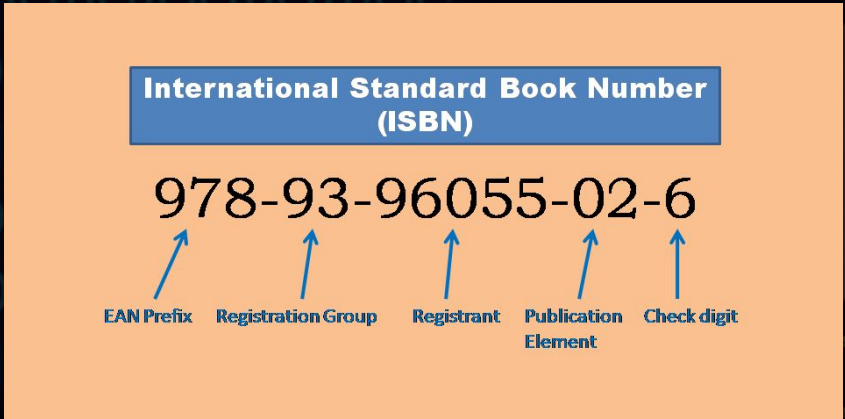
Data Cleaning Steps

- Duplicates
- String formatting
- ISBN format verification
- Handling of ratings
- Age filtering and corrections
- Data type conversions
- Missing values

ISBN Format Verification [2]

Filtering Results:

- Books dataset: 529 bad ISBNs filtered out (0.2%).
- Ratings dataset: 7069 bad ISBNs filtered out.



Handling Ratings

Filtering Implicit Ratings:

- Filtered out 716,109 implicit ratings (ratings=0), 62% of the ratings dataset.

Age Filtering in Users Dataset

Age Anomalies:

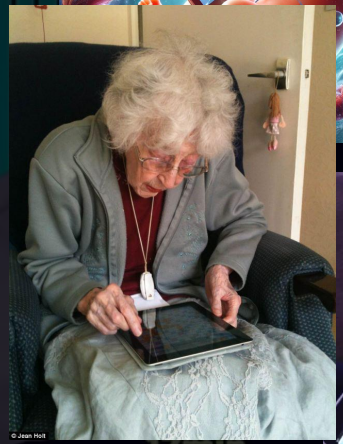
- Many impossible values

Filtering Criteria: [3]

- Ages 6-122 → NaN

Results:

- Significant reduction in anomalous age values.



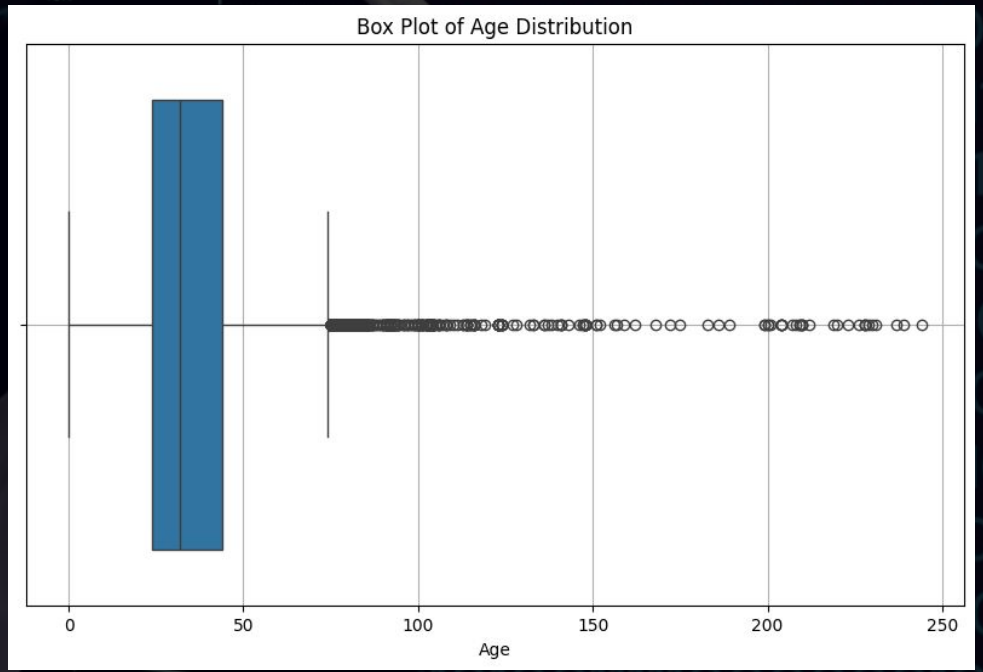
Handling Missing Values

Books Dataset: 4 rows
Ratings Dataset: 0 rows

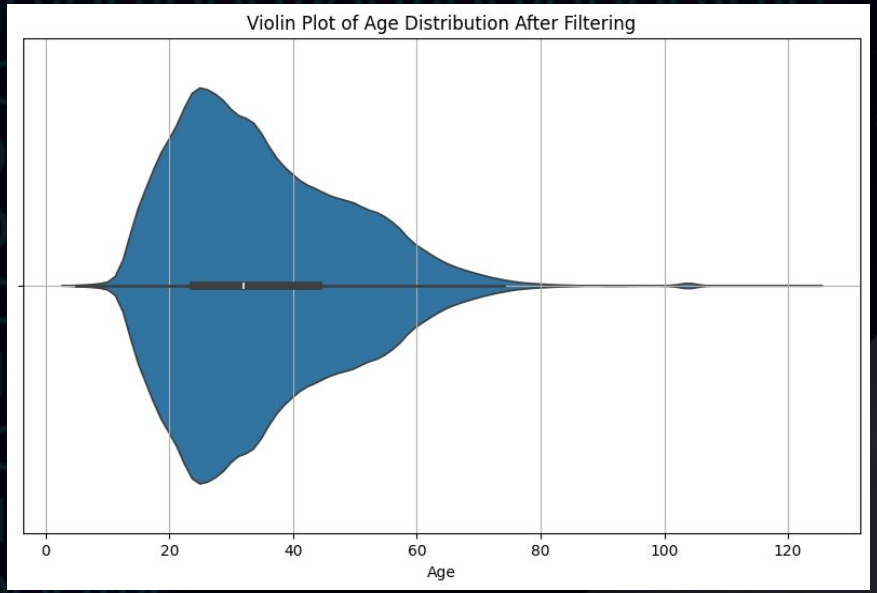
Users Dataset:

- 111,714 rows with missing age values (40%) → median.
- Empty location values → filtered out.

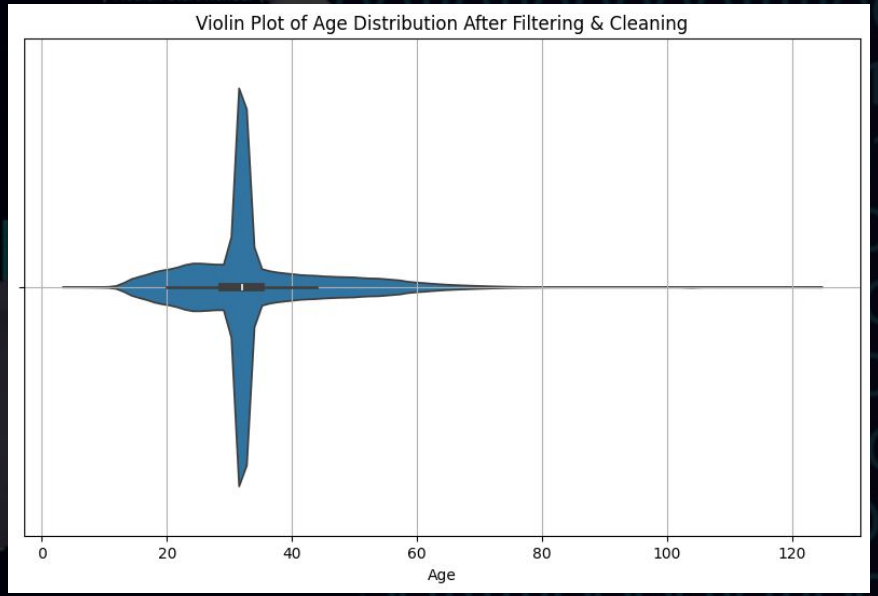
Box Plot of Age Distribution Before Cleaning:



Box Plot, Age Distribution After Filtering:

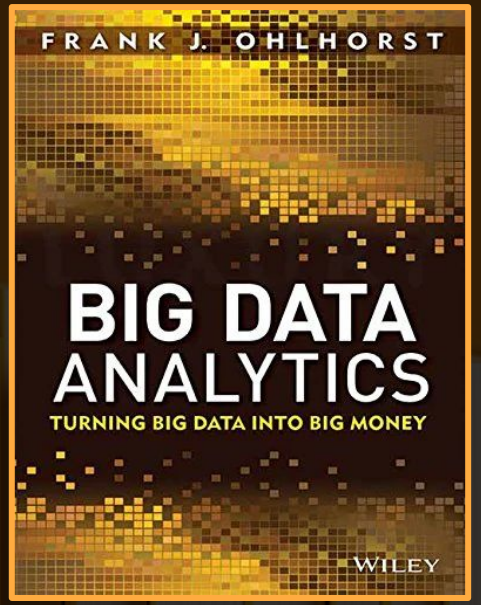


Box Plot, Age Distribution After Filtering & Cleaning:

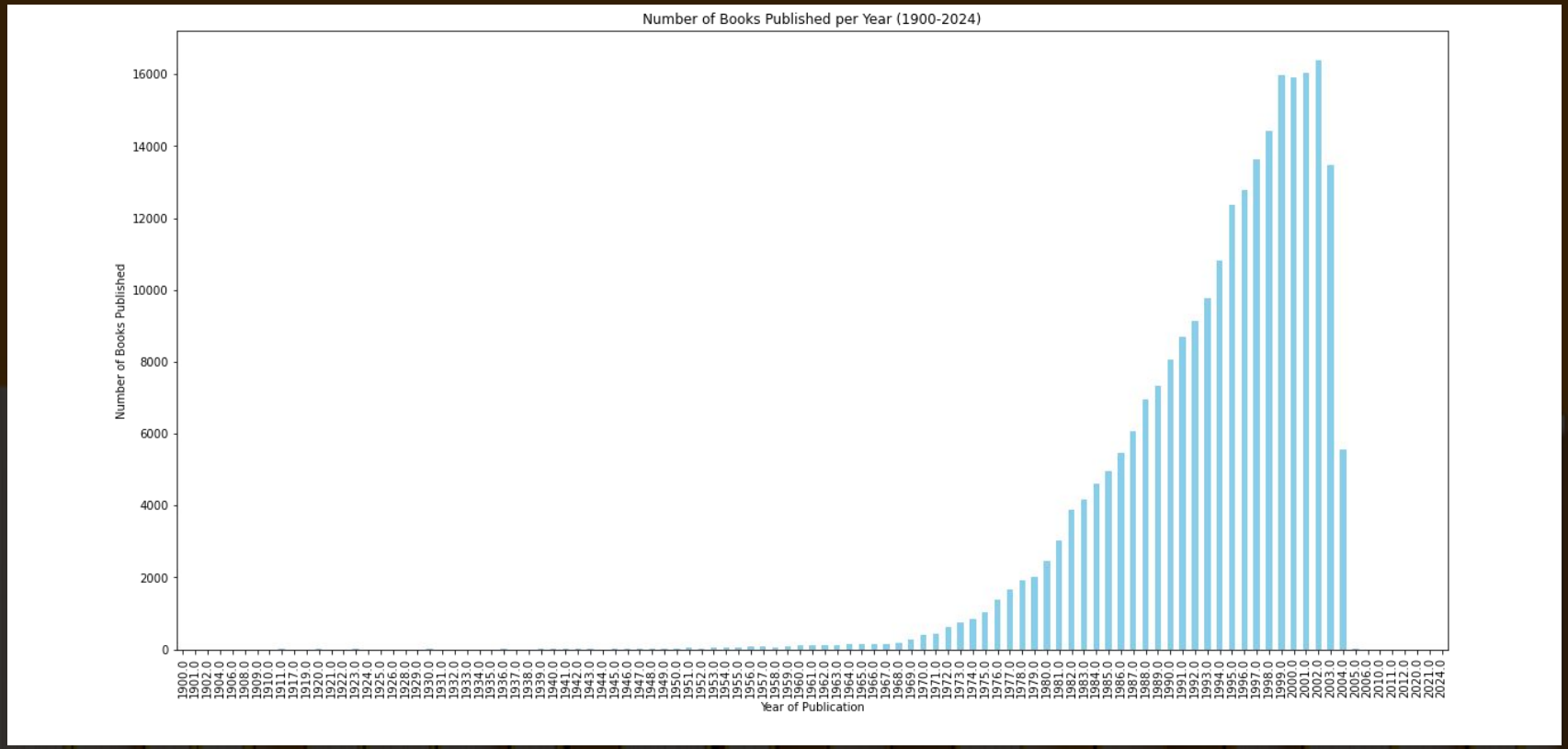


Motivation

- Overview Past
  - Best + Worst Rated Books & Publishers
  - Frequencies (ratings + publishings)
  - Gender Bias
  - Extremists
- Publishers
  - (Less) Popular Authors
  - Book & Author preferences (per Age Group)
  - Rating differences (ages & location)
  - Title Analysis
- Users
  - Book recommendation system



Finding: Frequencies of Books per Year



Finding: Frequencies of Book & Author Ratings

	Book	Ratings	Author	Ratings
Nr. 1	The Lovely Bones: A Novel	707	Stephen King	4491
Nr. 2	Wild Animus	581	Nora Roberts	2934
Nr. 3	The Da Vinci Code	494	John Grisham	2535
Nr. 4	The Secret Life of Bees	406	James Patterson	2182
Nr. 5	The Nanny Diaries: A Novel	393	J.K. Rowling	1743

Finding: Top 10 Best and Worst Rated Books

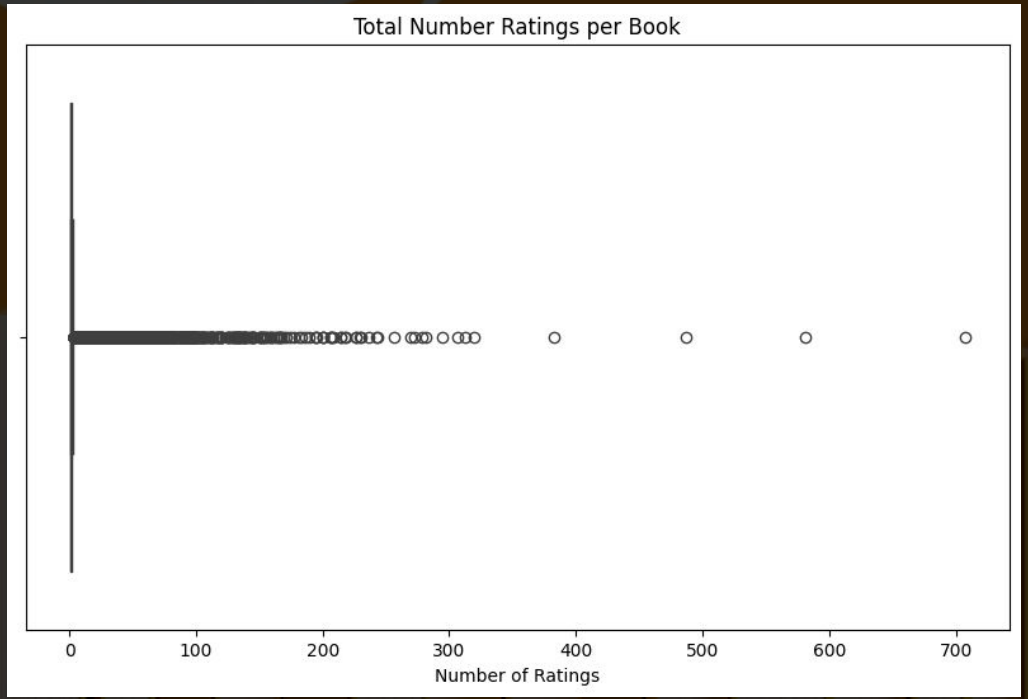
Motivation:

Understand which books are most and least appreciated by readers.

Identify trends in literature ratings.

Method: Grouping, filtering out based on rating counts, average rate calculations

Finding:



Finding: Top 10 Best and Worst Rated Books

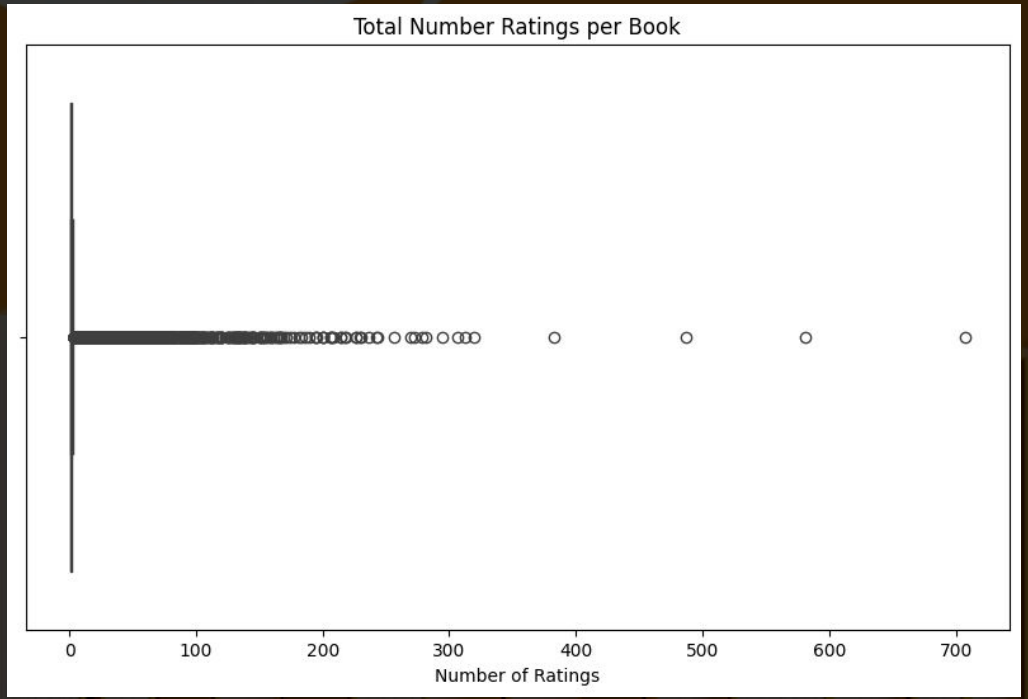
Motivation:

Understand which books are most and least appreciated by readers.

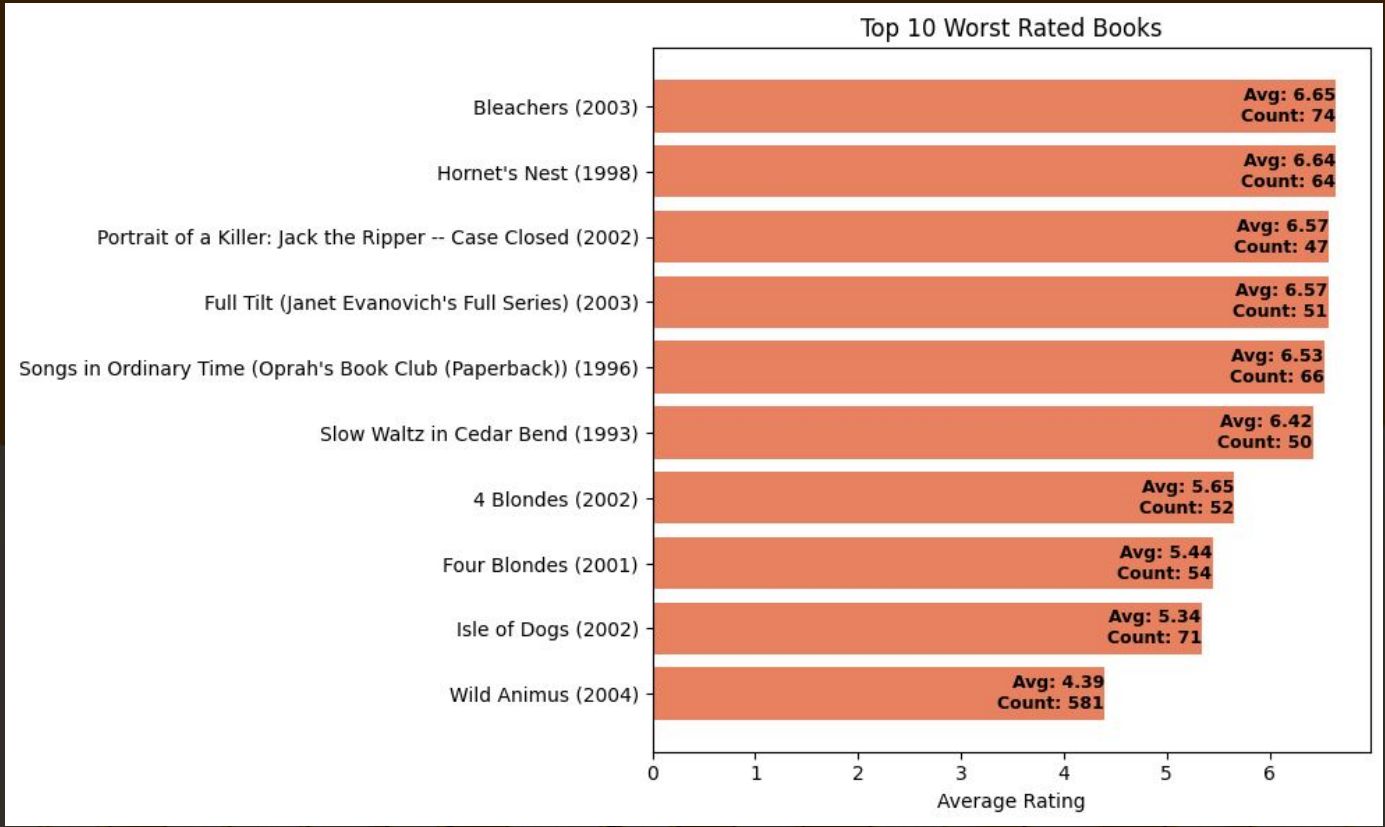
Acquire a base for identifying trends in literature ratings.

Method: Grouping, filtering out based on rating counts, average rate calculations

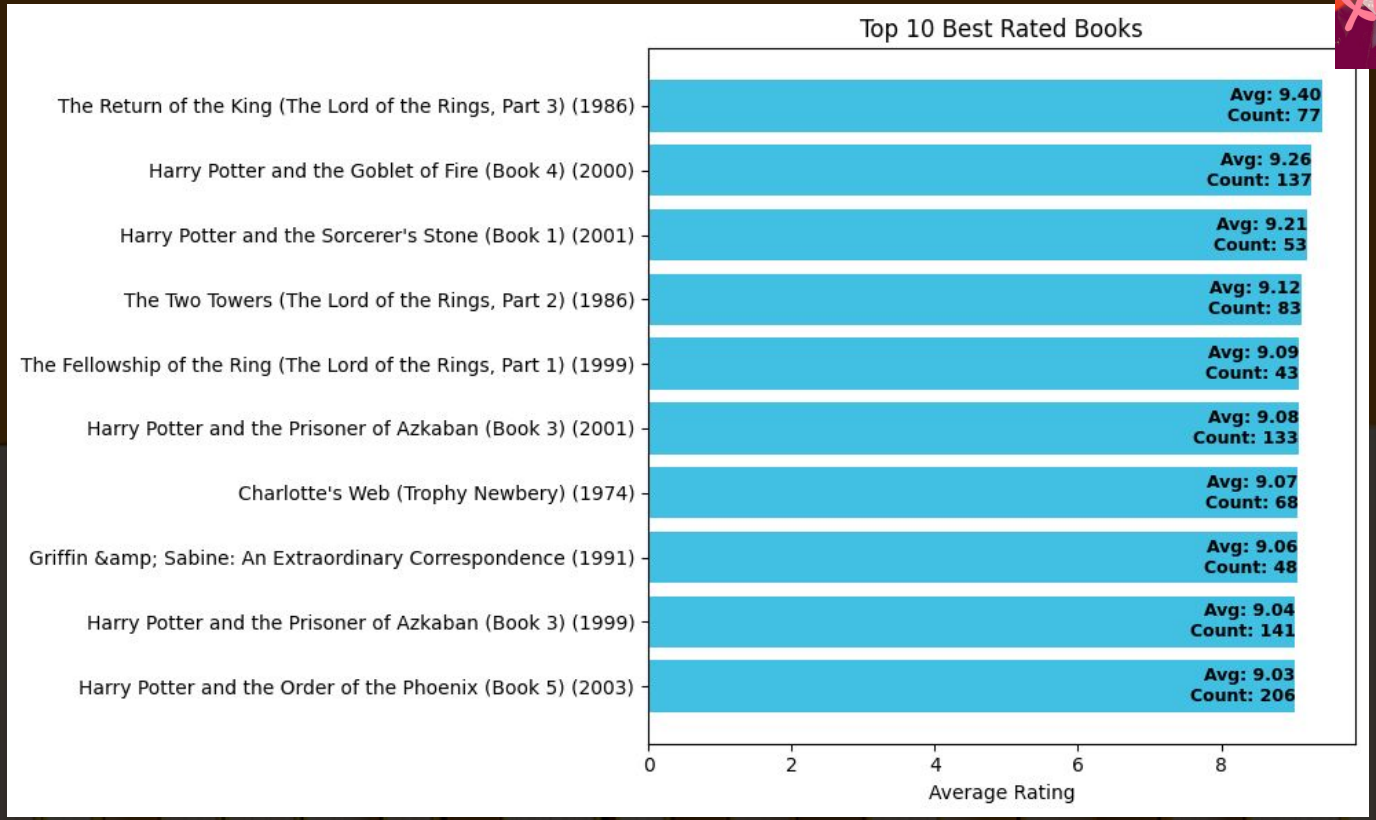
Finding:



Finding: Top 10 Worst Rated Books

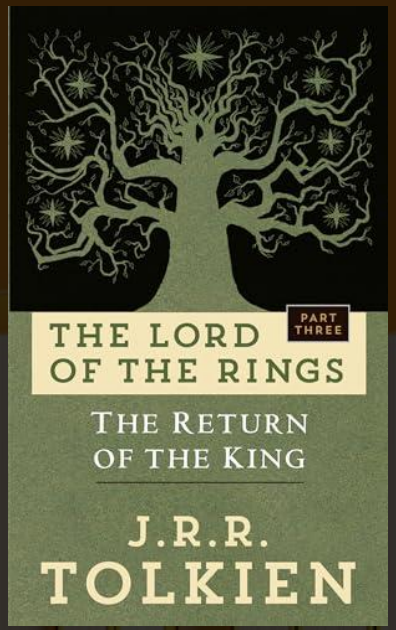


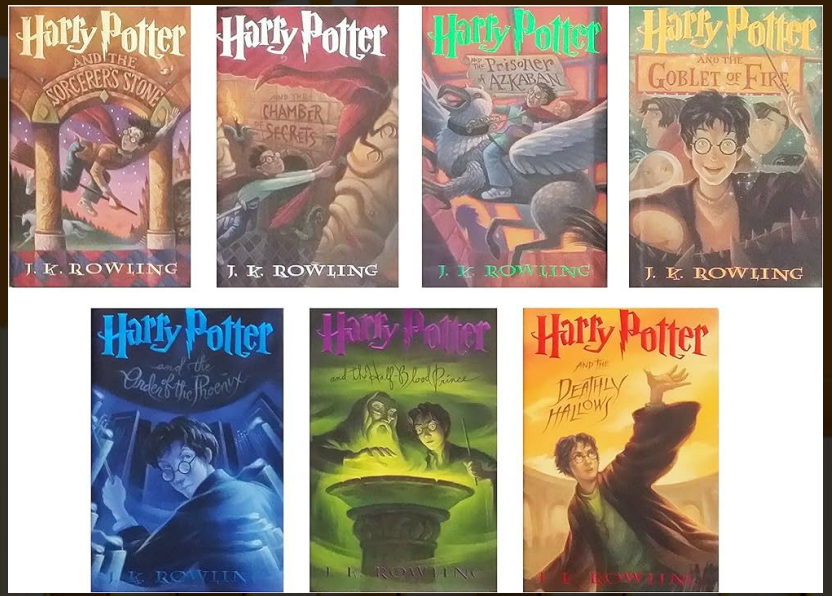
Finding: Top 10 Best Rated Books



Finding: Top 10 Best Rated Books

Best book:





Finding: Popular Books Among Users Who Dislike the Most Popular Book

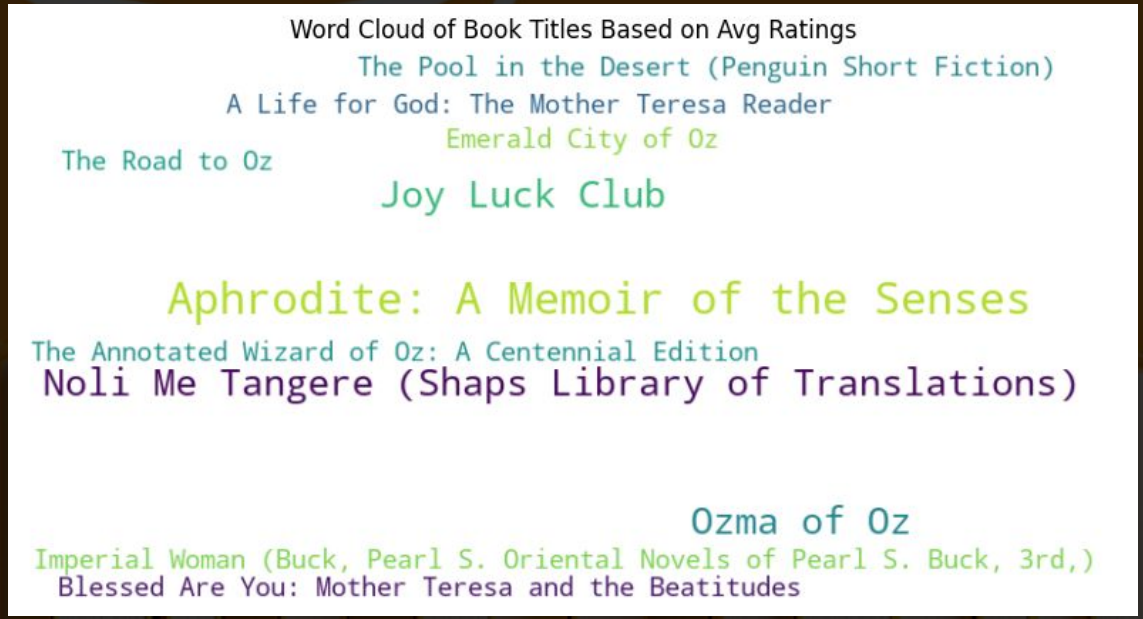
Motivation:

Explore the preferences of users who rated the most popular book poorly

and understand if there are common alternative preferences.

Method: Bad rating threshold, grouping, average rate calculations

Finding:



2 users, and 11 books with a 10 rating

Finding: Best and Worst Rated Book per Age Group

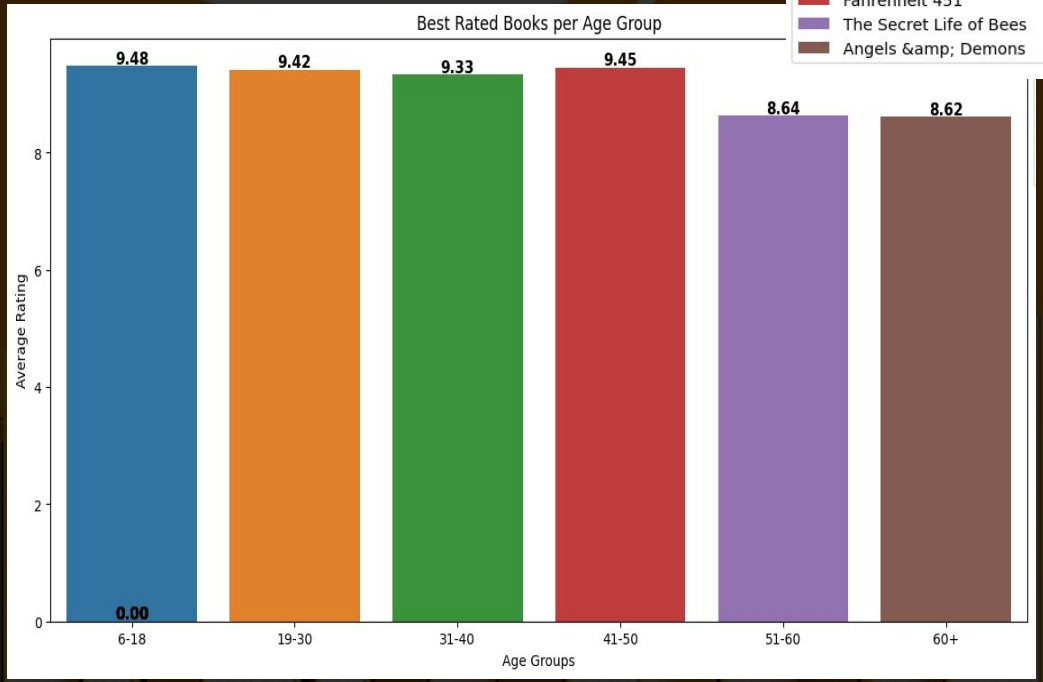
Motivation:

Understand how book ratings vary across different age groups and

identify the most and least preferred books for each age group.

Method: Merging, grouping, average rate calculation, rating count filtering

Finding:



Finding: Best and Worst Rated Book per Age Group

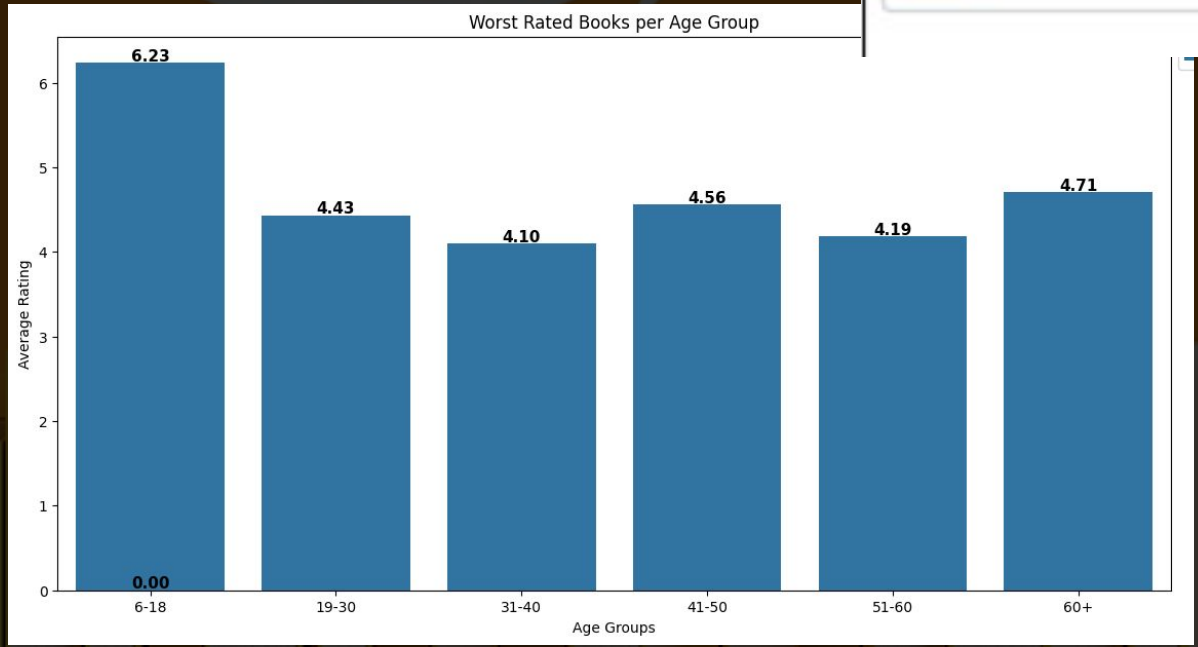
Motivation:

Understand how book ratings vary across different age groups and

identify the most and least preferred books for each age group.

Method: Merging, grouping, average rate calculation, rating count filtering

Finding:



Finding: Most and Least Liked Authors per Age Group

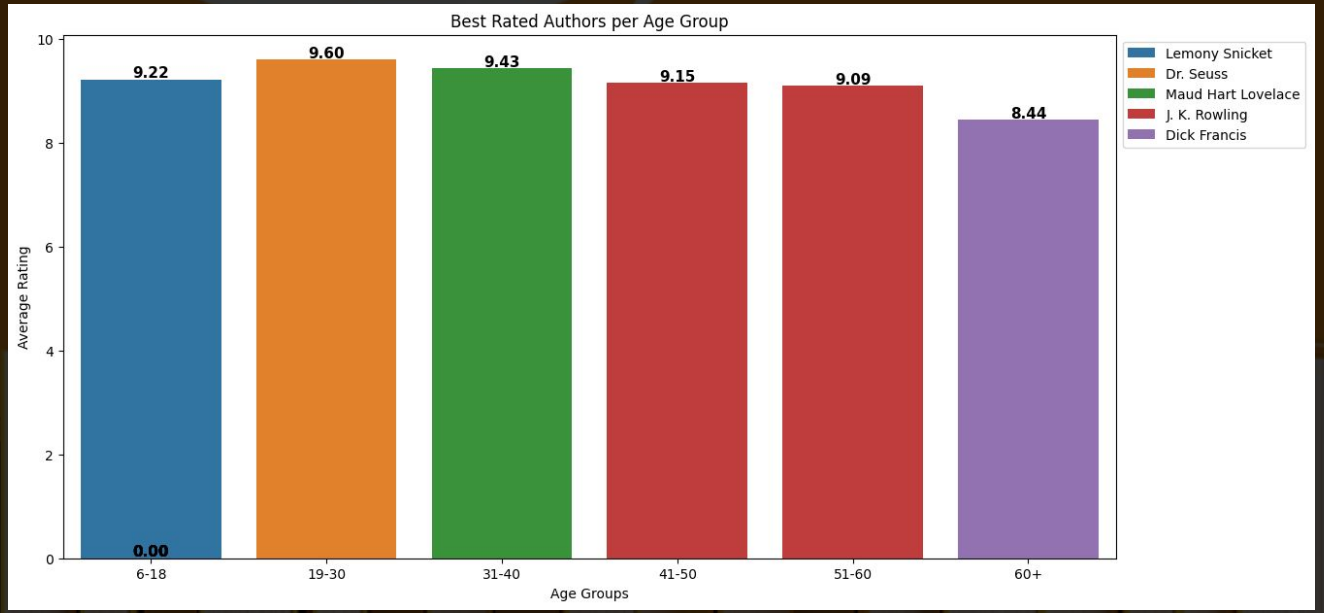
Motivation:

Explore how author preferences vary across different age groups.

Identify the most and least liked authors for each age group.

Method: Merging, grouping, average rate calculation, rating count filtering

Finding:



Finding: Most and Least Liked Authors per Age Group

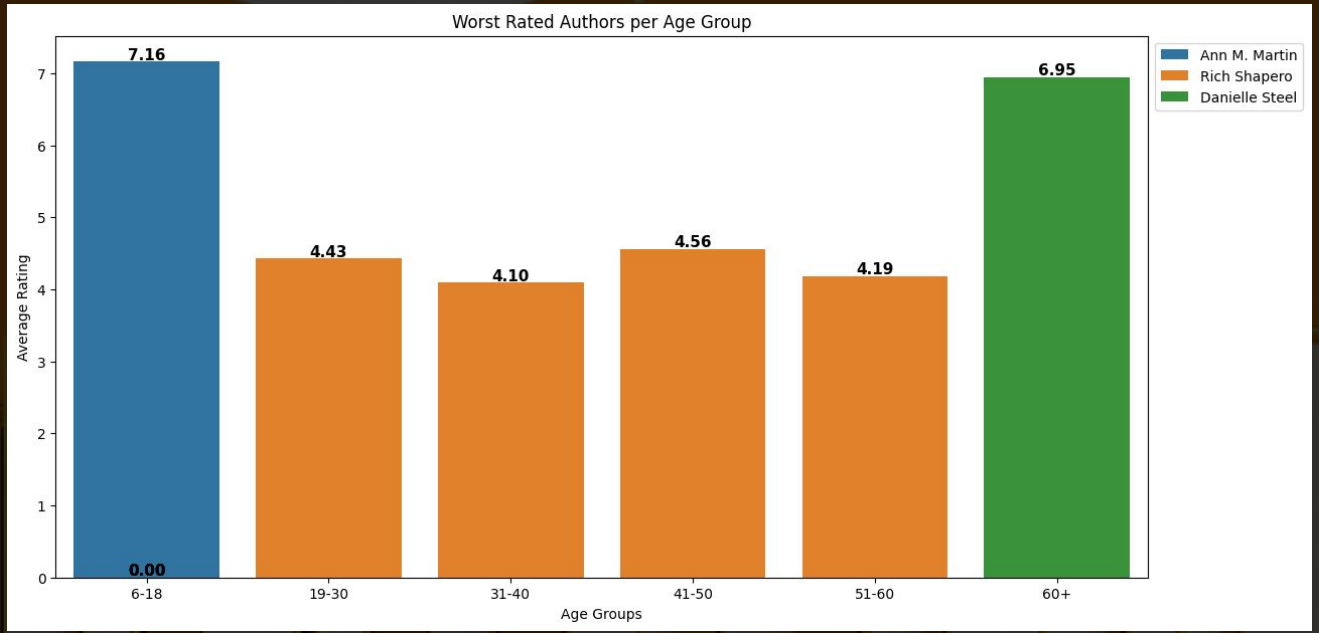
Motivation:

Explore how author preferences vary across different age groups.

Identify the most and least liked authors for each age group.

Method: Merging, grouping, average rate calculation, rating count filtering

Finding:



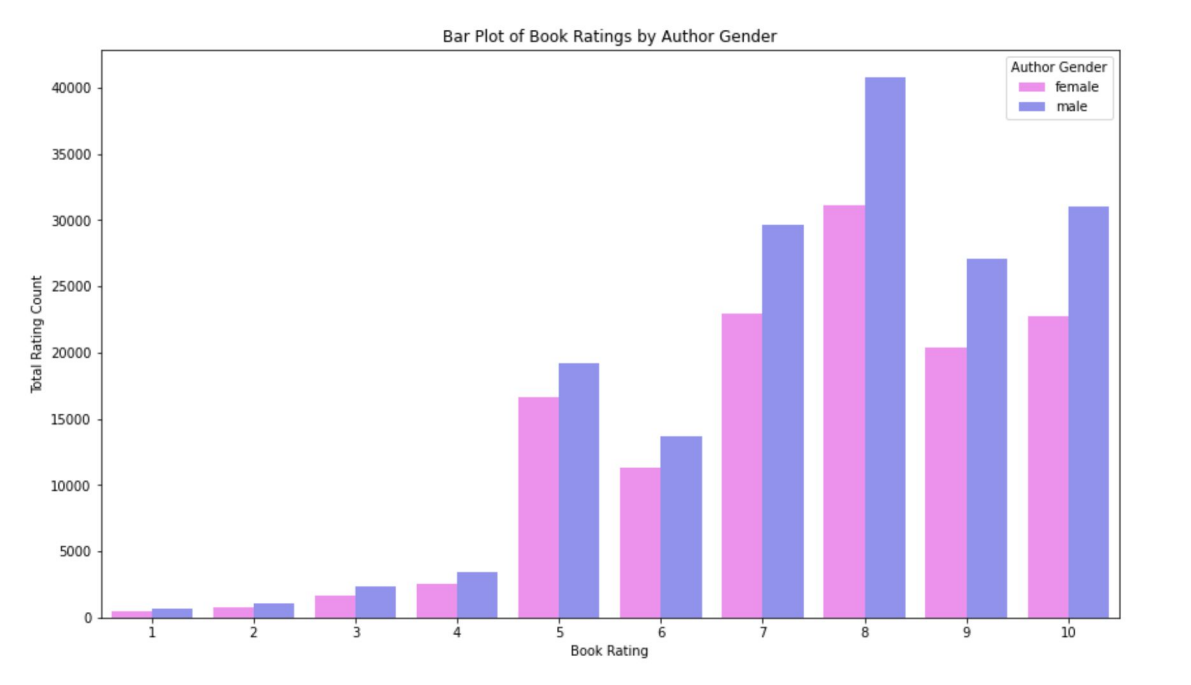
Finding: Female Authors are Lower Rated than Male Authors!

Motivation: Women face glass ceiling when climbing corporate ladder due to gender bias but is there a similar gender bias for female authors against male authors?

Method:

1. Genderize the first names of the authors.
2. Retrieve rating distribution per gender.
3. Conduct One-Sided Mann-Whitney U Test.

Finding:

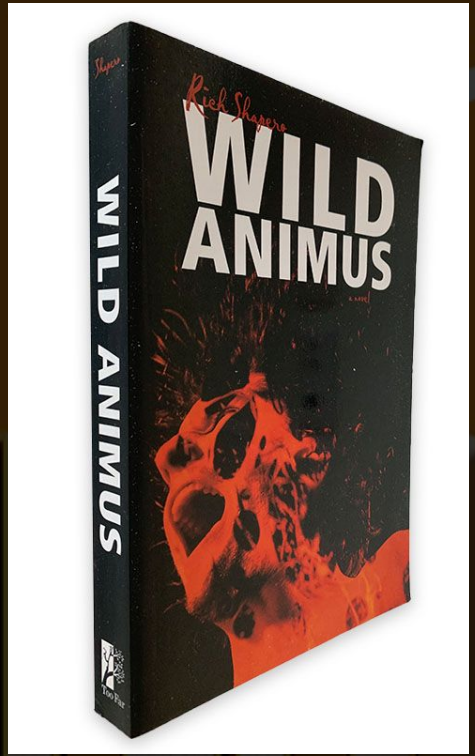


One-sided Mann-Whitney U test (alternative='greater'): U statistic = 11286741544.0, p-value = 0.0000000000

Finding: Female Authors are Lower Rated than Male Authors!

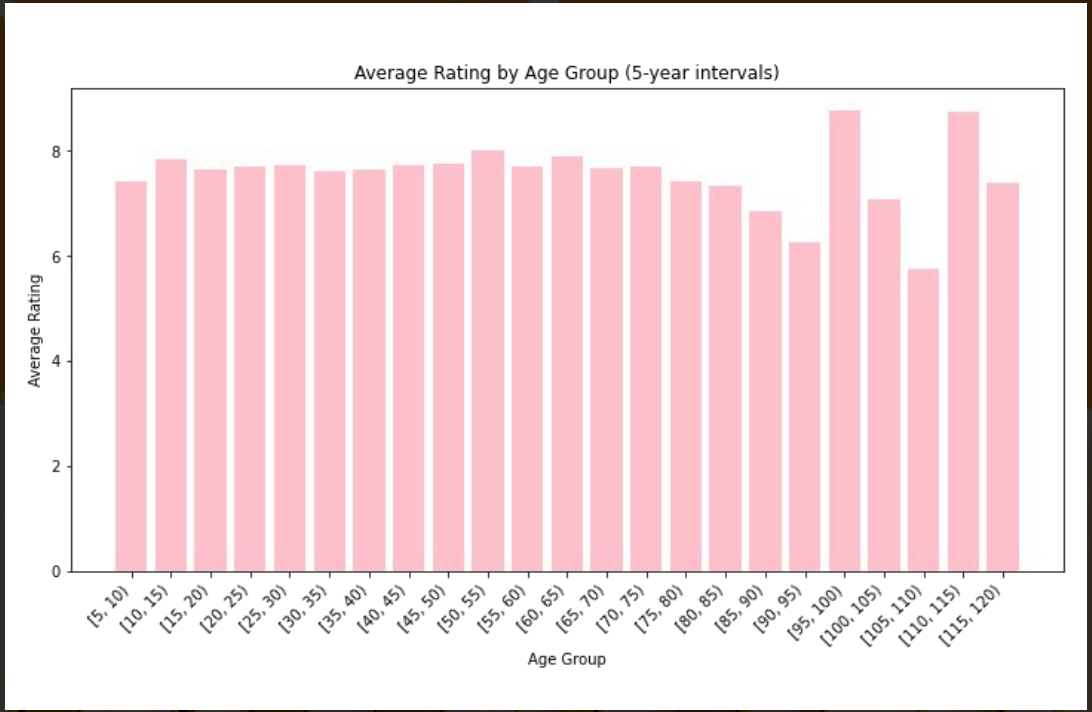
Finding: Publishers & Authors: Popularity

- Best and Worst



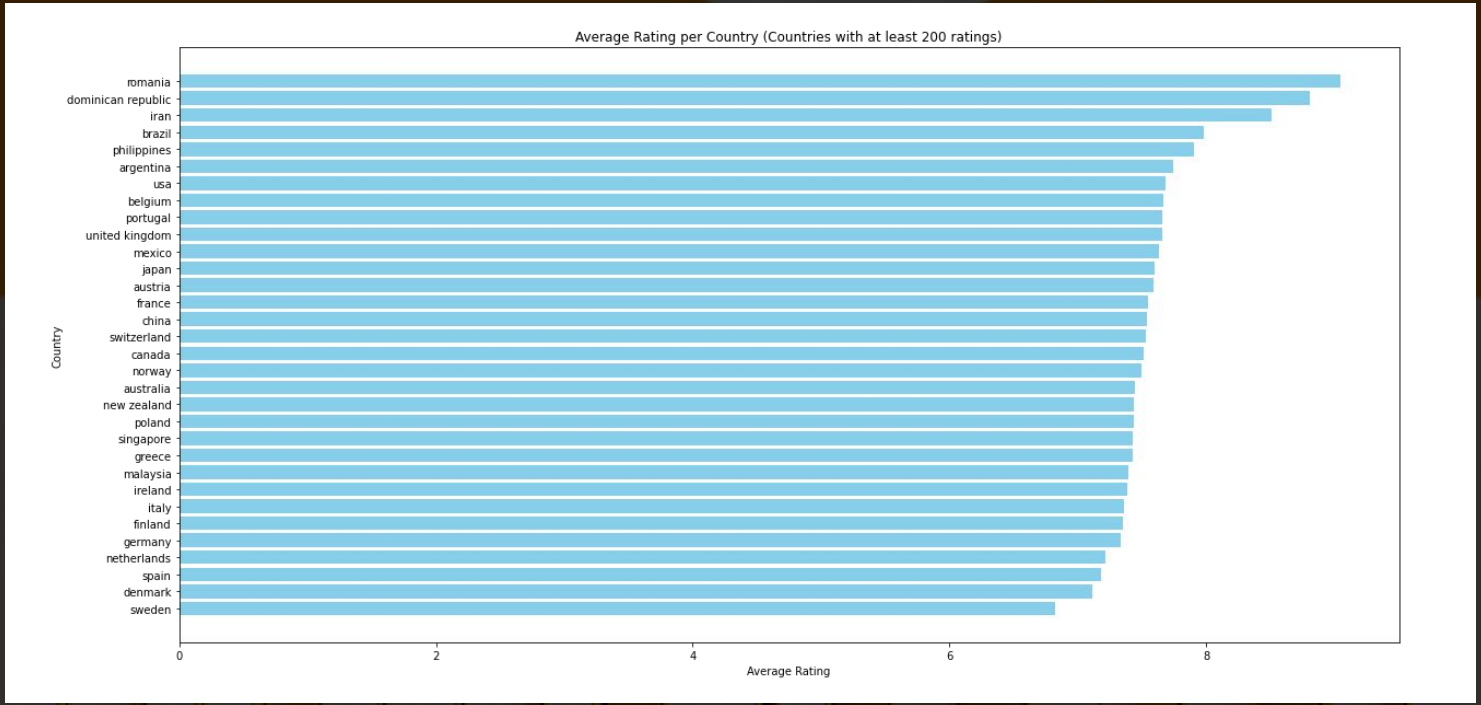
Finding: Publishers Rating Differences

→ Kruskal-Wallis test



Finding: Publishers Rating Differences

→ Kruskal-Wallis test



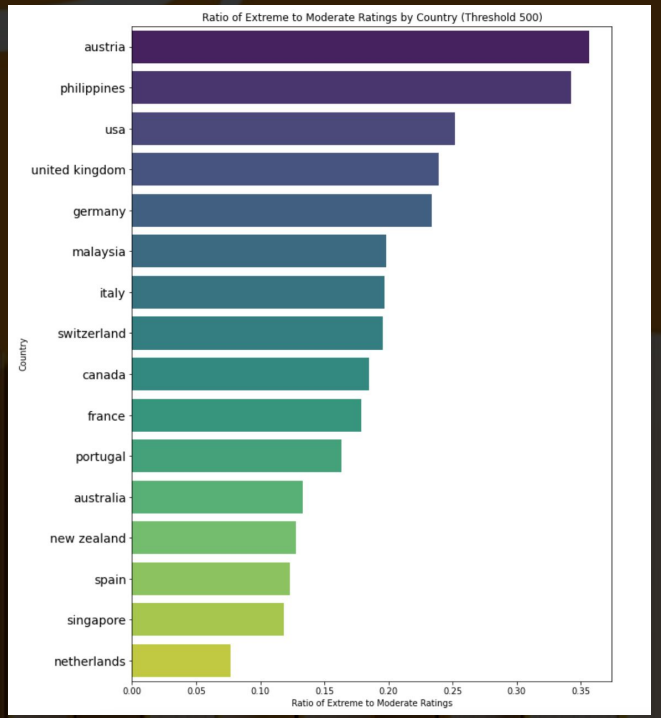
Finding: Country Indicates if a User is an Extremist Rater!

Motivation: Could where a user if from have an effect on them to drive them into being extremist raters (rating 1 or 10)?

Method:

1. Retrieve raters amongst users and categorize them into countries.
2. Retrieve the ratio of extreme rates (1 or 10) to moderate rates (2 to 9) per country.
3. Create a contingency table and perform the chi-square test.

Finding:



Chi-Square Test Statistic: 1468.214183509379, p-value: 3.4713748927200112e-304

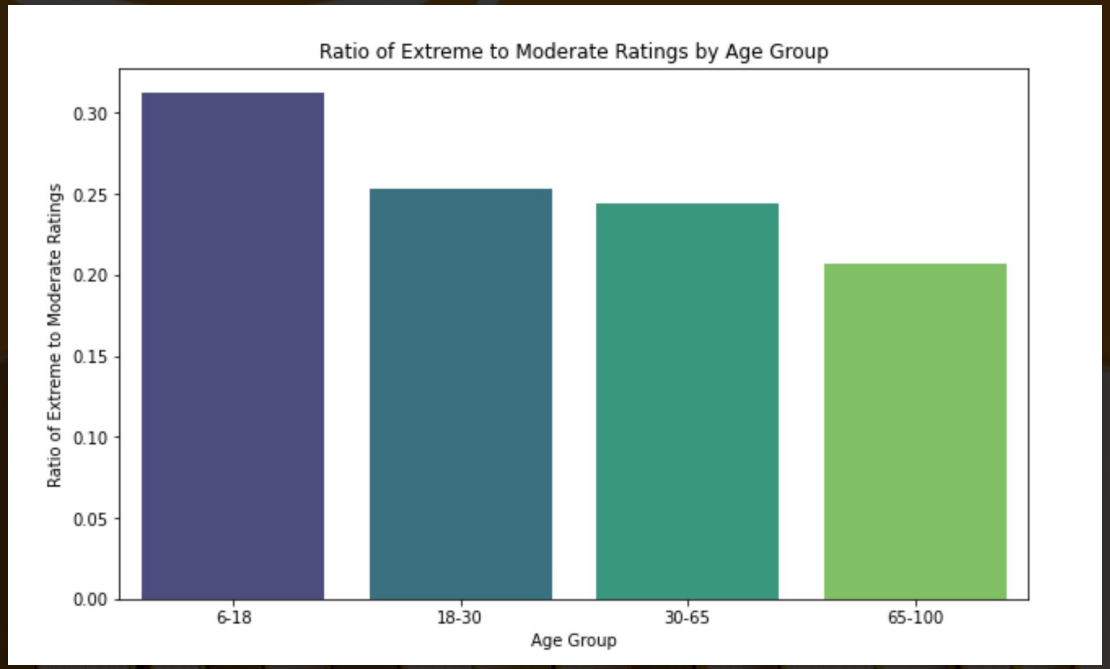
Finding: Age Group Indicates if a User is an Extremist Rater!

Motivation: Could the age group of a user have an effect on them to drive them into being extremist raters (rating 1 or 10)?

Method:

1. Retrieve raters amongst users and categorize them into age groups.
2. Retrieve the ratio of extreme rates (1 or 10) to moderate rates (2 to 9) per country.
3. Create a contingency table and perform the chi-square test.

Finding:



Chi-Square Test Statistic: 146.69790949501035, p-value: 1.3584276074351713e-31

Finding: No Correlation between Title Sentiment and Rating!

Motivation: Could striking titles with negative or positive sentiment have an effect on the rater to drive them to rate in high or low values by appealing to their emotions?

Method:

1. Assign sentiments to the titles of the books using VADER.
2. Categorize the sentiment values into Negative, Neutral and Positive.
3. Retrieve the number of books per rating and category.
4. Compute Spearman Correlation.

Finding:

Spearman correlation between sentiment and ratings: correlation = 0.0097, p-value = 0.0000000019

Finding: Age Prediction

Motivation:

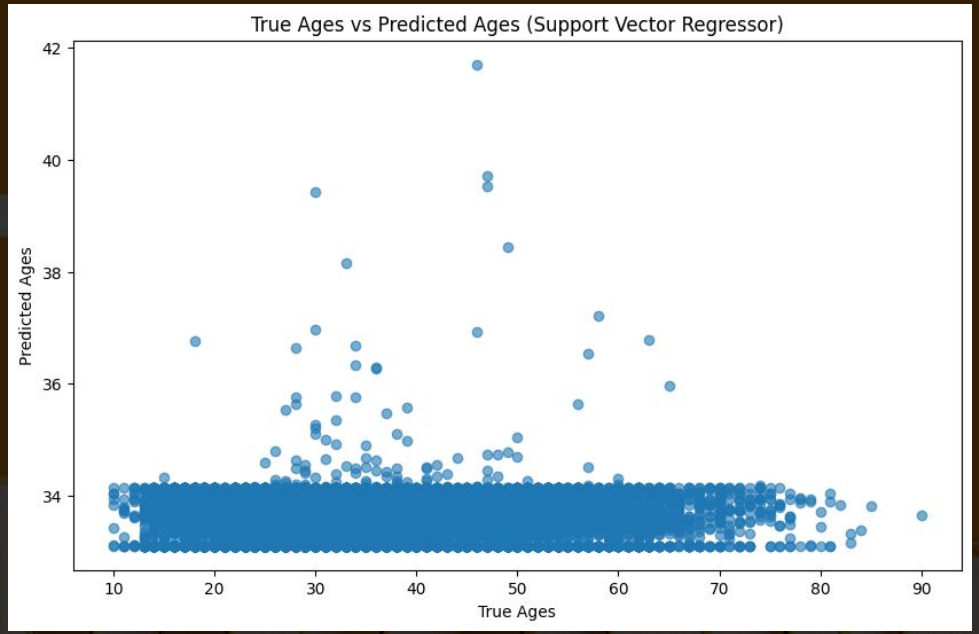
Predict User Age with Rating + Reading habis?

Issue: Missing/Invalid Values

Use: Fill in NaN values

Method: ML method (e.g. linear regression, SVP)

Mae of predicting age by avg rating of number of ratings with SVM: 11.09 !!!

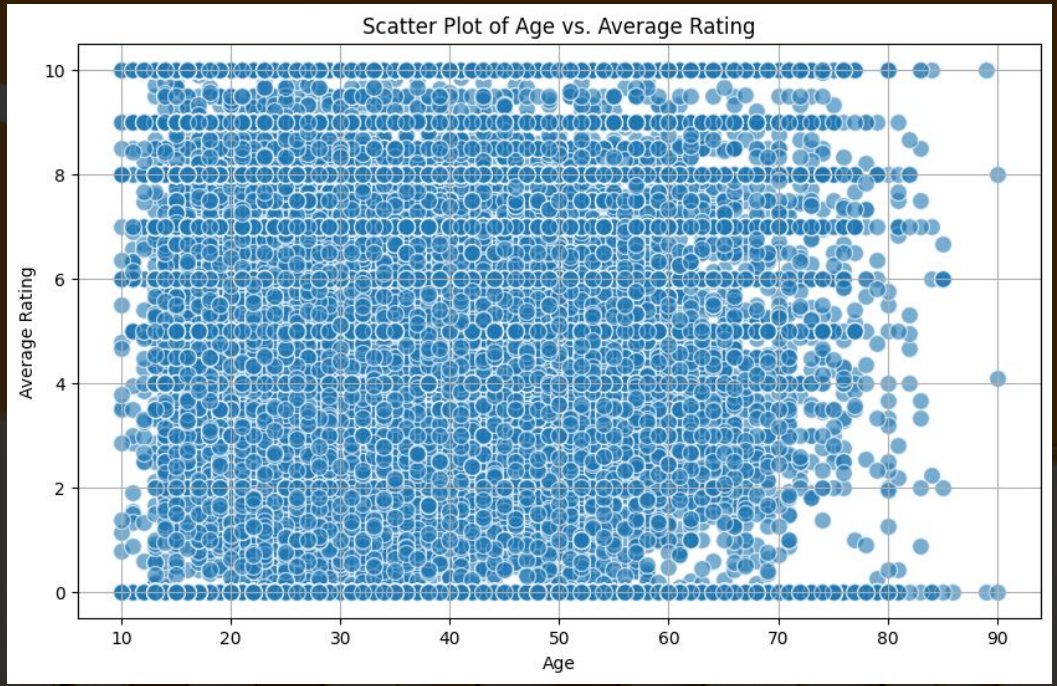


Finding: Age Prediction

Error: Very high

→ No relationship between Age and Average Rating (as Kruskal-Wallis showed)

→ Prediction not feasible



Further Work

- Book Recommendation Systems
- Color-Analysis on Book Covers
- Country prediction
- Analysis on Publishers Focusing on Certain Age Groups
- User Interaction Analysis

Further Work: Age Prediction

A POSSIBLE SOLUTION: EXTRACT MORE FEATURES

It is unreasonable to predict the age with a few features.

Possible features: location (country), preferred publisher(s), favorite book(s), etc.

Possible algorithms: Linear regression, SVM, Neural networks, Evolutionary algorithm, etc.

MAE of Predict ages by countries

and favorite publishers:

12.48 !!!!!!

Further Work Alternative: Predict location (country)

Challenge:

Naming of countries is not consistent throughout the dataset

Solution:

Eliminate invalid names + substitute informal names by names in *pycountry*

```
array(['usa', 'spain', '', 'canada', 'france', 'united kingdom',
       'portugal', 'belgium', 'india', 'philippines', 'germany', 'greece',
       'netherlands', 'finland', 'romania', 'australia', 'pakistan',
       'austria', 'switzerland', 'new zealand', 'italy', 'malaysia',
       'chile', 'ethiopia', 'taiwan', 'brazil', 'poland', 'argentina',
       'ireland', 'iran', 'saudi arabia', 'slovenia', 'turkey"',
       'hong kong', 'denmark', 'japan', 'turkey', 'norway', 'china',
       'rwanda', 'guatemala', 'mexico', 'sweden', 'thailand', 'colombia',
       'singapore', 'monaco', 'luxembourg', 'samoa',
       'trinidad and tobago', 'bermuda', 'england', 'jersey', 'slovakia',
       'south korea', 'israel', 'ghana', 'qatar', 'catalunya', 'albania',
       'moldova', 'españa', 'czech republic', 'brunei', 'nigeria',
       'costa rica', 'iceland', 'andorra', 'finland"', 'south africa',
       'lleida', 'u.s.a.', 'uruguay', 'lithuania', 'catalonia', 'egypt',
       'united states', 'netherlands"', 'bulgaria', 'zimbabwe', 'iraq',
       'thailand"', 'russia', 'polk', 'hungary', 'indonesia',
       'catalunya spain', 'switzerland"', 'philippines"', 'croatia',
       'papua new guinea', 'algeria', 'grenada', 'galiza', 'spain"',
       'antigua and barbuda', 'kenya', 'italia', 'jamaica', 'guernsey',
       'oman', 'cherokee', 'wales', 'burkina faso', 'nepal', 'vietnam',
       'urugua', 'n/a - on the road', 'afghanistan', 'euskal herria',
       'kuwait', 'yugoslavia', 'alderney', 'ouranos', 'belize', 'latvia',
       'here and there', 'united arab emirates', 'cyprus',
       'solomon islands', 'malta', 'zambia', 'ecuador', 'lebanon',
       'bangladesh', 'the', 'france"', 'syria', 'n/a', 'america',
...
       'saint lucia', 'u.a.e', 'tanzania', 'morocco', '"',
       'guinea-bissau', 'tobago', 'madrid', 'uganda', 'aruba',
       'tajikistan', 'puerto rico', '\\"n/a\\""', 'mongolia',
       'everywhere and anywhere', 'x', 'deutschland', 'galiza neghra',
       'mozambique', 'angola', 'hernando'], dtype=object)
```

References

[1] “Books Dataset,” www.kaggle.com. https://www.kaggle.com/datasets/saurabhbagchi/books-dataset

[2] “ISBN Information - Anatomy of a 10-digit ISBN,” isbn-information.com, 2021. https://isbn-information.com/the-10-digit-isbn.html

[3] “List of the verified oldest people,” Wikipedia, May 18, 2021. https://en.wikipedia.org/wiki/List_of_the_verified_oldest_people#:~:text=The%20oldest%20person%20ever%20whose


QUESTIONS?
Q&A
Q&A?
QUESTIONS
Q&A?
QUESTIONS
Q&A
QUESTIONS
Q&A
OA