goodreads

Authors & Books

An analysis using goodreads dataset and much more

Team 2: Anirudh Fatemeh Samantha Zhenwei

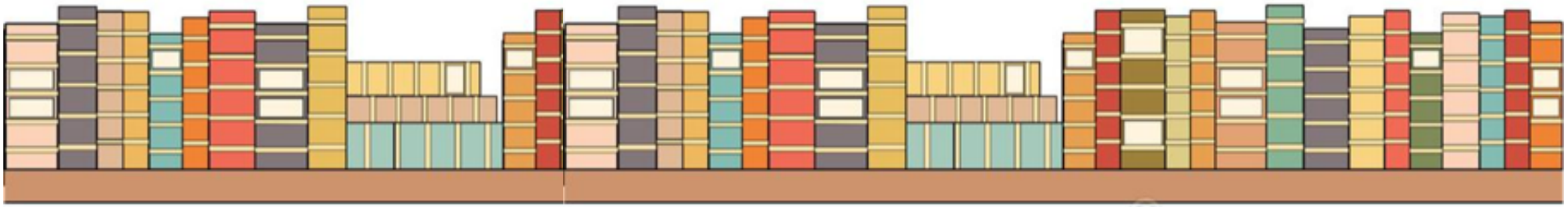
Overview

Authors:

- Ratings ⭐
- Influence & Popularity 🗣️
- Favoured Genres 📚

Books:

- Features & Genres
- Frequent Words in Titles
- Title Length
- Genres Spanned



Datasets | Methodology

- Data Pre-processing: Merge / Clean (omit null values, change and filter columns, …)
- Data Analysis: Queries
- Generating/Improving Plots
- Genre Prediction Model

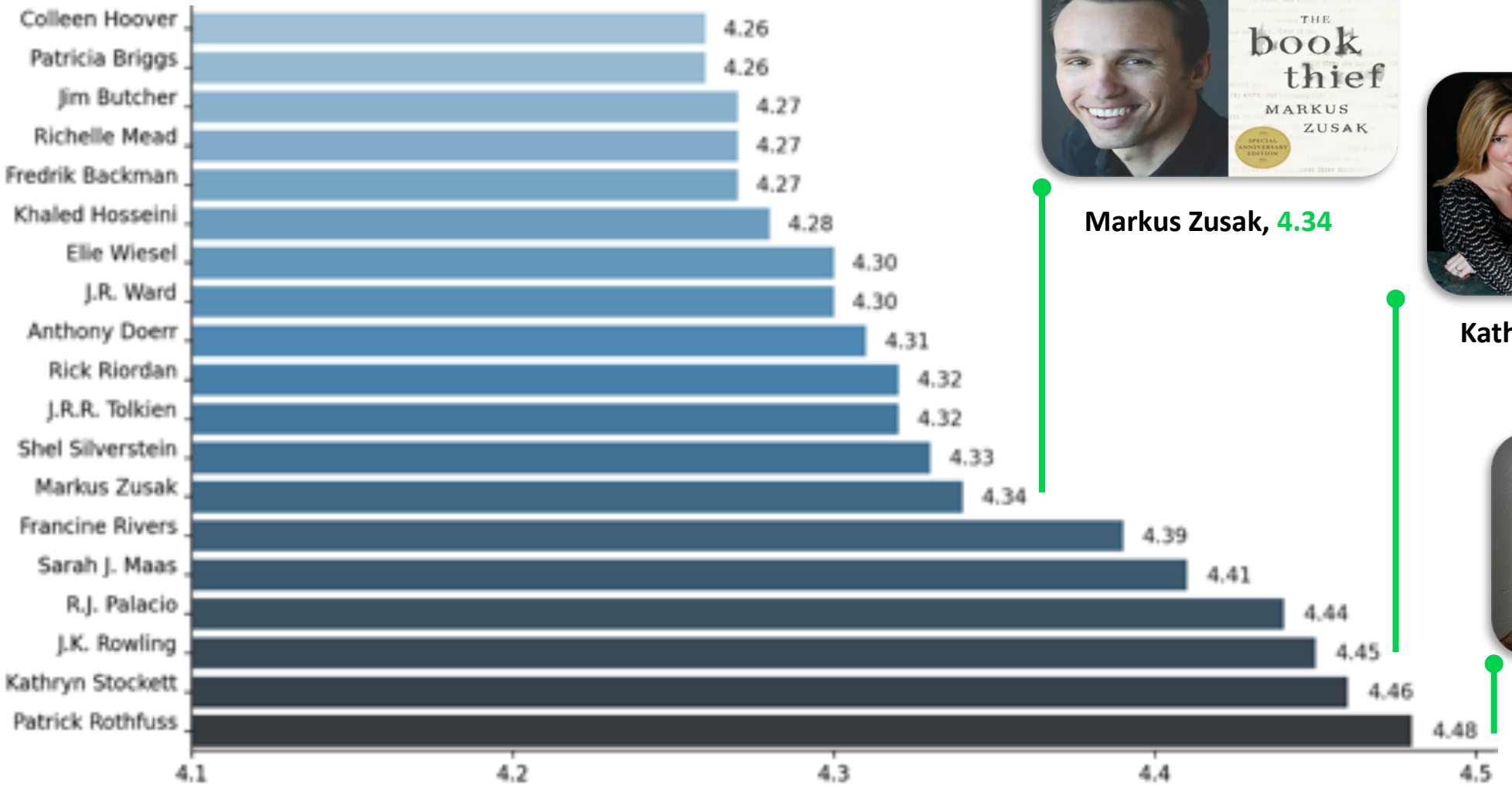
Dataset	Shape (rows, columns)	Main Columns Used	Main Preprocessing Tasks
Goodreads best books ever (Kaggle)	54301 rows, 12 columns	genres, title, etc.	Splitting genres column, Omitting null values
Authors dataset (Kaggle)	22892 rows, 20 columns	author, gender, country, latitude, longitude,	Splitting date to get year, Omitting null values
Authors dataset (Github)	209518 roes, 20 columns	most except isbn	Omitting null values, filtering, etc.
Goodreads books (Kaggle)	11123 rows, 12 columns	tile, rating+count, author, …	Omitting null values

Merged

	title	author(s)	genres	format	average_rating	rating_count	review_count	publication_date	num_pages
284	When We Were Orphans	Kazuo Ishiguro	Fiction|Historical|Historical Fiction|Mystery|...	Paperback	3.47	21291	2037	12/1/2007	336 pages
285	The Rachel Papers	Martin Amis	Fiction|Novels|Contemporary|European Literatur...	Paperback	3.59	8404	397	9/29/1992	240 pages
286	The Delta Star	Joseph Wambaugh	Fiction|Mystery|Mystery|Crime	Paperback	3.63	755	22	1/1/1984	291 pages
287	The Best Short Stories	J.G. Ballard|Anthony Burgess	Short Stories|Fiction|Science Fiction|Literature	Paperback	4.20	1441	92	2/13/2001	302 pages
288	Homegrown Democrat: A Few Plain Thoughts from ...	Garrison Keillor	Nonfiction|Politics|Humor|Autobiography|Memoir	Paperback	3.97	1266	154	8/29/2006	288 pages

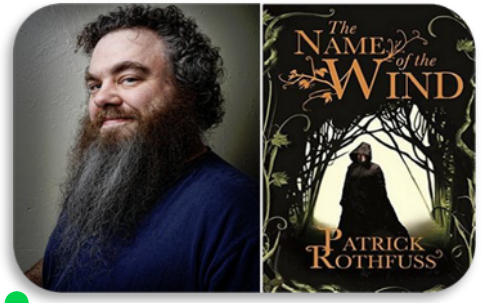
Part I: An Analysis of the Authors

Average Ratings of All the Authors Around Us



Markus Zusak, 4.34

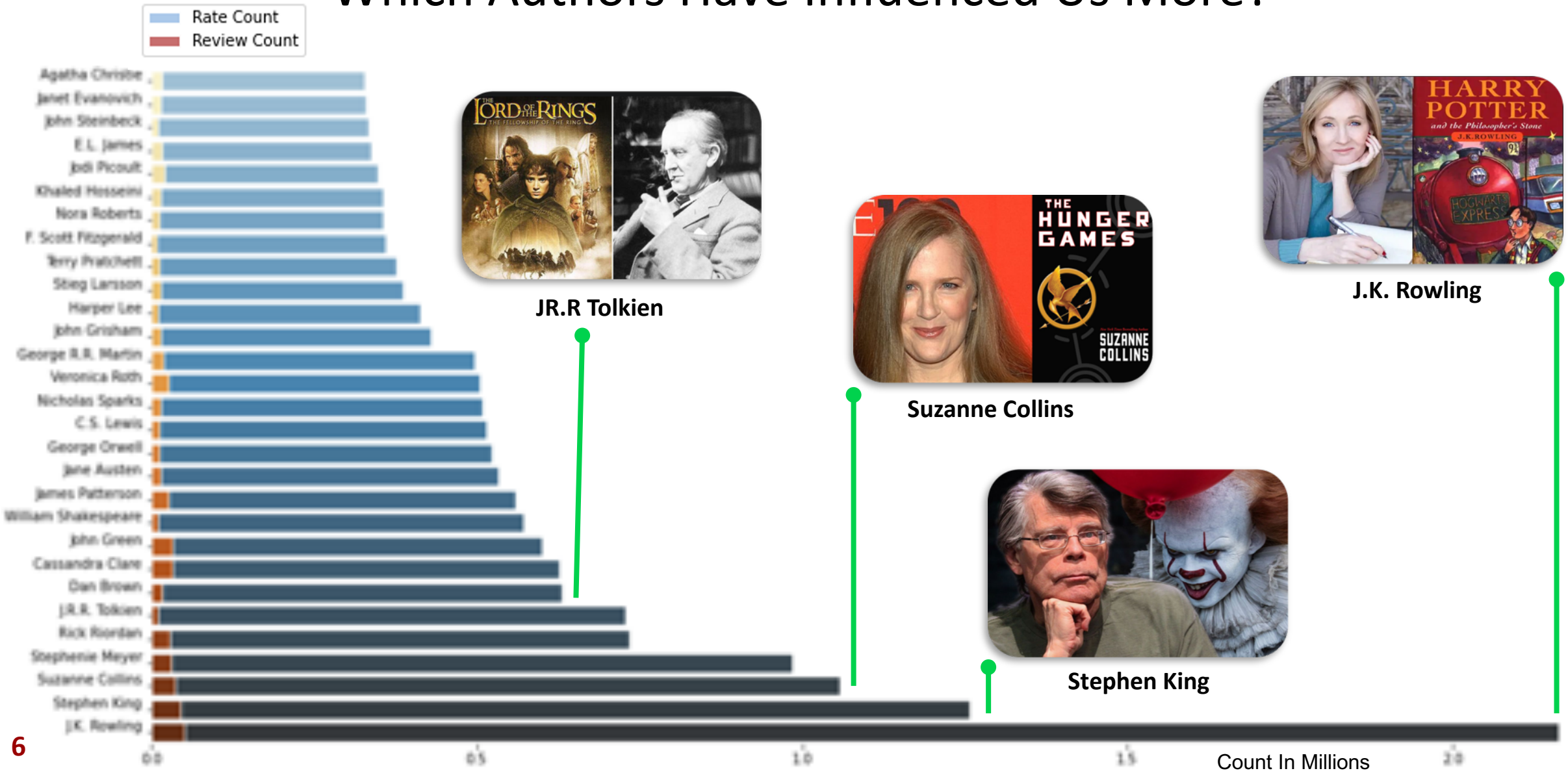
Kathryn Stockett, 4.46



Patrick Rothfuss, 4.48

Author's Rating

Which Authors Have Influenced Us More?



JR.R Tolkien

Suzanne Collins

Stephen King

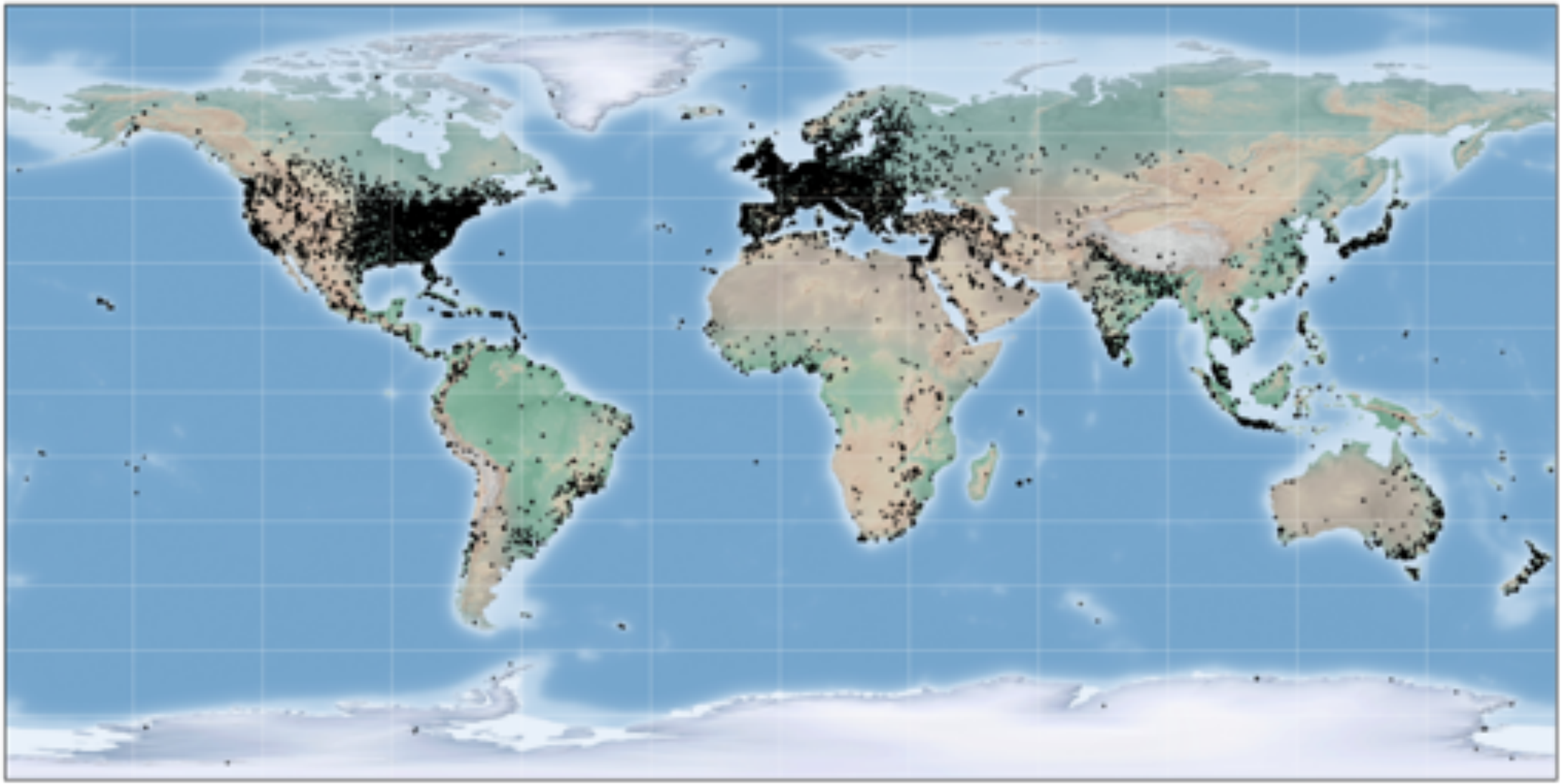
J.K. Rowling

Which Genres do the Authors Like to Write in 21st Century?

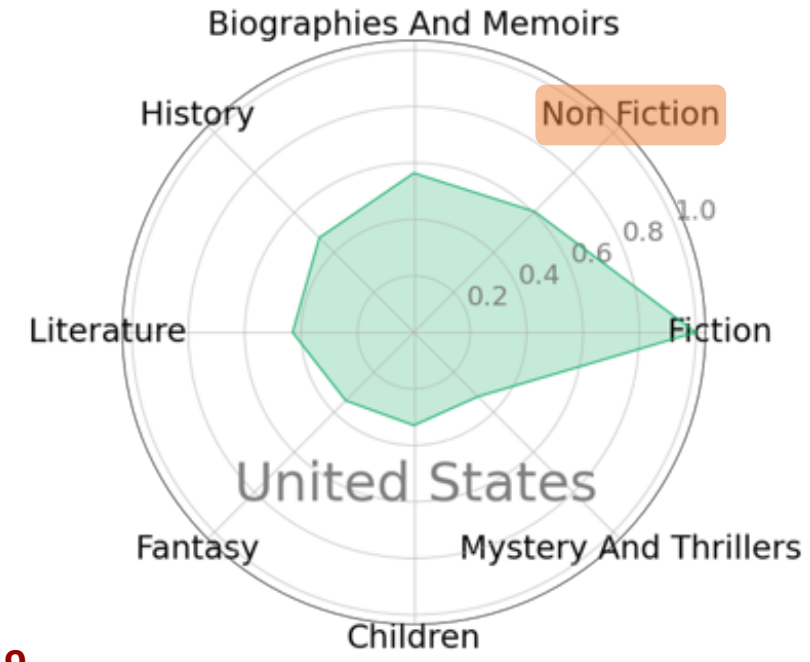
Mystery

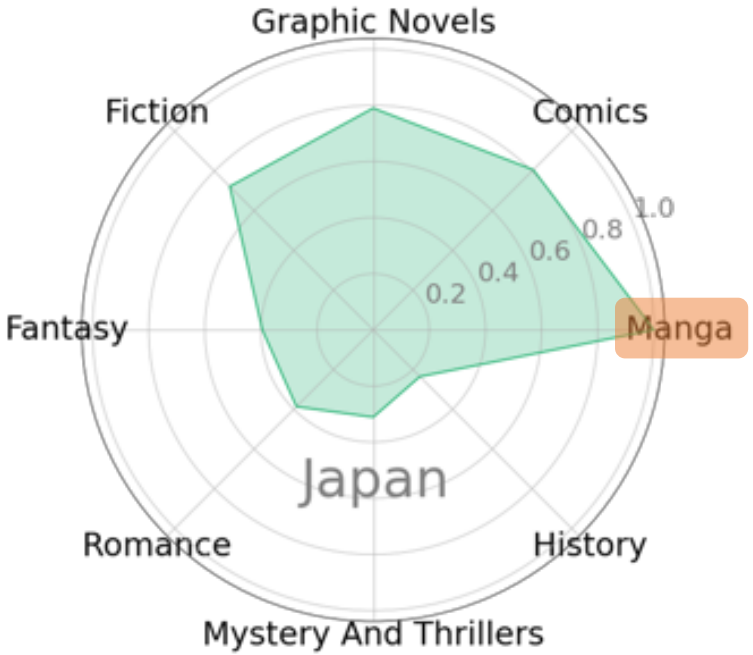
Fantasy

Where have the Authors Lived from the 17th to 21st Century?

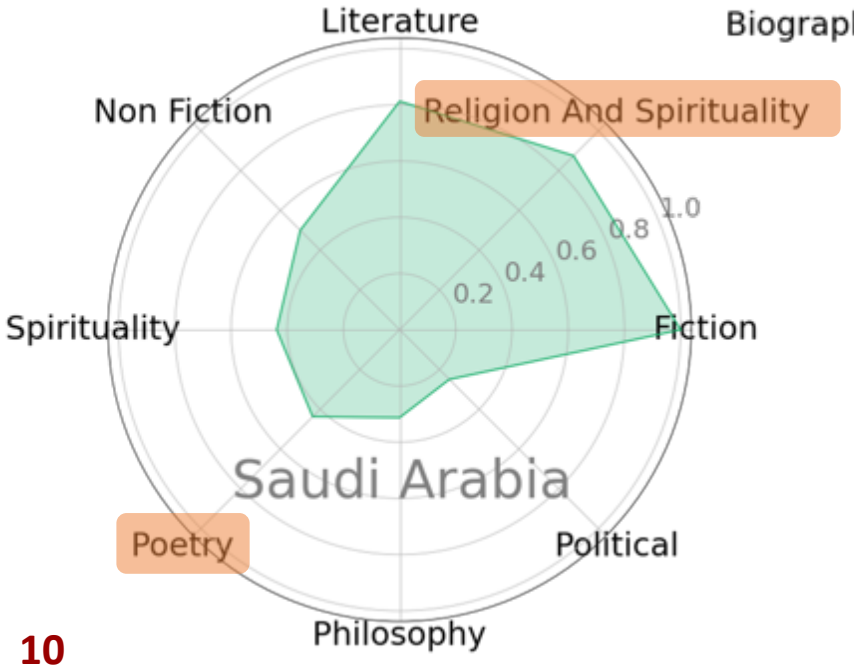


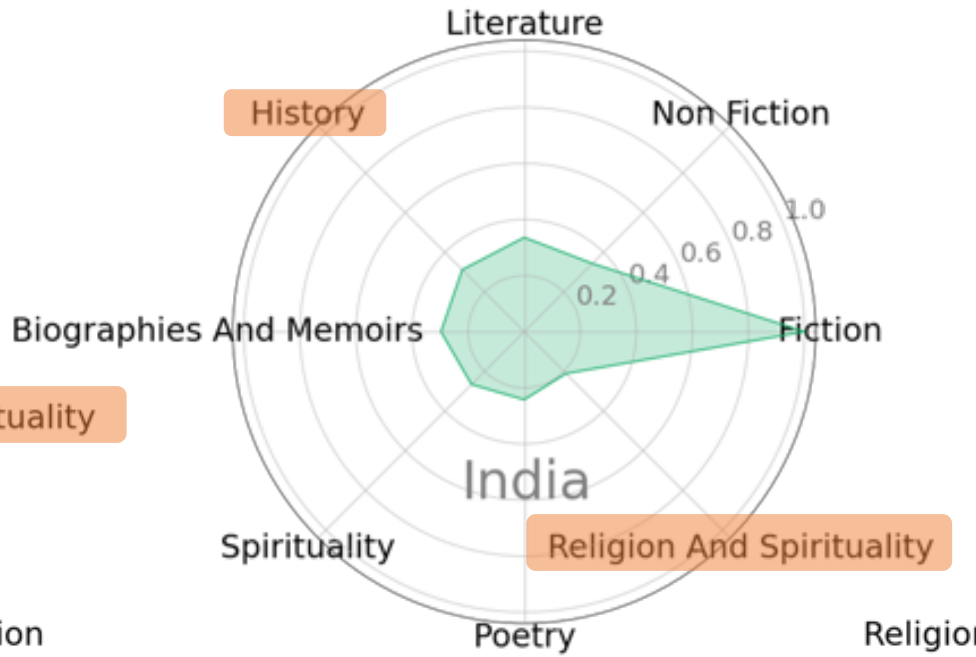
Distribution of Genres between Different Countries in 20th Century

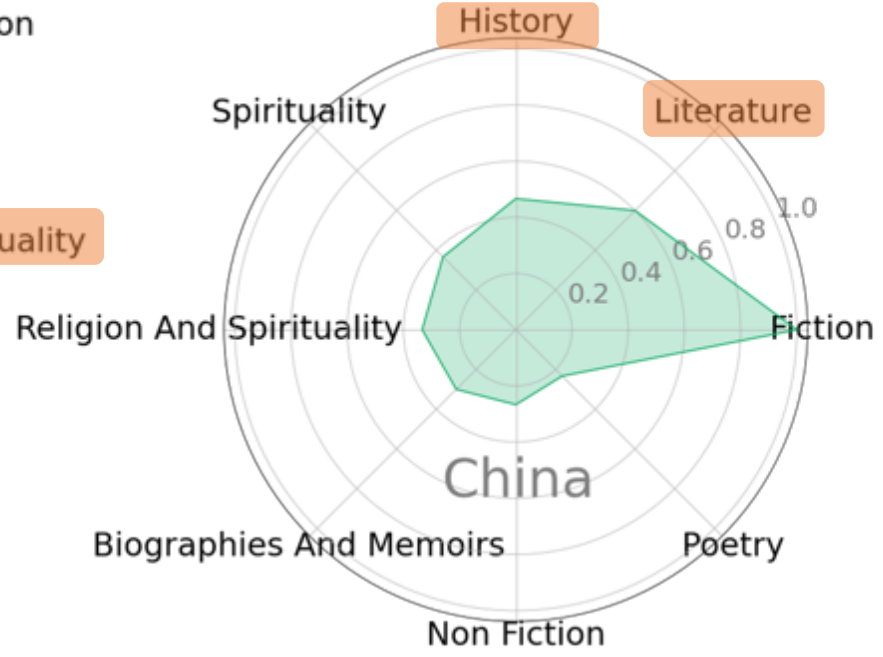




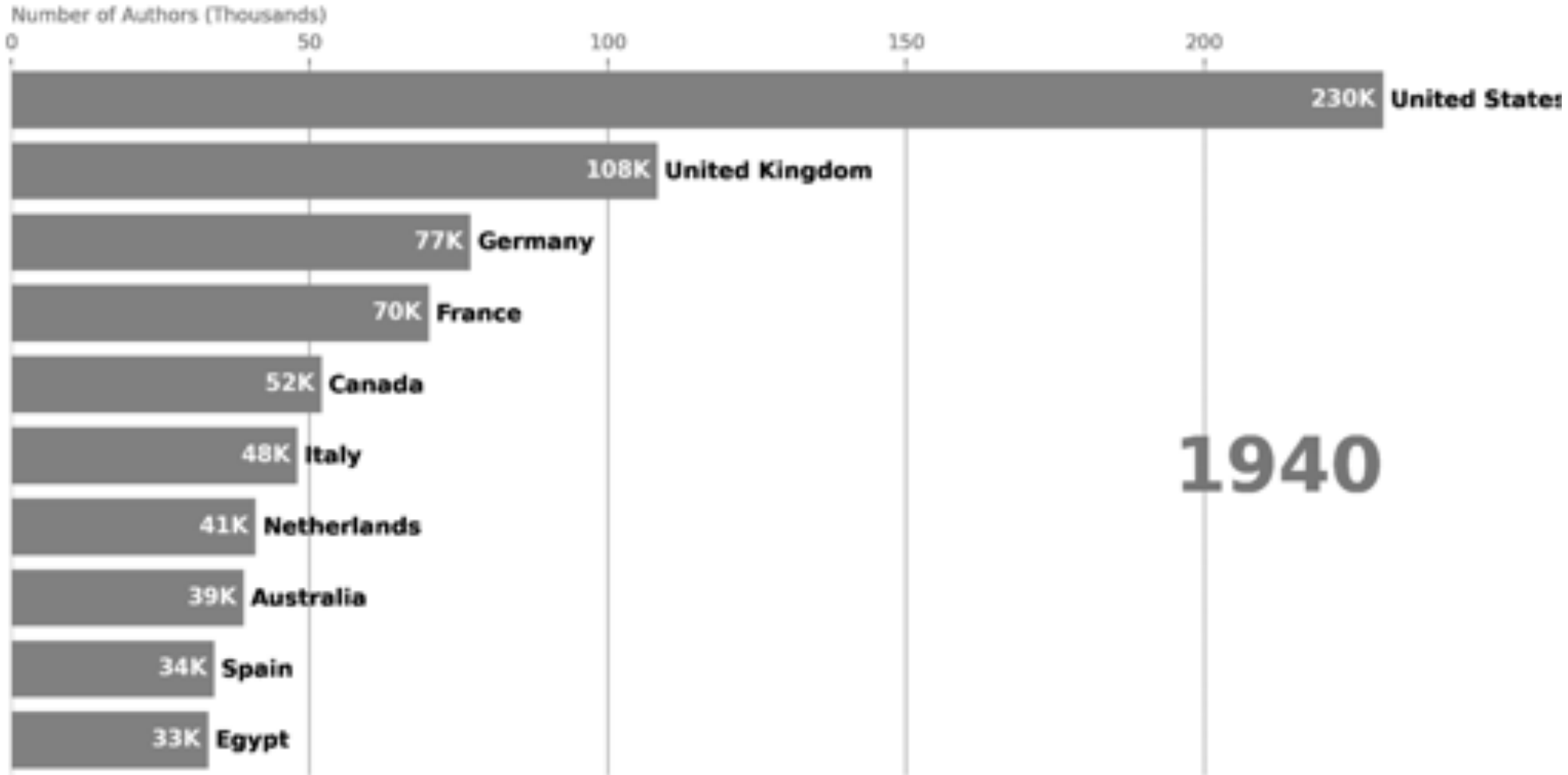
Distribution of Genres between Different Countries in 20th Century



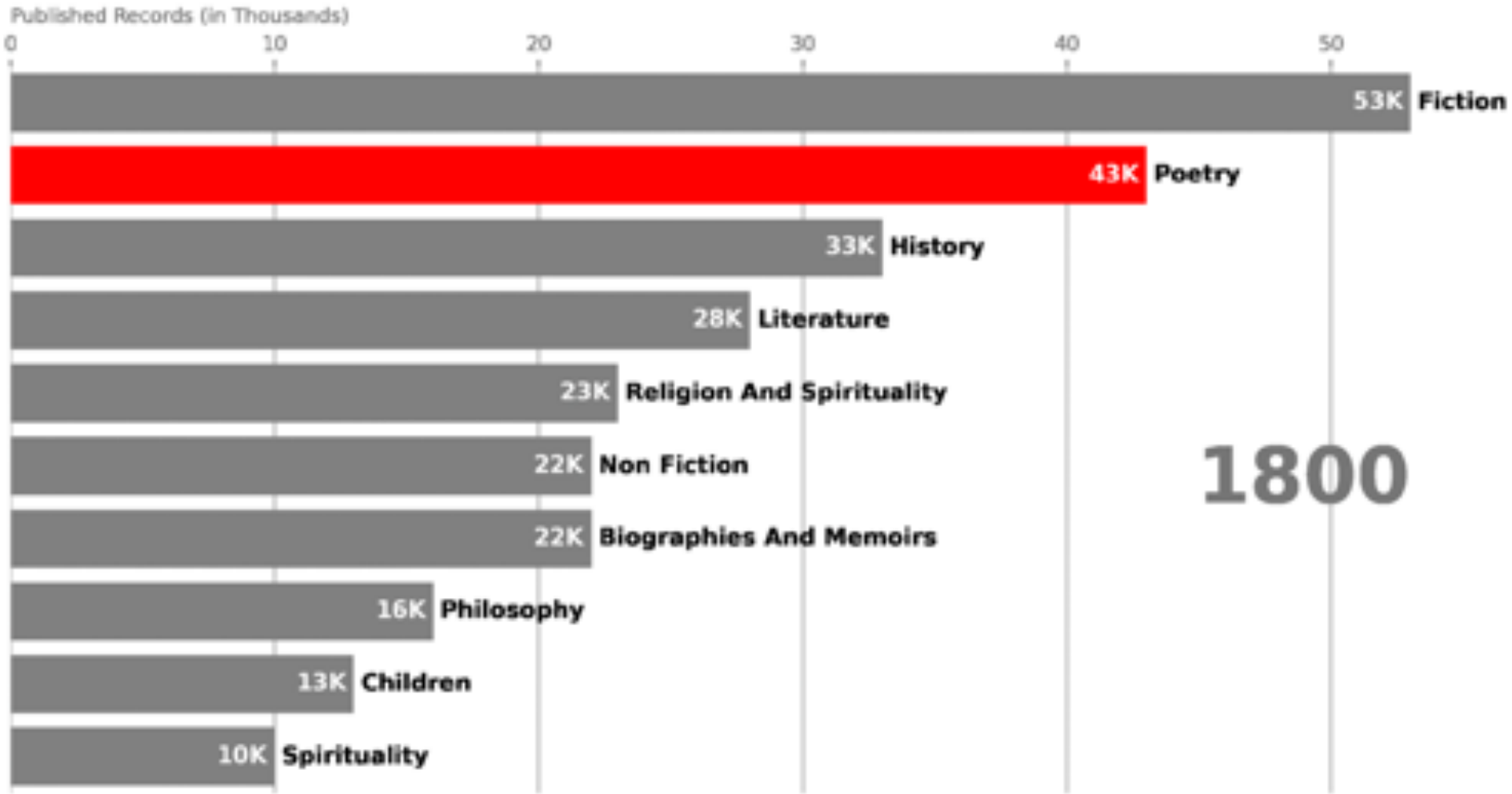




The Rise of Manga

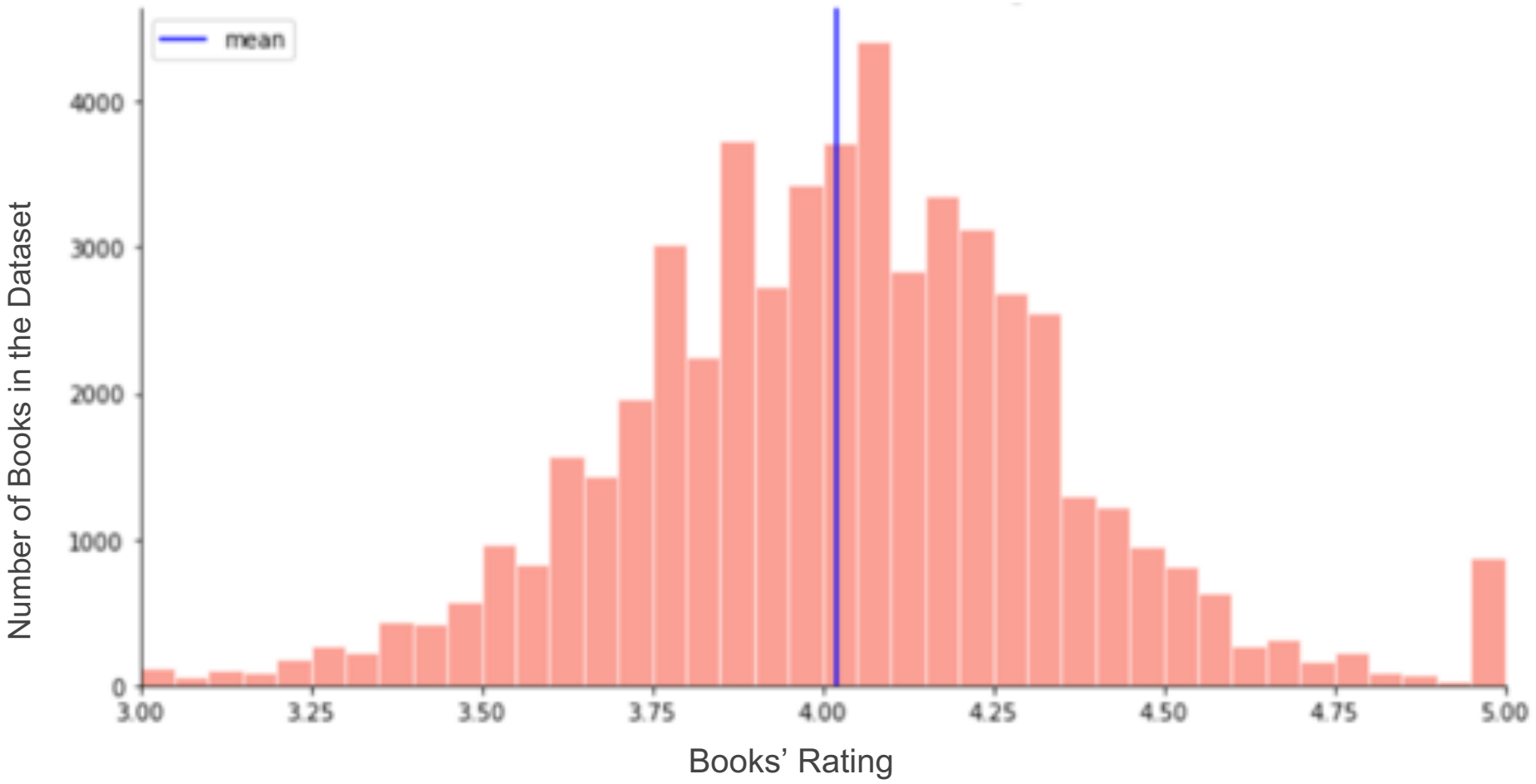


The Fall of Poetry

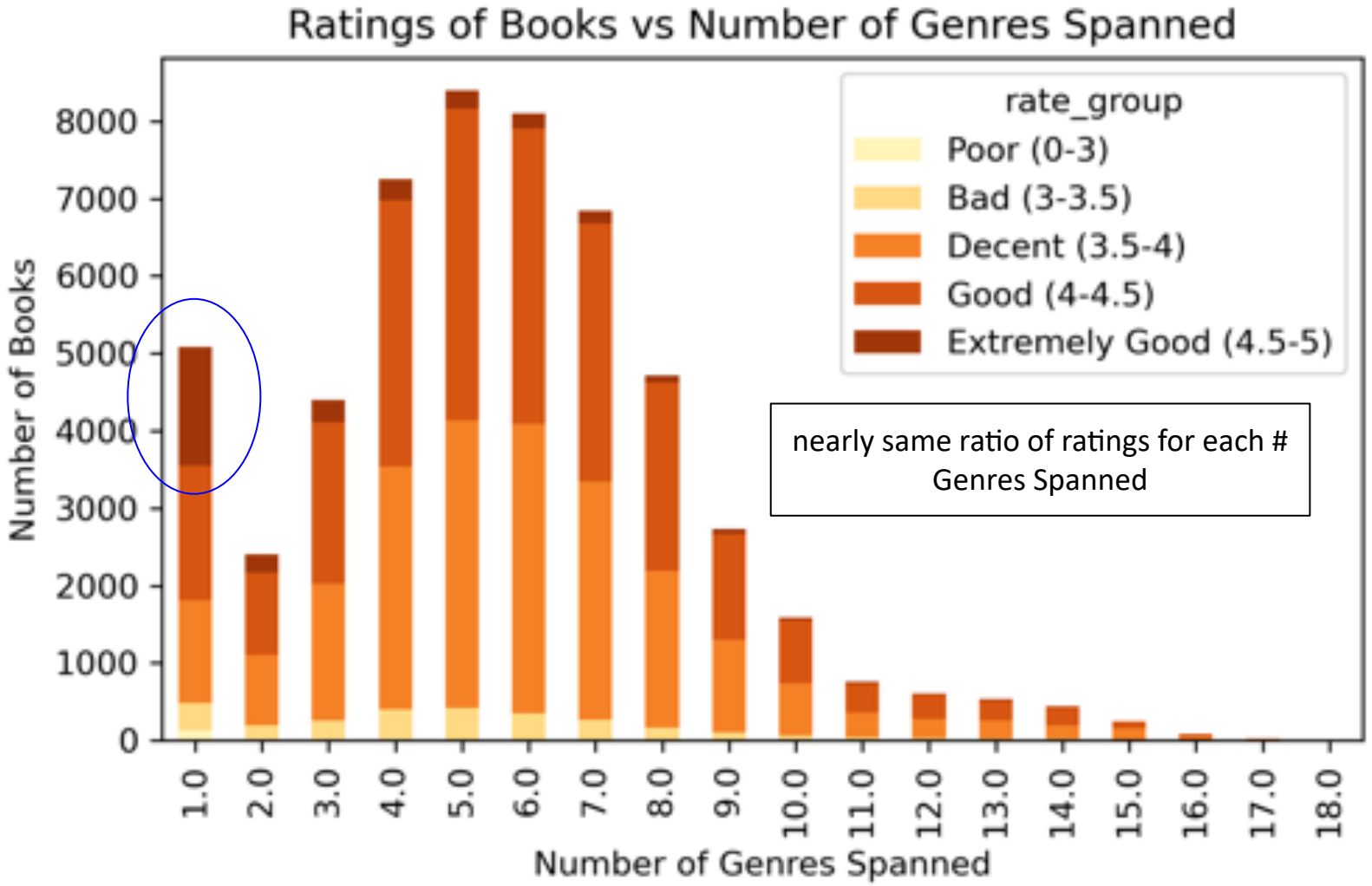


Part II: An Analysis of the Books

Distribution of Book Ratings



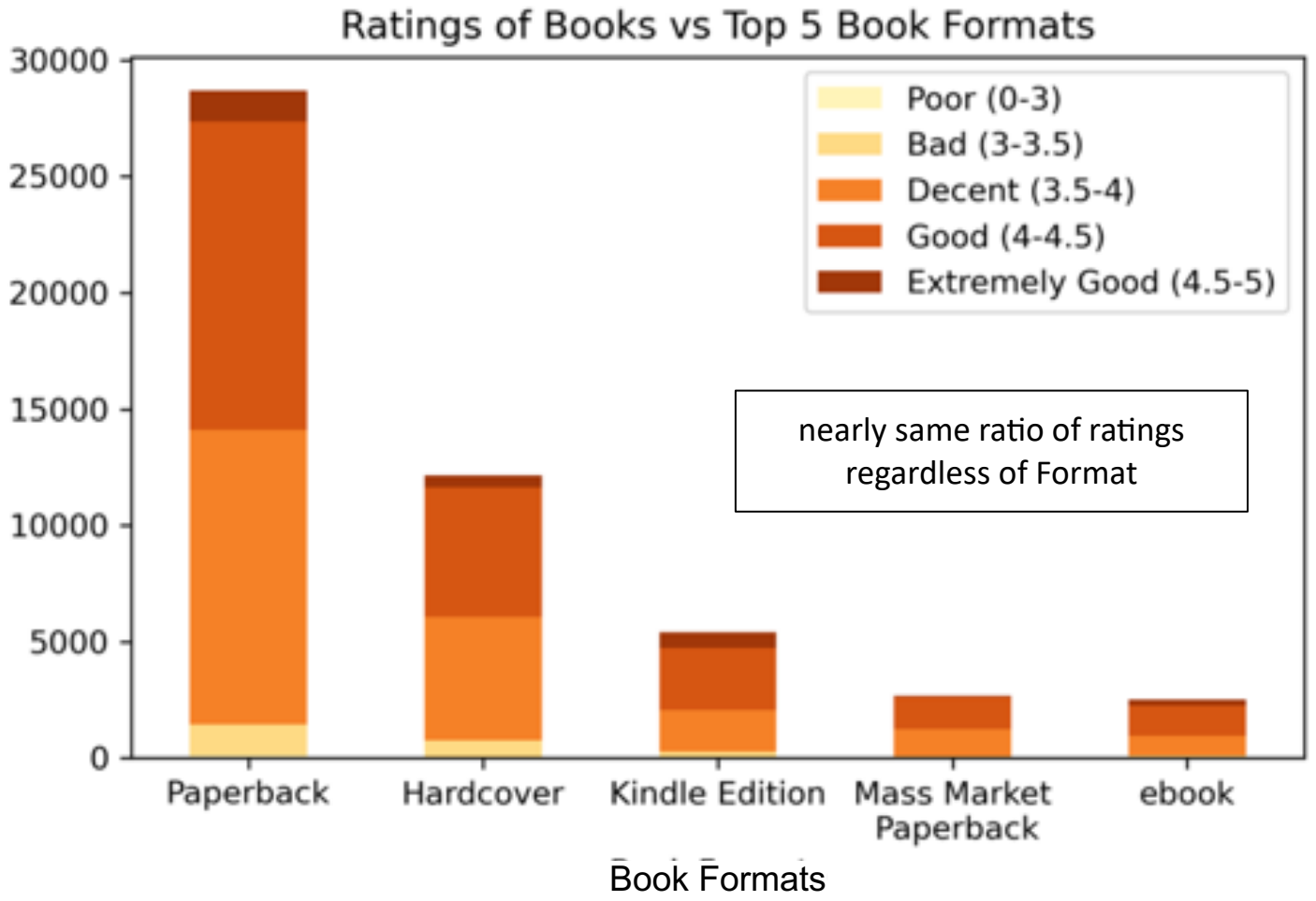
Does a Diverse book have a better Rating?



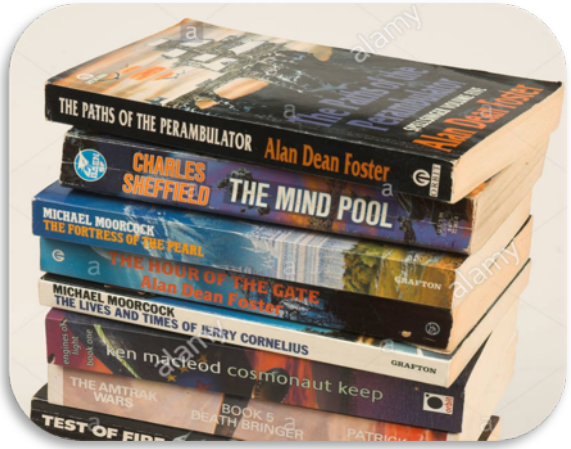
Large Print
Colour Code Index

- Action / Adventure
- Family Saga
- Classic
- Humour
- Crime
- Chick Lit
- Fantasy / Sci-Fi
- Historical
- Horror
- Romance
- General
- Short Stories
- Thriller
- War

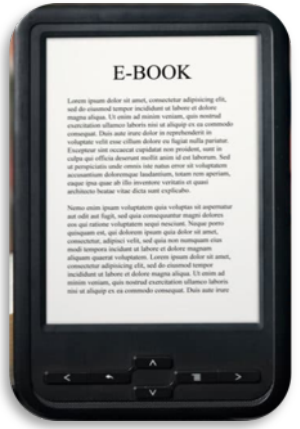
Does the Book Format influence a Books' Rating?



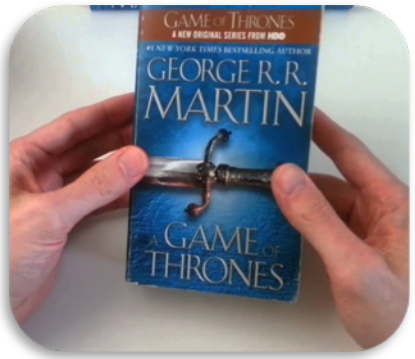
nearly same ratio of ratings regardless of Format



Paperback

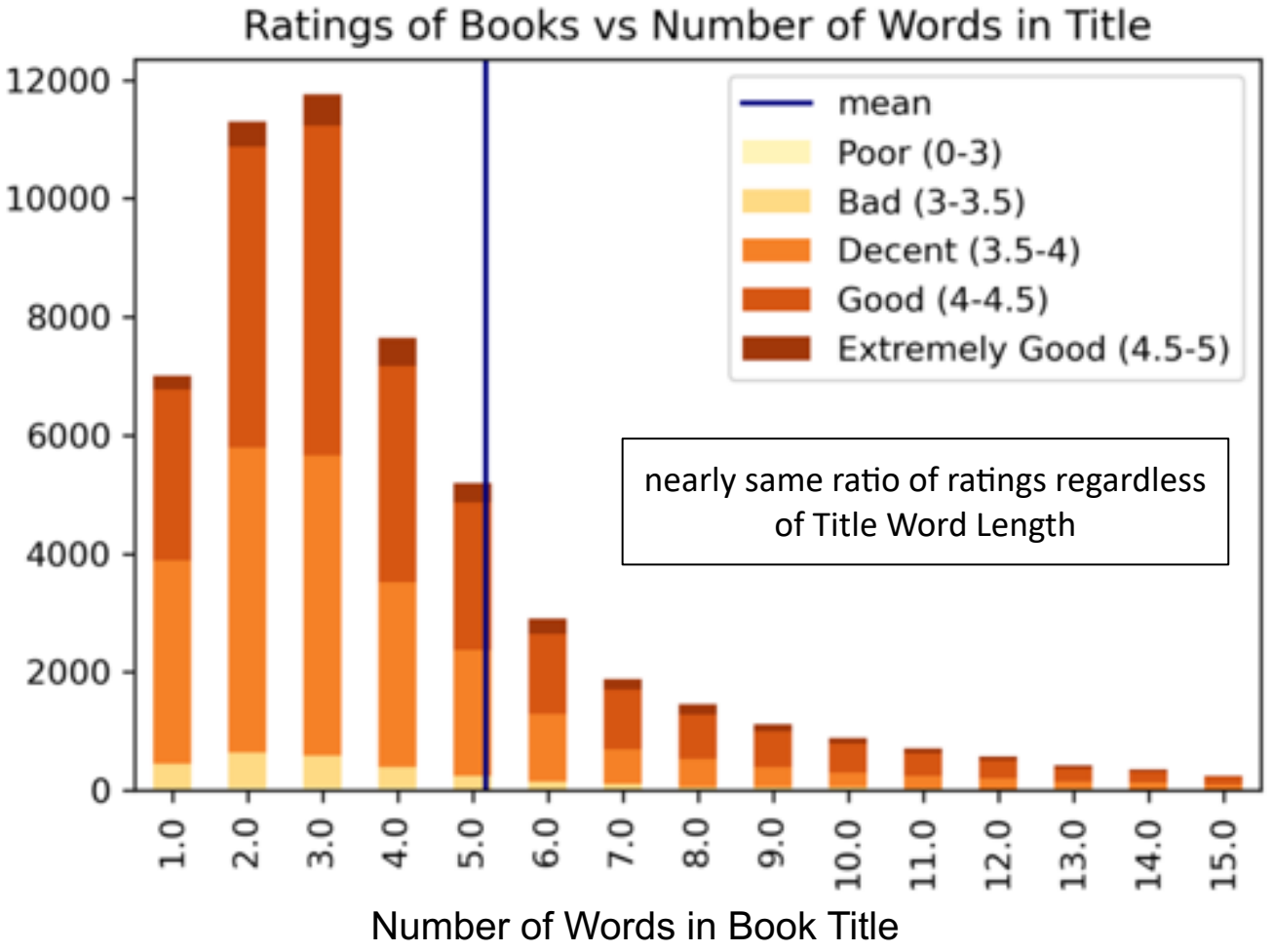


ebook

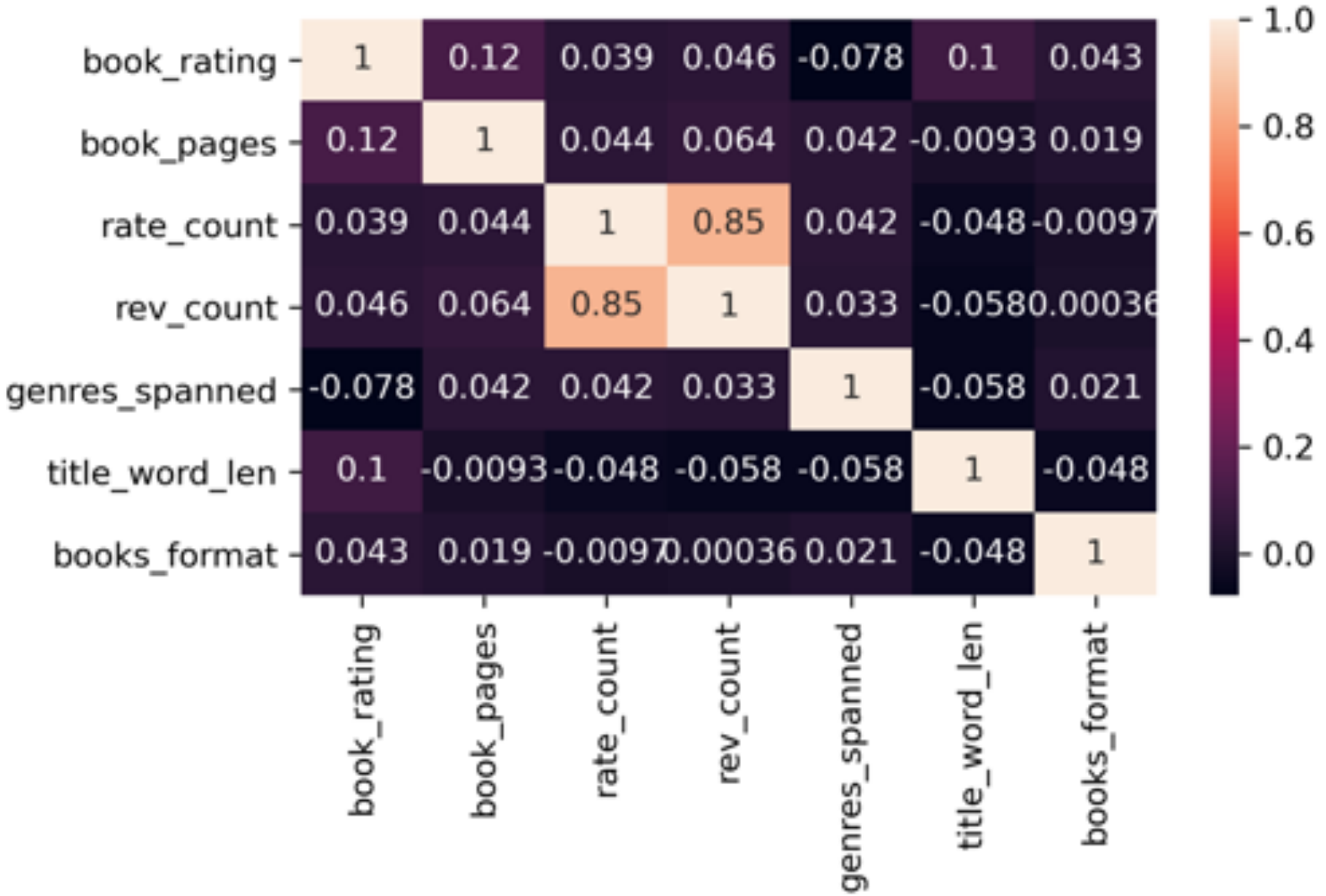


Mass Market Paperback

Does the Book's Title's Number of Words influence its rating?

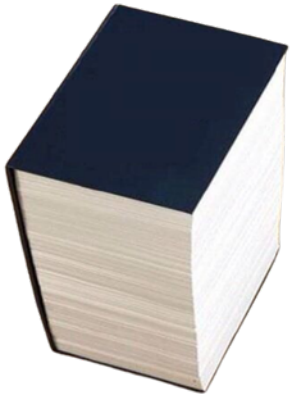


Books Features Correlation Matrix



Don’t judge a book by its’...

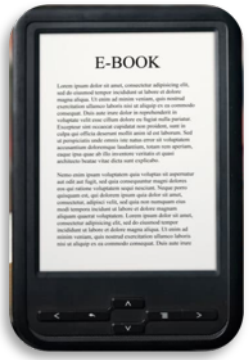
Length

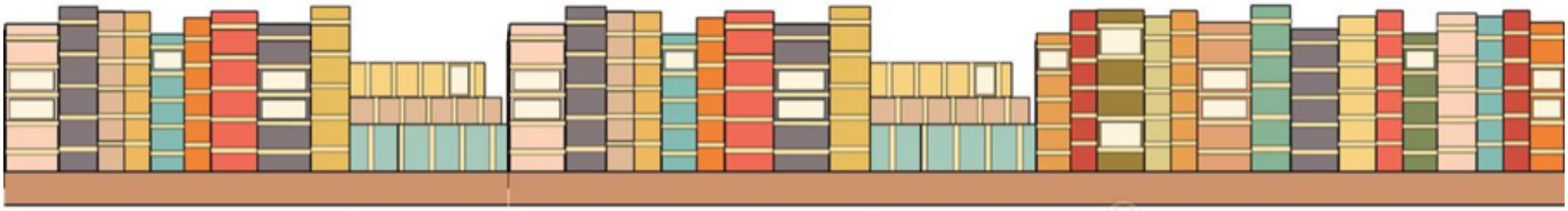


Number of genres

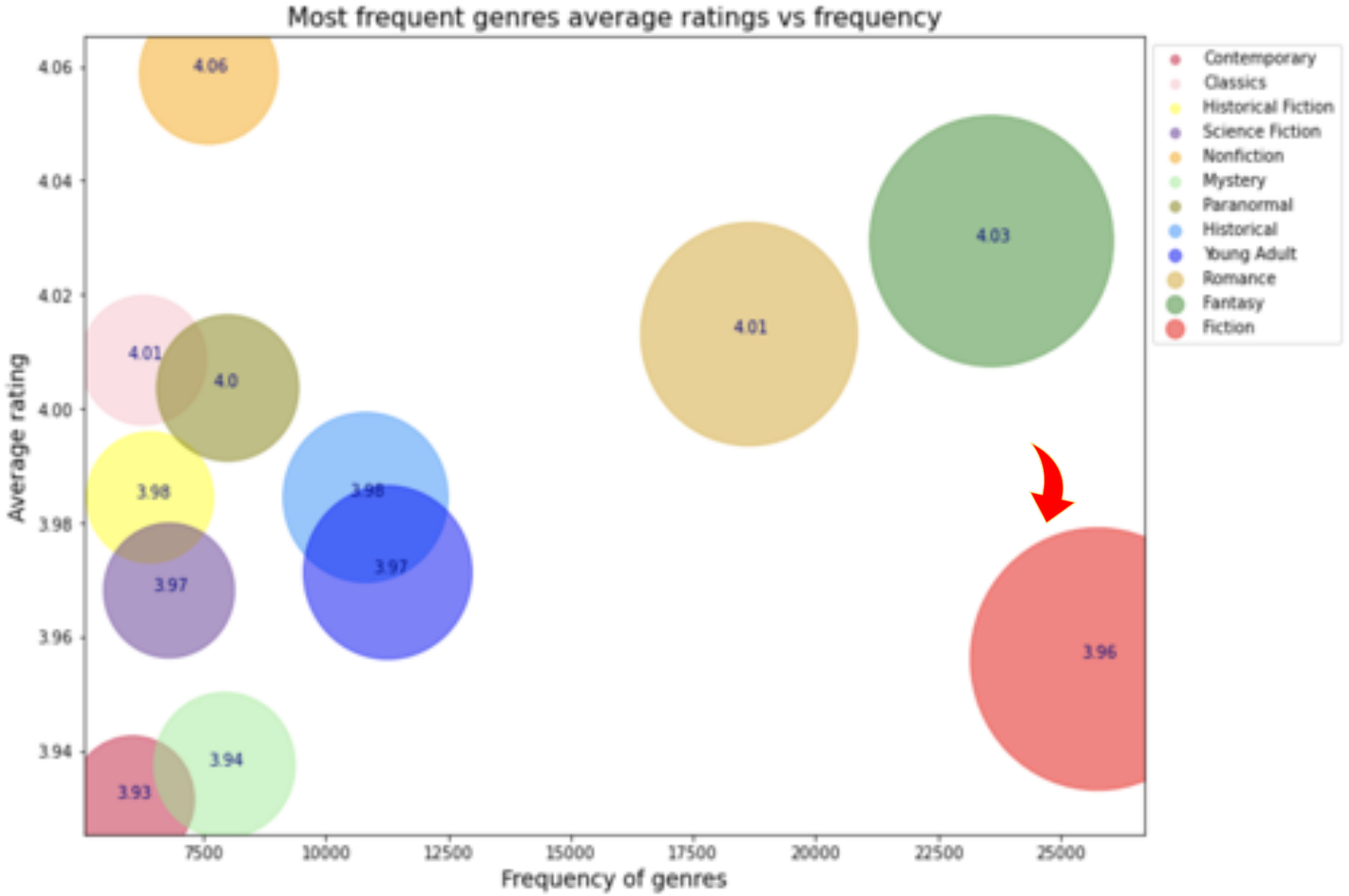
Title length

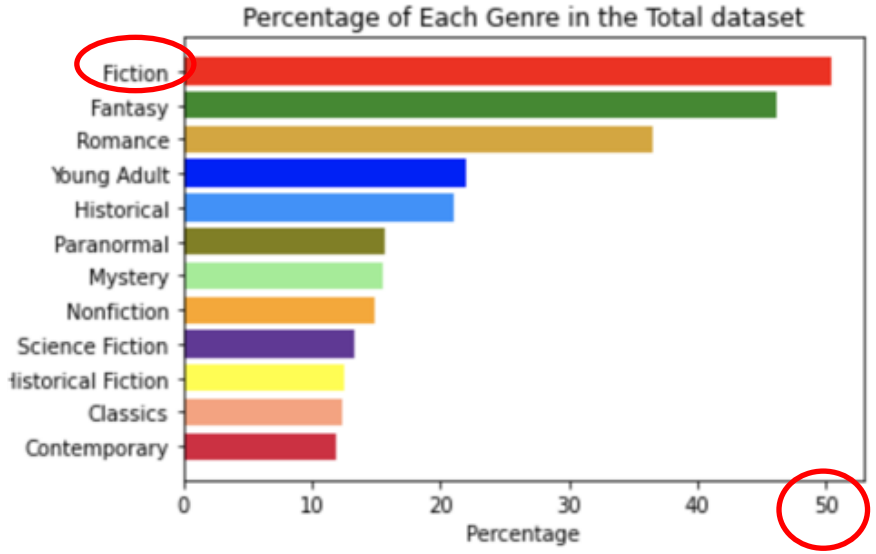
Format





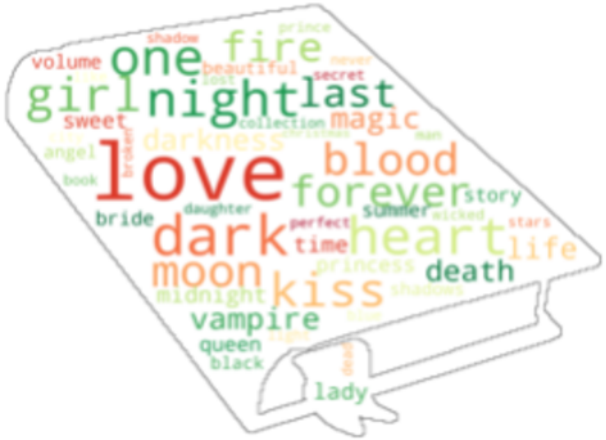
Count and Average Ratings of Most Frequent genres 1925-2019





Most Frequent Words in Books' Titles in Different Genres

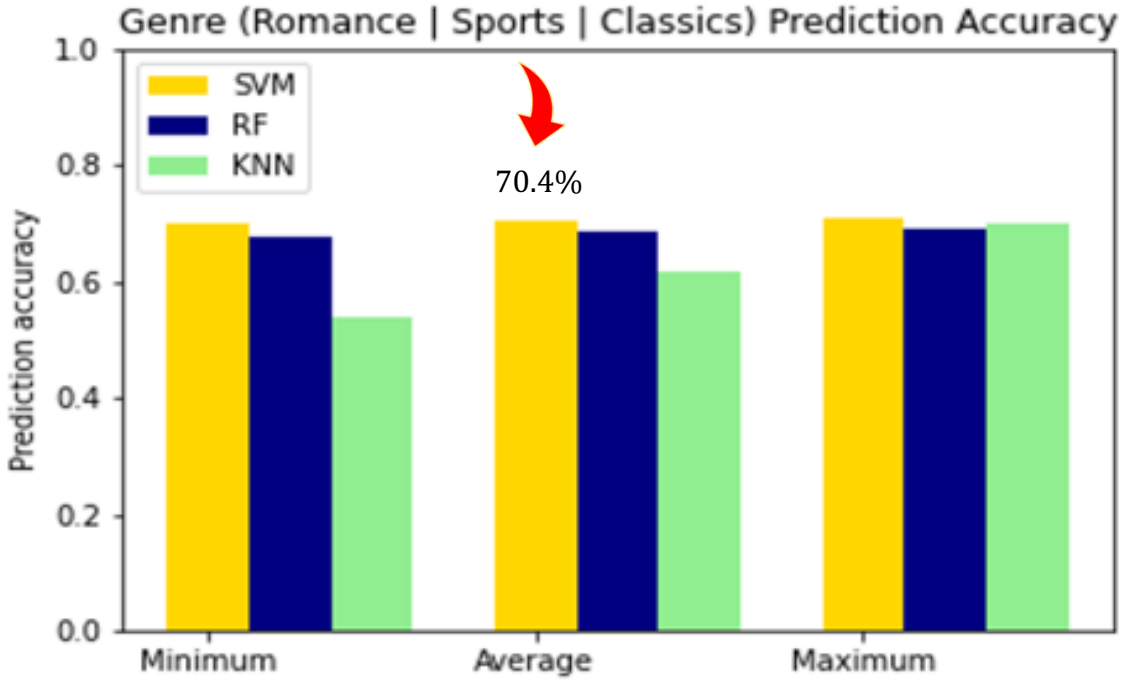
- Based on the most 50 frequent words for Classics, Sports and Romance genre
- Pre-processing: Omitted titles' punctuations, non-English and stop words



Romance | Classics | Sports

Can we predict a book's genre based on its title?

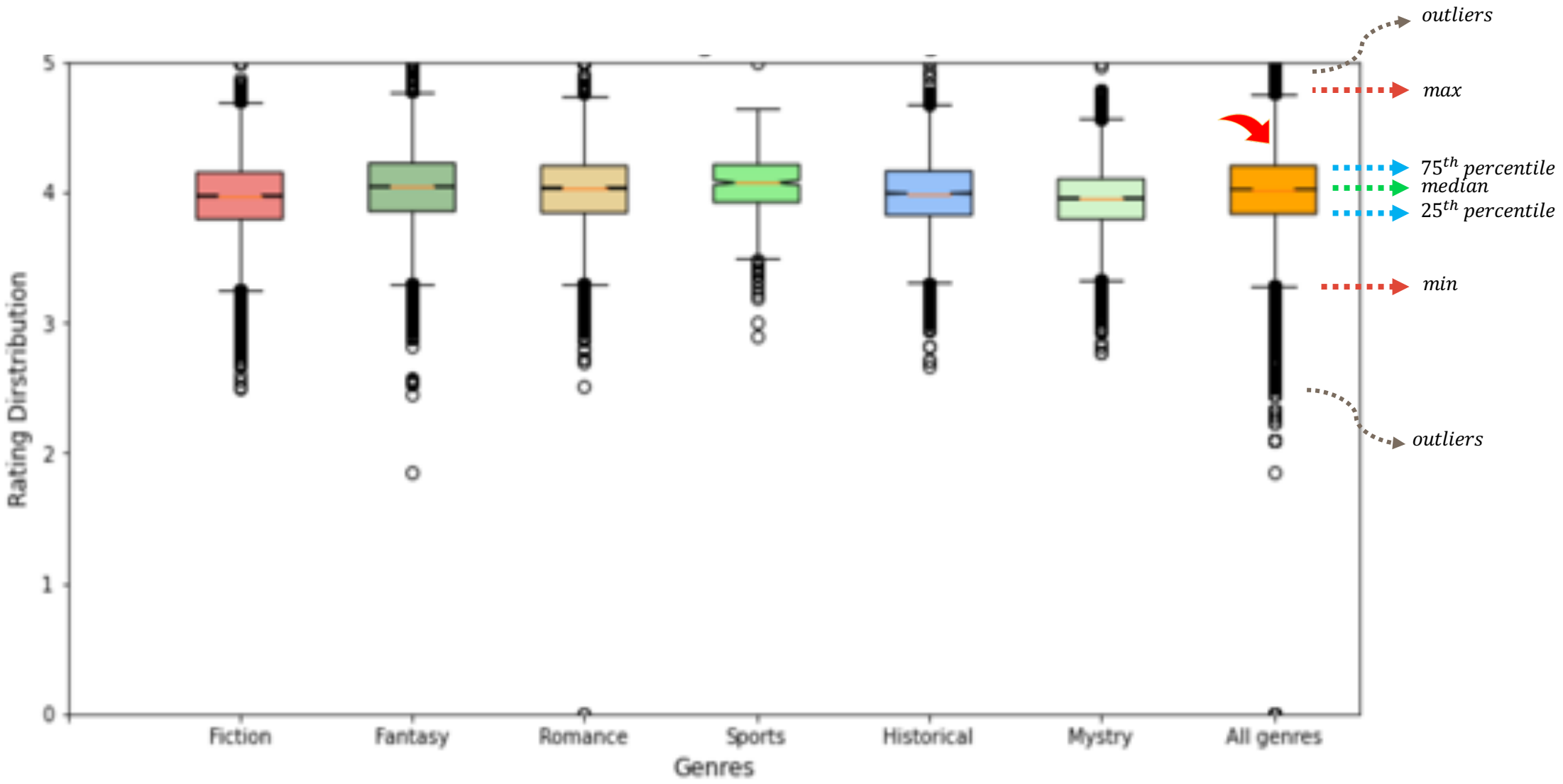
- Picked the most 150 frequent words for each Classics, Romance and Sports genres
- Pre-processing: Omitted titles':
  - Punctuations ('!', '?', …), Non-English , Stop words ("the", "it", …)
- Created word vectors* of length 346**
- Ran multi-label classifiers
  - SVM, RF, KNN (k=5)



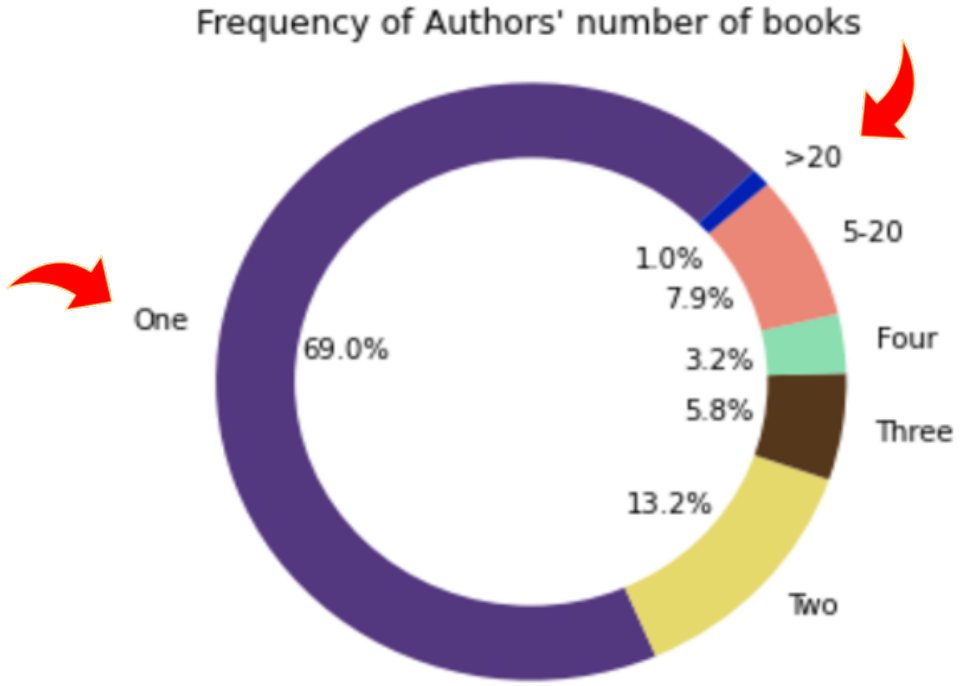
* Word Frequency vectors

** $150 \times 3 - 346$ duplicates of most frequent words in different genres

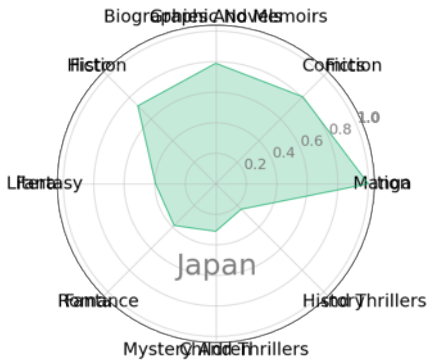
Distribution of Books Average Ratings for Each Genres

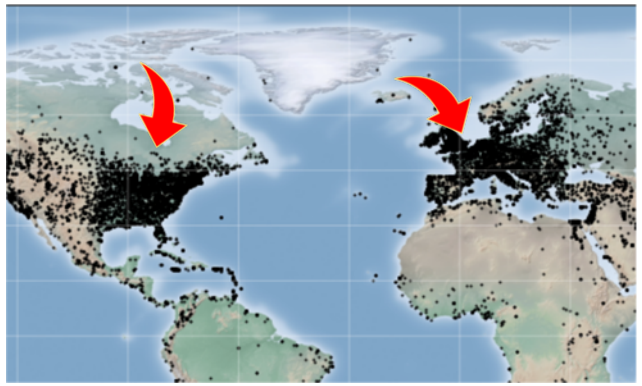


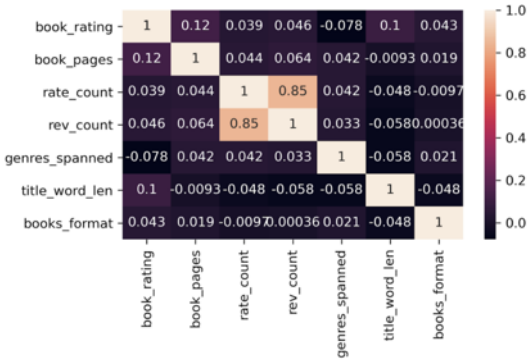
Distribution of Number of Books per Author

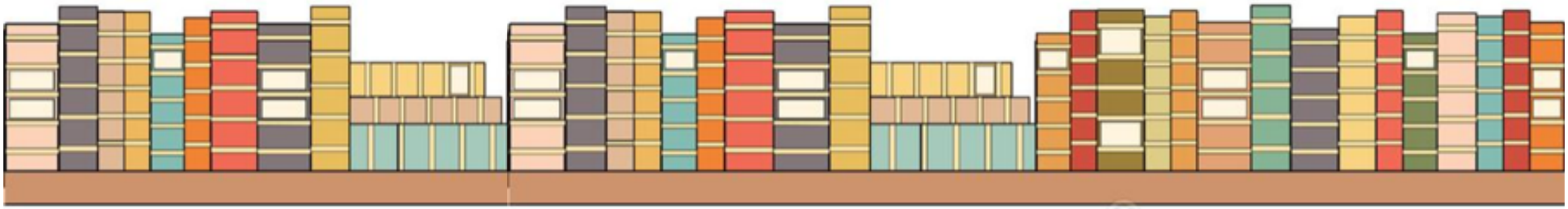


Conclusions & Key Takeaways









W3.CSS Template

web_143.html

ECE 143

An Analysis of Goodreads Dataset

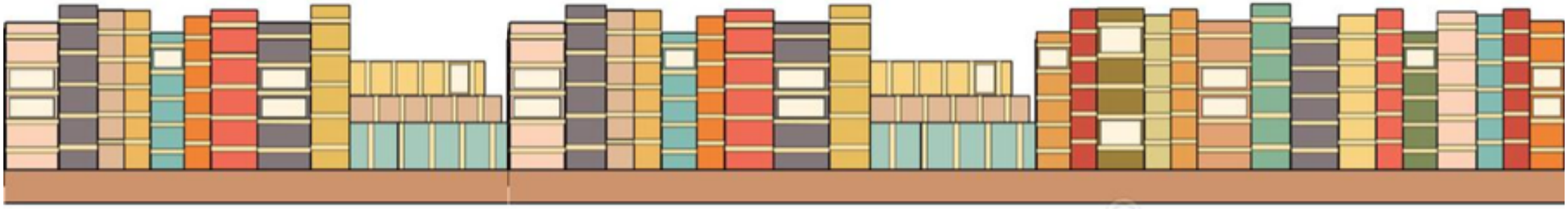
Datasets and Preprocessing | Authors | Genres | Ratings

Title-based Genre Prediction | Resources

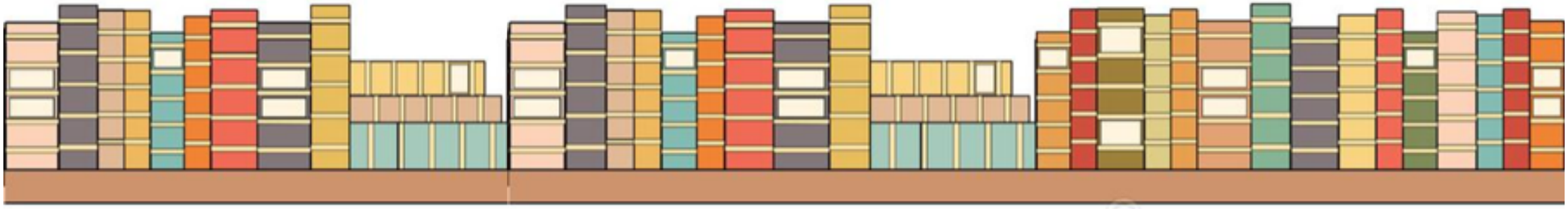
Dataset	Shape (rows, columns)	Main Columns Used	Main Preprocessing Tasks
Goodreads_books (Kaggle)	11123 rows, 12 columns	title, ratings_count, authors	Omitting null values
Goodreads best books ever (Kaggle)	54301 rows, 13 columns	genres	Splitting genres column, Omitting null values
Authors dataset (Kaggle)	52882 rows, 20 columns	author, gender, country, latitude, longitude,	Splitting date to get year, Omitting null values
Authors dataset (GitHub)	2053 rows, 25 columns	most except two	Omitting null values, Filtering etc.

References

- Kaggle Goodreads [link] (Dataset)
- Kaggle Best of Goodreads [link] (Dataset)
- Goodreads Authors/books dataset: Kaggle [link] (Dataset)
- Authors dataset: GitHub [link] (Dataset)
- A Kaggle analysis notebook [link]
- Towards data science: book rating [link]
- Towards data science: All authors around us [link]



Any Questions?



Thank You!