



# goodreads Authors & Books

An analysis using goodreads dataset and much more

Team 2: Anirudh Fatemeh Samantha Zhenwei



The Plague  
Albert Camus

# Motivation

VF Vanity Fair

## 31 Great Quarantine Reads, Chosen by the Vanity Fair Staff

Previously we presented a list of pandemic reads for those looking to order their ... This book is so ridiculous and fun to read out loud, though it ...

Mar 25, 2020



The Stanford Daily

## Classics for quarantine: Reading in Michel de Montaigne's 'Essays'

When I decided to take a gap year last summer, I set out to fill my spare time with writing about the books I have been reading. This was new for ...

1 day ago



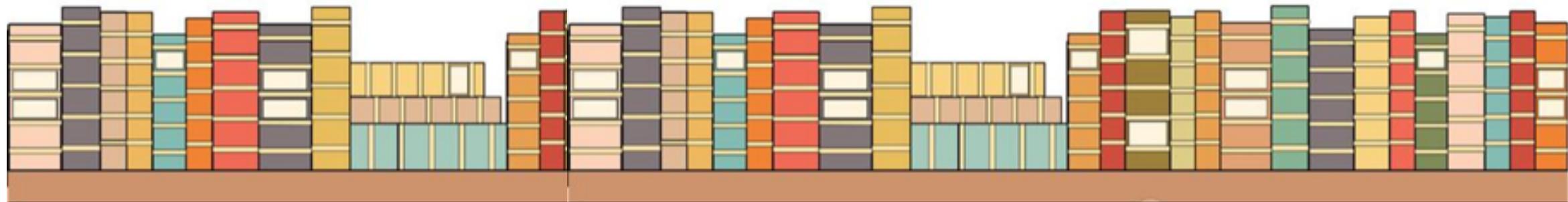
# Overview

## 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 rows, 20 columns	most except isbn	Omitting null values, filtering, etc.
Goodreads books (Kaggle)	11123 rows, 12 columns	title, rating+count, author, ...	Omitting null values



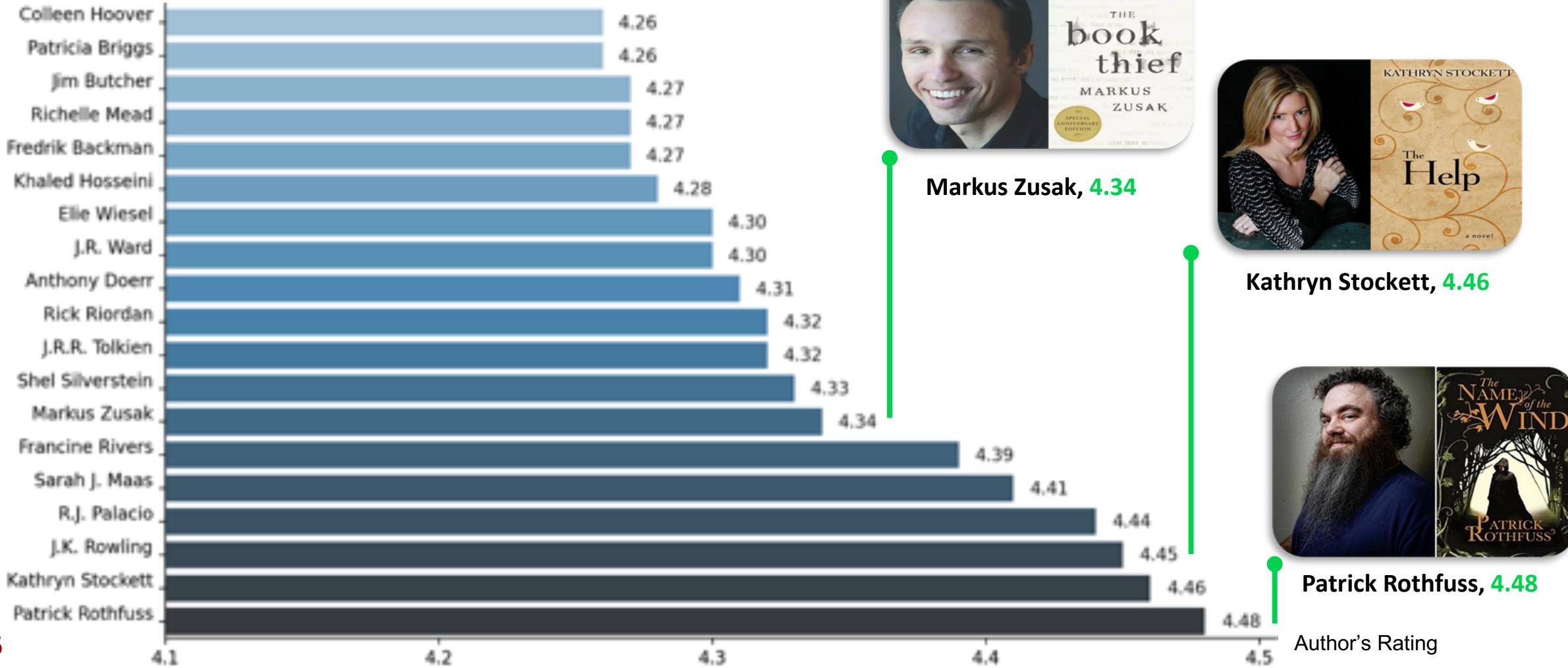
	title	author(s)	genres	format	average_rating	rating_count	review_count	publication_date	num_pages
284	When We Were Orphans	Kazuo Ishiguro	Fiction Historical Historical Fiction Mystery ...	Paperback	3.47	21291	2037	12/1/2007	336 pages
285	The Rachel Papers	Martin Amis	Fiction Novels Contemporary European Literature ...	Paperback	3.59	8404	397	9/29/1992	240 pages
286	The Delta Star	Joseph Wambaugh	Fiction Mystery Mystery Crime	Paperback	3.63	756	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

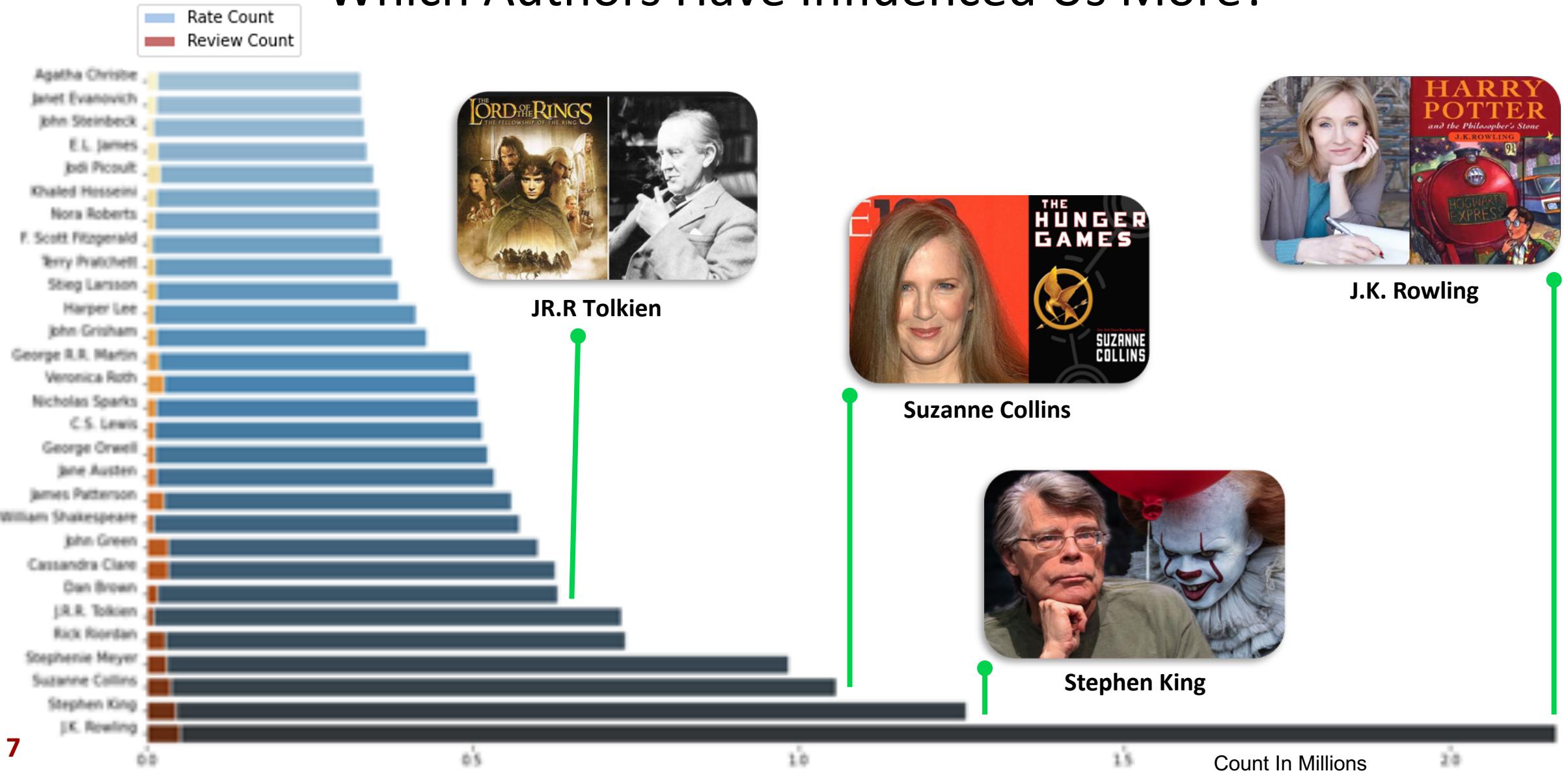
---



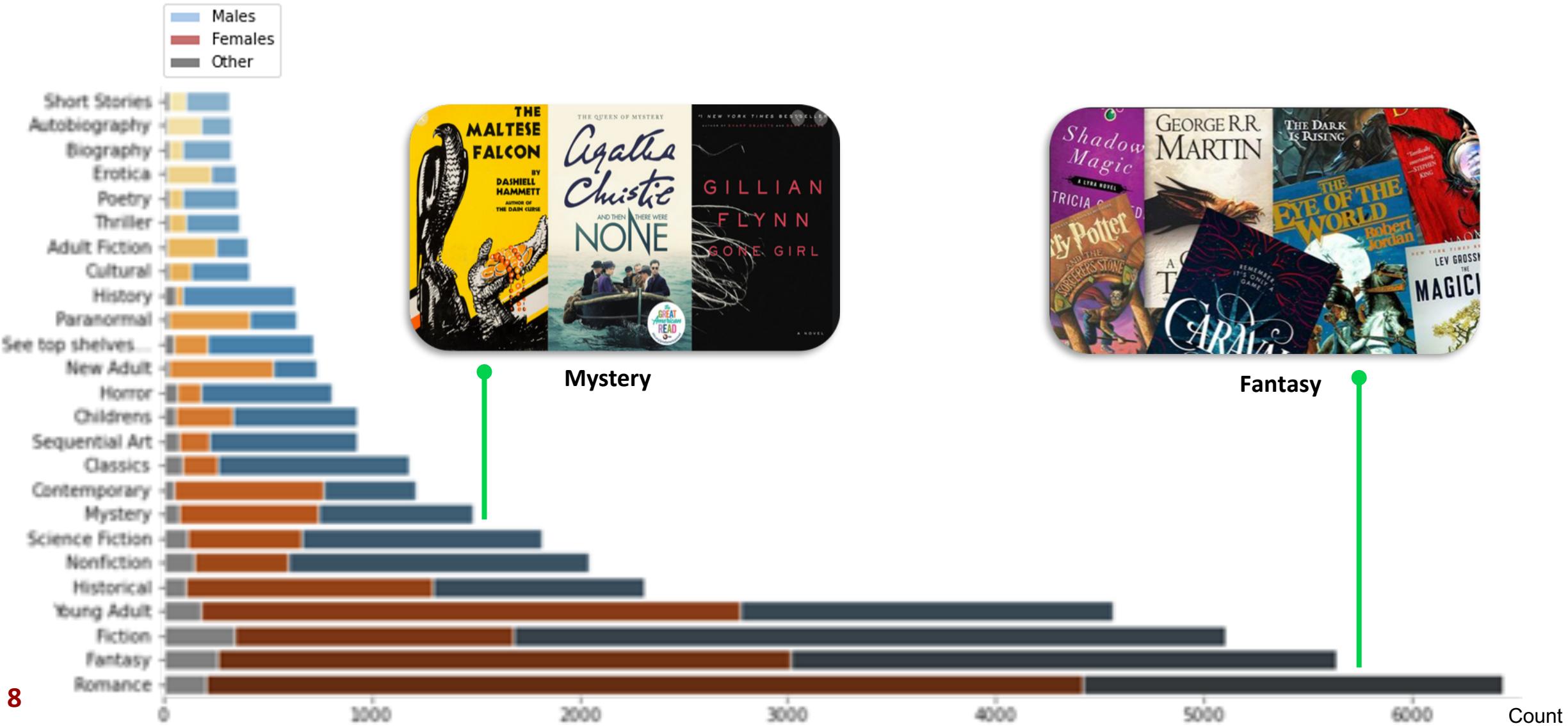
# Average Ratings of All the Authors Around Us



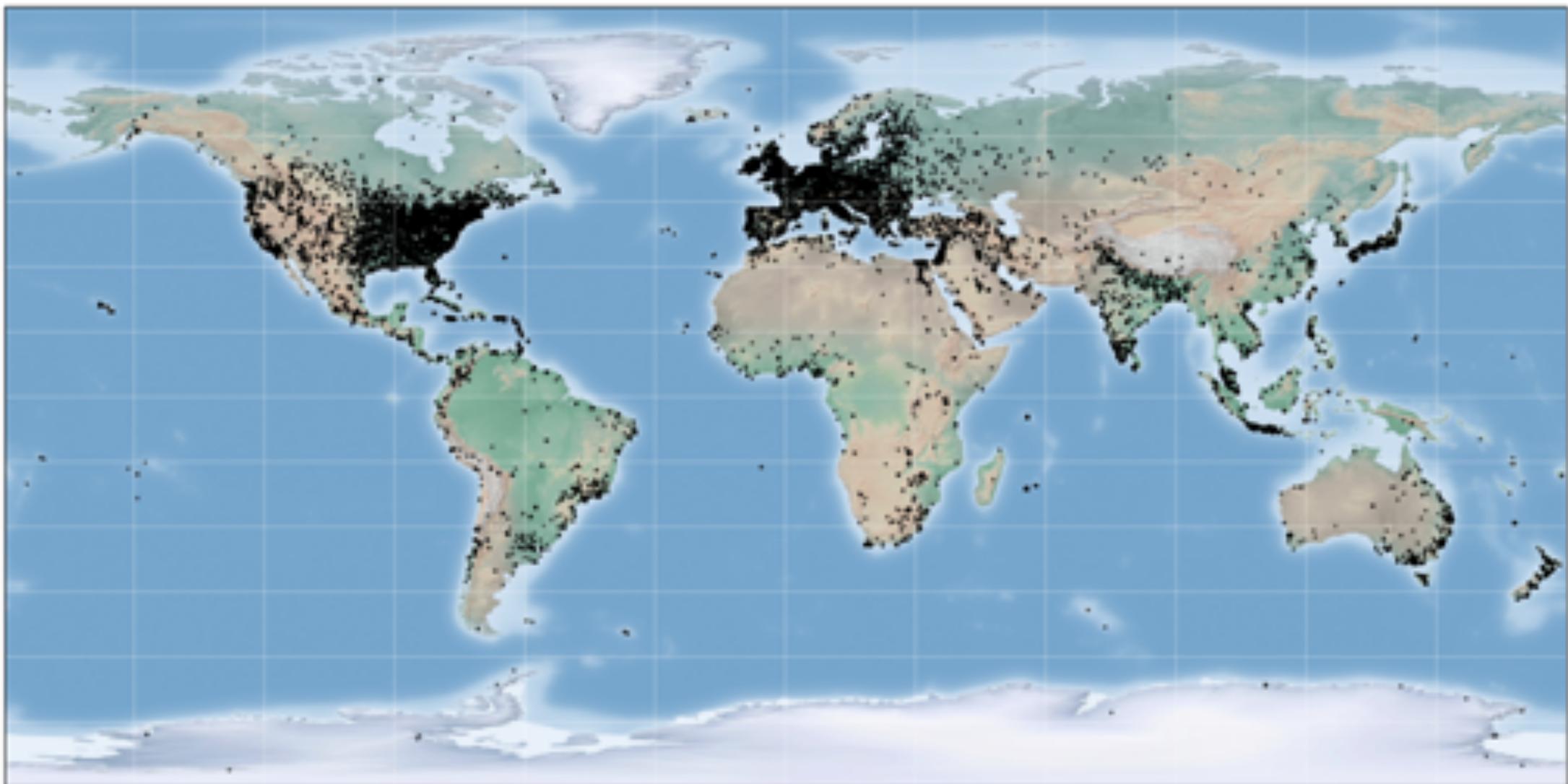
# Which Authors Have Influenced Us More?



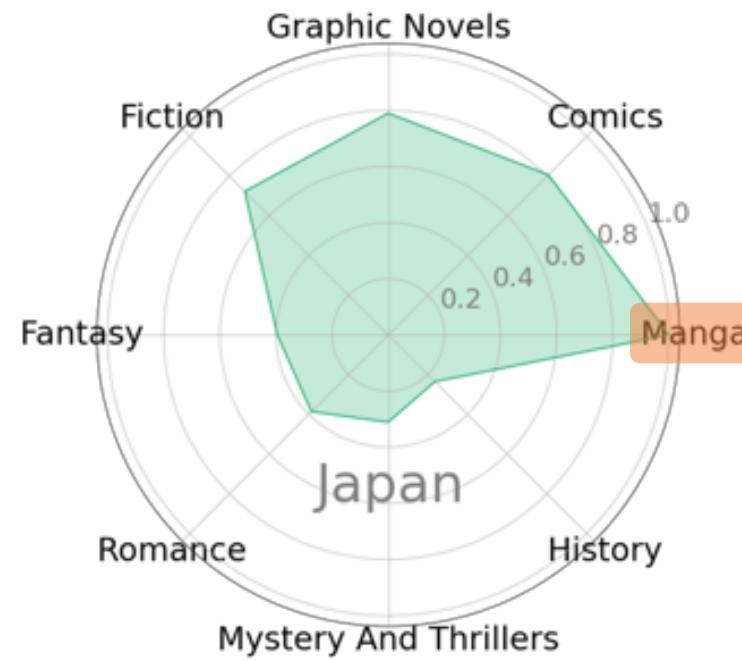
# Which Genres do the Authors Like to Write in 21<sup>st</sup> Century?



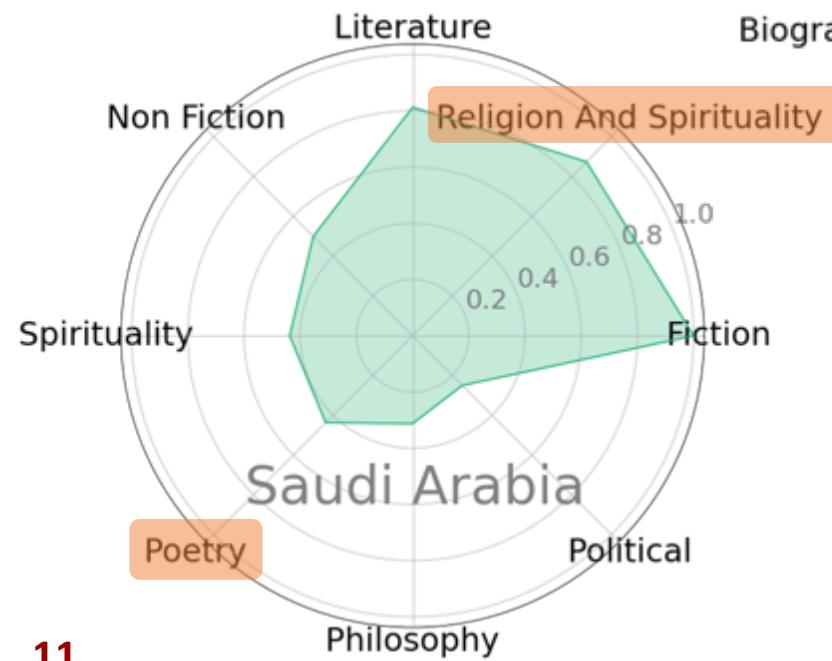
# Where have the Authors Lived from the 17<sup>th</sup> to 21<sup>st</sup> Century?



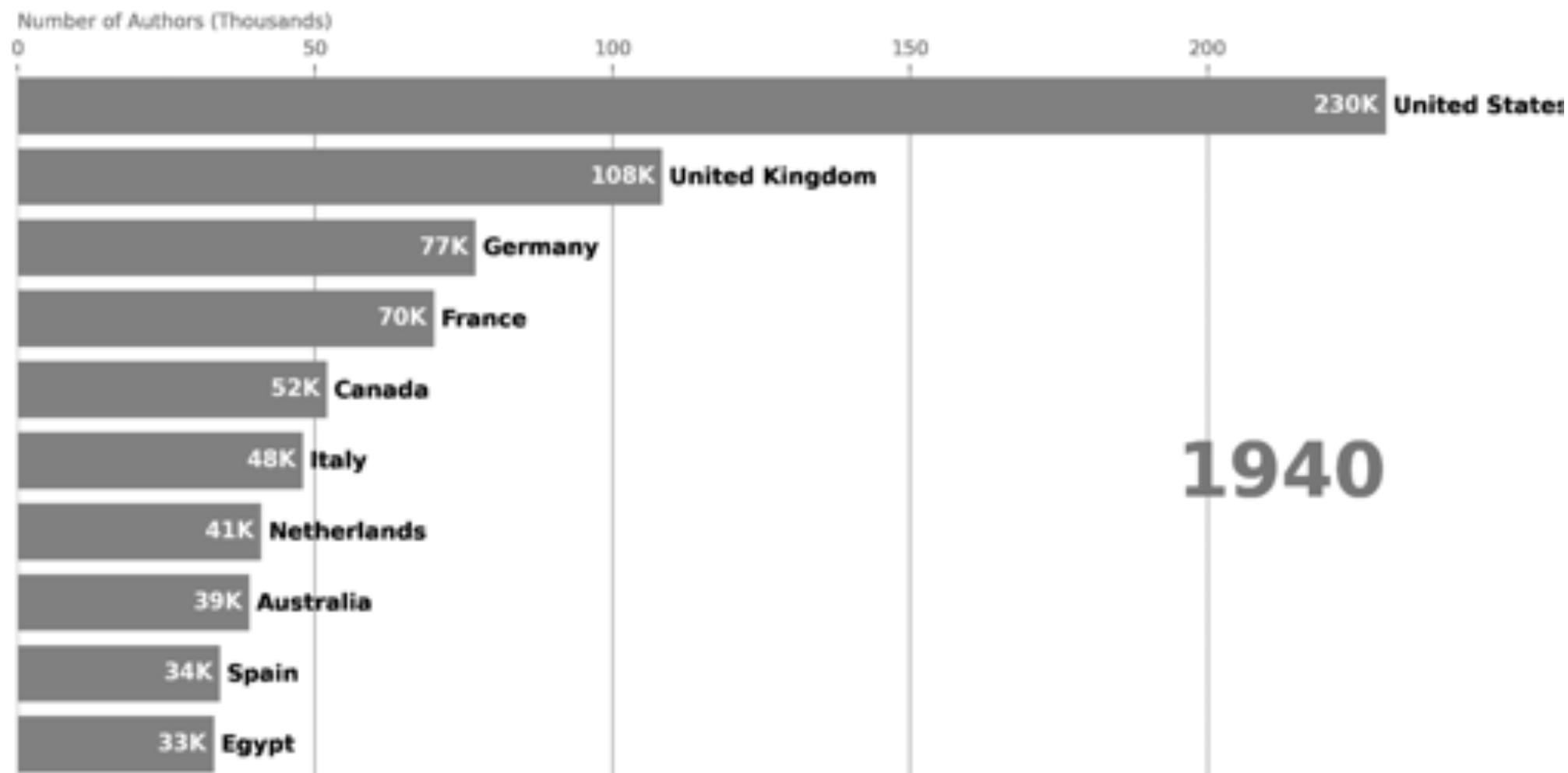
# Distribution of Genres between Different Countries in 20<sup>th</sup> Century



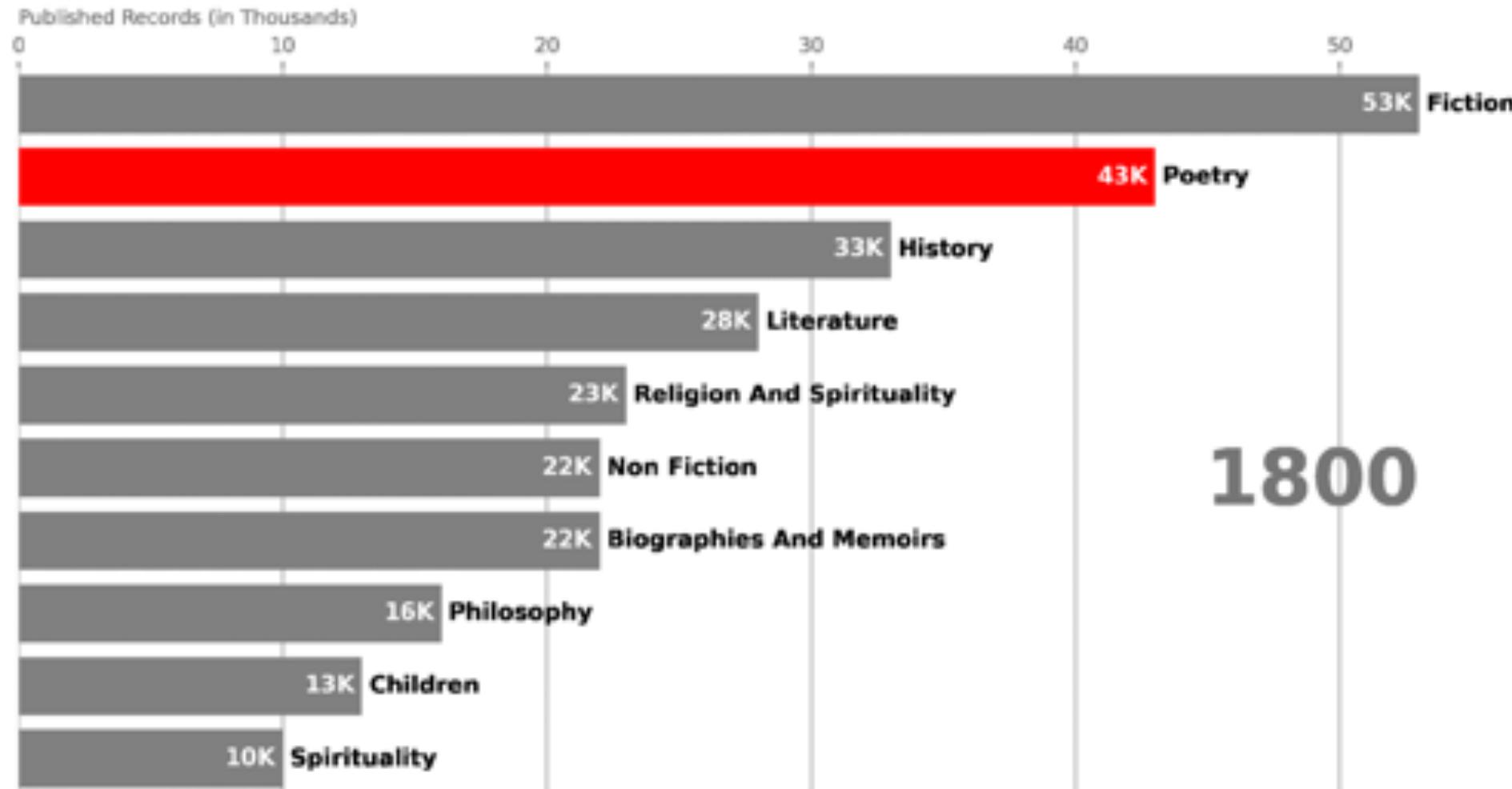
# Distribution of Genres between Different Countries in 20<sup>th</sup> Century



# The Rise of Manga



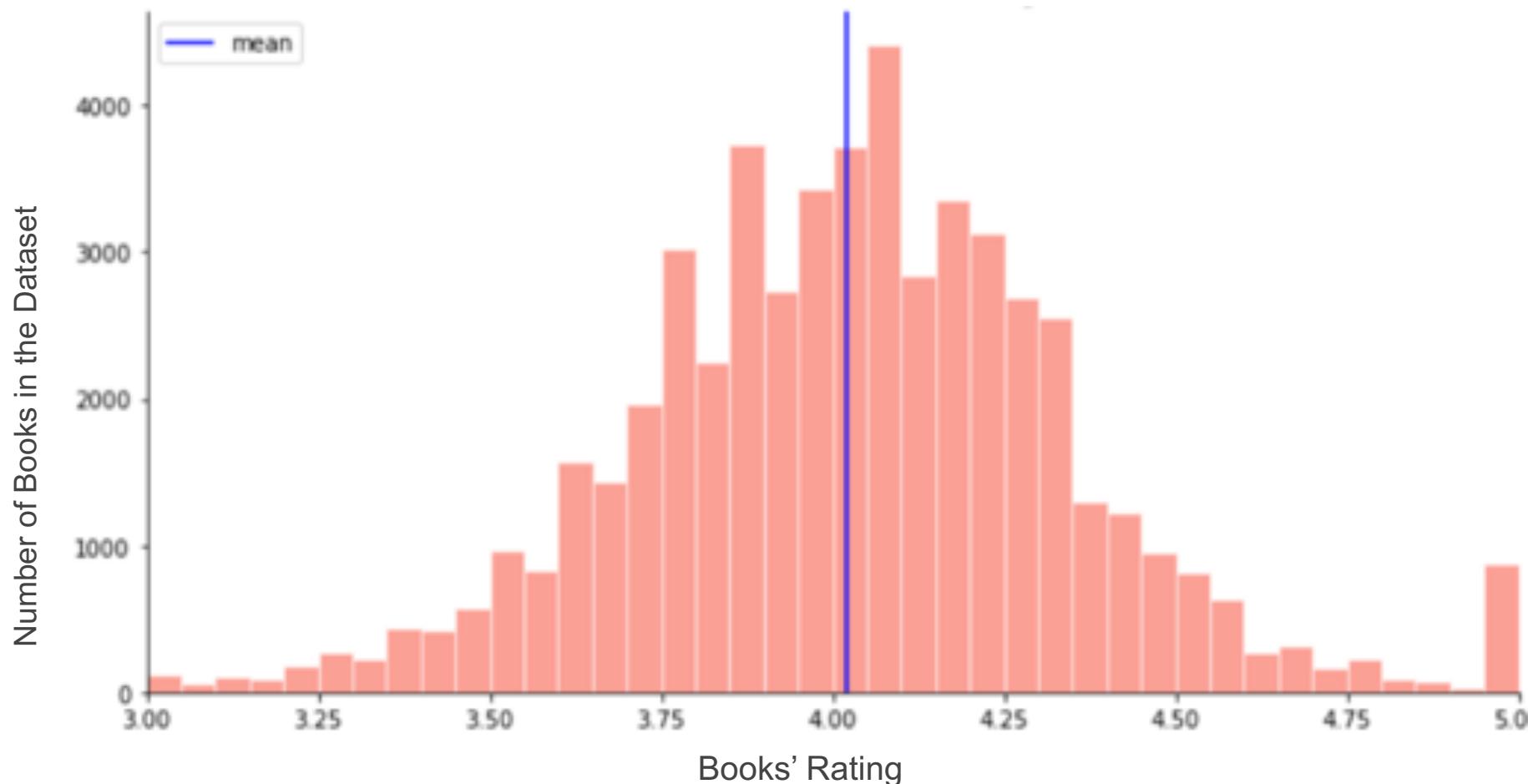
# The Fall of Poetry



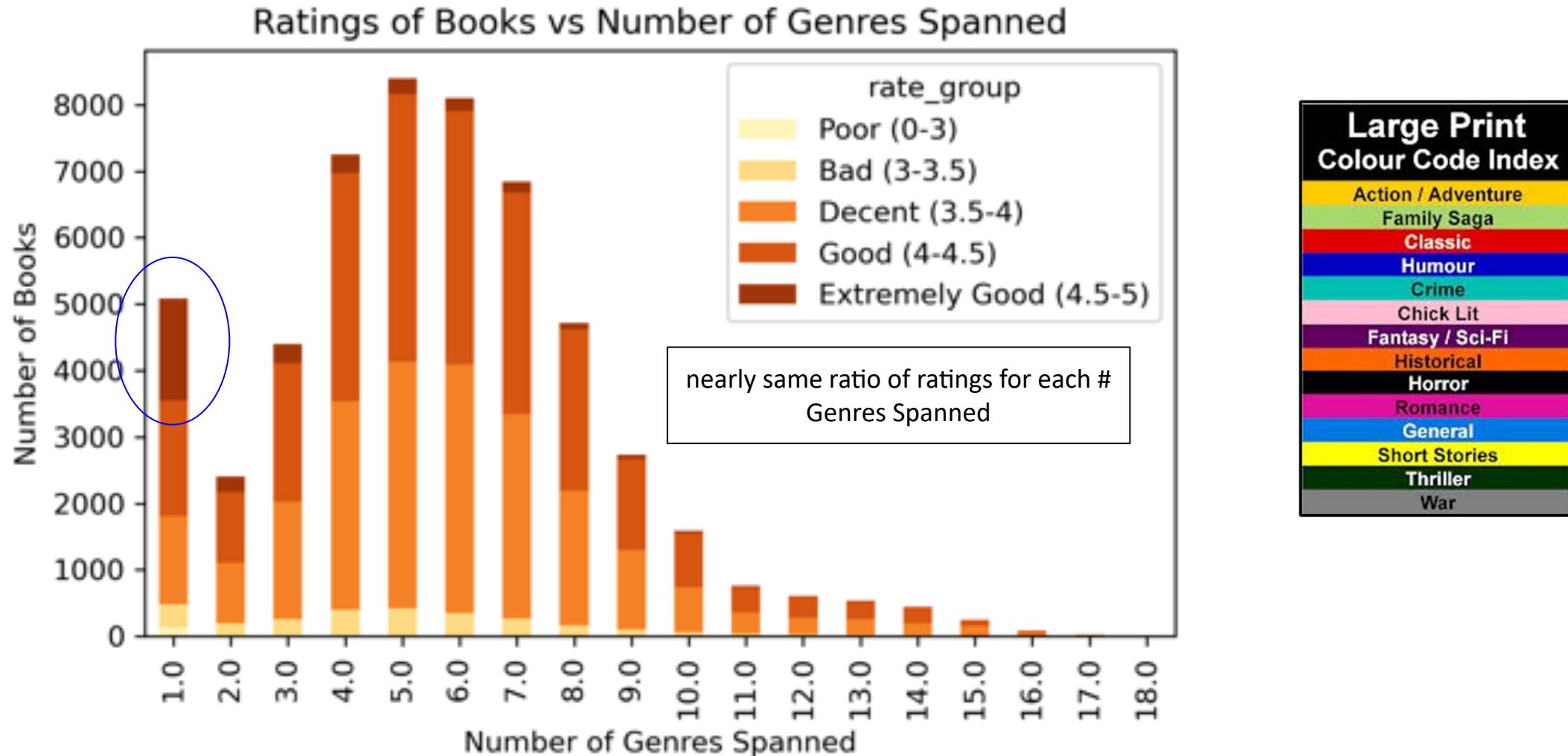
# Part II: An Analysis of the Books



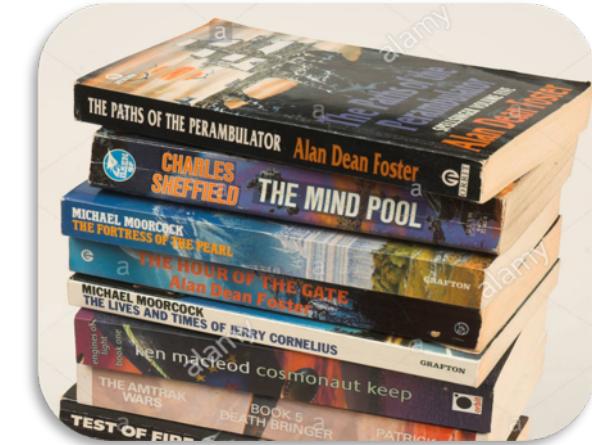
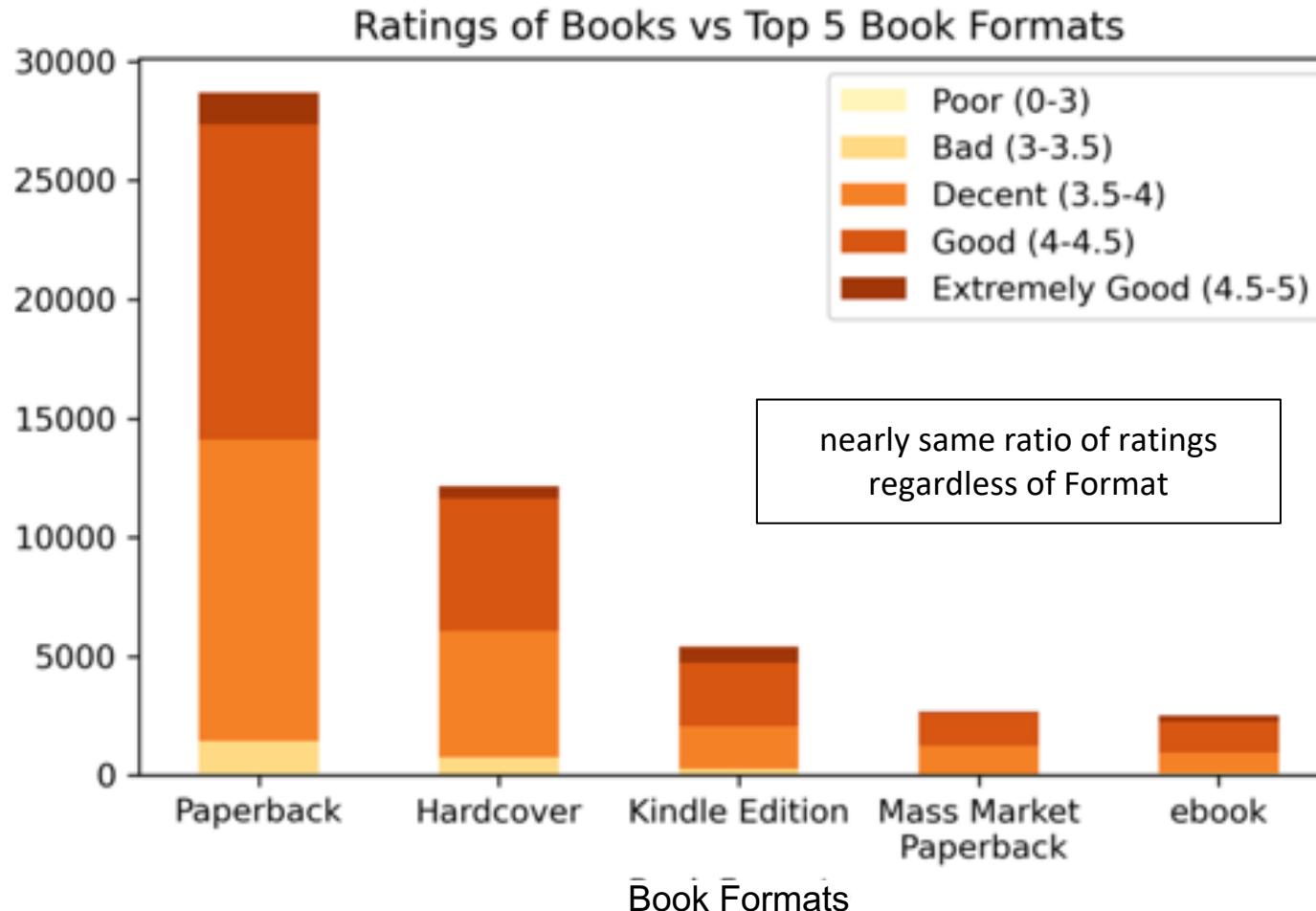
# Distribution of Book Ratings



# Does a Diverse book have a better Rating?



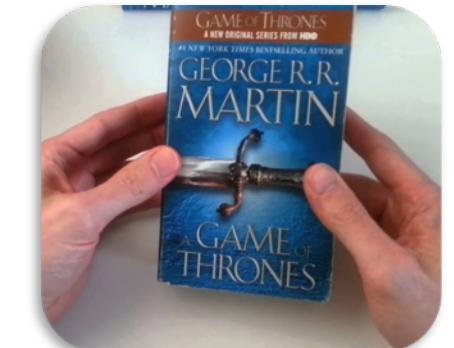
# Does the Book Format influence a Books' Rating?



Paperback

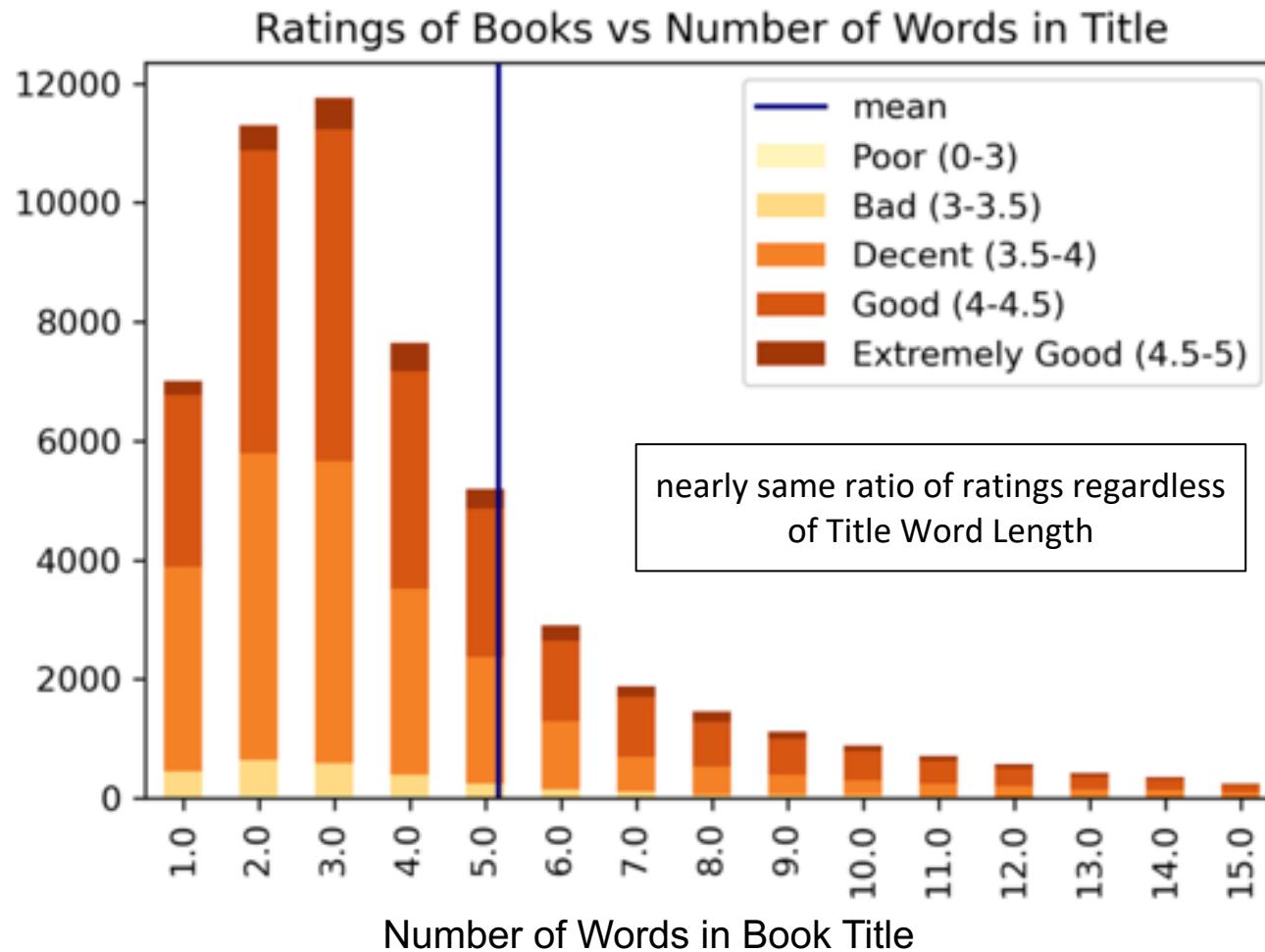


E-BOOK

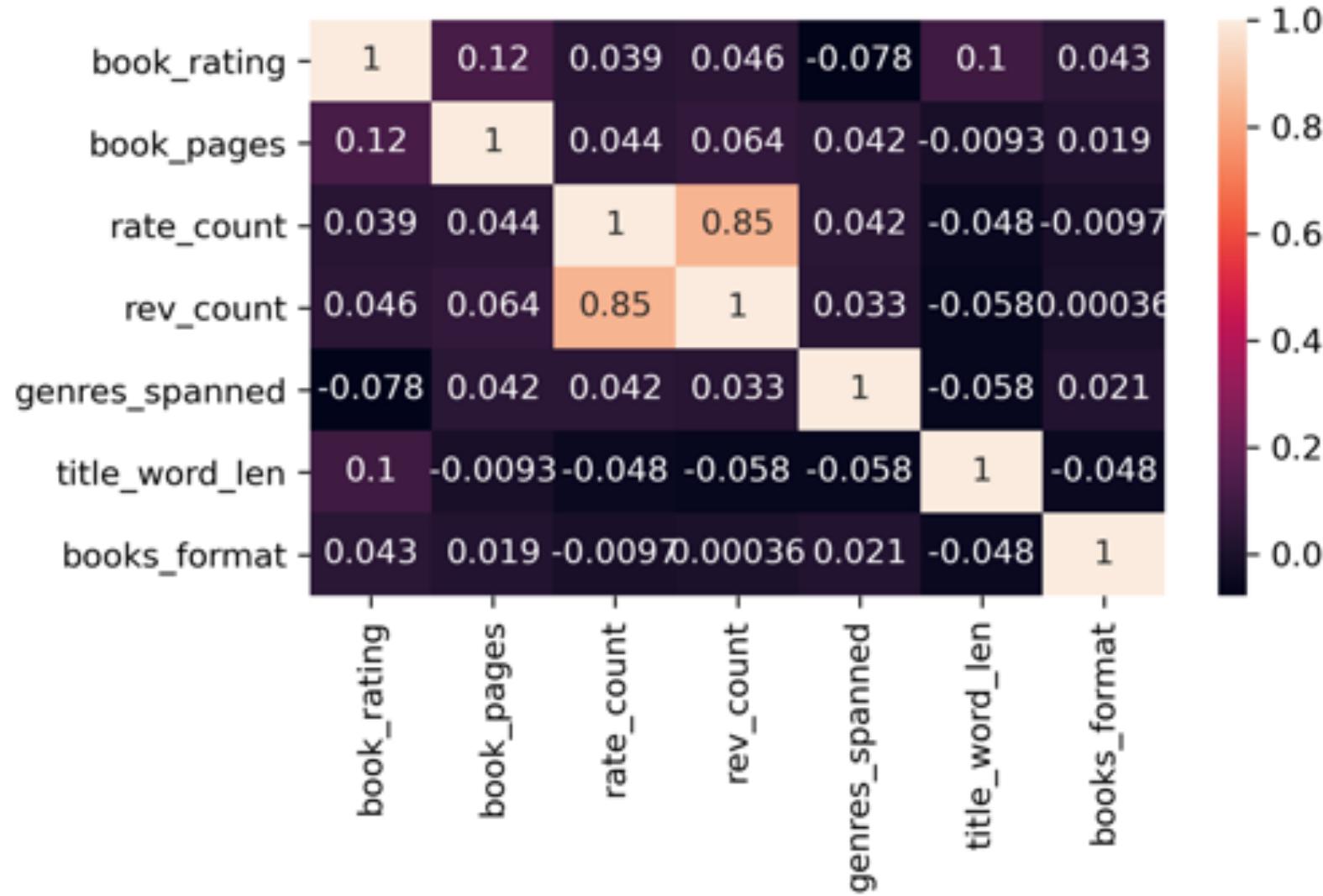


Mass Market Paperback

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

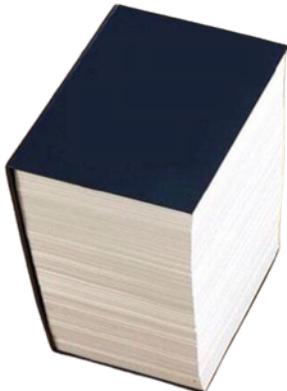


# Books Features Correlation Matrix



# Don't judge a book by its'...

Length



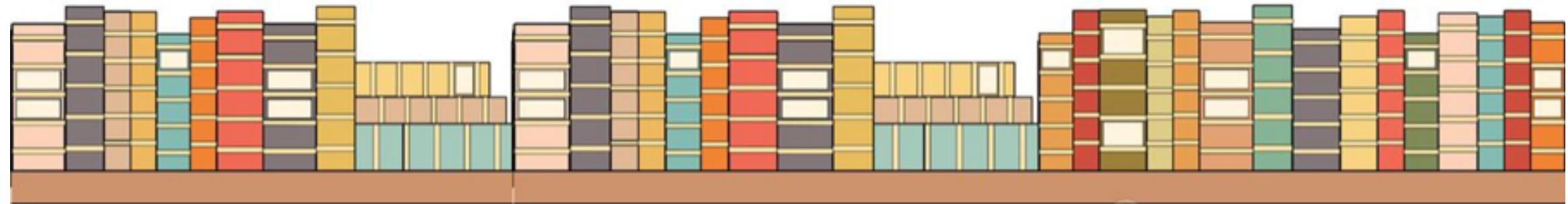
Number of genres



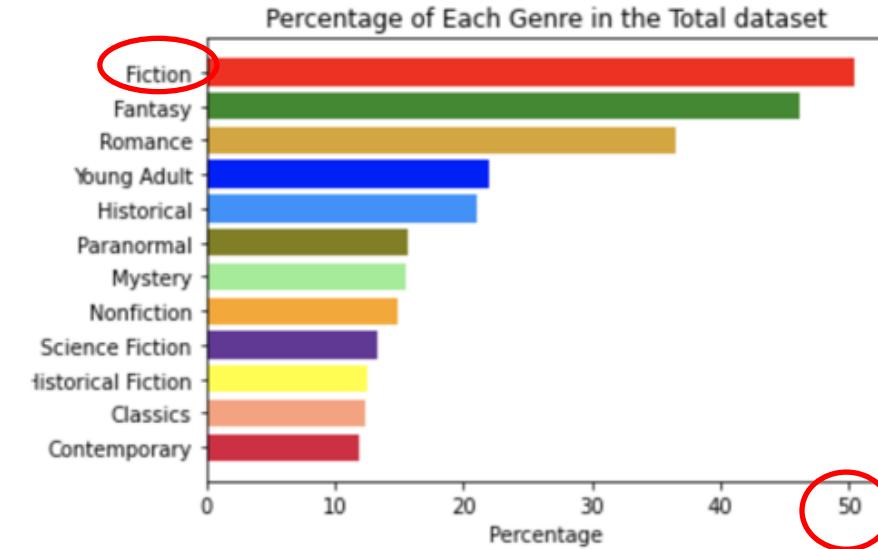
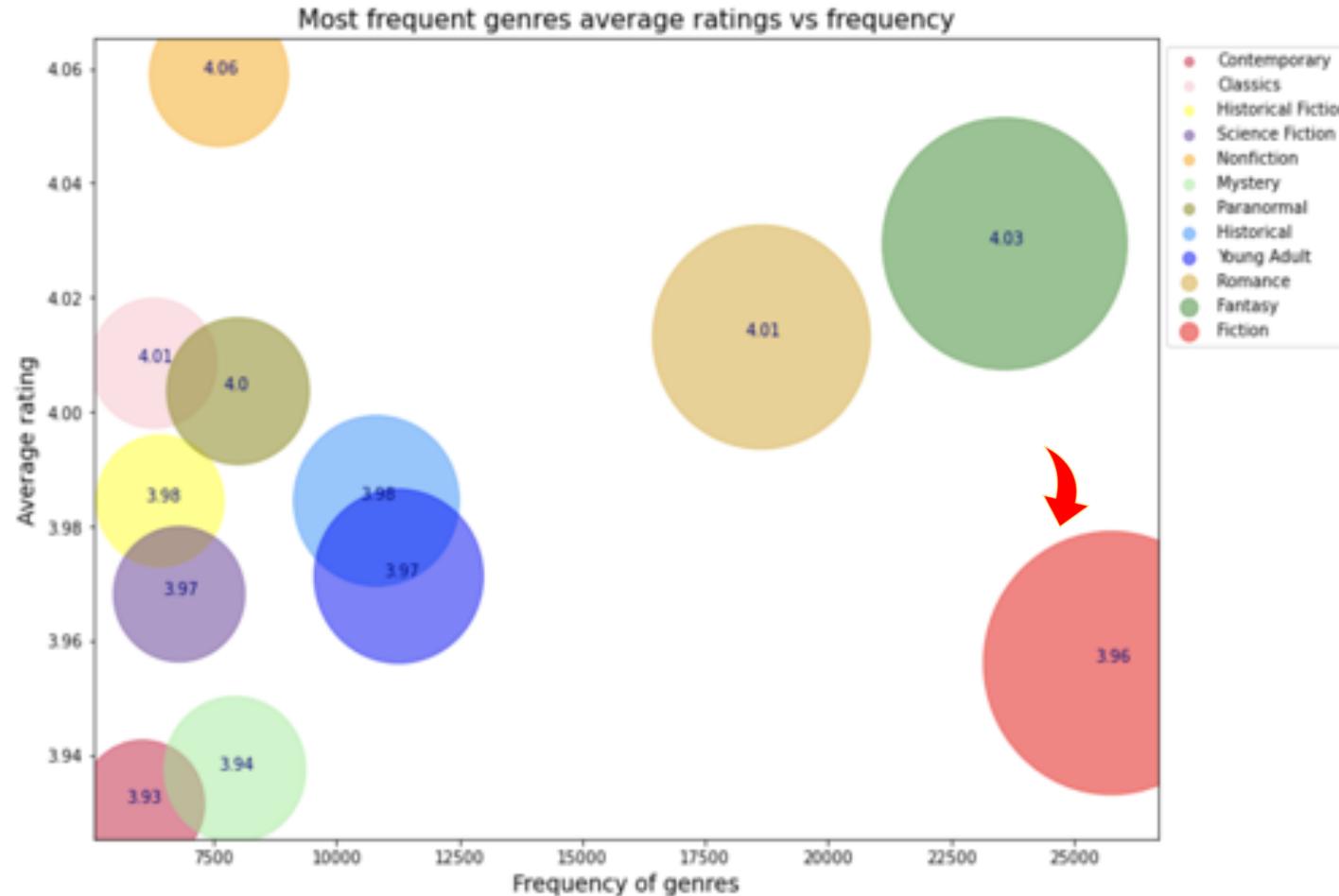
Title length



Format

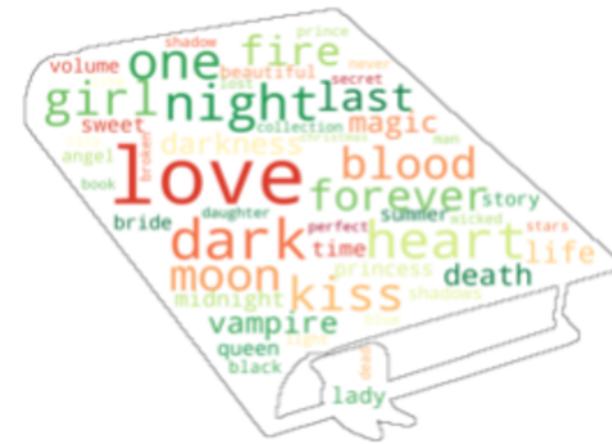


# Count and Average Ratings of Most Frequent genres 1925-2019



# Most Frequent Words in Books' Titles in Different Genres

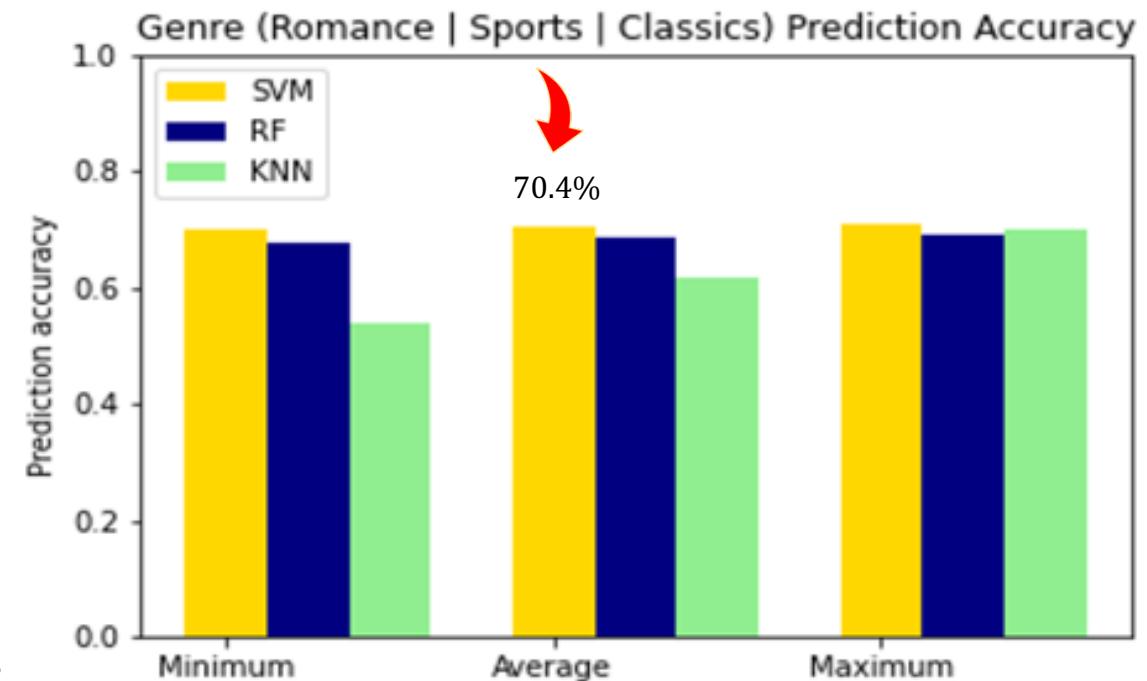
- Based on the most 50 frequent words for Classics, Sports and Romance genre
  - Pre-processing: Omitted titles' punctuations, non-English and stop words



Romance | Classics | Sports

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

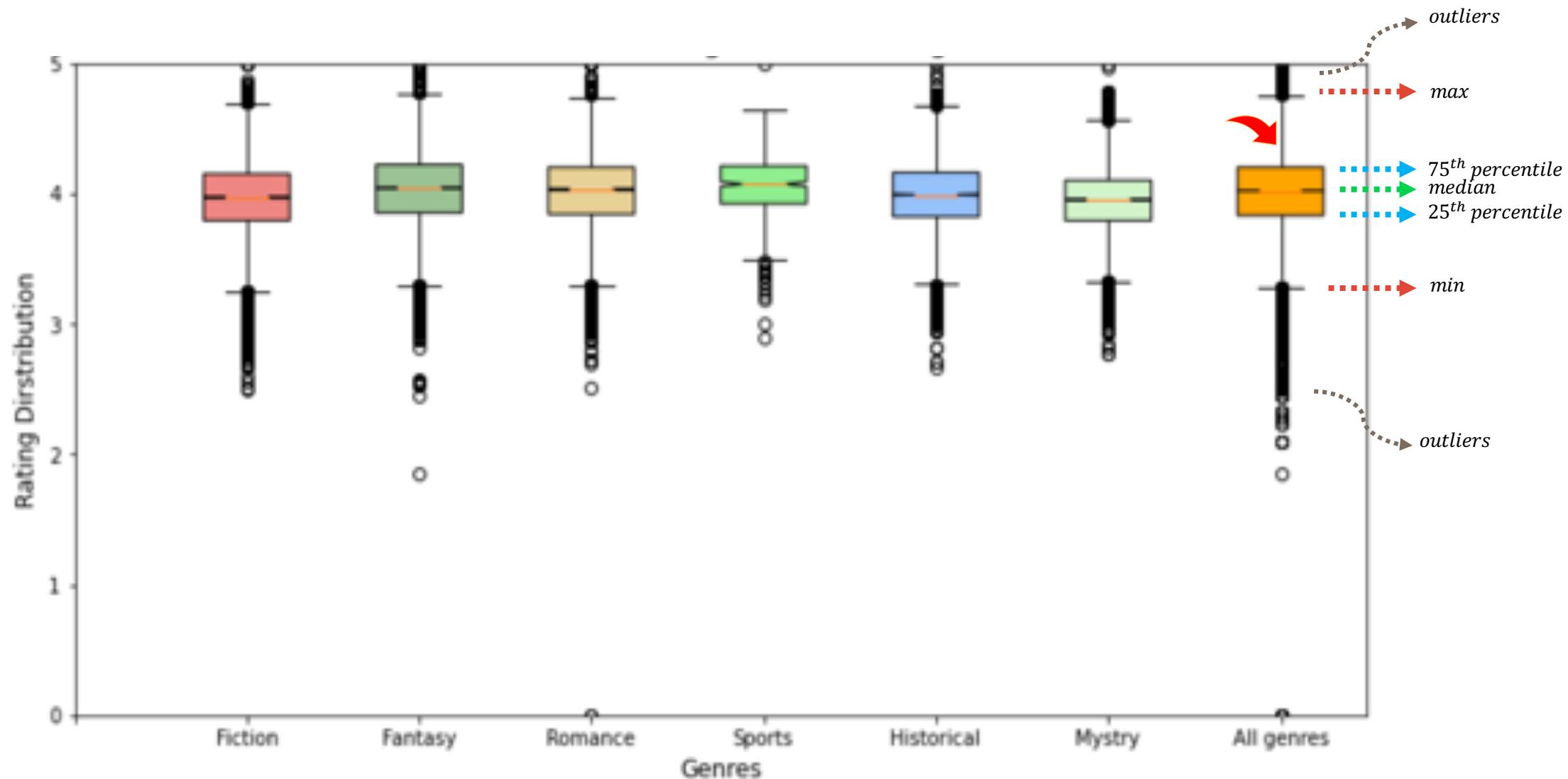
- Picked the most 150 frequent words for each Classics, Romance and Sports genres
- Pre-processing: Omitted titles':
  - Punctuations ('!', '?', ...), Non-English , Stop words ("the", "it", ...)
- Created word vectors\* of length 346\*\*
- Ran multi-label classifiers
  - SVM, RF, KNN (k=5)



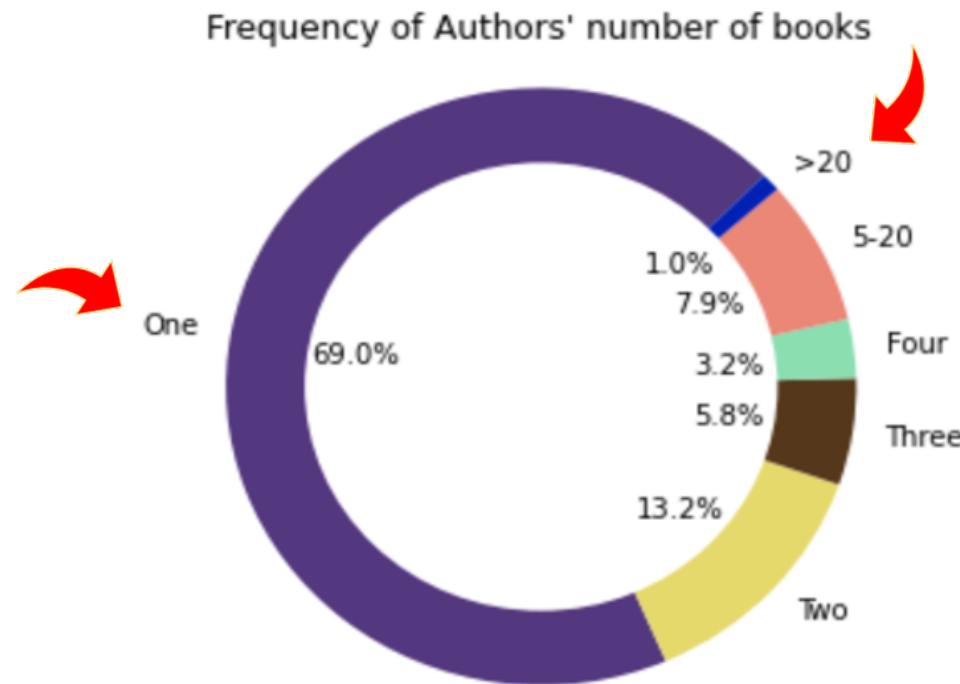
\* Word Frequency vectors

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

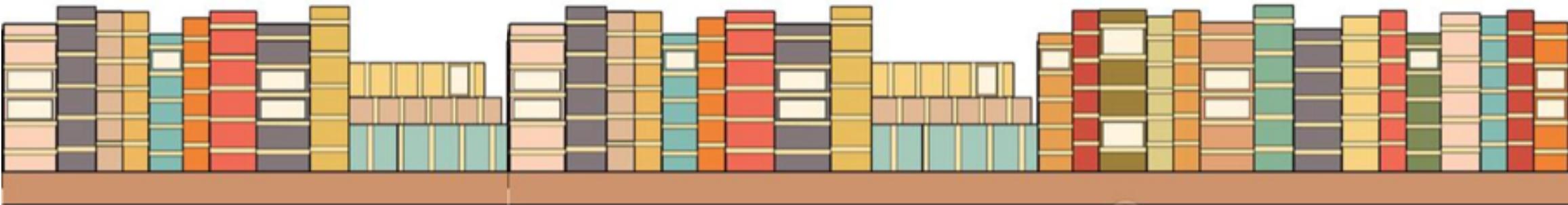
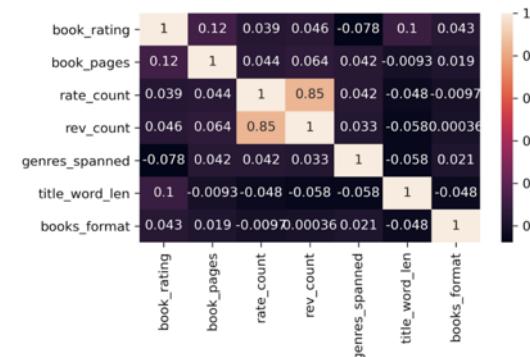
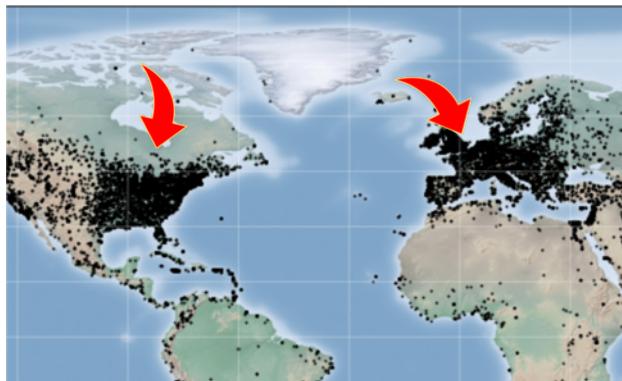
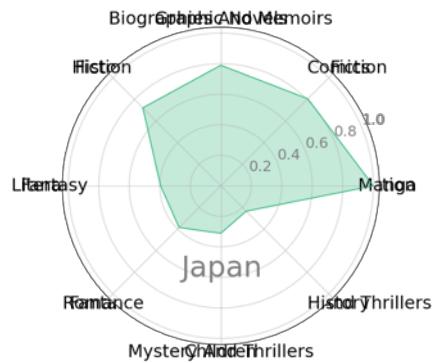
# Distribution of Books Average Ratings for Each Genres



# Distribution of Number of Books per Author



# Conclusions & Key Takeaways





# ECE 143

## An Analysis of Goodreads Dataset



Datasets and Preprocessing

Authors

Genres

Ratings

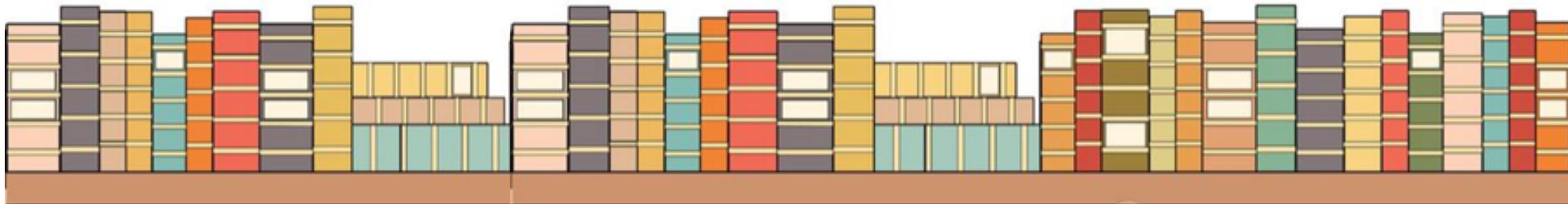
Title-based Genre Prediction

Resources

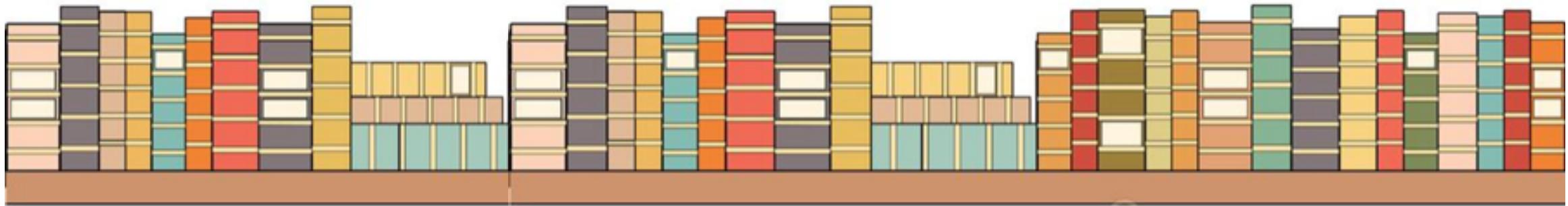
Dataset	Shape (rows, columns)	Most Common Used	Most Preprocessing Tasks
Goodreads_movies (Foglio)	11133 rows, 12 columns	title, rating, genre, author	Dropping null values
Goodreads_lowest_books (Foglio)	14001 rows, 12 columns	author	Building genre column, Dropping null values
Authors dataset (Foglio)	20000 rows, 20 columns	adult, gender, country, birthdate, biography	Splitting data to get year, Dropping null values
Authors dataset (Foglio)	20000 rows, 20 columns	adult except year	Dropping 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!

