And the Plot Thickens
What does book data tell us

Author: Laura L. Lamoureux
Publisher: IST 719

This "book" is dedicted to readers everywhere who are interested in data about the books they read.

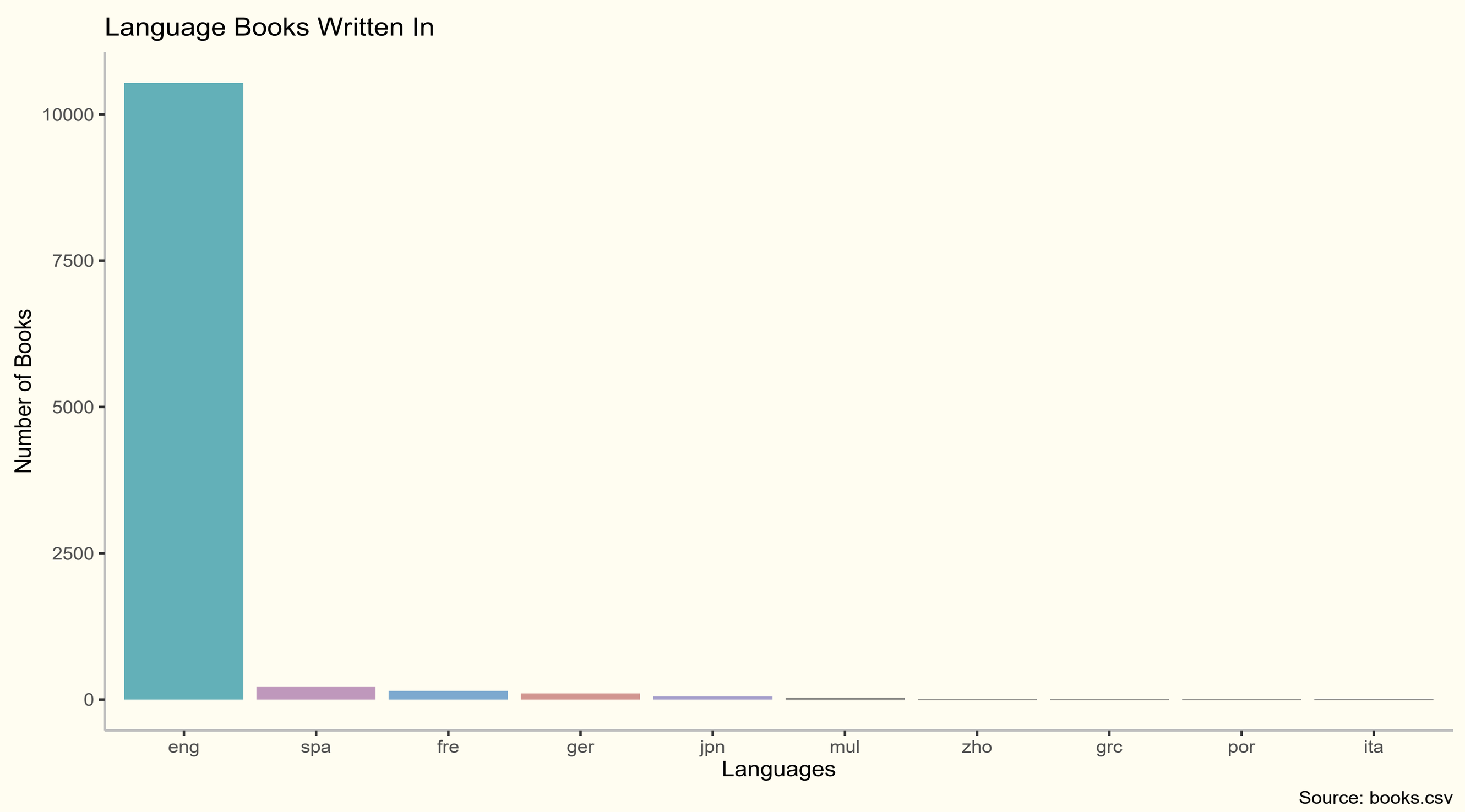
Chapter 1: Data and Sources

Data Source: https://www.kaggle.com/jealousleopard/goodreadsbooks?select=books.csv
Background book image: 123rf.com - Vector - Open book with bookmark icon
Book image: VectorStock.com 287226
Foot image: Vexels.com Right Foot Footprint Silhouette
12 columns and 11,127 rows
Column Names: *BookID, Title, Authors, Average_Rating, ISBN, ISBN13, Language_Code, Num_Pages, Ratings_Count, Text_Reviews_Count, Publication_Date, Publisher*
Cleaning included: renaming column names, dropping 4 rows due to no data

Chapter 2: Language

What languages are these books written in?

As we can see - most of the books in this particular dataset are written in English Data cleaning for this chart consisted of combining all versions of English (eng, eng-US, eng-GB, etc.; also limiting the chart to the top languages



According to Wikipedia (https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers)

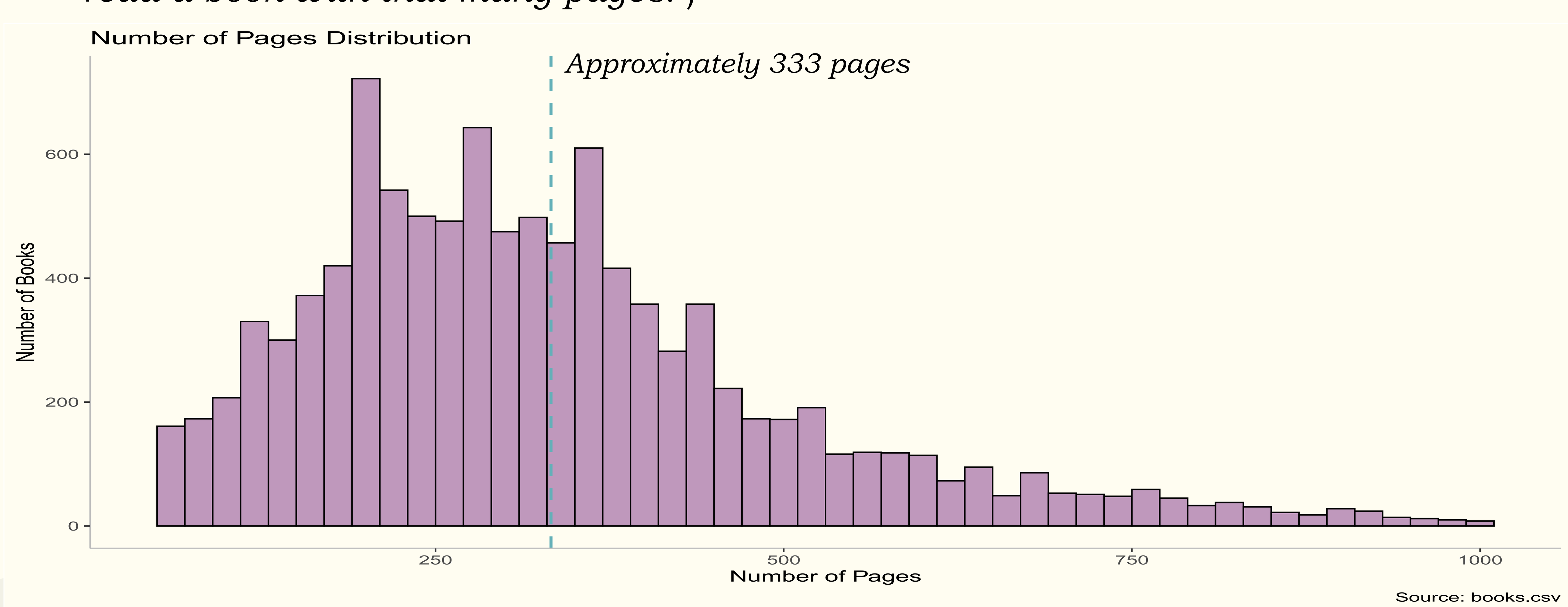
English is the third most spoken language in the world behind Mandarin Chinese and Spanish respectively.

Goodreads (original website where this book data comes from is an American built social-cateloging/media platform, so it's not surprising that most of the books are in English.

Chapter 3: Number of Pages

Of the books in this dataset, what is the average number of pages?

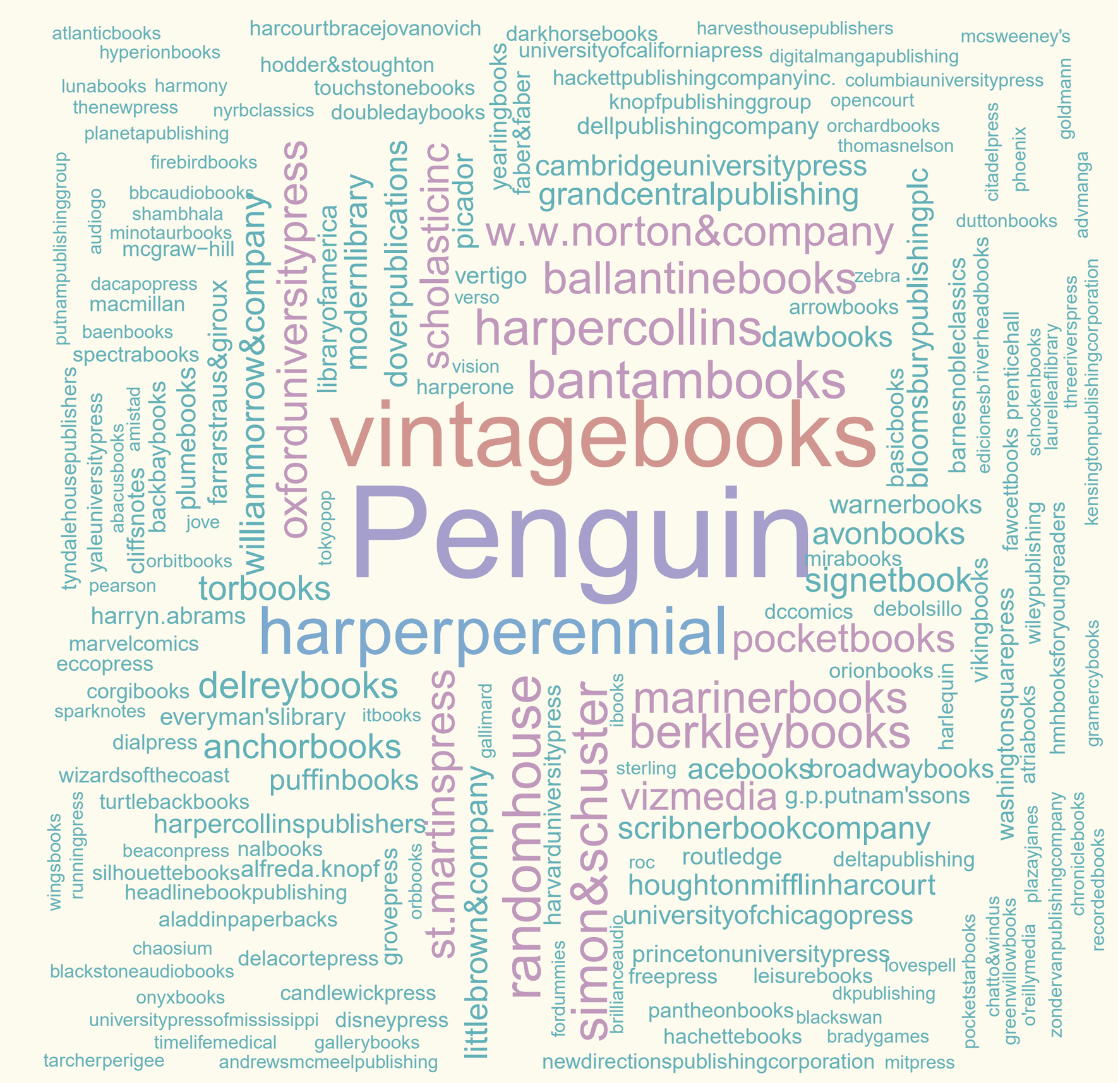
Data cleaning resulted in excluding all books with less than 50 pages; and all books with more than 1,000 pages (*don't know about anyone else, but I don't think I've ever read a book with that many pages!*)



Chapter 4: Most Books by Publisher

Of the books in this dataset, which publisher has the most published books?

Data cleaning resulted combining all the publishers that were similar in title/spelling i.e. bantam = bantambooks; also removed spaces and most punctuation to get the word cloud to look cleaner into a single text file.



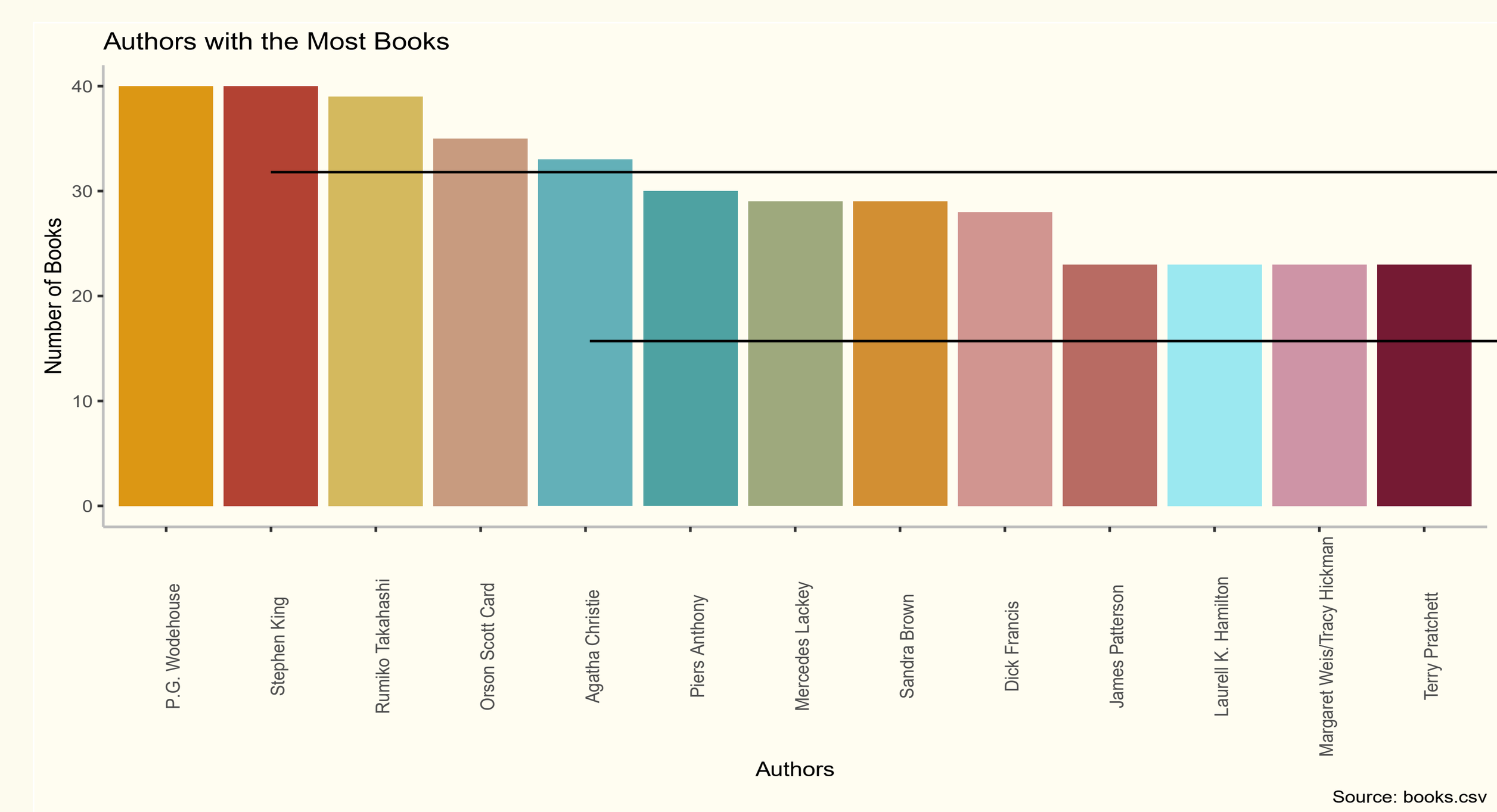
According to Publishers Weekly (https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/72889-ranking-america-s-largest-publishers.html)

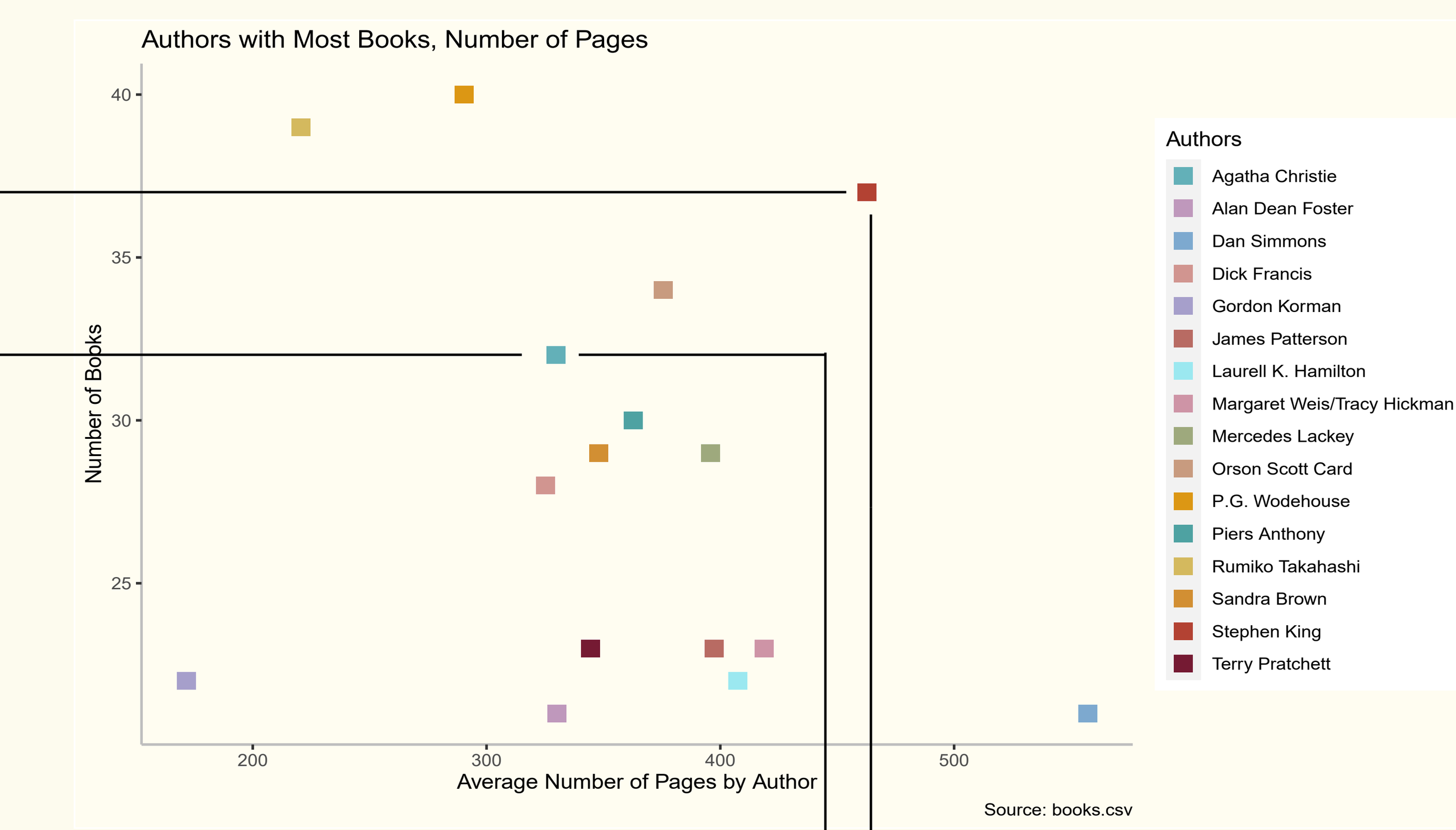
Penguin Random House, HarperCollins and Simon & Schuster are the three largest publishers in the United States (as of 2016).

Publishing houses today carry many imprints: for example: Penguin Random House has almost 275 publishing imprints (including Penguin and Random House separately). This dataset contained all of the individual imprints that may sit with a larger publishing house. Even so, it's easy to see Penguin is the leader

Chapter 5: All about the Authors

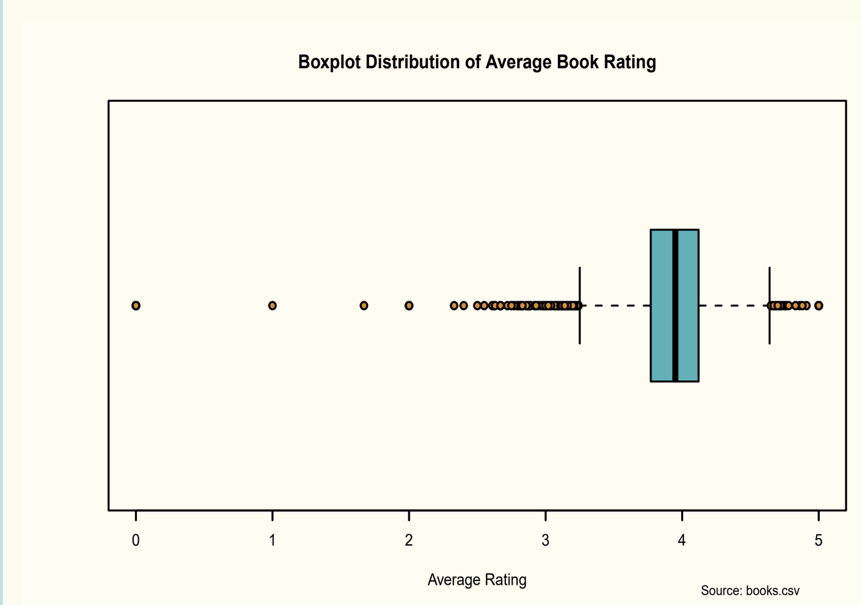
Of the books in this dataset, which authors have the most books and the most pages?

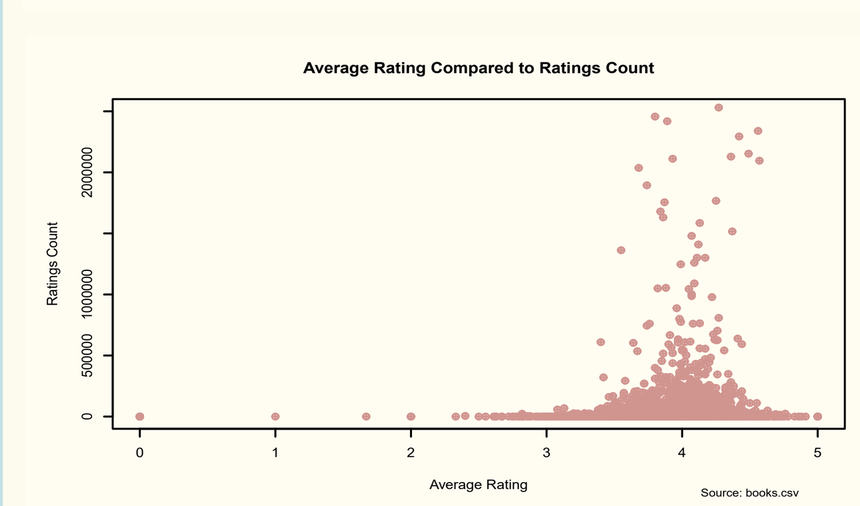




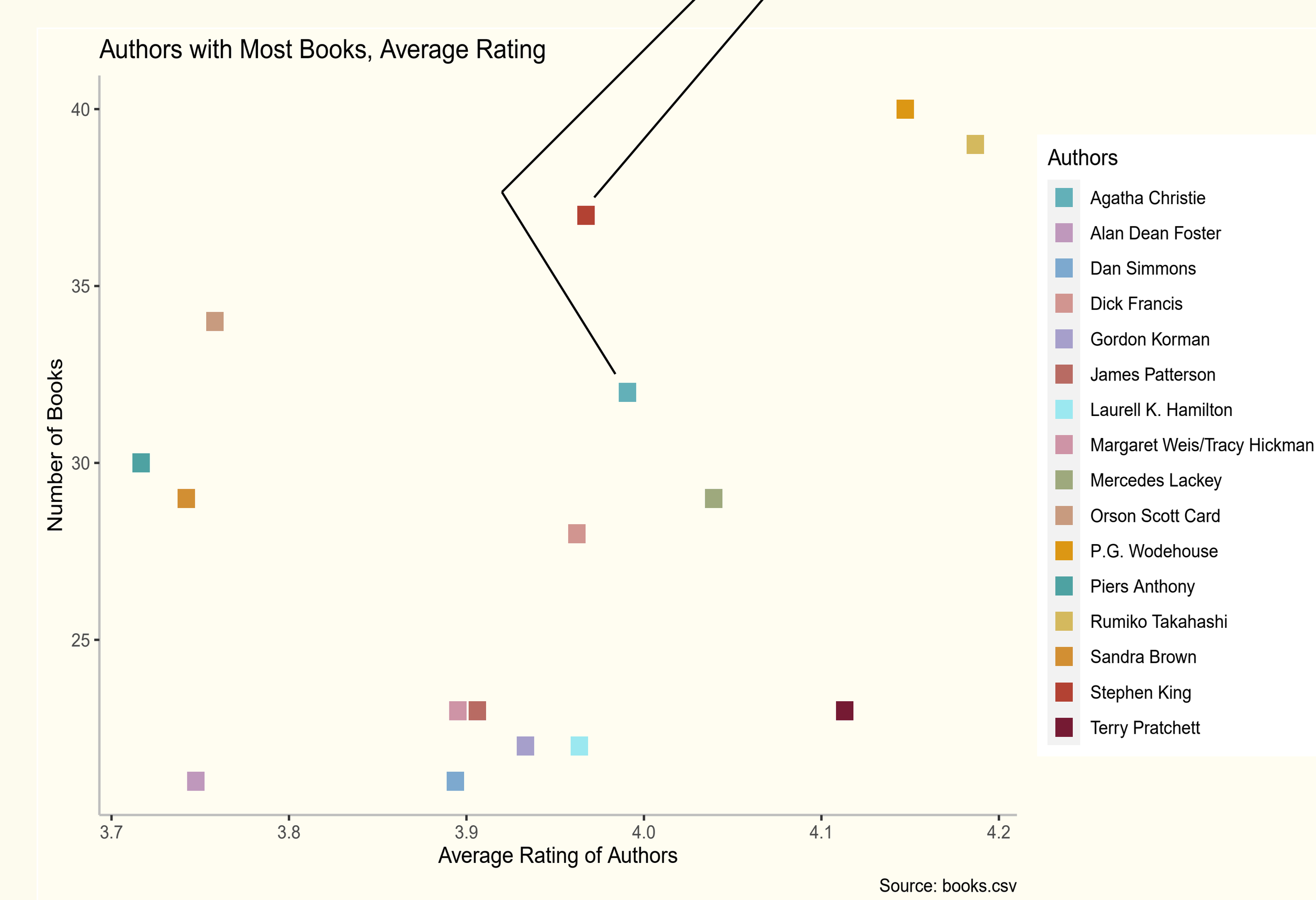
Chapter 6: It's all about those ratings!

What do ratings tell us about us (and the authors we love)?





As you can see - we readers either don't rate books we don't like (or perhaps we aren't finishing books we don't like so they aren't rateable); or we just feel bad about giving books a rating less than 3.



THE END