

# Top 500 “Greatest” Novels (1021-2015)

Anna Preus and Aashna Sheth

2024-04-10

## Table of contents

<b>Data Essay</b>	<b>2</b>
Introduction . . . . .	2
<b>HISTORY</b> . . . . .	3
<b>WHERE DID THE DATA COME FROM? WHO COLLECTED IT?</b> . . .	4
<b>WHY WAS THE DATA COLLECTED? HOW IS THE DATA USED?</b> . .	5
The Top 500 list: . . . . .	5
Our curated dataset: . . . . .	5
<b>HOW WAS THE DATA COLLECTED?</b> . . . . .	5
The top 500 list: . . . . .	5
Which libraries are represented? . . . . .	6
Which books are represented? . . . . .	6
How are editions and translations counted? . . . . .	7
Our curated dataset: . . . . .	7
<b>EXAMINING BIAS</b> . . . . .	8
The top 500 list: . . . . .	8
Our curated dataset: . . . . .	14
<b>POPULARITY VS CANONICITY</b> . . . . .	15
<b>USING FULL TEXT DATA</b> . . . . .	29
<b>CONCLUSION</b> . . . . .	29
<b>Explore the Data</b>	<b>30</b>
Library Top 500 . . . . .	30
<b>Exercises</b>	<b>31</b>
<b>R</b>	<b>31</b>
<b>Python</b>	<b>31</b>

<b>Discussion &amp; Activities</b>	<b>31</b>
Activity 1 . . . . .	31
Activity 2 . . . . .	32
Activity 3 . . . . .	32

## Data Essay

### Introduction

This dataset contains information on the top 500 novels most widely held in libraries, according to [OCLC](#), a library organization with over 16,000 member libraries in over 100 countries. The dataset includes information on authors' biographies, library holdings, and online engagement for each novel, as well as the full text for all works that are not currently under copyright (190 novels).

#### Brief Survey

If you use our materials in your class or another setting, we would love to [hear about it!](#)

```
import {viewof dataSummaryView, viewof dataUrl, viewof dataSet, tableContainer, table, table}
```

```
//|echo: false
viewof dataSet
viewof dataUrl
tableContainer
```

```
//|echo: false
//|output: false
table
html`<link href='https://unpkg.com/tabulator-tables@5.3.1/dist/css/tabulator.min.css' rel='s
tabulator.min.css`
```

```
//|echo: false
viewof dataSummaryView
```

#### Creative Commons License

This work is licensed under CC BY 4.0

This dataset is based on a list of the [Top 500 Novels](#) compiled by OCLC from information in their online database [WorldCat](#), the largest database of library records. The first section of

the list was published online with great fanfare as the [Library 100](#) in 2019, accompanied by the claim that for novels, “literary greatness can be measured by how many libraries have a copy on their shelves.”

We wondered about the implications of this claim and about what it means to base ideas of “literary greatness” on the number of libraries that hold a particular work. How do historical biases in systems of literary production and preservation figure into these kinds of claims? Which libraries’ records are included in the data? And how do we even define what counts as a novel?

To contextualize the initial list and dig into its claims about literary greatness, we collected information on each novel from a number of other databases, including [Wikipedia](#), [Goodreads](#), [Project Gutenberg](#), the [Virtual International Authority File \(VIAF\)](#), and [Classify](#) (a now-shuttered OCLC tool), which we have compiled here.

The dataset was created by Anna Preus and Aashna Sheth, who are also the authors of this data essay.

## HISTORY

To start, what is a novel? “Novel” is an umbrella term for works of longform fiction in a range of genres: romance, sci-fi, historical fiction, horror, detective fiction, westerns, etc. The word “novel” was first used in English to describe a “long fictional prose narrative” in the 1600s (OED), and the form increased in popularity across the 18th and 19th centuries. Interestingly, OCLC’s list of top 500 novels extends much further back than this. The oldest work on the list is *The Tale of Genji*, a classic work of Japanese literature written over 1,000 years ago. On the other end of the timeline, the list includes many contemporary best-sellers, including all the titles in the *Harry Potter*, *Twilight*, and *Hunger Games* series.

This long time span is one of the things that makes OCLC’s data, and the list specifically, so interesting. A key issue in literary studies is which works from the past we continue to read in the present, and which works from the present we’ll continue to read in the future. The vast majority of novels fall out of circulation shortly after they’re published, quickly becoming part of what Margaret Cohen has called “the great unread” (Cohen 61).<sup>1</sup> The Top 500 list, though, represents historical works that have achieved exceptional levels of attention and have entered what is often referred to as the literary “canon.” Ankhi Mukherjee defines the canon as “a set of texts whose value and readability have borne the test of time,” noting that this “involves not merely a work’s admission into an elite club, but its induction into ongoing critical dialogue and contestations of literary value” (Mukherjee). Canonical works continue to be read, taught, and discussed, and in popular terminology they’re often considered “classics.” These are works you might read in a high school or college English class: F. Scott Fitzgerald’s *The Great Gatsby*, for example, or Jane Austen’s *Pride and Prejudice*.

---

<sup>1</sup>Franco Moretti also uses this term, borrowing it from Cohen. We follow Cohen’s use of the term.

One of the things that defines a classic is the fact that it stays in print for a long period of time. When a book is published, it is issued in an edition with a specific number of physical copies. If the book is profitable, it may be re-issued in different editions over many years and edited repeatedly by different scholars across time. If it becomes canonical, it is likely to be issued in dozens or hundreds of editions even long after the author's death, leading to more physical copies of the book in circulation. Importantly, though, there is not just one canon or one stable set of classics. Canons are constructed and reinforced by people; they are socially and historically defined and are bound up in power relationships and in histories of exclusion and erasure. This is what makes OCLC's task of defining the top 500 greatest novels of all time so potentially problematic: their data reflects a history of canonization that has influenced library collections, and which has long been biased toward English-language texts, White male authors, and works produced in Europe and North America.

The newer works included on the list are books that have achieved immense popularity and widespread sales in recent years. These works, which were published during the period that Dan Sinykin has termed the "Conglomerate Era," are usually issued by publishers that operate as part of large, multinational corporations, and which have the resources to print and distribute millions of books around the world (Sinykin). Many of these novels have also been adapted into major films or TV series.

By focusing on books that librarians have chosen to continue to make available to readers, OCLC was able to create a list of widely read novels that includes both classic texts and more recent, popular works by living authors. The list, though, also reflects various forms of bias rooted in literary history, in library collections, and in the data itself. We wondered, whose conception of "literary greatness" is being represented? How does OCLC's data compare to other potential indicators of popularity or canonicity? And, for that matter, how was the list actually constructed?

## WHERE DID THE DATA COME FROM? WHO COLLECTED IT?

**###The Top 500 list** The initial list of Top 500 novels was collected by a team at OCLC, the non-profit organization that manages WorldCat. It was compiled based on analysis of data in WorldCat, which consists of catalog records created and entered by librarians at OCLC member libraries.

**###Our curated dataset** Building on this list, we compiled data from a number of other databases, including Project Gutenberg, VIAF, Wikipedia, and Goodreads—a process that is described in greater detail below.

## WHY WAS THE DATA COLLECTED? HOW IS THE DATA USED?

### The Top 500 list:

OCLC's goal in producing the Top 500 list seems to be to share information about an important set of texts based on the unprecedented amount of information in their database, as well as to encourage library patronage and reading. The website for the list includes a "[Librarians Kit](#)" with a variety of publicity materials—from printable bookmarks to Instagram tiles—that can help bring attention to books in the Top 500 list within libraries' collections.

Need to link to local image file or to image URL    ![alt\_text](top\_500\_kit.png "image\_tooltip")

### Our curated dataset:

Our goal as researchers was to collect data from additional sources in order to understand how the list was constructed and to contextualize and question its claims about literary greatness.

## HOW WAS THE DATA COLLECTED?

### The top 500 list:

The Top 500 list represents a massive data extraction and analysis effort on the part of OCLC. While they do not provide detailed information on how the list was compiled, they do offer a brief explanation of the process that went into creating the list on their [FAQ page](#) (written in the context of the top 100, but also applies to the top 500):

Materials in libraries are described and tracked in WorldCat in two ways. Any specific work of literature, music, art, history, etc., has an associated **catalog record**. This describes the item in a general sense. Every copy of the same book, for example, shares the same record. WorldCat also tracks library **holdings**, which indicate that a specific library has (or holds) at least one copy of that item.

The Library 100 is based on the total number of holdings for a specific novel across all libraries that have registered that information in WorldCat. When a library tells OCLC, "We have a copy of that book available," that counts as a holding, and in the case of The Library 100, counts as +1 toward its ranking on the list.

This process initially sounds straightforward: to create the Top 500 list, the OCLC team presumably searched the title of a work, counted the number of libraries that held each title, and published the first 500. But when we dug into the database, we found it was actually much more complicated than that. The list is influenced by a range of factors, including which

libraries' collections are represented, what kinds of books are considered, and how holdings are totalled across different editions and translations of individual titles.

### **Which libraries are represented?**

According to OCLC, "WorldCat holdings information represents the collective inventory of OCLC member libraries" (WorldCat Holdings). But who are these member libraries? And where are they? OCLC publishes some summary data about WorldCat, revealing, for example, that it currently holds over 548 million bibliographic records representing over 3.3 billion library holdings in 490 languages. But while OCLC stresses its position as "The worldwide catalog of library resources" and emphasizes the membership of libraries in over one hundred countries, it doesn't provide much specific information on where these libraries are located or what kinds of institutions they are (Inside WorldCat).

In order to get a general sense of the geographic distribution of OCLC member libraries, we dug into the organization's [directory](#) and conducted filtered searches for libraries in each country. We found that over 70% of OCLC's members are in the U.S., followed by 7% in Germany, 4% in Australia, 2.6% in Canada, and 1.5% in the U.K. Clearly, OCLC is most well represented in the U.S., where it is based, and the fact that three of the other top four countries in terms of membership have English as a national language helps to explain why English-language materials are disproportionately represented in the catalog and in the Top 500 List.

Need to link to local image file or to image URL - pie chart

We used a similar approach to look at what kinds of institutions are represented in WorldCat, this time filtering by "Library Type." We found that most OCLC members are school libraries (29%), public libraries (29%), or academic libraries (25%) and that membership is fairly evenly distributed across these categories. The prominence of school libraries and academic libraries raises the issue of which patrons have access to these libraries—and thus whose conception of popularity is being represented in the holdings data. It also points to the influence of educators on this picture of the Top 500 novels.

Need to link to local image file or to image URL - bar chart

### **Which books are represented?**

Since the list focuses specifically on *novels* in these libraries' collections, it is also narrowed by genre. OCLC discusses its process for identifying novels on its FAQ page, noting that they began with "everything in WorldCat that counts broadly as 'fiction'" and then winnowed the list down through the removal of known categories like "children's books, poetry, drama, folklore, comics," and "short stories." The final list was later "reviewed by an editorial team."

Importantly, the Top 500 List is also based only on holdings of physical books, and it "does not include e-books, audiobooks, children's adaptations, film adaptations, etc." This exclusive focus on print books puts emphasis on the choices of librarians, since libraries have limited

shelf space and periodically have to cull their print collections. As OCLC puts it, “libraries offer access to trendy and popular books. But, they don’t keep them on the shelf if they’re not repeatedly requested by their communities over the years.” By contrast, they suggest that ebooks are often incorporated via “automatic links to free collections on the web,” which do not “represent a specific decision to add a particular novel to a library’s collection” (FAQ). While this may be the case, given the popularity of eBooks (Zhang and Kudva), a focus on print must have influenced the overall makeup of the list, and, again, whose idea of popularity or “greatness” it represents.

### **How are editions and translations counted?**

One further complication is that in WorldCat, records are stored by edition, meaning that each edition of a particular novel has its own catalog record. An individual title like, say, Miguel de Cervantes’s *Don Quixote*, may have been released in hundreds or thousands of editions since its initial publication.

This means that when developing the list, the OCLC team actually had to find all the editions of a specific title and sum the number of libraries that hold that edition across all editions. **Thus the top 500 list is not only a representation of how many libraries carry the work, but a representation of how many times a book has been re-edited and re-issued; the more editions a book has, the more records are created and the more copies of a book a library may hold.** Often, there are duplicate records for individual editions, which may affect the overall count of copies tallied by OCLC. And when a work is translated into different languages, all the editions of all the translations are also recorded in WorldCat, which also figures into the count of total holdings for each novel.

The combined influence of these different factors can be seen in the representation of works in languages other than English, which make up around 14% of the list. The non-English-language texts that are at the top of the list—*Don Quixote*, *Crime and Punishment*, *Madame Bovary*, *The Three Musketeers*, and *War and Peace*—have all been widely translated into English, a trend that continues as you go down the list.

### **Our curated dataset:**

We chose to contextualize the Library Top 500 List by compiling additional information on each novel from a range of other sources. We focused on gathering three main categories of information: information that could help us understand what types of works—and whose works—were included on the list, data that could potentially provide alternate measures of popularity or canonicity, and the full text of each novel that was in the public domain. We collected information from the following sources:

**WorldCat:** we used the now-shuttered OCLC tool Classify to gather data from WorldCat based on an OWI (OCLC Work ID) for each of the 500 novels on the list.<sup>2</sup> We recorded total physical and eholdings for this work. The Top 500 list only considers physical holdings. The number of holdings in our curated dataset is not perfectly descending as the top 500 rank decreases, as one would expect. This is likely due to complications with the OWI number and with the inclusion of translations; the top 500 list uses multiple OWIs to calculate total holdings, while we only use one. Which OWIs the top 500 curators use for each work is unclear.

Need to link to local image file or to image URL `![alt_text](viaf_example.png "image_tooltip")`

**Wikipedia:** We used Wikipedia, the popular, free, volunteer-authored encyclopedia, to identify the year of first publication for each novel on the list.

**Goodreads:** Goodreads, which is owned by Amazon, is the largest social networking site related to books, with over 150 million members. It allows users to rate, review, and discuss a huge range of texts. We drew on data from Goodreads as a potential alternate indicator of texts' popularity, collecting total number of reviews, total number of ratings, and average overall rating for each novel on the list.

**Project Gutenberg:** We used Project Gutenberg to access the full-text of all novels on the list that are currently in the public domain, or in other words, out of copyright. We chose Project Gutenberg because their eBooks are edited by volunteers, whereas many larger content repositories, like Internet Archive and HathiTrust, only make available machine-generated transcriptions of historical texts, which tend to be less accurate.

Our work creating this dataset not only builds on the work of the OCLC team who compiled the Top 500 list, but on the labor of the thousands of librarians who created records held in WorldCat and VIAF, of the volunteers who transcribed texts for Project Gutenberg and wrote articles for Wikipedia, and of the social media users who reviewed and rated books on Goodreads.

## EXAMINING BIAS

### The top 500 list:

The OCLC's definition of "literary greatness" is biased based on the libraries that OCLC represents, the list's exclusive focus on physical books, and its emphasis on raw number of holdings, which is influenced by number of editions. OCLC acknowledges potential biases in

---

<sup>2</sup>For more on how editions of works are clustered in WorldCat see "Clustering WorldCat Discovery." **VIAF:** The Virtual International Authority File is an OCLC-run database that contains structured records—called "name authority files"—for individual authors and creators. We used VIAF to gather information on authors whose novels were included on the list, including their birth and death dates, nationalities, genders, and occupations.



their claims, noting that “The [top 500] list emphasizes many books that we tend to think of as ‘classics,’ because those are the novels most often translated, retold in different editions, taught and widely distributed in library collections. Because of this, the list tends to reflect more dominant cultural views.”

A key reason we decided to collect additional data related to the list was to explore what kinds of works, and especially whose works, it represents. Drawing on author data gathered from VIAF, we can calculate some overall descriptive statistics for the list.

Looking at the AUTHOR\_GENDER column, we can count the number of authors identified as male and the number identified as female (VIAF only includes options for binary genders, which is discussed further below), and we can see that over 70% of the novels were written by men.

```
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv("final_merged_dataset.tsv", sep='\t', header=0, low_memory=False)

df["author_gender"].value_counts(dropna=False)
```

```
author_gender
male      354
female    145
NaN        1
Name: count, dtype: int64
```

We can use a similar approach to look at the nationalities of authors whose works are represented on the list. Focusing on the AUTHOR\_NATIONALITY column, we can count how many times each country code appears, and see that over 80% of the novels were written by authors from the U.S. or the U.K.

```
df["author_nationality"].value_counts(dropna=False)
```

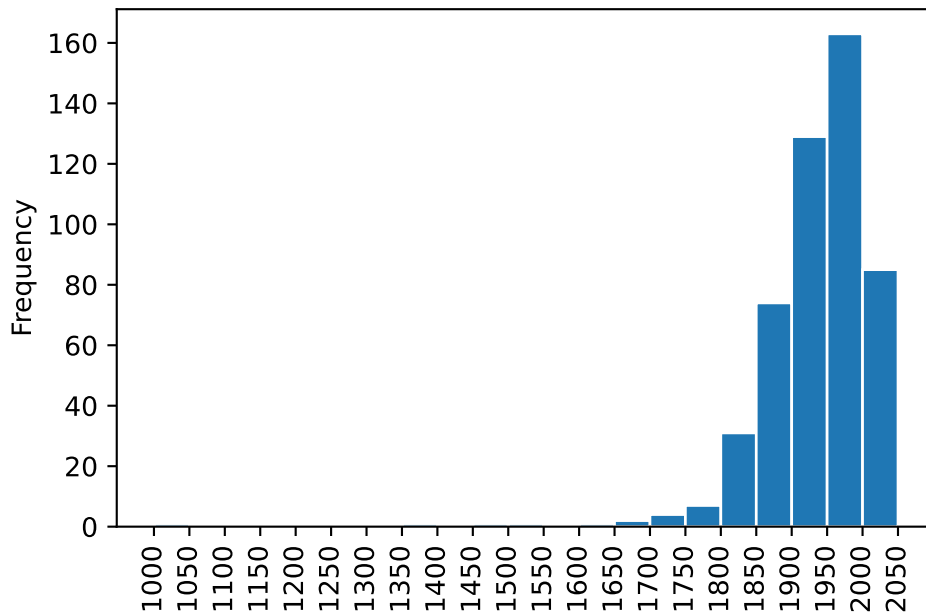
```
author_nationality
US      257
GB      148
FR       27
DE       10
RU       10
CA        8
IE        8
```

```
IT      5
SE      4
CZ      3
CO      3
AU      3
CH      2
CL      2
NG      1
NaN     1
MX      1
PL      1
ES      1
JP      1
CN      1
ZA      1
BR      1
IN      1
Name: count, dtype: int64
```

To find out what time period is most frequently represented on the list, we can look at the `PUB_YEAR` column and see that almost 50% of novels were first published between 1950 and 2000.

```
import numpy as np

bins = np.arange(1000, 2060, 50)
bars = df['pub_year'].plot.hist(bins=bins, edgecolor='w')
plt.xticks(rotation='vertical');
plt.xticks(bins);
```



We can also get a sense of the immense influence of individual authors who appear on the list numerous times. The most represented authors are John Grisham (19 novels) and Charles Dickens (15 novels).

```
df["author"].value_counts(dropna=False).head(10)
```

```
author
John Grisham          19
Charles Dickens       15
John Steinbeck         8
C.S. Lewis            8
J.K. Rowling           7
Nicholas Sparks        7
Stephen King           7
Laura Ingalls Wilder   7
Beverly Cleary         5
Thomas Hardy           5
Name: count, dtype: int64
```

Drawing on slightly more complex techniques, we can see that there is a strong positive correlation ( $p=1.1165e-73$ ,  $r=0.6985$ ) between the current ranking of the Top 500 List and a ranking based on the total number of editions for each novel. This suggests that the more editions a novel has, the more likely it is to be higher on the list, which is relevant because European

and American editing practices have long favored authors occupying dominant social positions. Historically, works by White authors and male authors are more likely to have been re-edited and re-issued and to be considered literary classics (Gates; Mandell).<sup>3</sup>

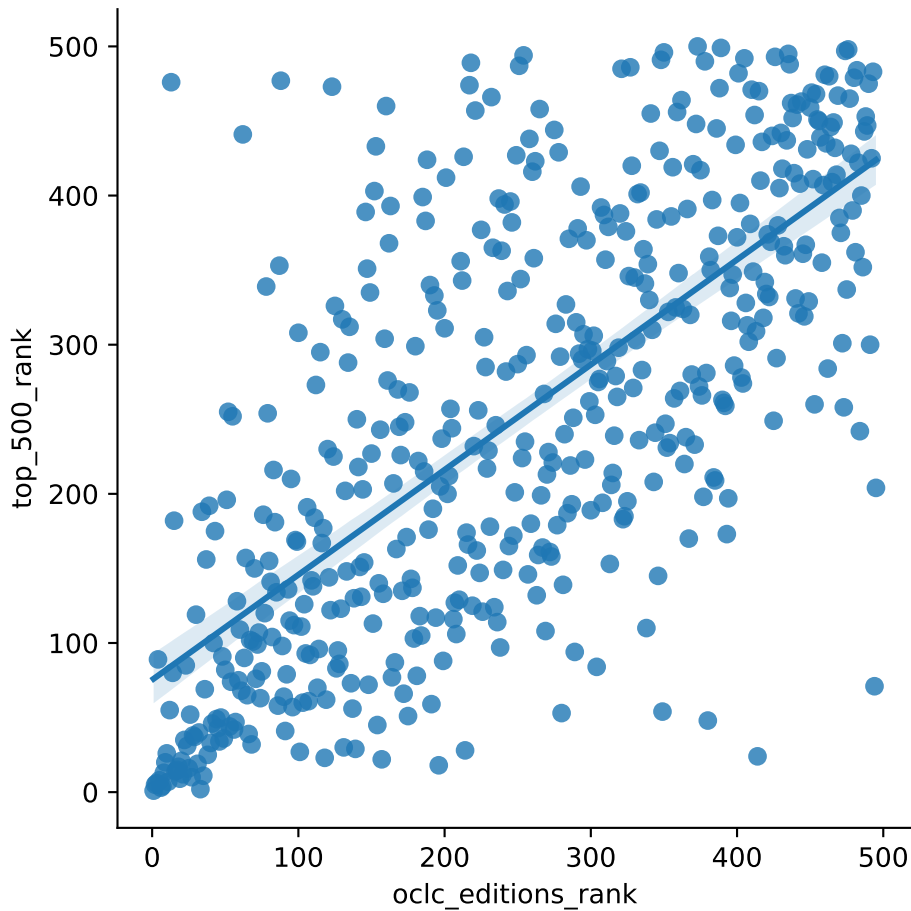
```
import pandas as pd
import seaborn as sns
from scipy import stats
# inspired by: https://www.sfu.ca/~mjbrydon/tutorials/BAinPy/08_correlation.html

sns.lmplot(x="oclc_editions_rank", y="top_500_rank", data=df)
dropped_df = df[df.oclc_editions_rank.notna()]
print(stats.pearsonr(dropped_df['oclc_editions_rank'], dropped_df['top_500_rank']))
```

```
PearsonRResult(statistic=0.6985608812420623, pvalue=1.1165447422670395e-73)
```

---

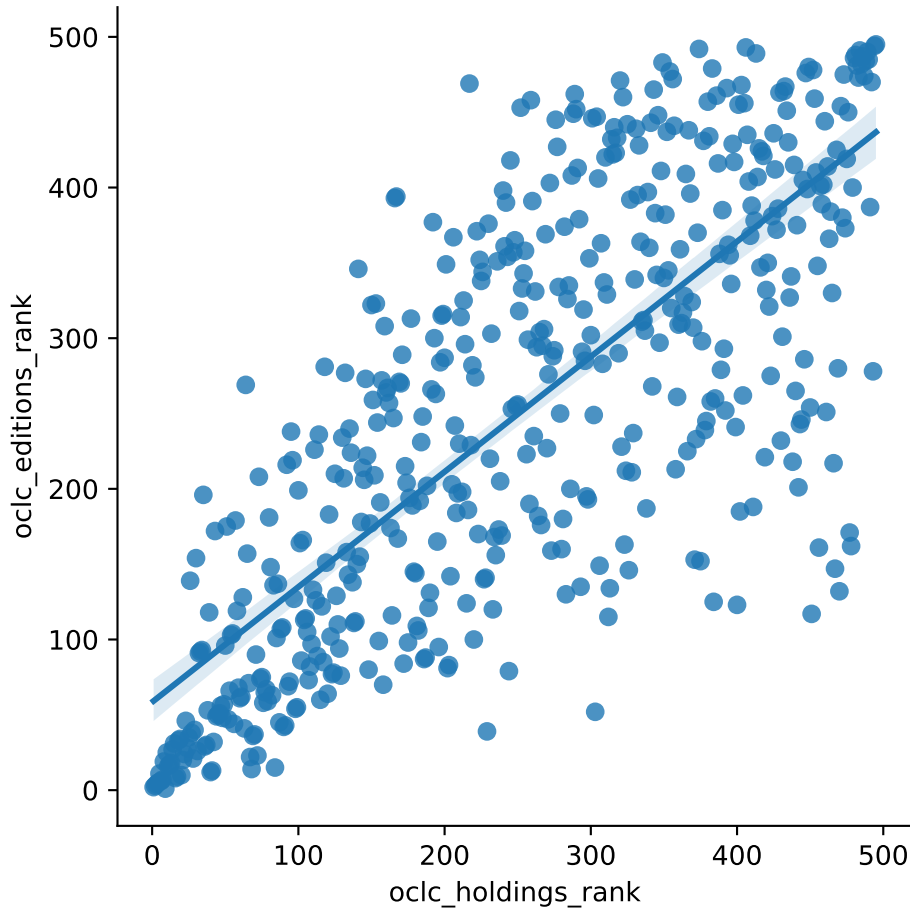
<sup>3</sup>Laura Mandell argues that “women writers are being recovered and forgotten in cycles, both in print and potentially in digital media,” pointing out that historically “works by men have been published and republished” while “women writers only appear in the materiality of the single print run” (Mandell). In his work on “What Makes a ‘Classic’ African American Text,” Henry Louis Gates Jr. discusses the historical exclusion of Black authors from the Penguin Classics series, as well as his work editing a new series of African American Classics for the imprint. He notes that “texts by people of color, and texts by women” are “still struggling, despite enormous gains over the last twenty years, to gain a solid foothold in anthologies and syllabi.” These kinds of biases in turn affect which works appear on library shelves.



Similarly, we confirm that there is a very strong positive correlation ( $p=5.6541e-96$ ,  $r=0.7642$ ) between number of editions and number of holdings of a novel; the more editions a book has, the more total holdings are reported in OCLC.

```
sns.lmplot(x="oclc_holdings_rank", y="oclc_editions_rank", data=df)
dropped_df = df[df.oclc_editions_rank.notna() & df.oclc_holdings_rank.notna()]
print(stats.pearsonr(dropped_df['oclc_holdings_rank'], dropped_df['oclc_editions_rank']))
```

```
PearsonRResult(statistic=0.7642639335763275, pvalue=5.65410769095468e-96)
```



### **Our curated dataset:**

Although the additional data we curated helps to contextualize the Top 500 List and to reveal some of its biases, the data we added also contains its own biases. For starters, as researchers, we both primarily work in English, and we are pursuing this project at a University in the U.S. These contexts have informed our areas of inquiry and the sources we’ve chosen to use. We primarily drew on widely used online databases created in English-language contexts (VIAF, Project Gutenberg, etc.). Further, we have limited our data collection to OCLC’s list of the Top 500 novels and did not attempt to expand to other rankings of literary greatness or to additional novels.

The sources we have used, of course, have biases of their own. VIAF relies on a standardized vocabulary, which can be helpful for data analysis and organization, but erases important nuances. For example, VIAF categorizes gender with the binary labels of “male” and “female,” with the only other option being “unknown.” This, of course, reinforces binary understandings

of gender and obscures the existence of non-binary people (Drabinski). Labels used in fields like “AUTHOR\_NATIONALITY,” “FIELD\_OF\_ACTIVITY,” and “OCCUPATION” also do not paint a complete picture. The entries in the latter two columns are based on Library of Congress data and may not be equally rich for all authors. And nationality labels from VIAF can obfuscate racial, political, ethnic, and tribal affiliations, and flatten the complexity of individual authors’ experiences.<sup>4</sup> For example, the nationality for Sherman Alexie, author of *The Absolutely True Diary of a Part-time Indian*, is listed as “U.S.A.,” but his identity as a member of the Spokane Tribe of Indians is not referenced. In another example, the first nationality listed for Khaled Hosseini, author of *The Kite Runner*, is “U.S.A.” followed by “Afghanistan.” This is not inaccurate but it is oversimplified, since Hosseini was born in Kabul, lived in Iran, France, and Afghanistan throughout his childhood, and then moved to California after his family sought political asylum in the U.S.

We urge researchers using this dataset to consider its biases when drawing conclusions, and to seek other sources to expand it, question it, and/or to fill in information that may be missing or lacking.

You can find more metadata analysis [here](#).

## POPULARITY VS CANONICITY

Because we were interested in whose opinions are represented on the list, we wanted to bring in an alternate measure of popularity, and we decided to use information from Goodreads. Goodreads was appealing because of its prominence online (over 130 million users), which we hoped might help us consider the opinions of a somewhat different set of readers than those theoretically represented through the physical holdings of libraries. Melanie Walsh and Maria Antoniak, for example, have drawn on Goodreads reviews to analyze how social media users define the “Classics.” Drawing on this work, we compare the ranking of novels on OCLC’s original list of Top 500 novels to the rankings of those same novels based on Goodreads ratings and number of reviews. Through this comparison we aim to consider how social media users engage with “classic” and “popular” novels and to interrogate the relationship between canonicity and popularity, using information from different data sources.

To unpack the differences between the Goodreads data and the Top 500 rankings, we first need to think about how we want to compare the two lists. Given that we have recorded Goodread rankings by average star rating and total number of ratings, which metric would be better to use? Would we want to create another metric?

---

<sup>4</sup>Safiya Umoja Noble argues that “information organization is a matter of sociopolitical and historical processes that serve particular interests,” tying library cataloging and classification systems to “the development of racial classification” in the 19th century (136-137). And Roopika Risam also highlights the role of public-sector knowledge institutions in perpetuating these structural biases, emphasizing “the failure to take into account the complicity of universities, libraries, and the cultural heritage sector in devaluing black and indigenous lives and perpetuating the legacies of colonialism in the cultural and digital cultural records alike” (14).

For our purposes, we decided to use total number of ratings instead of average rating, since it seemed most closely related to how OCLC measures popularity—by number of holdings, not how much patrons say they enjoy reading the books.

```
def top_5_comparison(col_name):
    print(df[["title", "author", "top_500_rank", col_name]].head(5))

    sorted = df.sort_values(by=[col_name])
    print(sorted[["title", "author", "top_500_rank", col_name]].head(5))

top_5_comparison("gr_num_ratings_rank")
```

	title	author	top_500_rank	\
0	Don Quixote	Miguel de Cervantes	1	
1	Alice's Adventures in Wonderland	Lewis Carroll	2	
2	The Adventures of Huckleberry Finn	Mark Twain	3	
3	The Adventures of Tom Sawyer	Mark Twain	4	
4	Treasure Island	Robert Louis Stevenson	5	

	gr_num_ratings_rank
0	211
1	133
2	68
3	88
4	145

	title	author	top_500_rank	\
44	Harry Potter and the Sorcerer's Stone	J.K. Rowling	45	
172	The Hunger Games	Suzanne Collins	173	
131	Twilight	Stephenie Meyer	132	
28	To Kill a Mockingbird	Harper Lee	29	
33	The Great Gatsby	F. Scott Fitzgerald	34	

	gr_num_ratings_rank
44	1
172	2
131	3
28	4
33	5

Above you can see that the GoodReads rankings and the top 500 rankings aren't very aligned! What factors might affect popularity on Goodreads compared to OCLC?



```

import math

def print_rankings(d, col_name):
    rank_B = d[col_name]
    rank_A = d["top_500_rank"]
    title = d["title"]
    points_moved = 0
    if (math.isnan(rank_B)):
        points_moved = 501
    else:
        if rank_B > int(rank_A):
            points_moved = rank_B - rank_A
            print(f"\u001b[31m  -{int(points_moved)} {title}")
        elif rank_B < rank_A:
            points_moved = rank_A - rank_B
            print(f"\u001b[32m  +{int(points_moved)} {title}")
        else:
            print(f"\u001b[30m   {int(points_moved)} {title}")
    d["points_moved"] = int(points_moved)
    return d

df = df.apply(lambda d: print_rankings(d, "gr_num_ratings_rank"), axis=1)

```

```

-210 Don Quixote
-131 Alice's Adventures in Wonderland
-65 The Adventures of Huckleberry Finn
-84 The Adventures of Tom Sawyer
-140 Treasure Island
-2 Pride and Prejudice
-39 Wuthering Heights
-32 Jane Eyre
-125 Moby Dick
-85 The Scarlet Letter
-197 Gulliver's Travels
-266 The Pilgrim's Progress
-85 A Christmas Carol
-214 David Copperfield
-71 A Tale of Two Cities
-22 Little Women
-86 Great Expectations
+8 The Hobbit, or, There and Back Again
-35 Frankenstein, or, the Modern Prometheus

```

-149 Oliver Twist  
-209 Uncle Tom's Cabin  
-72 Crime and Punishment  
-159 Madame Bovary: Patterns of Provincial life  
-69 The Return of the King  
-42 Dracula  
-160 The Three Musketeers  
-16 Brave New World  
-155 War and Peace  
+25 To Kill a Mockingbird  
-122 The Wizard of Oz  
-73 Les Misérables  
-43 The Secret Garden  
+21 Animal Farm  
+29 The Great Gatsby  
-4 The Little Prince  
-124 The Call of the Wild  
-444 20,000 Leagues Under the Sea  
-59 Anna Karenina  
-193 The Wind in the Willows  
-17 The Picture of Dorian Gray  
-50 The Grapes of Wrath  
-32 Sense and Sensibility  
-279 The Last of the Mohicans  
-159 Tess of the d'Urbervilles  
+44 Harry Potter and the Sorcerer's Stone  
-193 Heidi  
-242 Ulysses  
-192 The Complete Sherlock Holmes  
-41 The Count of Monte Cristo  
-27 The Old Man and the Sea  
+22 The Lion, the Witch, and the Wardrobe  
-184 The Hunchback of Notre Dame  
-293 Pinocchio  
-28 One Hundred Years of Solitude  
-274 Ivanhoe  
-259 The Red Badge of Courage  
-24 Anne of Green Gables  
-146 Black Beauty  
-120 Peter Pan  
-127 A Farewell to Arms  
-349 The House of the Seven Gables  
+35 Lord of the Flies

- 233 The Prince and the Pauper
- 209 A Portrait of the Artist as a Young Man
- 367 Lord Jim
- +55 Harry Potter and the Chamber of Secrets
- 287 The Red & the Black
- 11 The Stranger
- 116 The Trial
- 224 Lady Chatterley's Lover
- 298 Kidnapped: The Adventures of David Balfour
- +56 The Catcher in the Rye
- +38 Fahrenheit 451
- 164 A Journey to the Center of the Earth
- 213 Vanity Fair
- 75 All Quiet on the Western Front
- +6 Gone with the Wind
- 201 My Ántonia
- +47 Of Mice and Men
- 405 The Vicar of Wakefield
- 235 A Connecticut Yankee in King Arthur's Court
- 164 White Fang
- 255 Fathers and Sons
- 242 Doctor Zhivago
- 324 The Decameron
- +79 Nineteen Eighty-Four
- 187 The Jungle
- +51 The Da Vinci Code
- 26 Persuasion
- 88 Mansfield Park
- 114 Candide
- 107 For Whom the Bell Tolls
- 178 Far from the Madding Crowd
- +66 The Fellowship of the Ring
- 319 The Return of the Native
- 294 Sons and Lovers
- +52 Charlotte's Web
- 214 The Swiss Family Robinson
- 210 Bleak House
- 278 Père Goriot
- 252 Utopia
- 327 The History of Tom Jones, a Foundling
- +94 Harry Potter and the Prisoner of Azkaban
- 314 Kim
- 150 The Sound and the Fury

+92 Harry Potter and the Goblet of Fire  
-278 The Mill on the Floss  
+36 A Wrinkle in Time  
-72 The Hound of the Baskervilles  
+27 The Two Towers  
-78 The War of the Worlds  
-152 Middlemarch  
-146 The Age of Innocence  
-6 The Color Purple  
-50 Northanger Abbey  
-24 East of Eden  
-45 On the Road  
+19 Catch-22  
-105 Around the World in Eighty Days  
-244 Hard Times  
-37 Beloved  
-71 Mrs. Dalloway  
-131 To the Lighthouse  
-14 The Magician's Nephew  
+108 Harry Potter and the Order of the Phoenix  
-29 The Sun Also Rises  
-96 The Good Earth  
-212 Silas Marner  
-15 Love in the Time of Cholera  
+5 Rebecca  
-230 Jude the Obscure  
+129 Twilight  
-215 A Passage to India  
-84 The Plague  
-266 Nicholas Nickleby  
-93 The Pearl  
-155 Ethan Frome  
-339 The Tale of Genji  
+105 The Giver  
+116 The Alchemist  
-146 The Strange Case of Dr. Jekyll and Mr. Hyde  
-52 Robinson Crusoe  
-138 Tender is the Night  
-112 The Idiot  
-22 Hatchet  
+124 The Kite Runner  
+36 One Flew Over the Cuckoo's Nest  
-199 The Portrait of a Lady

+84 The Outsiders  
-272 Ben-Hur  
-222 The Mayor of Casterbridge  
-204 Cry, The Beloved Country  
-53 The Last Battle  
-308 Captains Courageous  
-219 The Castle  
+76 The Metamorphosis  
-237 The Magic Mountain (Der Zauberberg)  
+10 James and the Giant Peach  
-18 The Horse and His Boy  
+140 Angels & Demons  
+12 The Voyage of the Dawn Treader  
+77 The Bell Jar  
-268 Women in Love  
-279 The Yearling  
-223 O Pioneers!  
+125 The Handmaid's Tale  
-165 The Moonstone  
-292 The Old Curiosity Shop  
-229 Little Dorrit  
+14 Prince Caspian: The Return to Narnia  
-237 Sister Carrie  
-26 The Silver Chair  
+171 The Hunger Games  
-183 This Side of Paradise  
-282 Eugénie Grandet  
-206 Of Human Bondage  
-320 Dream of the Red Chamber  
+127 Life of Pi  
+166 Harry Potter and the Deathly Hallows  
-68 Invisible Man  
-70 Steppenwolf  
-104 The Sorrows of Young Werther  
+46 Bridge to Terabithia  
-60 The Invisible Man  
+112 Holes  
+81 Siddhartha  
+37 A Tree Grows in Brooklyn  
-94 Through the Looking-Glass, and What Alice Found There  
+66 In Cold Blood  
-25 The House of the Spirits  
-259 Adam Bede

-280 The Betrothed  
+162 The Book Thief  
+14 Their Eyes Were Watching God  
-106 One Day in the Life of Ivan Denisovich  
-239 The Sea Wolf  
+182 Catching Fire  
-97 Roll of Thunder, Hear My Cry  
-220 Death Comes for the Archbishop  
-123 The House of Mirth  
-174 Light in August  
-237 The Pickwick Papers  
-292 Remembrance of Things Past  
-295 Barchester Towers and the Warden  
-219 The Bridge of San Luis Rey  
+176 The Help  
+80 Murder on the Orient Express  
+172 The Lovely Bones  
-171 The Appeal  
-261 Dombey And Son  
+149 Slaughterhouse-Five  
-209 An American Tragedy  
-9 The Bluest Eye  
+1 Little House In the Big Woods  
-22 Pippi Longstocking  
-201 Germinal  
-89 The Heart Is a Lonely Hunter  
-52 The Woman In White  
-183 Absalom, Absalom!  
-111 A Painted House  
+200 The Girl With the Dragon Tattoo  
-31 A Room With a View  
+76 Watership Down  
+182 Memoirs of a Geisha  
-215 Our Mutual Friend  
-229 Babbitt  
-159 The Red Pony  
-143 All the King's Men  
+59 Things Fall Apart  
-240 Lorna Doone  
-164 Johnny Tremain  
-10 Anne of Avonlea  
+26 Tuck Everlasting  
+88 The BFG

-45 Cannery Row  
+117 The Joy Luck Club  
+37 The Silmarillion  
-30 Roots  
+38 Little House on the Prairie  
-80 Native Son  
-52 Stuart Little  
-181 Cross Fire  
-169 The Power and the Glory  
+130 A Clockwork Orange  
+19 The Phantom of the Opera  
+27 The Martian Chronicles  
+155 The Road  
-239 The Way of All Flesh  
-251 Diary of a Wimpy Kid: The Long Haul  
-108 Villette  
+191 The Curious Incident of the Dog In the Night-Time  
-135 The Mysterious Island  
-50 Song of Solomon  
-198 Nana  
-160 Quo Vadis  
-192 Main Street  
+170 Matilda  
+162 Lolita  
+196 Paper Towns  
-176 Sounder  
+34 Are You There God? It's Me, Margaret  
+212 The Notebook  
+29 From the Mixed-Up Files of Mrs. Basil E. Frankweiler  
+96 Atlas Shrugged  
+81 The Fountainhead  
+134 Number the Stars  
+141 The Firm  
-108 Swann's Way  
+208 Ender's Game  
+98 The Name of the Rose  
+169 A Time to Kill  
+220 Water for Elephants  
+131 The Time Machine  
+226 Eragon  
+231 The Hitchhiker's Guide to the Galaxy  
-161 Buddenbrooks  
+221 A Thousand Splendid Suns

+6 The Witch of Blackbird Pond  
+215 And Then There Were None  
+49 A Separate Peace  
+232 Breaking Dawn  
+20 As I Lay Dying  
+194 The Girl Who Played With Fire  
+121 Where the Red Fern Grows  
-131 Le Morte D'Arthur  
+267 Mockingjay  
+181 The Pillars of the Earth  
-202 Persian Letters  
+136 The Client  
-34 Sula  
+15 Tales of a Fourth Grade Nothing  
-78 The Merry Adventures of Robin Hood of Great Renown In Nottinghamshire  
-91 Tortilla Flat  
-179 Look Homeward, Angel  
-185 The Mystery of Edwin Drood  
-6 Brideshead Revisited  
+138 The Pelican Brief  
+157 Atonement  
-157 Washington Square  
+129 Like Water for Chocolate  
+246 The Golden Compass  
+236 The Secret Life of Bees  
+297 The Fault In Our Stars  
-164 Nostromo  
-173 Finnegans Wake  
-22 The Brethren  
+189 Coraline  
+165 Heart of Darkness  
-8 On the Banks of Plum Creek  
-115 Rebecca of Sunnybrook Farm  
-168 The Ambassadors  
-146 The Secret Agent  
+66 The House on Mango Street  
-51 Go Tell It on the Mountain  
+18 The Testament  
+102 The Clan of the Cave Bear  
-87 Cranford  
+98 Because of Winn-Dixie  
-33 My Side of the Mountain  
+125 The Runaway Jury



-23 The Mouse and the Motorcycle  
+193 The Lost Symbol  
-141 The Forsyte Saga  
+301 Gone Girl  
+300 The Lightning Thief  
-170 The Last Days of Pompeii  
+92 The Reader  
-63 Caddie Woodlawn  
+88 The Tale of Despereaux  
+220 The Girl Who Kicked the Hornet's Nest  
-76 Dear Mr. Henshaw  
-10 The Killer Angels  
+88 Chronicle of a Death Foretold  
+222 The Five People You Meet In Heaven  
+160 The Master and Margarita  
-90 Winesburg, Ohio  
-107 P Is for Peril  
+268 My Sister's Keeper  
-143 Barnaby Rudge  
+4 Howards End  
+14 The Broker  
+8 The Camel Club  
-120 The Rainbow  
-23 The Man In the Iron Mask  
+62 Mary Poppins  
+210 Artemis Fowl  
+216 Dear John  
+123 Cold Mountain  
+228 Flowers for Algernon  
-31 The Dark Is Rising  
-102 Resurrection  
+22 Fearless Fourteen  
-139 A Sentimental Journey Through France and Italy  
+11 The King of Torts  
+216 The Graveyard Book  
-16 The Quiet American  
+82 The Chamber  
+74 The English Patient  
+110 Snow Falling on Cedars  
+21 The Long Winter  
+20 Sarah, Plain and Tall  
-44 Cross Country  
+56 The Spy Who Came In from the Cold

+331 A Game of Thrones  
+189 The Thorn Birds  
+45 Old Yeller  
+7 Ramona Quimby, Age 8  
-15 Death In Venice  
+19 By the Shores of Silver Lake  
+235 Inferno  
+104 Schindler's List  
+151 Jonathan Livingston Seagull  
+266 The Stand  
+55 The Last Juror  
+30 Shiloh  
+267 Girl With a Pearl Earring  
+167 The Murder of Roger Ackroyd  
+300 It  
+136 The Rainmaker  
+272 The Poisonwood Bible  
+68 The Indian in the Cupboard  
+71 The Maltese Falcon  
-84 The Warden  
+35 The Summons  
-26 Encyclopedia Brown: Boy Detective  
+339 The Time Traveler's Wife  
-5 The Incredible Journey  
+103 Daughter of Fortune  
-38 Shirley  
+85 Bud, Not Buddy  
+12 The Horse Whisperer  
+93 The Street Lawyer  
+95 Nausea  
-36 To Have and Have Not  
+70 The Bridges of Madison County  
+136 Anne of the Island  
0 The Winter of Our Discontent  
+339 The Shining  
+99 The Tenant of Wildfell Hall  
-3 First Family  
+111 The Partner  
+376 The Girl on the Train  
-62 The Black Arrow: A Tale of the Two Roses  
-90 The Rise of Silas Lapham  
+153 The Choice  
-82 The Virginian: A Horseman of the Plains

+307 A Walk to Remember  
+350 The Maze Runner  
+176 The Westing Game  
+11 Misty of Chincoteague  
+142 Diary of a Wimpy Kid: The Last Straw  
+19 King Solomon's Mines  
-56 The Princess of Cleves  
-14 Jacob Have I Loved  
+158 Mrs. Frisby and the Rats of NIMH  
+300 Misery  
+167 The Cider House Rules  
-28 King of the Wind  
+109 The Once and Future King  
+254 The Witches  
+264 The Subtle Knife  
+118 When You Reach Me  
+310 Carrie  
-30 The Moon and Sixpence  
-51 The Higher Power of Lucky  
-65 Looking Backward, 2000-1887  
-39 The Wings of the Dove  
-55 The Summer of the Swans  
+40 Dangerous Liaisons  
+346 Jurassic Park  
+219 The Absolutely True Diary of a Part-time Indian  
+19 The Grey King  
+13 The Leopard  
+75 The Mammoth Hunters  
+84 The Trumpet of the Swan  
+263 The Lucky One  
+82 These Happy Golden Years  
-51 Arrowsmith  
+62 Julie of the Wolves  
+286 The Screwtape Letters  
+127 The Fall  
+226 The No. 1 Ladies' Detective Agency  
+5 Worst Case  
-15 Lost Horizon  
+317 The Gunslinger  
-38 The Slave Dancer  
+429 Harry Potter and the Half-Blood Prince  
+287 Inkheart  
+16 Ramona and her Father

+159 Inkspell  
+85 Ramona the Pest  
+189 Walk Two Moons  
+384 Miss Peregrine's Home for Peculiar Children  
+54 The Chocolate War  
+120 Sophie's Choice  
+403 Looking for Alaska  
+240 Breakfast at Tiffany's  
+62 The Razor's Edge  
+201 Dreamcatcher  
+127 Orlando  
+270 The Things they Carried  
+125 Little Town on the Prairie  
+202 Nights in Rodanthe  
+290 The Amber Spyglass  
+157 The Miraculous Journey of Edward Tulane  
+103 Flatland  
+350 Diary of a Wimpy Kid  
+338 The Memory Keeper's Daughter  
+203 The Wedding  
+278 Fried Green Tomatoes at the Whistle-Stop Cafe  
+103 The Cricket in Times Square  
+270 The Phantom Tollbooth  
-13 Rob Roy  
+209 The Death of Ivan Ilych  
+34 Alex Cross's Trial  
-22 Kenilworth  
+16 The Life and Opinions of Tristram Shandy  
+282 The Remains of the Day  
-14 M.C. Higgins, The Great  
+5 Call It Courage  
+272 Go Set a Watchman  
+77 Bleachers  
+9 Elijah of Buxton  
+37 Swimsuit  
+321 Cat's Cradle  
+35 The Caine Mutiny  
+45 The Heart of the Matter  
+170 Harriet, the Spy  
+55 Darkness at Noon  
+302 A Prayer for Owen Meany  
+294 The God of Small Things  
+130 The Associate

- +369 The Shack
- +45 The Naked and the Dead
- +419 The Sea of Monsters
- +306 Stranger in a Strange Land
- +220 Vision in White
- +53 The Whipping Boy
- +398 Room
- +378 Deception Point

You can find more metadata analysis [here](#).

## USING FULL TEXT DATA

In addition to the contextual information we gathered, we also collected the full text of all novels on the list that were out of copyright and available on Project Gutenberg. We have provided some ideas for analysis in this [Colab notebook](#), but we hope this full-text data will also offer opportunities for users to explore these novels on their own and to combine full-text and metadata analysis in new ways.

## CONCLUSION

The Top 500 List is presented in a straightforward manner. It is just a list of 500 novels that are widely held in library collections along with their authors. But when you start to dig into the data underlying the list, it gets much, much more complicated.

The list draws on hundreds of millions of library records representing billions of library holdings. This is such a vast amount of information that it may appear to provide opportunities to draw comprehensive conclusions. However, the data overwhelmingly represents the holdings of libraries in the U.S.A., the majority of which are also connected to some sort of educational institution. Though it claims to represent great novels from around the world, the list primarily includes English-language novels and novels popular in English translation.

The list also represents the disproportionate influence of academics and publishers, who chose to re-edit and re-issue certain texts and not others. The correlation we found between number of editions and number of holdings is likely to make intuitive sense to library users—especially users of academic libraries, which tend to hold many editions of classic texts, and which often continue to purchase these texts as they are re-edited and re-issued. Histories of canonization in the U.S. and Europe have long been biased toward works by White, male, middle and upper class authors—a fact which clearly influenced the composition of the list.

In pointing out these biases we do not intend to criticize OCLC for producing the list, which provides a useful snapshot of some of the most widely held works in their database and represents a tremendous data curation and analysis effort. We do, however, want to question the

notion that “literary greatness” can be measured by the number of physical copies of a book held on library shelves. It is important to dig into data that is used to make universal claims, especially when it evidences such strong biases toward a single linguistic tradition, toward particular geographic regions, and toward individual authors. John Grisham’s work appears nineteen times on this list, Charles Dickens’s work appears fifteen times, and John Steinbeck and C.S. Lewis’s work each appears eight times. Do we really want to posit that these four men wrote ten percent of the greatest novels across all languages and cultures across all time?

While each of these works deserves individual attention, looking at literary data in aggregate can help to reveal some of these biases and trends across a larger number of texts, and across library collections. We hope this dataset provides fruitful opportunities for exploration, and we have included a few more suggestions for analysis here.

## Explore the Data

```
//| echo: false
//| output: false
library_data = d3.csv("https://raw.githubusercontent.com/melaniewalsh/responsible-datasets-in-computer-science/master/data/library_data.csv")

use_data = d3.csv("https://raw.githubusercontent.com/melaniewalsh/responsible-datasets-in-computer-science/master/data/use_data.csv")
```

```
//| echo: false
//| output: false

filtered = library_data.filter(function(penguin) {
  return bill_length_min < penguin.bill_length_mm &&
    islands.includes(penguin.island);
})
```

```
//| echo: false
color = d3
  .scaleLinear()
  .domain([0, 100, 300])
  .range(["#cafcc2", "#fce7c2", "#eb9494"])
```

## Library Top 500

```

//| echo: false
viewof search = Inputs.search(library_data, {
  placeholder: "Search"
})

```

```

//| echo: false

/*Inputs.table(search, data)*/

Inputs.table(search, {
  layout: "fixed",
  rows: 50,
  sort: "top_500_rank",
  reverse: false,
  format: {
    /*RecreationVisits: x => d3.format('.2s')(x),*/
    pub_year: x => d3.timeFormat(x),
    author_birth: x => d3.timeFormat(x),
    author_death: x => d3.timeFormat(x),
    gr_num_ratings_rank: x => html`<div style='background:${color(x)}'>${d3.format('.2s')(x)}
  }
})

```

## Exercises

### R

### Python

## Discussion & Activities

### Activity 1

It is inevitable that the devices that the National Park Service uses to count visits to the parks — like induction loop counters installed on the road — will break. But they will also get *fixed* at different rates, in different locations, as we could see in the case of Crater Lake National Park (where a counter was fixed quickly) and Carlsbad Caverns National Park (where a broken counter from 2019 still has not been fixed).

There are many reasons for these disparities, but some of the big ones might be geography and resources. The more remote a park, the harder it is to get a repair team to it. The less-resourced a park, the lower the likelihood they have on-site repair teams, or are prioritized by the repair teams that can be dispatched.

With this in mind, look at the locations of the following parks. Suppose that each one has an outage in their induction loop counter: which ones would you expect to be fixed first, and why? Research the parks, and rank them on a scale of 1 to 5 (1 being highest, and 5 being lowest) of which would be fixed quickest.

Park	Priority (1-5)	Reason
Acadia NP		
Lassen Volcanic NP		
Saguaro NP		
Yosemite NP		
Mammoth Cave NP		

## Activity 2

The National Park Service sometimes fills in missing data with hard numbers or approximates data by applying special mathematical formulas. This is necessary work, but it is also under-explained work.

To see this in action, go to [the NPS page that documents park reports](#) and down the “Visitor Use Counting Procedures” PDF for three different parks.

How are the procedures for these three parks similar or different? What kind of effect do you think this has on the resulting data? What do you think is the best of documenting this information and communicating it to users of the data?

## Activity 3

In 2014 and 2015, Kobuk Valley National Park reported that there were zero visitors to the park.

Use publicly available internet data - Twitter posts, Flickr photos, etc - to try and find evidence of people visiting the park (there is existing evidence!).

Based on your findings, how do you think, differently, if at all, about Kobuk Valley’s decision to record zero visits and about alternative methods for counting visits?