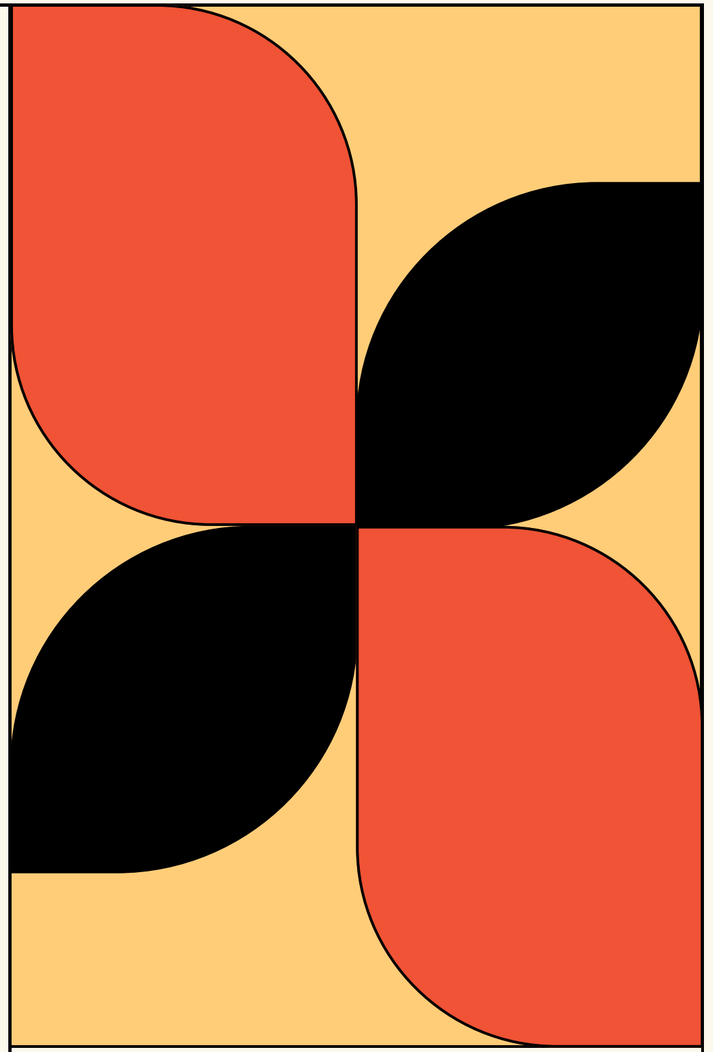
STANFORD
UNIVERSITY

Can You Judge a Book by Its Cover?



Final Project by: Serena Lee and Sunny Yu

DATASCI 112 - WINTER 2023

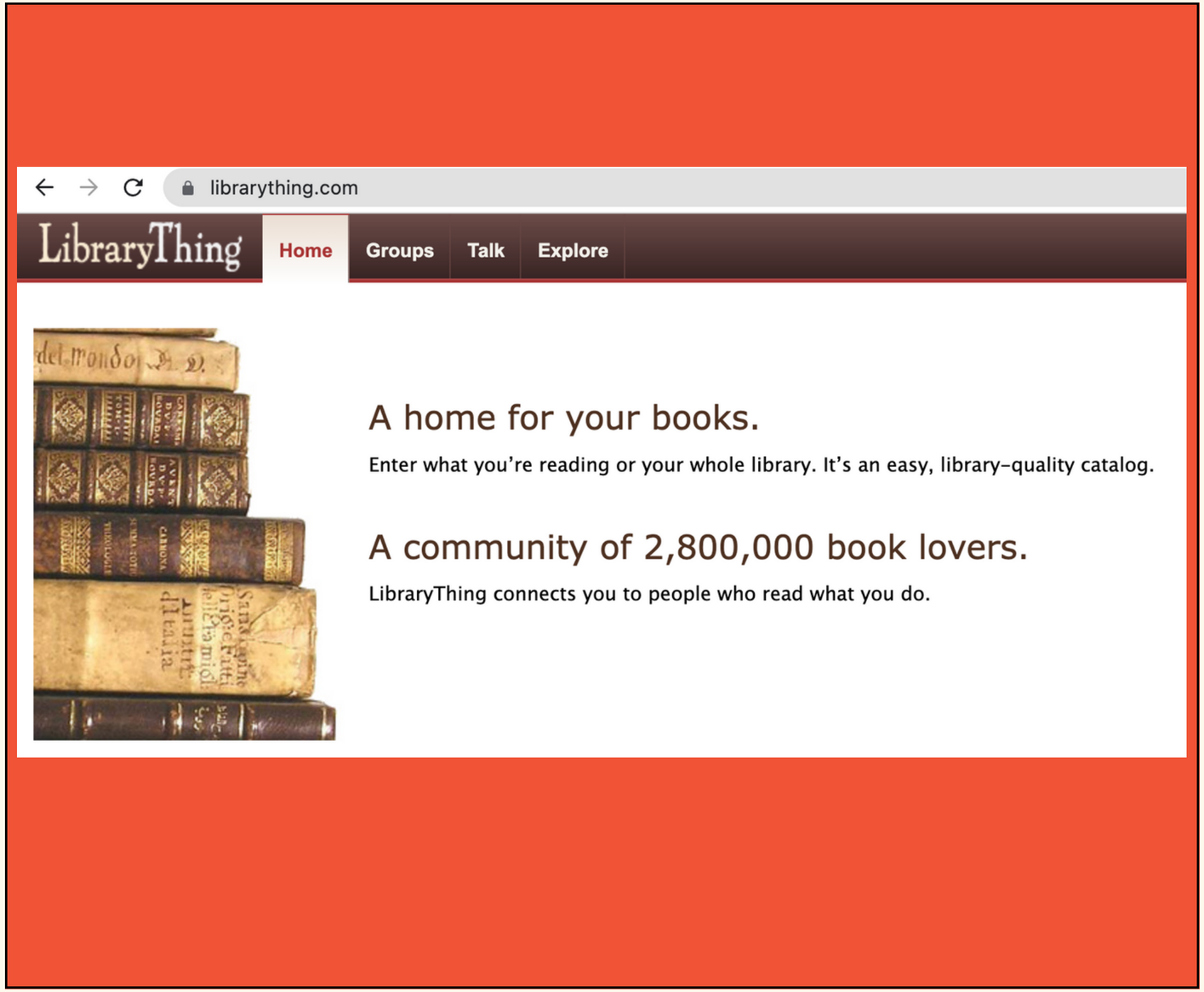
Agenda

1. Project Goals
2. Data Collection & Cleaning
3. Data Exploration
4. Machine Learning

Aim 1:
train a regression model to predict a book's rating based on surface-level features

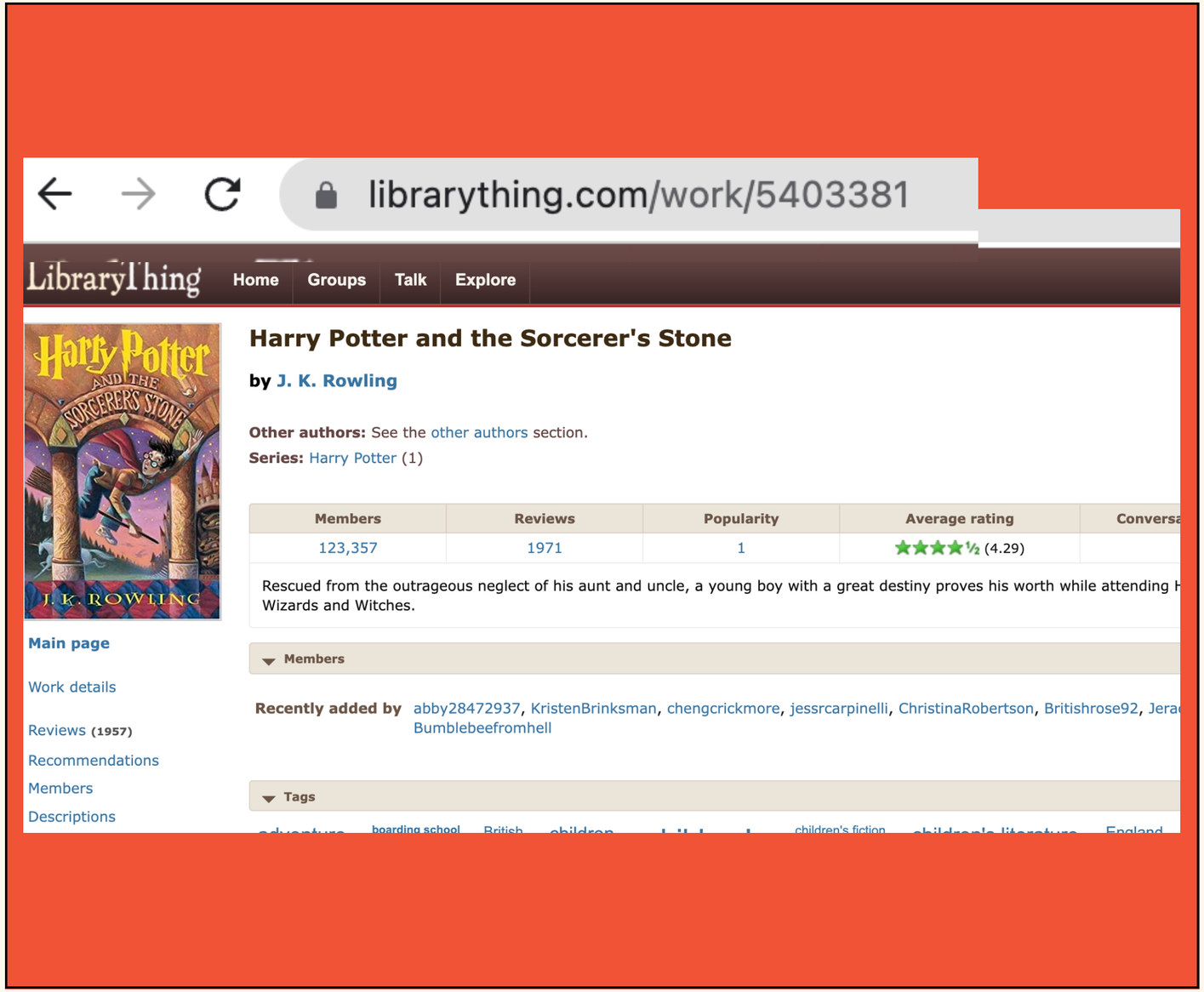
- Book cover
- Blurb
- First words
- Genre
- etc...

Aim 2:
train a clustering model to offer book recommendations



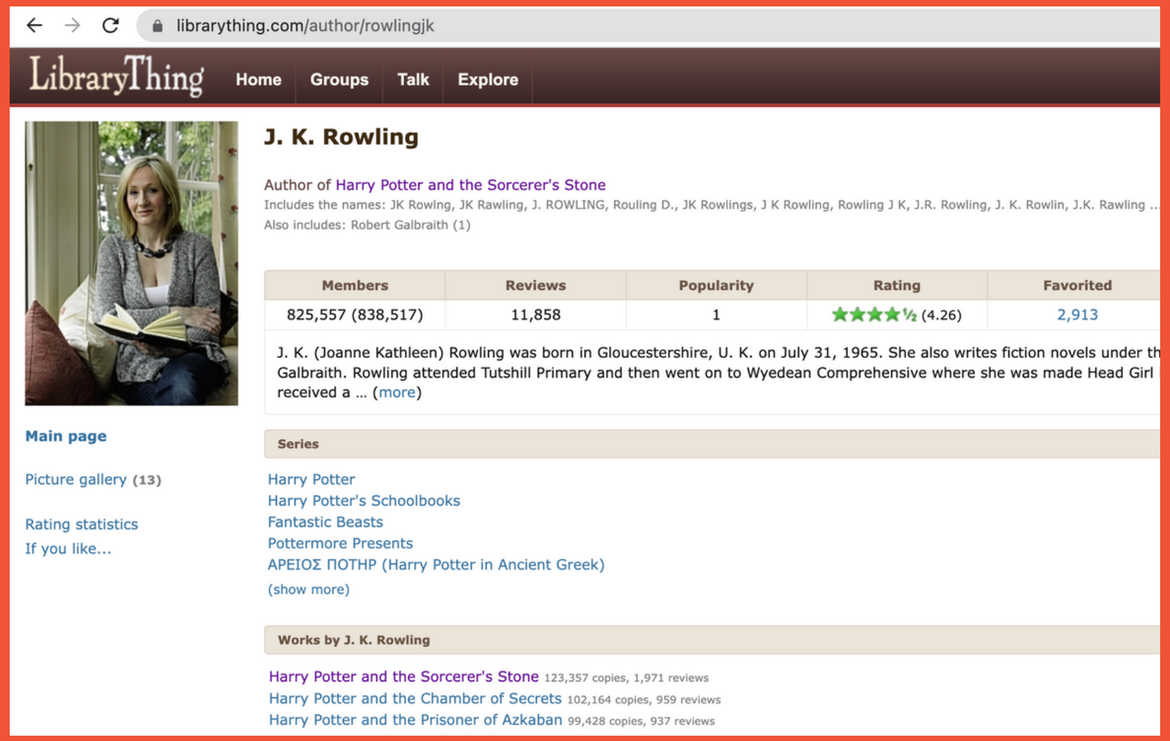
DATA COLLECTION

- **Primary source: LibraryThing**
- Web scraping (x4)
  - Author names (Wiki, Guardian, LibraryThing)
  - Book page URLs
  - Book metadata
  - Book covers
- Merged multiple dataframes



DATA COLLECTION

- Primary source: LibraryThing
- **Web scraping (x4)**
  - Author names (Wiki, Guardian, LibraryThing)
  - Book page URLs
  - Book metadata
  - Book covers
- Merged mutliple dataframes

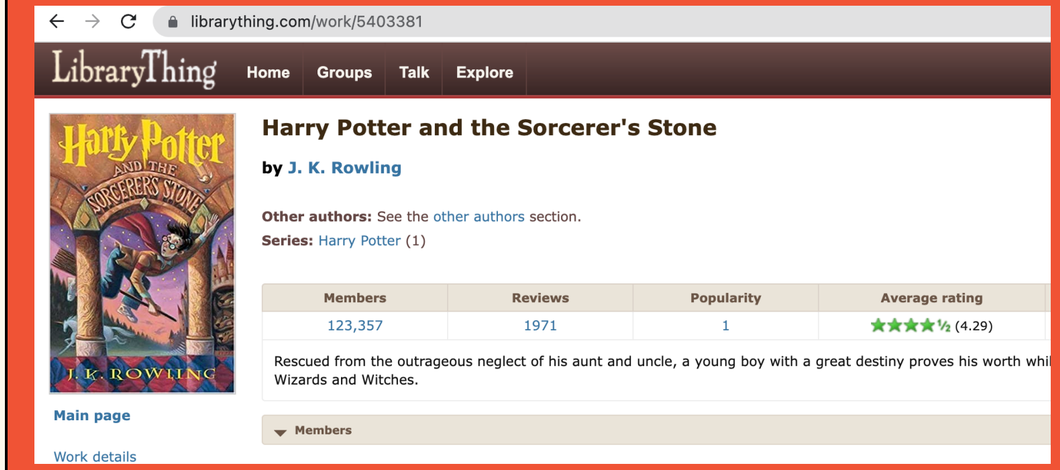


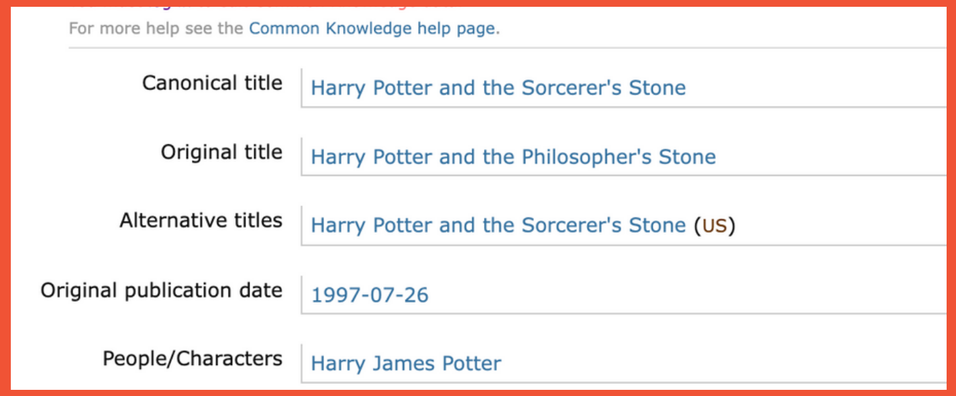
step 1: scrape 1613 author names from Wikipedia, The Guardian, and LibraryThing to get "author_names"

step 2: append author names to LibraryThing dir, loop through the author URLs to scrape their books to get "book_urls"

DATA COLLECTION

- Primary source: LibraryThing
- Web scraping (x4)
  - **Author names (Wiki, Guardian, LibaryThing)**
  - **Book page URLs**
  - Book metadata
  - Book covers
- Merged multiple dataframes





step 3: loop through book_urls to turn book metadata into a data frame

DATA COLLECTION

- Primary source: LibraryThing
- Web scraping (x4)
  - Author names (Wiki, Guardian, LibaryThing)
  - Book page URLs
  - **Book metadata**
  - **Book covers**
- Merged multiple dataframes

title	author	book_index	book_url	avg_rating	genre	description	publication_year	first_sentence	number_of_reviews
And Then There Were None	Agatha Christie	0	https://www.librarything.com//work/7962202	4.14	[Fiction and Literature, Mystery]	Ten houseguests, trapped on an isolated island...	1939	In the corner of a first-class smoking carriag...	500
Murder on the Orient Express	Agatha Christie	1	https://www.librarything.com//work/2742	4.07	[Fiction and Literature, Mystery]	Agatha Christie's most famous murder mystery, ...	1934	It was five o'clock on a winter's morning in S...	394
The Murder of Roger Ackroyd	Agatha Christie	2	https://www.librarything.com//work/3011	4.06	[Fiction and Literature, Mystery]	Agatha Christie's most daring crime mystery -	1926	Mrs Ferrars died on the night of the 16th-17th...	291

step 4: after looping through 119,215 URLs, we collected all the book data in separate runtimes, so we concatenated multiple data frames. We also merged the data with additional info we got from Kaggle (author identities).

step 5: scraped, read, resize image data from image URLs into a 1-D array for ML

DATA EXPLORATION

popularity of genres, authors, books over time

author information

common first words of a book and its trend

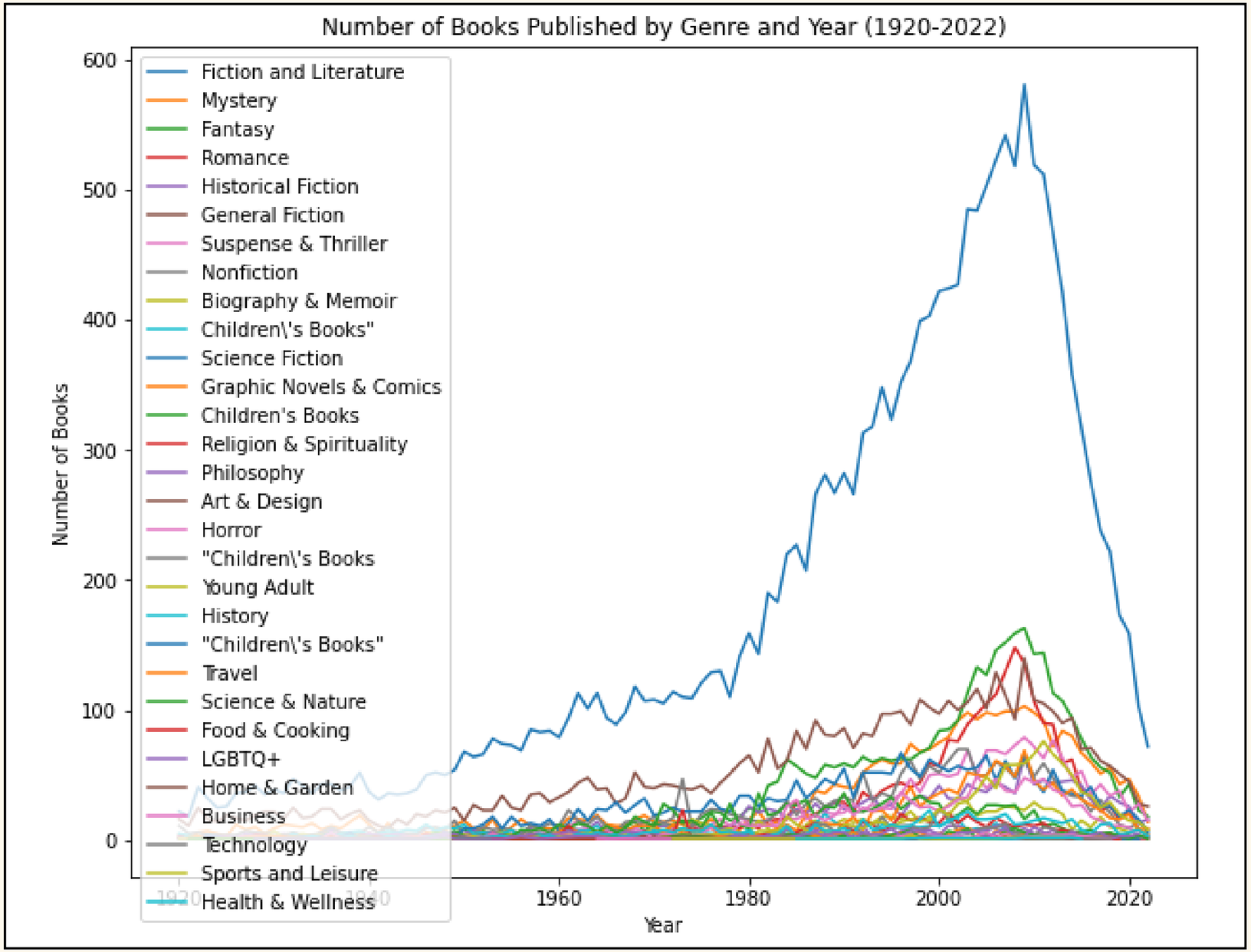
Our data comprised **24,971 books** published between **120AD and 2023** and spanned authors from **67 countries**. We wanted to explore trends across time, country, and genre.

Our exploration led us to these questions:

1. Where are the authors in our dataset from? Which countries produced the most best-selling authors?
2. Which genres were trending at what times?
3. What were the most common first words in books with ratings above 4.0+?
4. What are the common first word/words that begin a book, and how do they change over time?

DATA EXPLORATION

POPULARITY OF GENRES



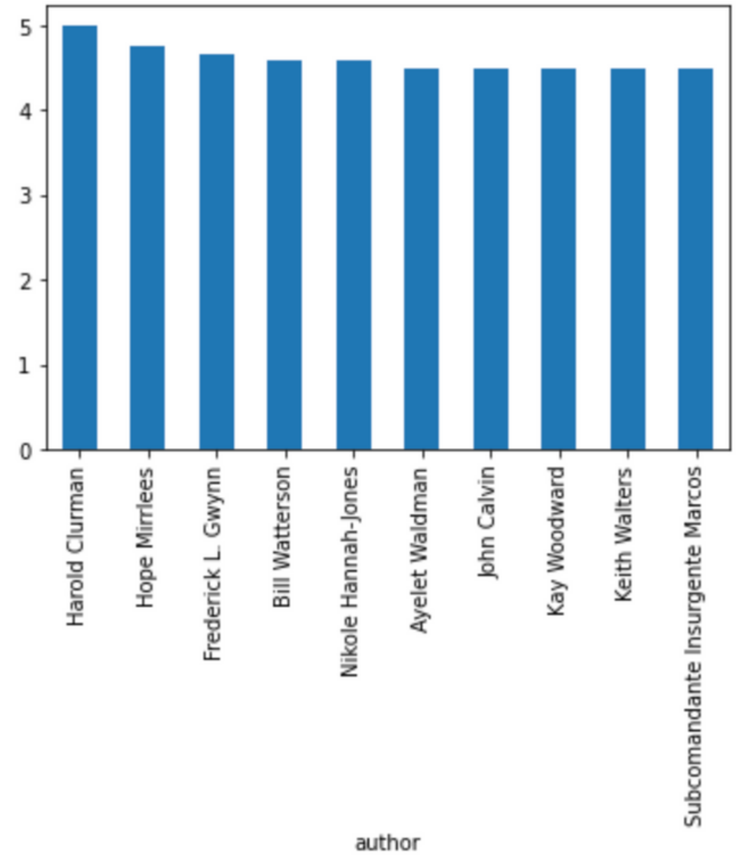
DATA EXPLORATION

POPULARITY OF AUTHORS

```
df_books = df_books[df_books['number_of_reviews'] != "None"]
df_books.loc['number_of_reviews'] = df_books['number_of_reviews'].astype(int)
df_filter = df_books[df_books['number_of_reviews']>100]
df_books.groupby('author')['avg_rating'].mean().nlargest(10).plot.bar()
```

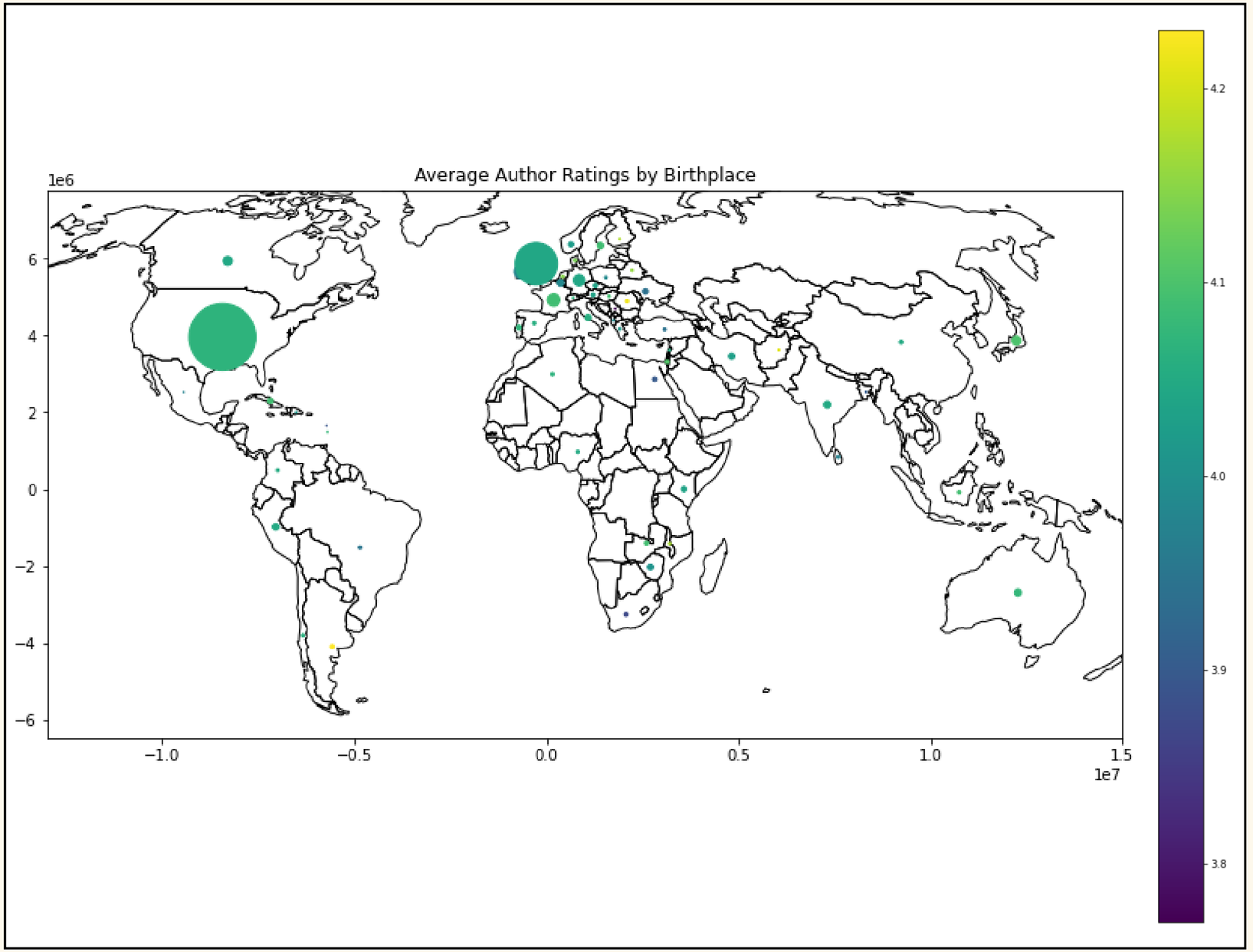
1s

```
<Axes: xlabel='author'>
```



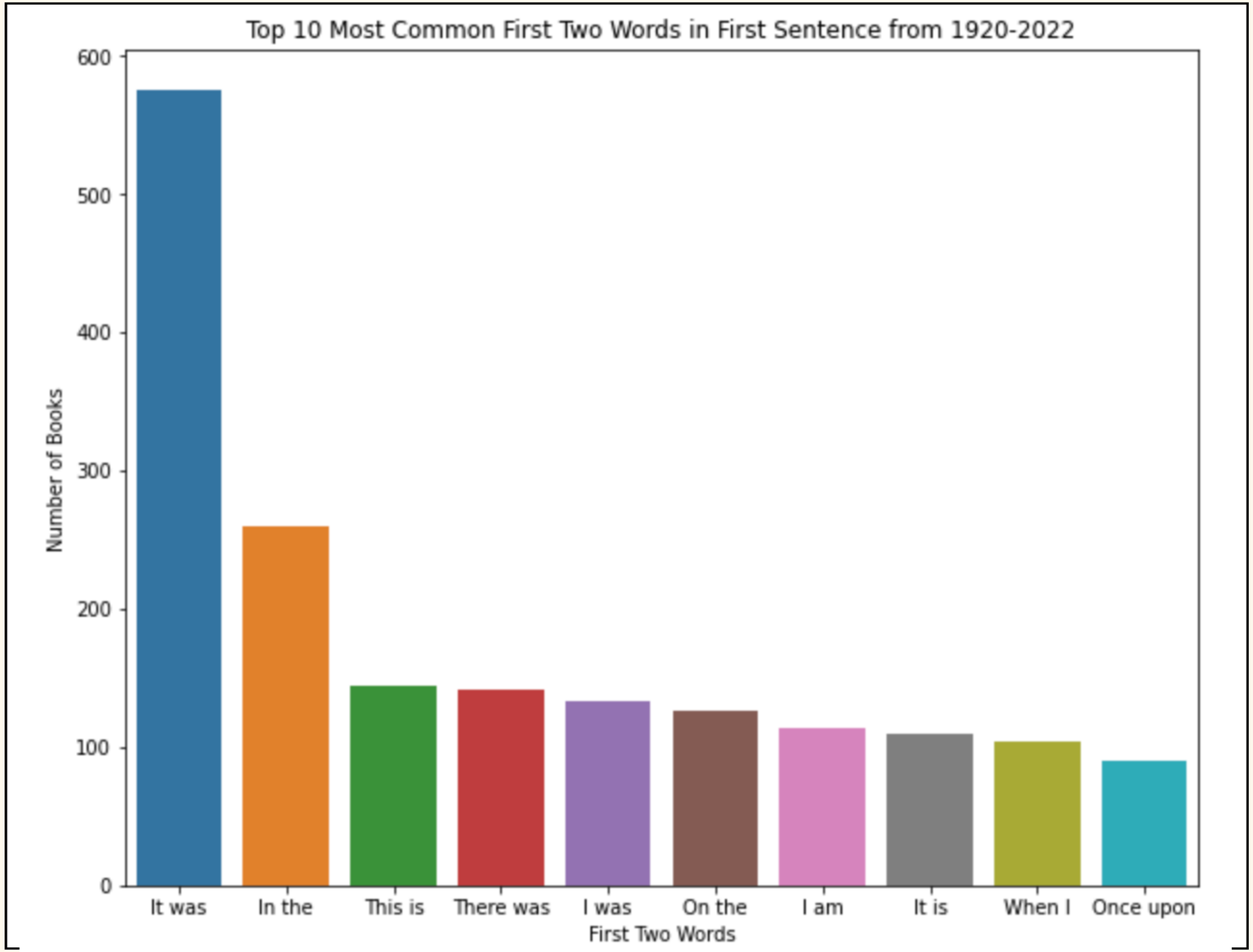
DATA EXPLORATION

AUTHOR BIRTHPLACE



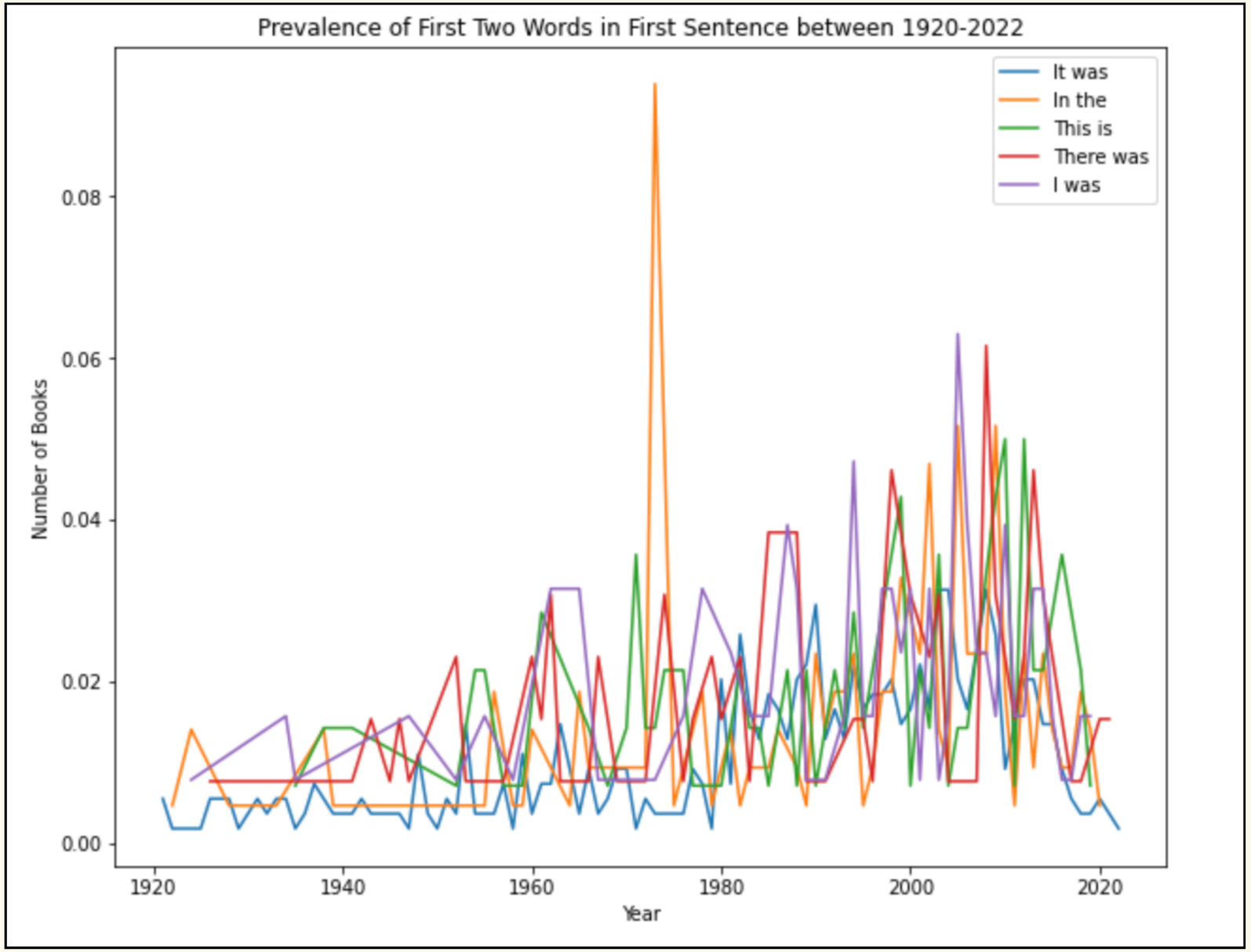
DATA EXPLORATION

FIRST TWO WORDS: FREQUENCY



DATA EXPLORATION

FIRST TWO WORDS: TRENDS



Machine Learning

1. **Predicting Book Ratings (regression)**

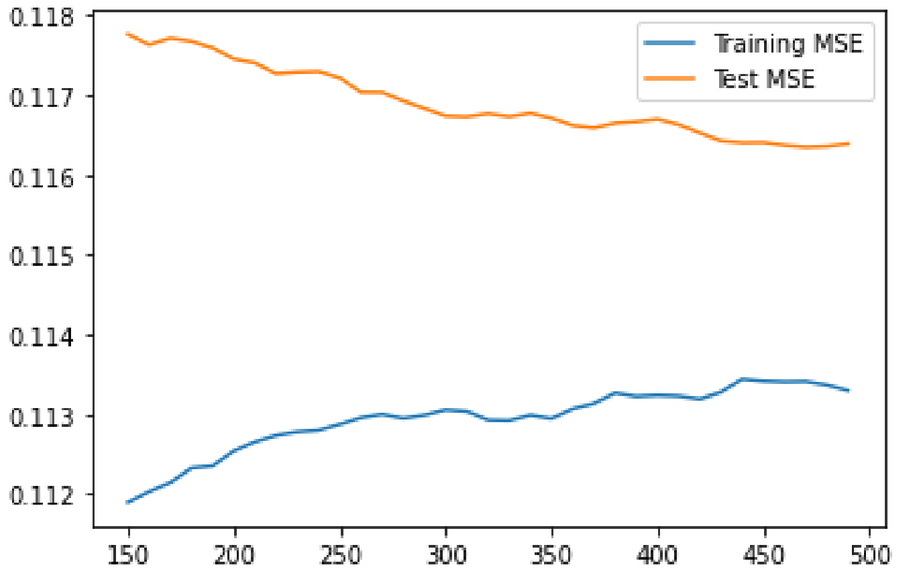
2. Clustering similar books

Machine Learning

1. **Predicting Book Ratings (regression)**

Using book covers

- KNN model of RGB pixel arrays




```
euclidean 0.11302739879961377
manhattan 0.11291609145348043
```

- 64x64 images with x3 (RGB) values
- 12288 columns (+2 for book title & rating)

```
df_images.shape

(23975, 12290)
```

Machine Learning

COMPARISON:
Different sets of **non-image** features

Best model: 0.0789

- *number of reviews*
- *description*
- *title*

1

ColumnTransformer
standardscaler
['number_of_reviews']
StandardScaler

2

ColumnTransformer
standardscaler | tfidfvectorizer-1 | tfidfvectorizer-2
['number_of_reviews'] | description | title
StandardScaler | TfidfVectorizer | TfidfVectorizer

3

ColumnTransformer
onehotencoder | tfidfvectorizer
['author'] | title
OneHotEncoder | TfidfVectorizer

4

ColumnTransformer
onehotencoder | tfidfvectorizer-1 | tfidfvectorizer-2 | tfidfvectorizer-3
['author'] | first_sentence | description | title
OneHotEncoder | TfidfVectorizer | TfidfVectorizer | TfidfVectorizer

```
Error of model #1: 0.0831082605045265 1
Error of model #2: 0.0789037170645755
Error of model #3: 0.0844344592238364 8
Error of model #4: 0.0812555221082089 7
```

Machine Learning

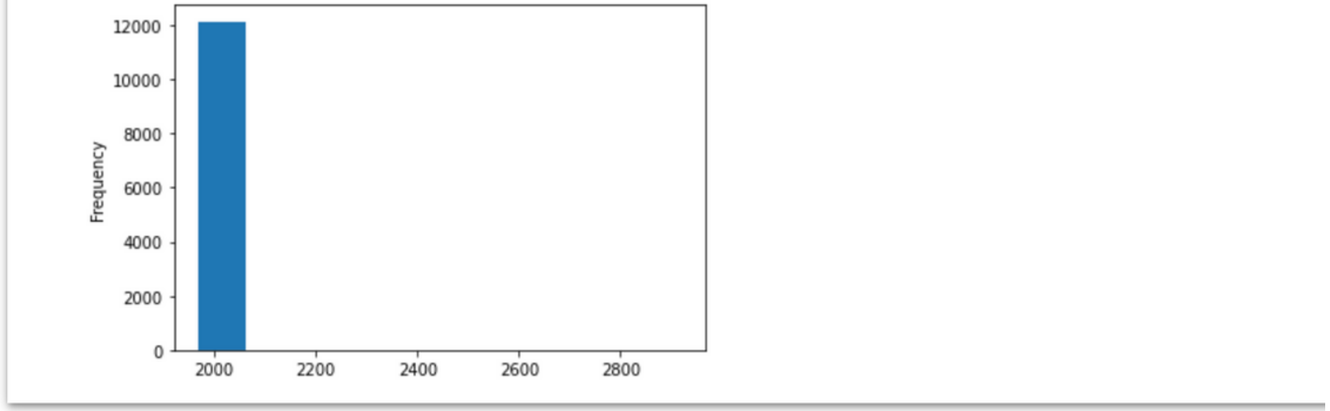
2. Clustering similar books

Finding clusters

- K-means
- Hierarchical

cluster 0: modern fantasy books/fiction

```
X_train[model.labels_== 0]['publication_year'].plot.hist()
```




```
[69] X_train[model.labels_== 0][['author','genre']].value_counts()
```

author	genre	
Clive Cussler	['Fiction and Literature', 'Suspense & Thriller']	84
Stan Berenstain	["Children's Books"]	84
Margaret Weis	['Fiction and Literature', 'Fantasy']	84
M. C. Beaton	['Fiction and Literature', 'Mystery']	48
Robert B. Parker	['Fiction and Literature', 'Mystery']	47
		..
Jennifer Weiner	['Biography & Memoir']	1
	['Fiction and Literature', 'General Fiction', 'Romance', 'Mystery']	1
	['Fiction and Literature', 'Romance', 'General Fiction']	1
	['Fiction and Literature']	1
Åsne Seierstad	['Nonfiction', 'History', 'Biography & Memoir']	1

Length: 3716, dtype: int64

Machine Learning

2. Clustering similar books

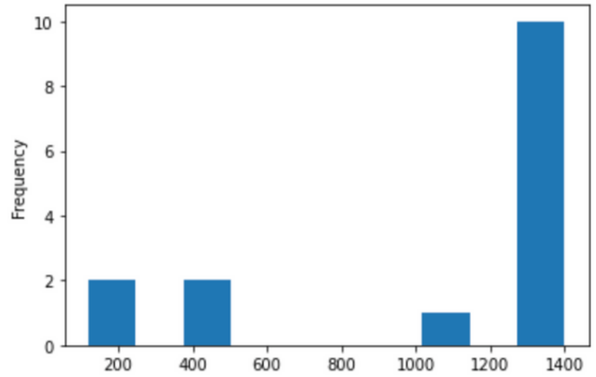
Finding clusters

- K-means
- Hierarchical

cluster 3: ancient and early modern poets

```
X_train[model.labels_== 3]['publication_year'].plot.hist()
```

```
<Axes: ylabel='Frequency'>
```




```
[75] X_train[model.labels_== 3][['author','genre']].value_counts()
```

```
author            genre
Dante Alighieri   ['Fiction and Literature']                                                                  5
Geoffrey Chaucer  ['Fiction and Literature']                                                                  4
Apuleius          ['Fiction and Literature', 'General Fiction', 'Fantasy']                                   1
Geoffrey Chaucer  ['Fiction and Literature', 'Romance']                                                       1
Omar Khayyam      ['Fiction and Literature']                                                                  1
Saint Augustine   ['Philosophy', 'Nonfiction', 'Religion & Spirituality', 'History', 'Biography & Memoir']  1
                  ['Philosophy', 'Nonfiction', 'Religion & Spirituality']                                     1
Suetonius         ['Nonfiction', 'History', 'Biography & Memoir']                                             1
dtype: int64
```

Machine Learning

2. Clustering similar books

Using clusters

- Recommend a list of similar books

Input: Any Book Name

Output: 10 Recommended Books!

```
book_recommender('Madame Bovary')
```

	title	author	book_index
2825	Willy's Pictures	Anthony Browne	2825
1927	Notes on Grief	Chimamanda Ngozi Adichie	1927
4818	Der König verneigt sich und tötet	Herta Müller	4818
9534	The Three Musketeers, Vol 1 (of 2)	Alexandre Dumas, pere	9534
4116	Madame Bovary	Gustave Flaubert	4116
2191	Invisible	Paul Auster	2191
2423	The Shut Eye	Belinda Bauer	2423
3188	The Wisdom of Father Brown	G. K. Chesterton	3188
5085	Time must have a stop	Aldous Huxley	5085
2687	Stars and Bars	William Boyd	2687

Customizable recommender functions

1. similar authors

2. similar descriptions (TF-IDF)

3. similar writing style

4. holistic recommendation (all features)


Simple
Aida Sohm
WOMEN TALKING
A NOVEL
MIRIAM TOEWS
The Book of
ICHIGO ICHIE
The Art of Making the Most of Every Moment, the Japanese Way
SUSAN CHOI
NORMAL PEOPLE
Kiss
LEN HOANG

Thank you!

Book Recommender: K-Means Clustering

Function

```python
#function for book recommendation
def get_similar_books(title):
    book_index = df_books[df_books['title'].str.strip() == title].index[0]
    cluster = clusters[book_index]
    books_in_cluster = df_books[clusters == cluster]
    return books_in_cluster.sample(10)

get_similar_books('The Great Gatsby')
```

	title	author	book_
21734	Imaginary Numbers	Seanan McGuire	
12525	The Butlerian Jihad	Brian Herbert	
14399	Perfume: The Story of a Murderer	Patrick Süskind	
6319	Fall Down 7 Times Get Up 8: A Young Man's Voic...	Naoki Higashida	
24005	The Black Lyon	Jude Deveraux	
18421	Soul Harvest: The World Takes Sides	Tim LaHaye	
24601	Fever	Robin Cook	
23861	The Temptation of Forgiveness	Donna Leon	
12129	Fangirl: The Manga, Vol. 1	Rainbow Rowell	
14973	Deception Point	Dan Brown	

Function Input: Any Book Name!

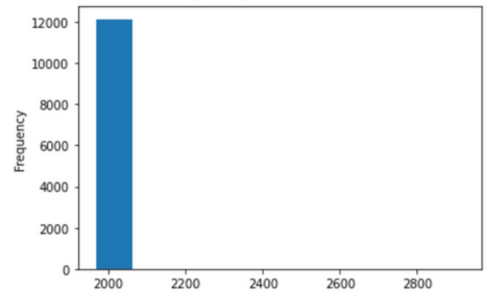
Function Output: A DataFrame With 10 random recommended books from the same cluster

Book Recommender: K-Means Clustering

cluster 1: 19th century classic literature
cluster 2: mystery books

```
X_train[model.labels_== 0]['publication_year'].plot.hist()
```

```
<Axes: ylabel='Frequency'>
```




```
[69] X_train[model.labels_== 0][['author','genre']].value_counts()
```

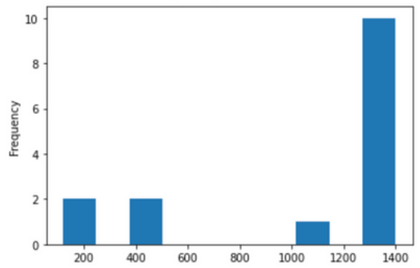
author	genre	
Clive Cussler	['Fiction and Literature', 'Suspense & Thriller']	84
Stan Berenstain	["Children's Books"]	84
Margaret Weis	['Fiction and Literature', 'Fantasy']	84
M. C. Beaton	['Fiction and Literature', 'Mystery']	48
Robert B. Parker	['Fiction and Literature', 'Mystery']	47
		..
Jennifer Weiner	['Biography & Memoir']	1
	['Fiction and Literature', 'General Fiction', 'Romance', 'Mystery']	1
	['Fiction and Literature', 'Romance', 'General Fiction']	1
	['Fiction and Literature']	1
Åsne Seierstad	['Nonfiction', 'History', 'Biography & Memoir']	1

Length: 3716, dtype: int64

cluster 3: ancient Green literature

```
X_train[model.labels_== 3]['publication_year'].plot.hist()
```

```
<Axes: ylabel='Frequency'>
```




```
[75] X_train[model.labels_== 3][['author','genre']].value_counts()
```

author	genre	
Dante Alighieri	['Fiction and Literature']	5
Geoffrey Chaucer	['Fiction and Literature']	4
Apuleius	['Fiction and Literature', 'General Fiction', 'Fantasy']	1
Geoffrey Chaucer	['Fiction and Literature', 'Romance']	1
Omar Khayyam	['Fiction and Literature']	1
Saint Augustine	['Philosophy', 'Nonfiction', 'Religion & Spirituality', 'History', 'Biography & Memoir']	1
	['Philosophy', 'Nonfiction', 'Religion & Spirituality']	1
Suetonius	['Nonfiction', 'History', 'Biography & Memoir']	1

dtype: int64

cluster 0: modern fantasy books/fiction

cluster 4: children's books