

BookRec

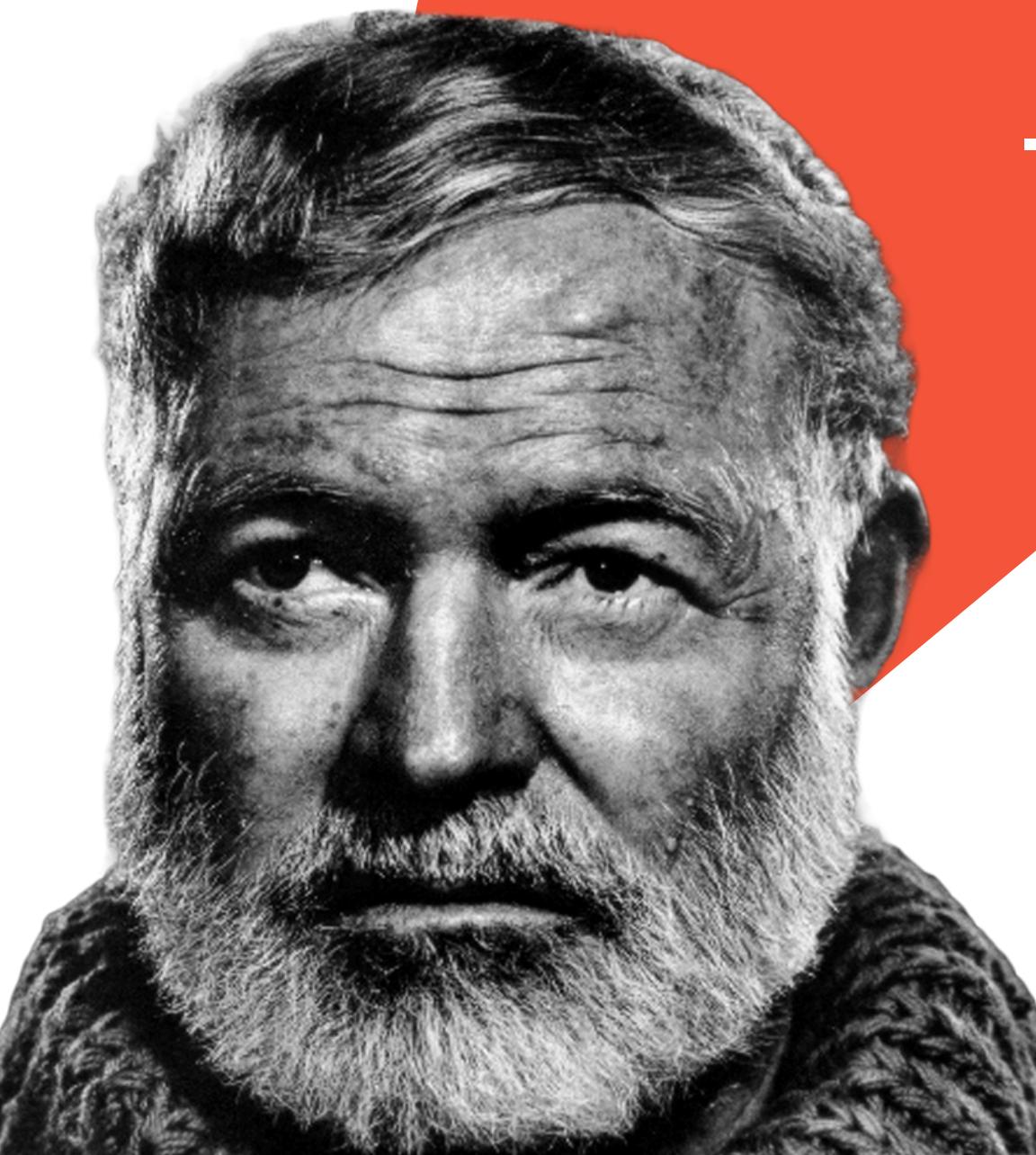
Mike Pearce



”

**“There is no friend as
loyal as a book.”**

— Ernest Hemingway



“Write drunk. Edit sober.”

— Also (maybe) Hemingway



Table Of Contents



**Project
Overview**



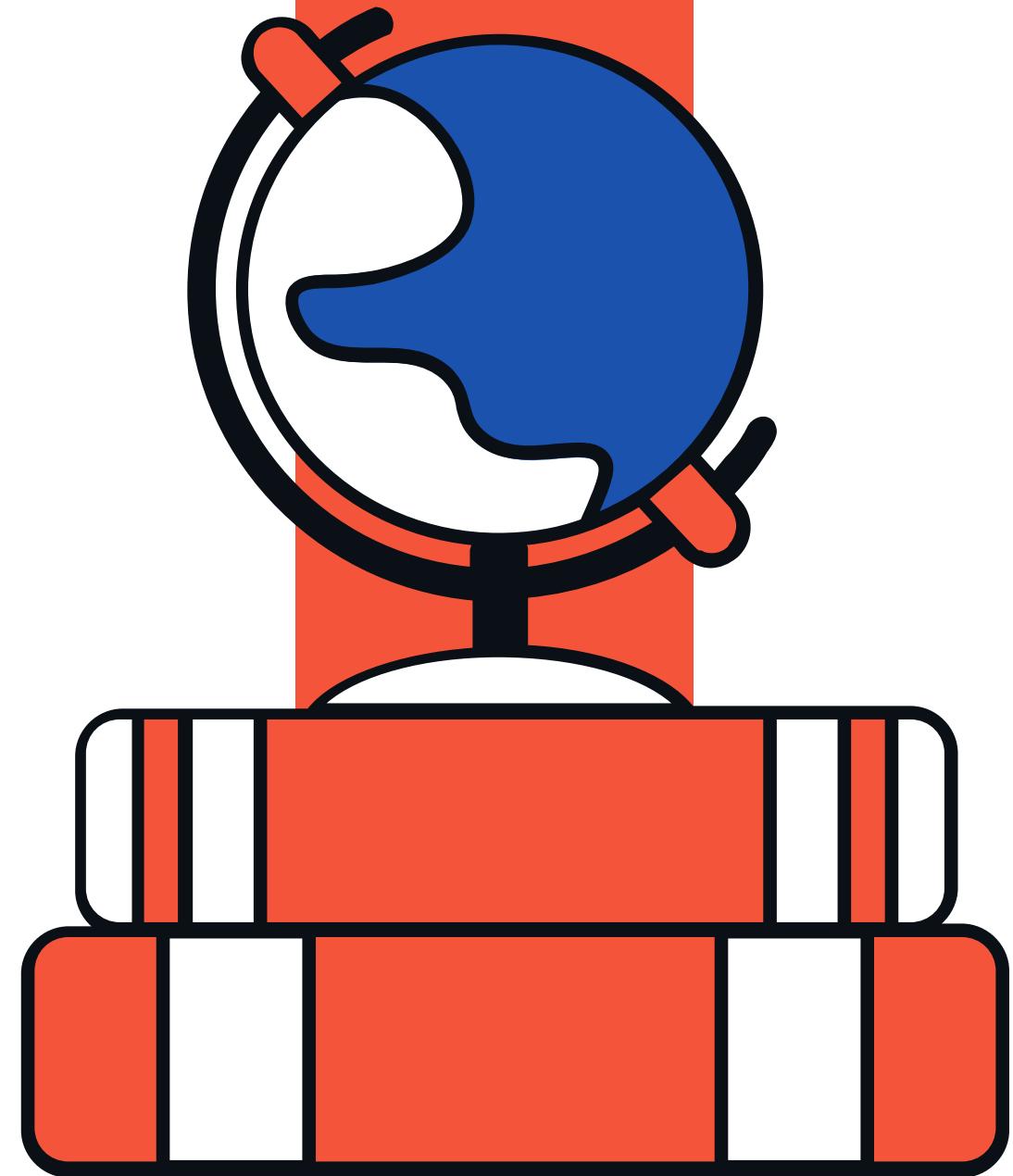
Data Sources



Clustering



App Demo



Project Goals



- 📖 Recommend books based on user preferences (age, favorite titles, etc.)
- 🌐 Highlight globally trending books — not just bestsellers
- 🧠 Suggest thematically similar books beyond genre tags
- 📊 Build a system using book metadata, themes, and clustering

Approach

Use APIs & web scraping to gather:

- Titles, authors, genres, ratings, reviews
- Similar books & writing styles

Data Sources

- Google Books API
- GoodReads.com



1K+
Books Catalog

400+
Authors

30+
Genres

100M+
User Reviews

GoodReads.com

Strengths:

- Lots of quality user reviews
- Large assortment of books
- Good data coverage/quality

Weaknesses:

- Time intensive to pull data via web scraping (Selenium)

The screenshot shows the Goodreads website interface. At the top, there is a navigation bar with links for 'Home', 'My Books', 'Browse ▾', and 'Community'. A search bar is also present. Below the navigation, the title 'Listopia' is displayed in green. Underneath it, the section 'Best Books Ever' is shown in bold black text. A note below states, 'The best books ever, as voted on by the general Goodreads community.' and 'Note to librarians: do not edit this list's description.' There are two tabs at the bottom: 'All Votes' (which is selected) and 'Add Books'. The first book listed is 'The Hunger Games (The Hunger Games, #1)' by Suzanne Collins, with a small thumbnail image of the book cover.

GoodReads.com - Best Books Ever List

The screenshot shows the GoodReads website with the search bar at the top. Below it, the 'Best Books Ever' list is displayed with three entries:

- 1 The Hunger Games (The Hunger Games, #1)** by Suzanne Collins. Rating: 4.34 avg rating – 9,320,534 ratings. Want to Read | Rate this book.
- 2 Harry Potter and the Order of the Phoenix (Harry Potter, #5)** by J.K. Rowling. Rating: 4.50 avg rating – 3,617,453 ratings. Want to Read | Rate this book.
- 3 Pride and Prejudice** by Jane Austen. Rating: 4.29 avg rating – 4,518,487 ratings. Want to Read | Rate this book.

Each entry includes a 'Vote For This Book' button and a score: 4,062,564, and 41,317 people voted for The Hunger Games; 3,212,611, and 32,817 people voted for Harry Potter; and 2,834,191, and 29,058 people voted for Pride and Prejudice.

Contains

- Popularity
- Average Rating
- Book Title
- Book Author
- Book Details Page (URL)

Missing:

- Genres***
- Other details (ISBN, Page #, etc)

Book Details Page



The Hunger Games #1

The Hunger Games

Suzanne Collins

★★★★★ 4.34

9,320,557 ratings · 233,975 reviews

Could you survive on your own in the wild, with everyone else trying to kill you? You don't live to see the morning?

In the ruins of a place once known as North America, the powerful Capitol rules the twelve outlying districts. The Capitol is cruel and keeps the districts in line by forcing them all to participate in the Hunger Games, a fight to the death on live TV.

Want to read

Contains

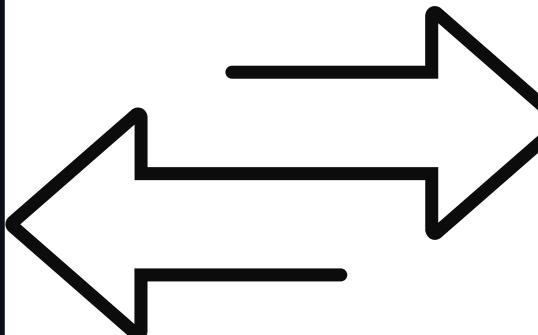
- Genres
- ISBN
- Page #'s
- Reviews



Web scraping

A screenshot of the Goodreads website showing the top three books. The first book is "The Hunger Games" by Suzanne Collins, ranked #1. It has a 4.34 average rating from 9,320,534 ratings and a score of 4,062,564. The second book is "Harry Potter and the Order of the Phoenix" by J.K. Rowling, ranked #2, with a 4.50 average rating from 3,617,453 ratings and a score of 3,212,611. The third book is "Pride and Prejudice" by Jane Austen, ranked #3, with a 4.29 average rating from 4,518,487 ratings and a score of 2,834,191.

500x



A detailed view of the "The Hunger Games" book page. The book cover features the title "THE HUNGER GAMES" in large white letters, a golden Mockingjay logo, and the author's name "SUZANNE COLLINS". Below the cover, there is a green "Want to read" button. The page also includes the book's description: "Could you survive on your own in the wild, with everyone else? You don't live to see the morning?" and a summary: "In the ruins of a place once known as North America, the people living in the capital city are very rich and powerful. They are surrounded by twelve outlying districts that are much poorer and less powerful. The capital city is very cruel and keeps the districts in line by forcing them to send one boy and one girl between the ages of twelve and eighteen to participate in the Hunger Games, a fight to the death on live TV."

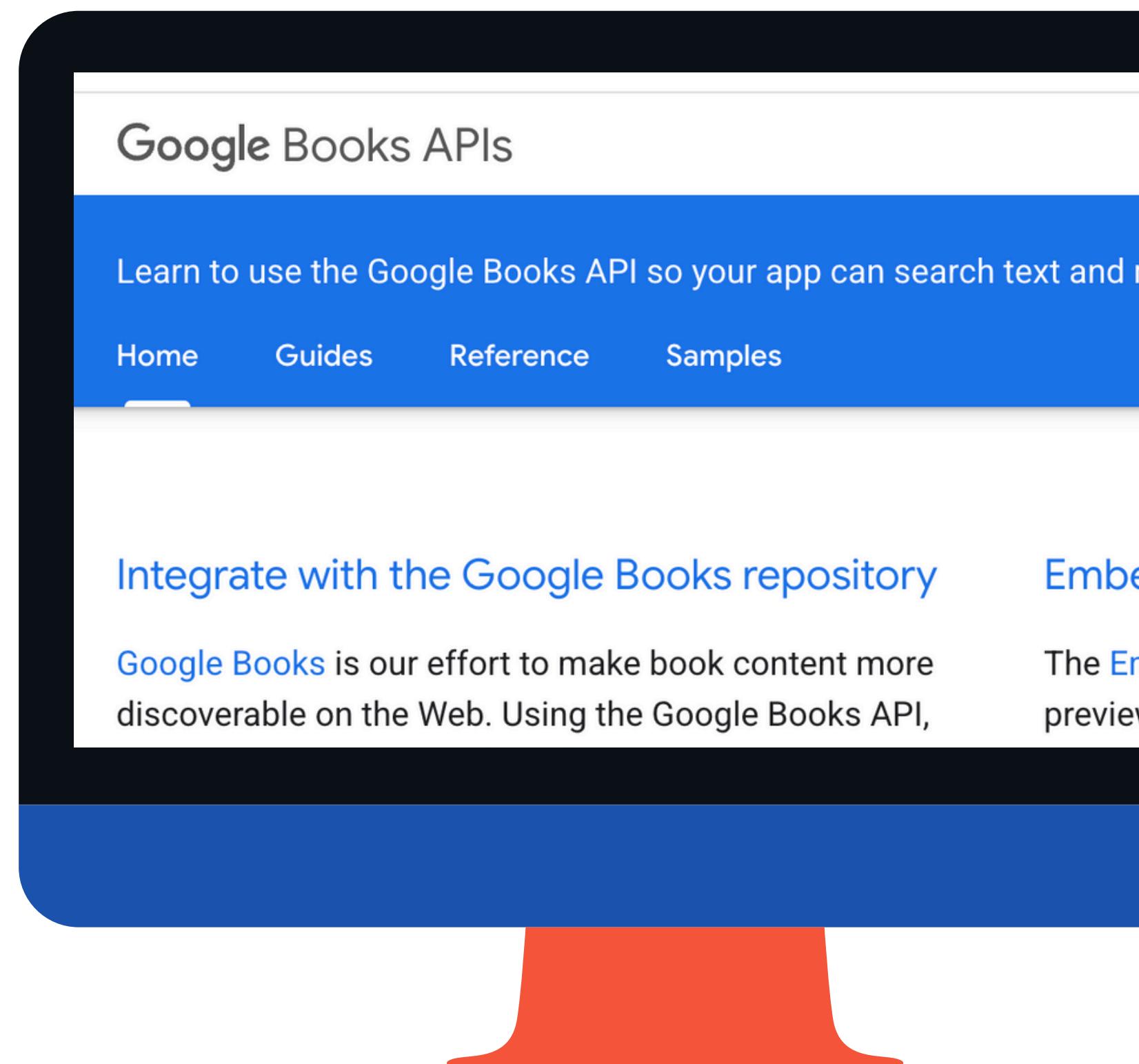
Google Books API

Strengths:

- Very easy and quick to use
- Contains a broad range of books (new, old classic)

Weaknesses:

- Lots of missing data
- Very few ratings and data on popularity/quality



Data Processing



Cleaned GoodReads Data

- Removed duplicates, extracted series, cleaned genres.

Processed Google API Data

- Normalized ratings, converted types, etc.

Aligned Columns

- Matched structures for seamless merging.

Merged Datasets

- Combined 1,000+ books into a unified dataset



Analysis - Ratings & Popularity

1 Popularity* vs. Avg Rating

- Weak negative correlation (-0.10)
- **Takeaway:** popular ≠ always highly rated (e.g Twilight)

1 Popularity* vs. Num Ratings

- Moderate negative correlation (-0.40)
- **Takeaway:** more ratings = higher popularity*

Avg Rating vs. Num Ratings

- Weak positive correlation (0.06)
- **Takeaway:** more ratings ≠ better ratings

* Popularity (low number = high popularity)

Analysis - Balanced Quality Score

Goal:

- Balance popularity and reviews
- Create one normalized score

Inputs:

- Avg_Rating : Weight 0.5
- Popularity : Weight 0.25
- Num_Ratings : Weight 0.25



Note: All scores were normalized and scaled first

Challenge - Missing Quality Data

⚠ Challenge:

- Many books from the Google API lacked ratings, reviews, or popularity data.

🛠 Solution:

- Filled missing scores using the **average score of other books by the same author.**
- KNN Imputation (Backup)

For remaining gaps, used **K-Nearest Neighbors (K=3)** to estimate scores based on similar books.

Analysis - Book Genres



Keywords by Age Group

Create a new **numerical column** representing age group



Ages 0–8:

- "Childrens", "Juvenile Fiction"



Ages 8–18:

- "School", "Middle Grade", "Young Adult"

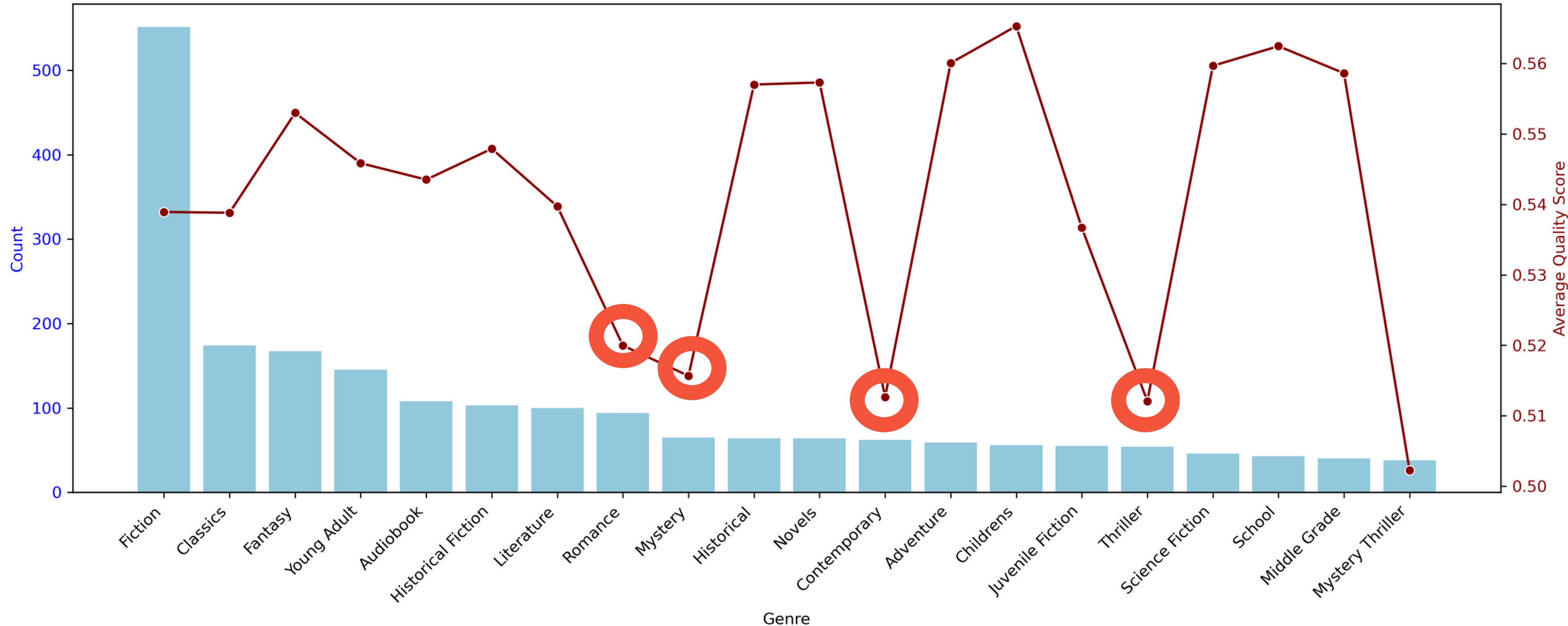


Ages 18–100 (default):

- "Romance", "Thriller", "Mystery", "Horror" (and anything else not matched above)

Analysis - Book Genres & Quality

Top 20 Genres by Count and Their Average Quality Score



Most Controversial Genres

Genres with the most mixed ratings and quality scores – often loved or loathed:

-  Romance
-  Mystery
-  Contemporary
-  Juvenile Fiction
-  Thriller
-  Mystery Thriller
-  Paranormal

"Controversial" = high variance in reader ratings and quality perception

Book Clustering Overview

We used **unsupervised machine learning** to group similar books based on content and reader data.

Step 1: Feature Preparation

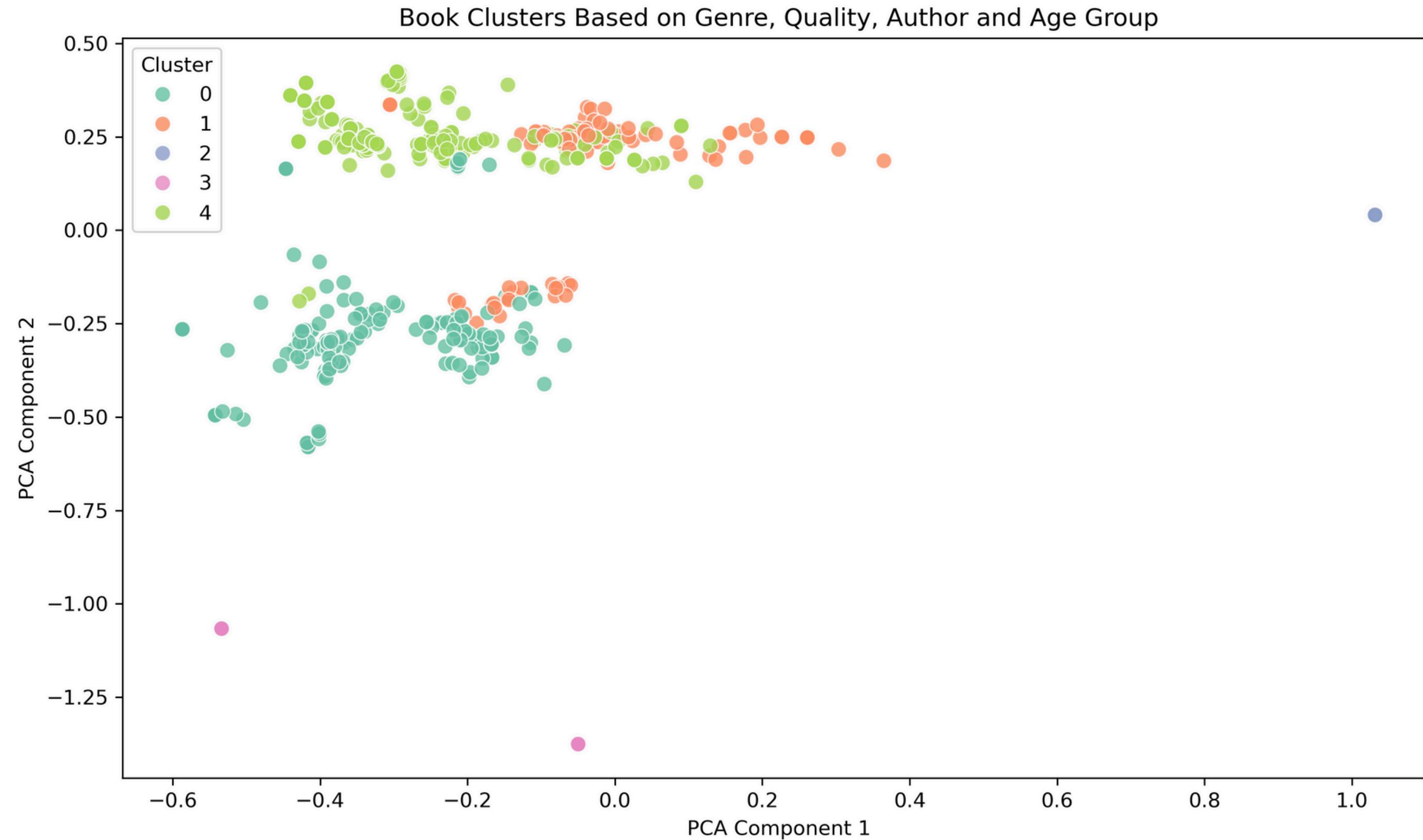
- Converted **genres** and **authors** into text features using **TF-IDF***
- Normalized numeric features: **quality score** and **age group**

Step 2: Clustering

- Combined all features and applied **KMeans** clustering--
- Grouped books into 5 distinct clusters based on content and metadata

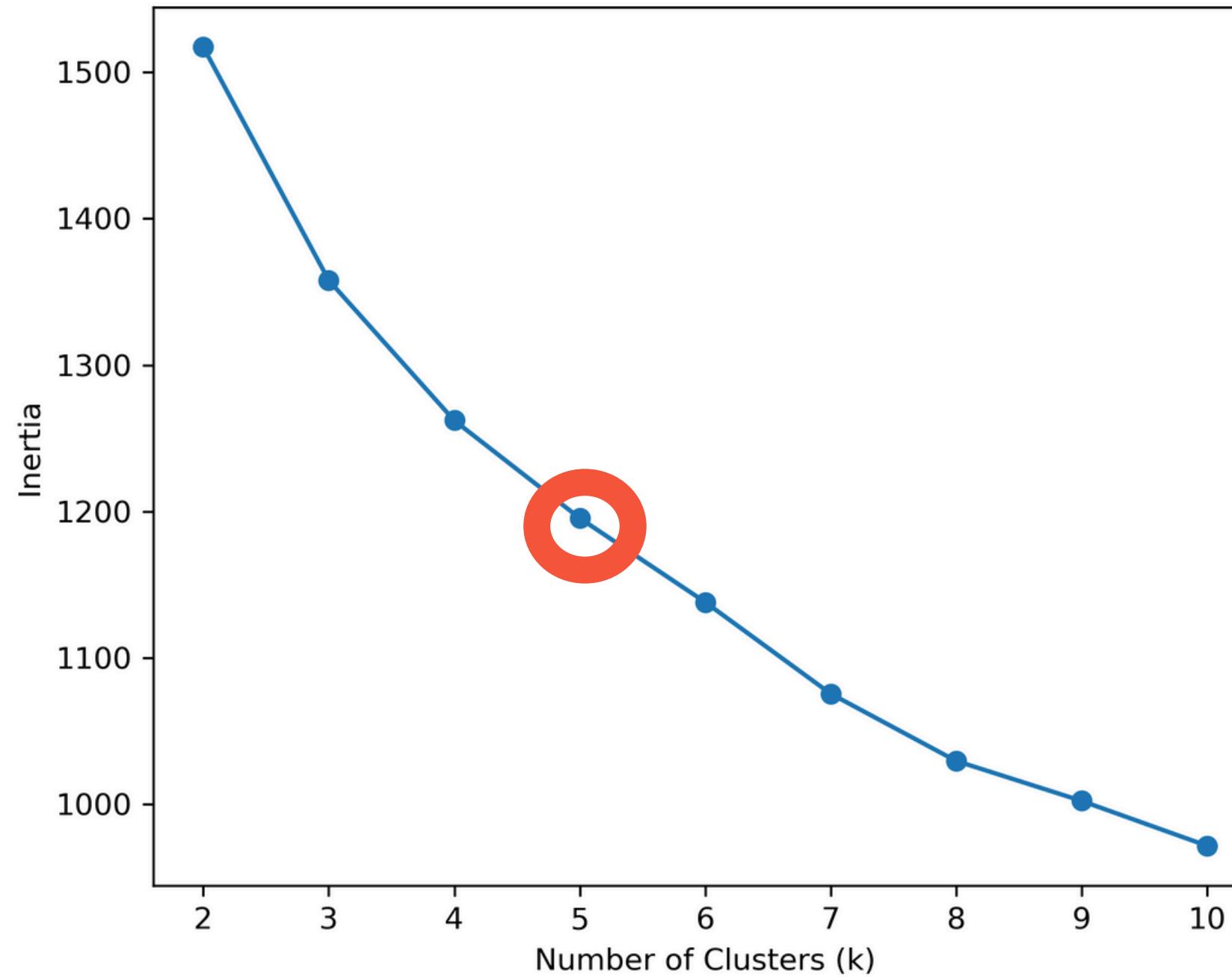
**TF-IDF: Term Frequency–Inverse Document Frequency*

Book Clusters Visualized

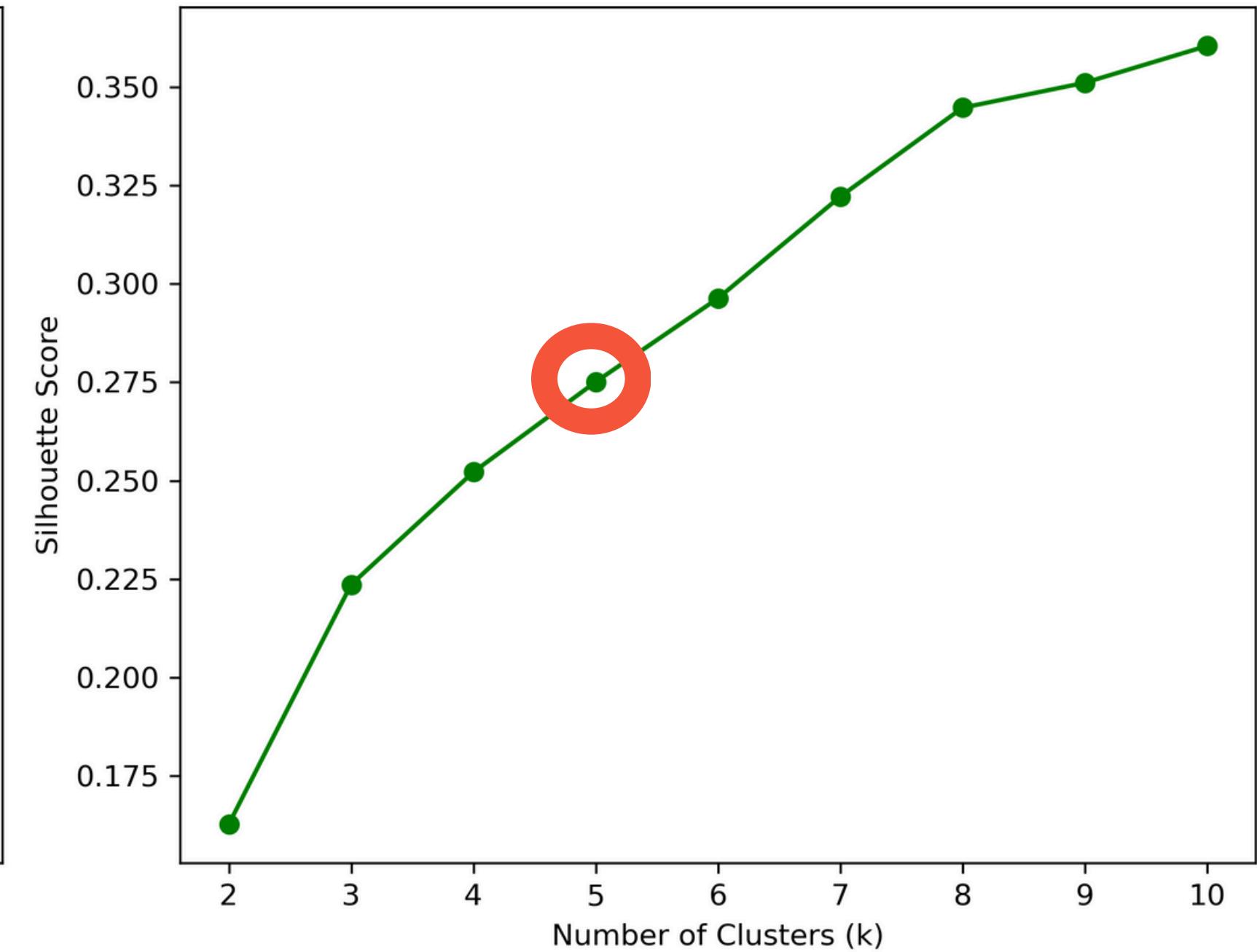


Optimizing Clusters

Elbow Method: Inertia vs Number of Clusters



Silhouette Score vs Number of Clusters



📌 Takeaway:

Both the elbow method and silhouette score indicate that **5 clusters** offer the best balance of cohesion and separation for grouping similar books.

Final Recommendations

Primary Data Source

→ Prioritize scraping from **GoodReads.com**

Enhance Recommendation Criteria

→ Incorporate **Genre, Author, Age Group, and Popularity**

Deeper Analysis Needed

→ Investigate **low-popularity** books to understand gaps and improve filtering or labeling.

→ More grouping and cleaning of **genre** category to remove duplicates

App Demo

