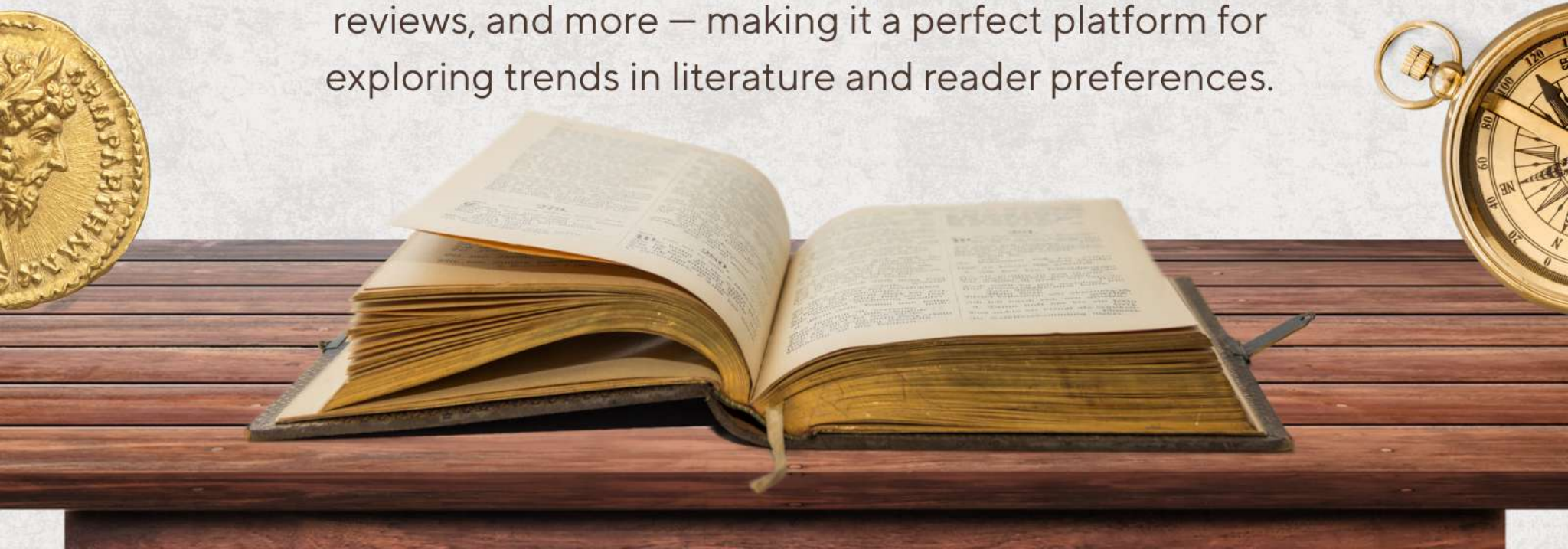
National Telecommunication Institute

Goodreads Data Analysis
From Scraping to Insights

Goodreads

Goodreads is the world's largest social platform for book lovers. It allows users to discover, review, rate, and organize books they've read or want to read. With millions of titles and user-generated data, it serves as a rich source of insights into reading trends, book popularity, and author reach. For data analysts, Goodreads offers valuable public data such as book titles, authors, average ratings, number of reviews, and more — making it a perfect platform for exploring trends in literature and reader preferences.



goodreads

Scrapping Technique

Goodreads Scraping Process "Overview"

1. Define Search Queries

We choose a list of keywords (e.g., "data science", "psychology", "business") to target specific book topics.

2. Send Search Requests

For each keyword, we automate the browsing of Goodreads search results — collecting data from the top 5 pages per query.

Goodreads Scraping Process "Overview"

3. Parse Book Details

From each book in the search results, we extract:

- Title of the book
- Author name
- Average Rating (e.g., 4.2 stars)
- Number of Ratings (e.g., 10,000+ users)
- Book URL (link to the book's page)

4. Fetch Book Detail Pages

For every unique book link, we send a separate request to access more detailed information.

Goodreads Scraping Process “Overview“

5. Extract work_id

Inside each book’s HTML, we search for a unique identifier called work_id, which links to the underlying "work" on Goodreads (used for database integrity).

6. Save to Excel

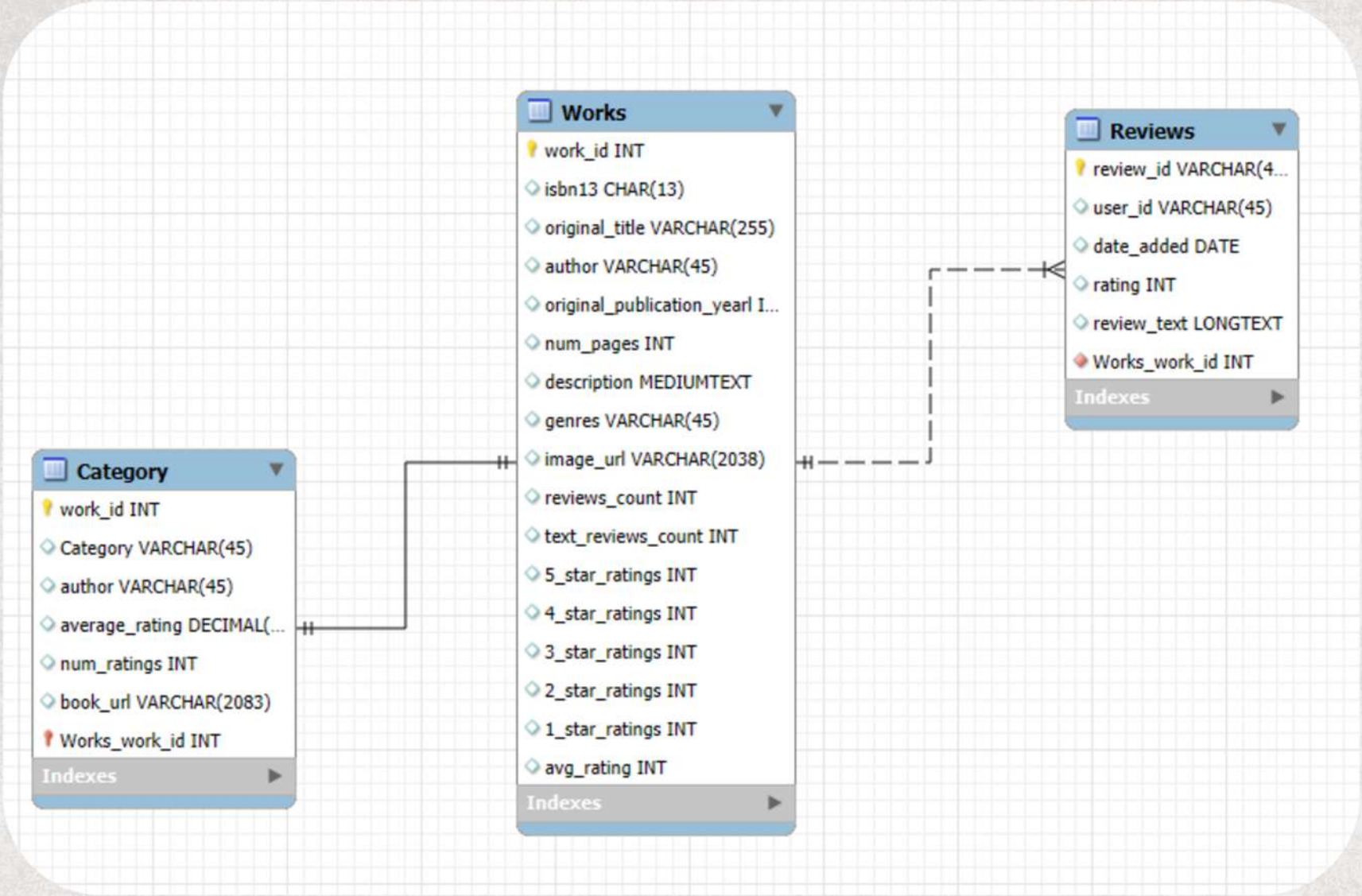
The final dataset — including all categories and their book details — is saved into an organized Excel file, with separate sheets per category and a combined sheet for analysis.

Data Integration

We integrated **two** complementary data sources to enrich our analysis:

- **Scraped Data from Goodreads**
- **Maven Analytics Goodreads Dataset**
- **A dataset containing over 1M+ user reviews for 13K+ books, collected in late 2017.**
- **The "works" table includes detailed book metadata (author, genres, publication date, ratings, etc.).**
- **The "reviews" table contains English text reviews with spoiler tags and user interactions.**

Database Schema - Goodreads Review Integration



goodreads

Cleaning Procedure

Data Cleaning Using Python

Our Dataset:

- Reviews File
- Works File
- Scrapped File

Python Notebook Used:

- Google Colab: cloud based python IDE

Reviews Sheet Cleaning Procedure

- Import Libraries (**Pandas** & **Numpy**)
- Read File as **CSV**
- Dataset Exploration
- Data Cleaning Phase
- Cleaned Dataset Exportation

Cleaning Phase

```
[14] df['rating'] = df['rating'].astype('Int64')
```

```
[15] df['started_at'] = pd.to_datetime(df['started_at'], errors='coerce')
     df['read_at'] = pd.to_datetime(df['read_at'], errors='coerce')
     df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
```

```
[16] df['work_id'] = df['work_id'].astype(str)
```

```
[18] df.drop(columns=['started_at'], inplace=True)
```

```
[19] df.drop(columns=['read_at'], inplace=True)
```

```
[20] df.drop(columns=['n_comments'], inplace=True)
     df.drop(columns=['n_votes'], inplace=True)
```

Cleaned Dataset Exportation

```
[21] !pip install openpyxl
```

Show hidden output

```
[22] df.to_excel("cleaned_reviews_sheet.xlsx", index=False)
```

```
[23] from google.colab import files
     files.download("cleaned_reviews_sheet.xlsx")
```

Works Sheet Cleaning Procedure

- Import Libraries (**Pandas** & **Numpy**)
- Read File as **CSV**
- Dataset Exploration
- Data Cleaning Phase
- Cleaned Dataset Exportation

Cleaning Phase

```
[7]  df["similar_books"]=df["similar_books"].fillna("undefined")
```

```
[8]  df["description"]=df["description"].fillna("no description")
```

```
[9]  df['num_pages'] = df['num_pages'].astype('Int64')
     df['work_id'] = df['work_id'].astype(str)
```

```
[11] df["original_publication_year"] = df["original_publication_year"].astype(str)
```

```
[12] df['num_pages'] = df['num_pages'].fillna(0).astype('Int64')
```

```
[13] print(df['original_publication_year'].dtype)
```

Show hidden output

```
[14] df['original_publication_year'] = df['original_publication_year'].str.replace('.0', '', regex=False)
```

Cleaning Phase

```
[15] #applying lamda function and convert isbn13 from float to string
     df["isbn13"] = df["isbn13"].apply(lambda x: str(x).split(".")[0] if pd.notnull(x) else x)
```

```
[16] print(df["isbn13"].head())
```

Show hidden output

```
[17] #convert isbn10 to isbn13
     def convert_isbn10_to_isbn13(isbn10):
       #returns none if the value is null
         if pd.isna(isbn10):
             return None
       #remove - and space from the value as they're included in some isbn10 (0-123-45678-9)
         isbn10 = str(isbn10).replace("-", "").strip()
       #checking that it's 10 digits and not ending with x
         if len(isbn10) != 10 or not isbn10[:-1].isdigit():
             return None
         isbn13_body = "978" + isbn10[:-1]
       #if even add the value itself , else multiply by 3
         total = sum((int(d) if i % 2 == 0 else int(d) * 3) for i, d in enumerate(isbn13_body))
         check_digit = (10 - (total % 10)) % 10
       #return isbn13 as str
         return isbn13_body + str(check_digit)

     #apply on data
     df["isbn13"] = df["isbn13"].fillna(df["isbn"].apply(convert_isbn10_to_isbn13))
```

```
[18] df.drop(columns=["isbn"], inplace=True)
```

Cleaning Phase

```
[19] print("number of null before:", df["isbn13"].isnull().sum())
```

Show hidden output

```
[20] #calculate column length
     df["isbn13"].dropna().apply(lambda x: len(str(x))).value_counts()
```

Show hidden output

```
[21] df = df[df["isbn13"].apply(lambda x: len(str(x)) == 13 if pd.notnull(x) else True)]
```

```
[22] df["isbn13"].dropna().apply(lambda x: len(str(x))).value_counts()
```

Show hidden output

```
[23] #remove all rows having null isbn13
     df = df[df["isbn13"].notna()]
```

```
[24] print("number of null after:", df["isbn13"].isnull().sum())
```

Cleaned Dataset Exportation

```
[25] !pip install openpyxl
```

```
Requirement already satisfied: openpyxl in /usr/local/lib/python3.11/dist-packages (3.1.5)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.11/dist-packages (from openpyxl) (2.0.0)
```

```
[26] df.to_excel("cleaned_work_sheet.xlsx", index=False)
```

```
[27] from google.colab import files
     files.download("cleaned_work_sheet.xlsx")
```

goodreads

Overview Dashboard Using Excel

goodreads

Number of Count
11873

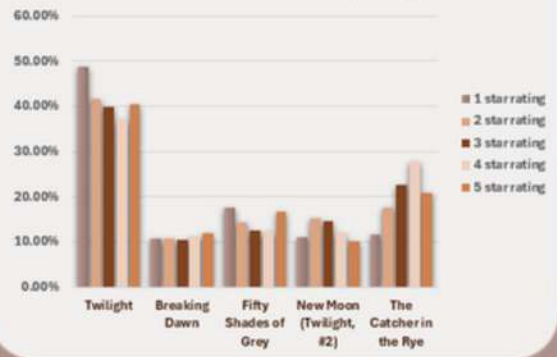
Total Reviews
5420

Average Rating
4.05

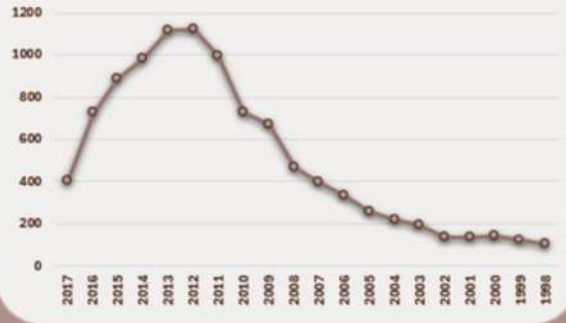
Quarter
- Qtr1
- Qtr2
- Qtr3
- Qtr4

Month
- Jul
- Aug
- Sep
- Oct
- Nov

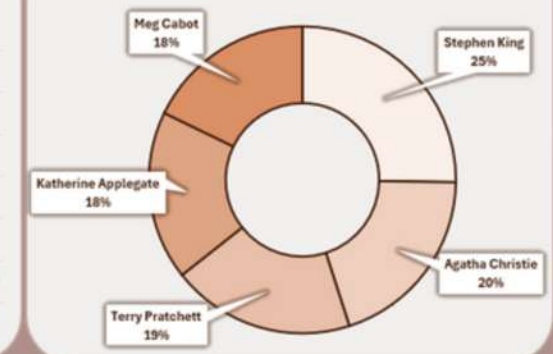
Star Ratings Distribution per top books



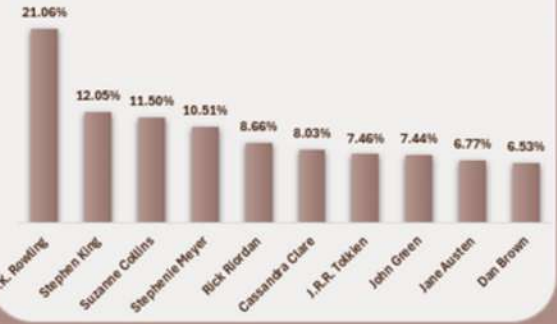
Number of Published Books Changing Over Time



AUTHORS WHO HAVE THE MOST BOOKS



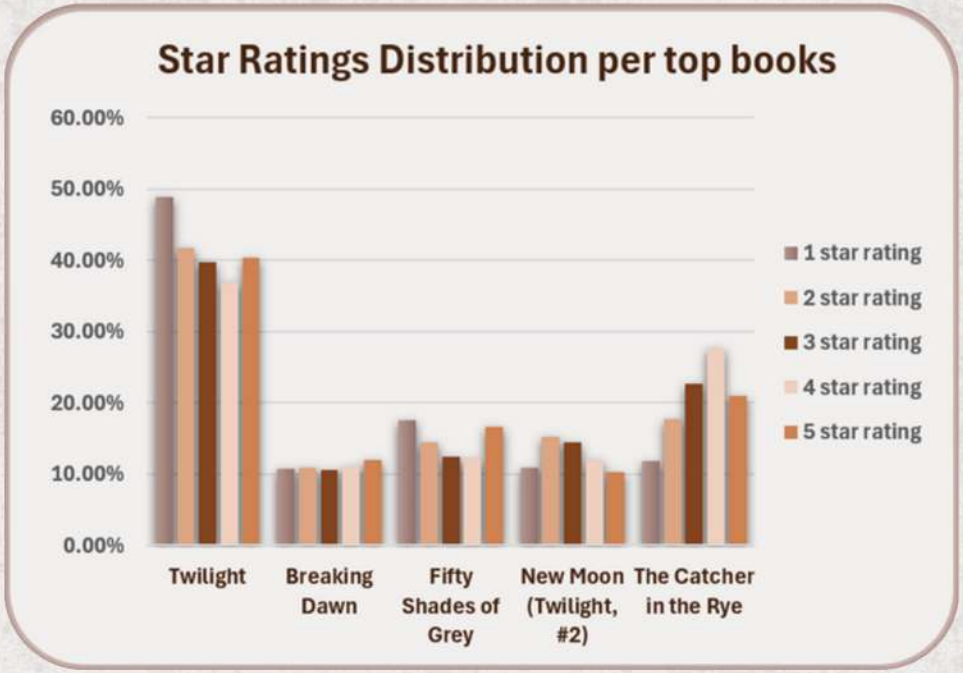
Authors Who Have The Most Reviews



Top 5 books Reviews Count

Top 5 books Average Rating

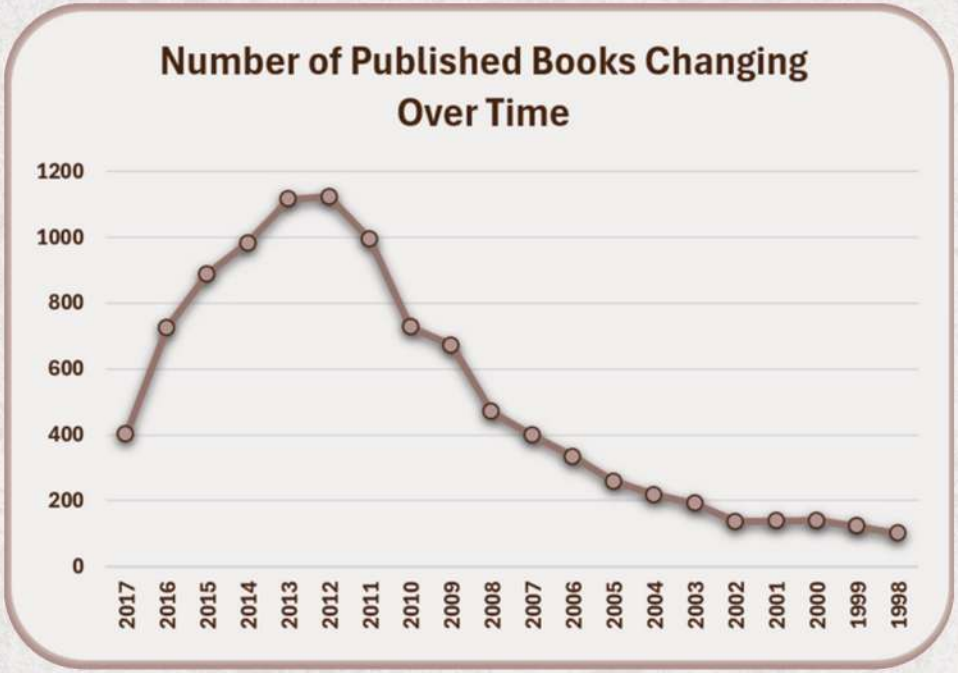
Star Ratings Distribution



Among the top books shown, Twilight dominates with the highest concentration of 5-star ratings (nearly 50%), indicating strong fan loyalty and overwhelmingly positive reception, while other titles like Breaking Dawn, New Moon, and The Catcher in the Rye have more balanced or mixed rating distributions, suggesting a more diverse reader sentiment.

This highlights Twilight as a standout in terms of reader satisfaction compared to its peers.

Published Books Over Time

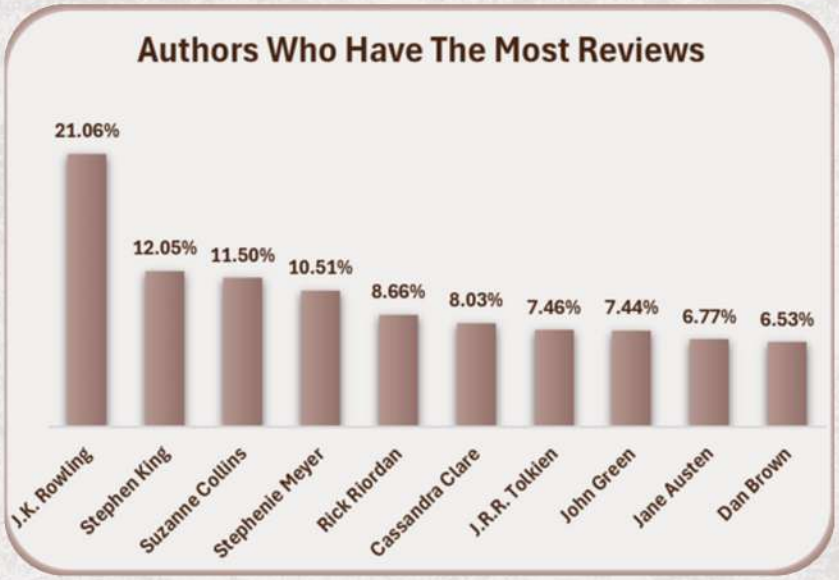


Peak publishing between 2012–2014
Sharp decline after 2015

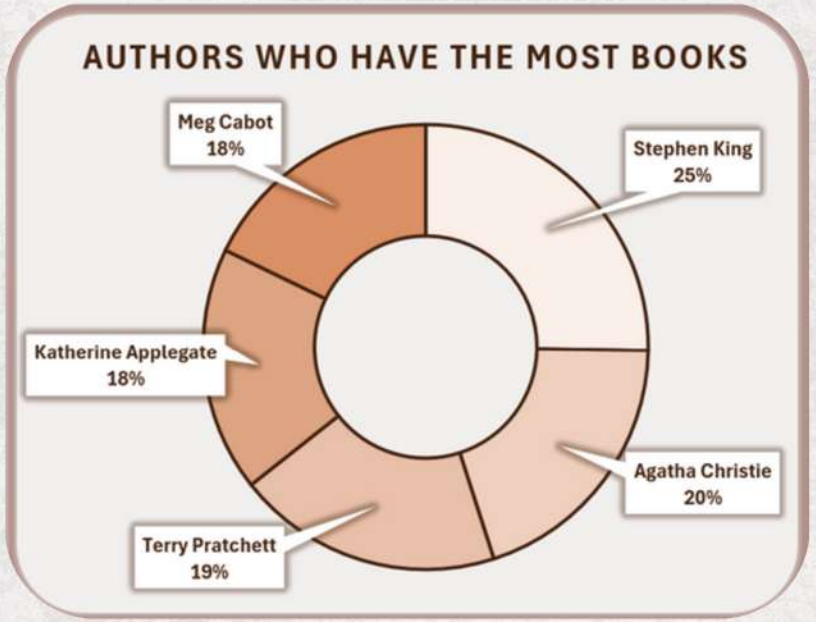
Possible reasons:

- Market saturation
- Shift to digital/indie platforms

Author Popularity

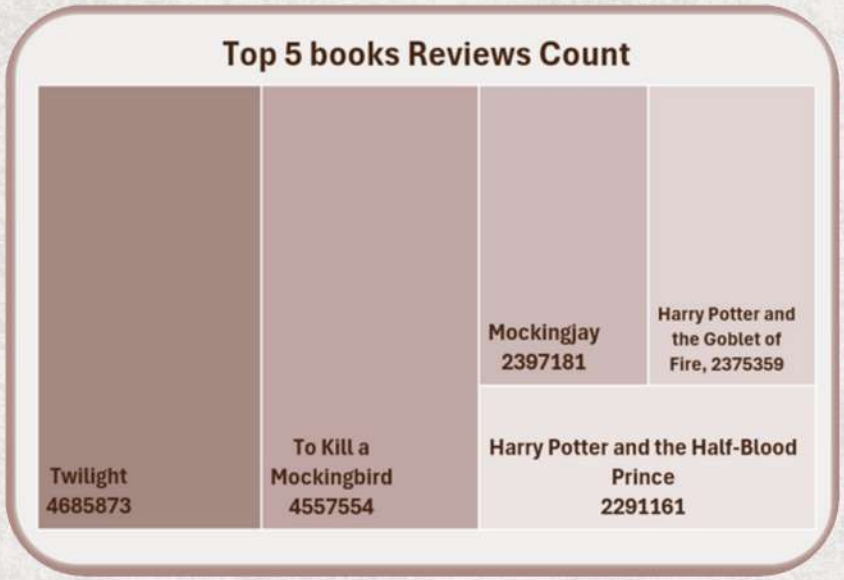


J.K. Rowling holds the highest share of reviews **(~21%)**, demonstrating a high engagement-per-book ratio



Stephen King and Agatha Christie lead in book count
(**25%** and **20%** respectively)

Top Books
(Engagement & Rating)



Twilight **dominates** in both reviews count and average rating, likely due to a large, vocal fanbase

Mockingjay and To Kill a Mockingbird also have high engagement, though with slightly lower ratings

goodreads

Report Using PowerBi

goodreads

User Engagement & Review Behavior

NTI

category

- [ ] Select all
- [ ] artificial intelligence
- [ ] business
- [ ] data science
- [ ] history
- [ ] machine learning
- [ ] psychology

average_rating

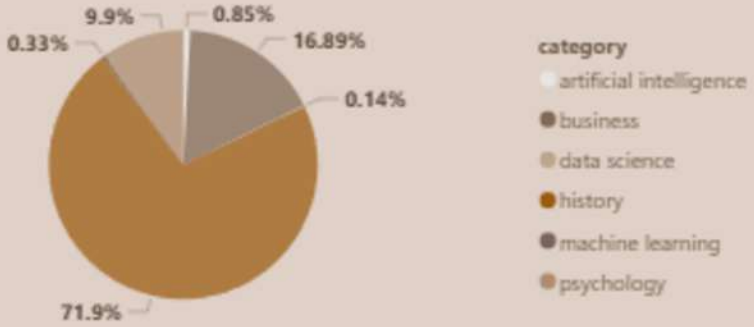
3.25 4.83

Quarter

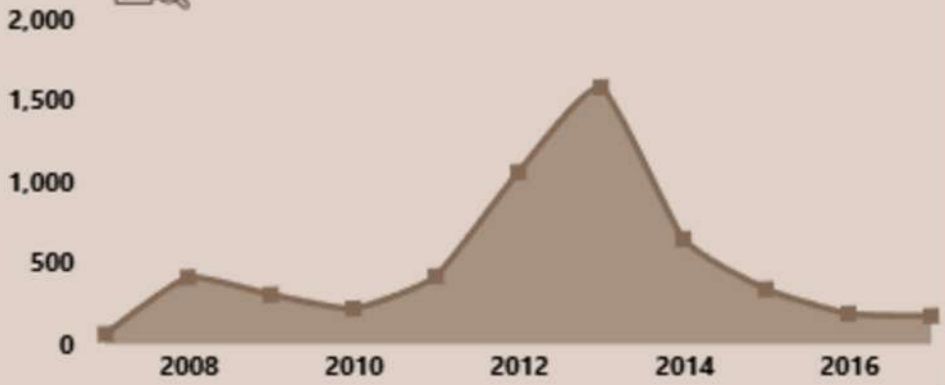
- [ ] Select all
- [ ] Qtr 1
- [ ] Qtr 2
- [ ] Qtr 3
- [ ] Qtr 4

Top authors according to ratings

Categories VS Ratings



Reviews Over Time

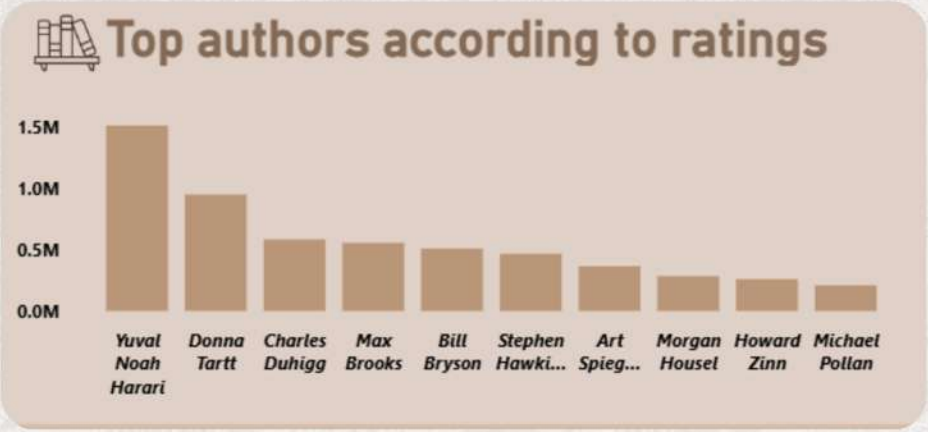


Dashboard 1: User Engagement and Review Behavior

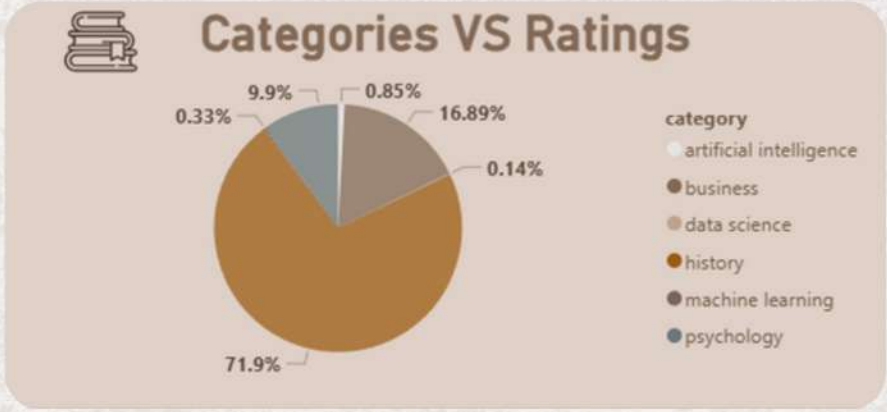
Overall Dashboard Insights:

This dashboard shifts the focus from the books themselves to the users who are reviewing them. It is highly effective at revealing patterns in user behavior, such as their rating tendencies, engagement levels, and activity trends over time. The insights here are crucial for understanding the health of the user community.

Closer view

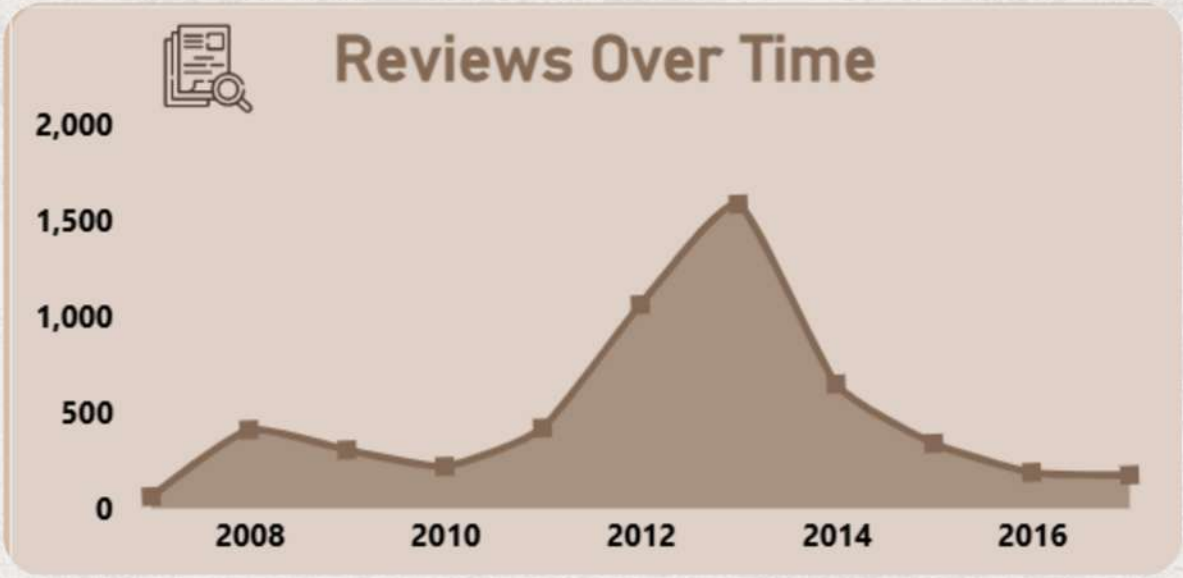


The steep drop-off after the top few authors indicates that a small number of authors are responsible for a significant portion of the total reviews.



This chart provides a proportional view of genre popularity. It visually confirms that "History" is the most dominant genre by a large margin. The smaller slices for other genres, such as "Psychology" and "business," show that while they have an audience, they make up a much smaller percentage of the overall review activity compared to the top genres.

User Review Activity Over Time



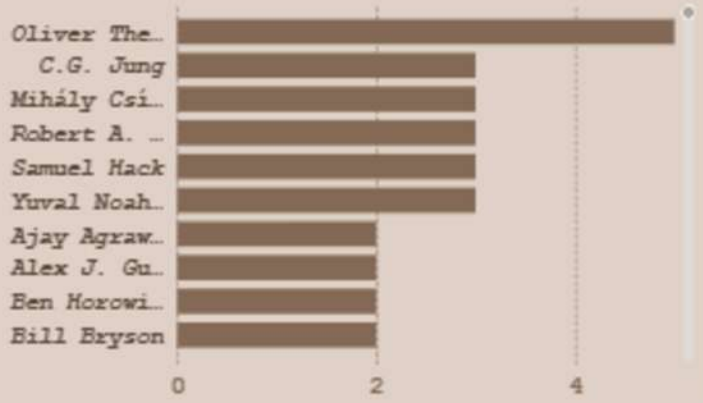
This chart provides a valuable historical perspective on user engagement. The line plot shows a significant and steady decline in review activity from a peak in early 2013 to a very low point in 2014. **This is a critical finding, as it suggests a major drop-off in user engagement or new content being added to the platform over that period.**

goodreads

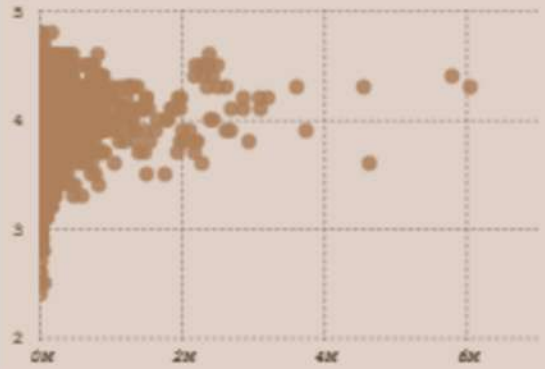
Author & Genre Analysis

NTI

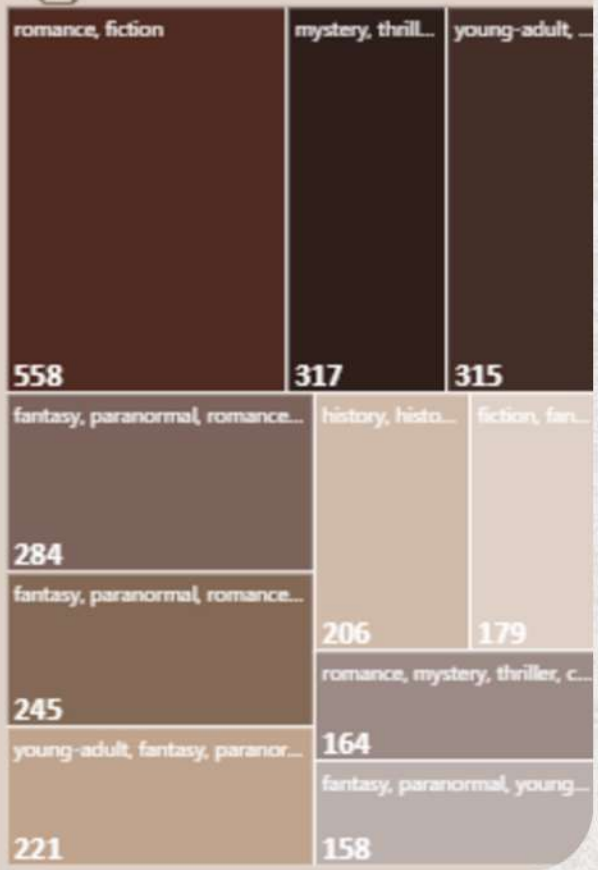
Authors & Num of works



Author Performance vs. Popularity



Top 10 Geners by reviews



Hidden Gems

original_title	author	avg_rating	reviews_count
Finding Center	Katherine Locke	4.20	520
Home Fires	Kate Sherwood	4.20	454
Infected: Epitaph	Andrea Speed	4.30	729
Light Of A Thousand Stars	Siobhan Davis	4.20	474
Memory of Scorpions, book 2: Lying with Scorpions	Aleksandr Voinov	4.20	742
Purple and Black	K.J. Parker	4.20	783
Rock N Soul	Lauren Sattersby	4.20	966
Star Trek: Harlan Ellison's The City on the Edge of Forever: The Original Teleplay	Harlan Ellison	4.20	935
Striker	Lexi Ander	4.30	870

Dashboard 2:
Author & Genre Analysis

Overall Dashboard Insights:

This dashboard provides a foundational overview of Authors' data, focusing on the Authors & Books themselves rather than the users. The insights are centered around author popularity, genre trends, and the general quality of books.

Hidden Gems Table

Hidden Gems

original_title	author	avg_rating	reviews_count
Home Fires	Kate Sherwood	4.20	454
Light Of A Thousand Stars	Siobhan Davis	4.20	474
Finding Center	Katherine Locke	4.20	520
Infected: Epitaph	Andrea Speed	4.30	729
Memory of Scorpions, book 2: Lying with Scorpions	Aleksandr Voinov	4.20	742
Purple and Black	K.J. Parker	4.20	783
Striker	Lexi Ander	4.30	870
Star Trek: Harlan Ellison's The City on the Edge of Forever: The Original Teleplay	Harlan Ellison	4.20	935
Rock N Soul	Lauren Sattersby	4.20	966

- The table provides a direct, actionable list of highly-rated books that are not yet widely known. These titles, such as **"Home Fires"** and **"Finding Center"** are excellent candidates for a "recommended reads" section of a dashboard or a book discovery feature.

Author Performance and Popularity

Popularity is not a perfect indicator of high ratings: The scatter plot shows that while authors like Stephen King and J.K. Rowling have a massive number of reviews, many other authors, some with a fraction of the reviews, have a higher average rating. This indicates that a book's popularity doesn't always reflect its quality or the high esteem it's held in by its readers.

Key Recommendations

- Highlight top performers (Twilight, Harry Potter) in marketing
- Investigate post-2015 publishing decline
- Focus on Power Users: The dashboard shows that a small number of users contribute a large portion of the reviews. You should focus on nurturing this group.
- Actionable Step: Create a "Top Contributor" program that offers badges, recognition, or exclusive content to your most active reviewers. This will incentivize their continued engagement.

Key Recommendations

- Promote Reviewing: The distribution of ratings is heavily skewed towards positive reviews.

Actionable Step: Make the review submission process more prominent and user-friendly. Consider adding a call-to-action button like "Write a Review" on book pages to encourage more feedback.

- Invest in Popular Genres: The dashboard highlights as the most popular genres.

Actionable Step: Focus on curating and promoting new books within these genres to capitalize on existing reader interest. Consider creating featured lists or special campaigns for these categories.

Scrapped dataset link:

goodreads dataset from here

Repo link:

check this out

Thank You