Executive Summary

The motivation behind creating a book recommendation system based on sentimental analysis on book reviews for publishing companies stems from our team's emphasis on contributing higher book sales to the companies as well as catering a personalized reading experience to book enthusiasts. By developing this book recommendation system, we aim to enhance user engagement, personalize reader experiences, and ultimately boost literacy rate and book sales. The project specifically processes and analyzes the data from amazon book review. Despite challenges of cleaning and processing big data through AWS tools, we gathered insights on the top 10 books and its categories, publishers, authors reviewed by helpful review count or rating scores. Ultimately, we hope our insights can give the company a guide on which book reviews they should be running, so the system can guide readers toward books that match their needs as well as higher sales for demanded books in the market.

Introduction

As most people from this group consist of Gen Z, we discovered that the present and upcoming generations have a shorter attention span and a trend of lack of reading. We aimed to analyze the Amazon book review dataset in order to analyze reviews on why certain books are demanded over others. With these insights, we hope the publishing company can build a book recommendation system by specifically analyzing book reviews in order to create a personalized reading recommendation system. This would be valuable to publishing companies and book retailers to create demanded books that will return higher book sales as well as give value to libraries by increasing reading interest in the upcoming generation.

Data wrangling

Summary statistics, Wrangling Process & Reasons

The dataset consisted: books_data, books_ratings. It consisted of both numeric (ex. score) and categorical (ex. title) variables but mostly consisted of text values due to the nature of a book review dataset, which explains the data-wrangling process. Our team stored the data using S3 buckets due to the tool being enabled to store 1GB+ data, then ingested/processed through HDFS due to the operations being a batch process (it's optimal to analyze multiple reviews instead of one by one for the business and the customers usually don't leave reviews for every book one by one) and not a real-time querying process. HDFS was also suitable for initial heavy lifting of data processing. Thus, we cleaned and queried the data using Jupyter Hub and Pyspark from the HDFS' Apache Spark ecosystem so it can read and write to HDFS natively while being less buggy than other applications such as Hue.

Challenges of the Data during the Storing and Processing Steps

We faced a series of challenges while processing a large CSV file for data analysis. Initially, we tried using Hue and JupyterHub to query the data on HDFS, but encountered issues with Hue's inability to accurately read the heavy text data in the CSV file. This led us to consider converting the file to JSON, which was unsuccessful due to the large file size. Our attempt to process the data on RDS before HDFS also failed due to the file's large size. Thus, we decided to load and process the data on HDFS using PySpark in Apache Spark's ecosystem. When attempting to save the cleaned data into a new CSV file, we encountered an error indicating that we needed to install Hadoop locally, which led us to switch to running PySpark on JupyterHub on AWS. Working with certain Python libraries, such as visualization tools like Matplotlib and Seaborn, which weren't functioning directly in PySpark, was quite challenging too. Resolving this issue took considerable time. To fix it, we installed the visualization libraries through the EMR cluster.

Insights & analytics

We decided to query business insights focusing on metrics like score, NHelpfulReviews, Ratings, etc by book title, publisher, and author because we think these metrics tell us a story on which book reviews the company should prioritize in doing sentimental analysis. Thus, we queried 13 insights that can not only help companies achieve higher book sales but also increase reading engagement for customers:

Exhibit 2 highlights the top 10 books by helpful reviews and total score, with 'Mere Christianity', 'Pride and Prejudice', and 'The Hobbit' leading across all queries. This uniformity suggests high market demand for these titles, reinforced by the top author status of 'The Hobbit's creator. This trend indicates that publishers could streamline sentiment analysis costs by focusing on these top-performing books, likely to be recommended and drive sales.

Exhibit 3 shows that 2000-2016 had the most book reviews followed by total helpful reviews. The publishing companies can focus more on running the sentimental analysis for reviews between 2000-2016 to analyze why people left the most reviews - maybe it was the boom of digital media during that time, or maybe the media led to people trying to engage in online communities such as online book clubs, etc.

Exhibit 4 shows top 10 categories by total number of helpful reviews with fiction, religion, history, etc being listed in descending order. Fiction’s helpful review count is doubled the other categories, which aligns with Exhibit 5 by showing the top few author writing fictions. When the company is running the sentimental analysis, it can focus on analyzing reviews for categories such as ‘fiction’, ‘religion’, ‘history’, and more to gain insight into why people think these categories are helpful, as well as will build a better recommendation system for users seeking for books with a specific reason.

Exhibit 5 insight shows the top 10 authors' names by total number of helpful reviews and total rating score. This insight aligns with Exhibit 4 by showing the top few authors writing top 10 categories such as fictions. When the company is running the sentimental analysis, it can focus on analyzing reviews for authors such as 'J. R. R. Tolkien', 'Jane Austen', 'Charles Dickens', and more to analyze if people think these reviews are helpful due to the authors' book career or the book itself (but the author's name was just mentioned in the review). This would help the publishing company to build an accurate recommendation system for users seeking books for a particular reason.

Exhibit 6 insight shows the top 10 publishing companies' names by the total number of helpful reviews and the total number of ratings. When the company is running the sentimental analysis, it can focus on analyzing reviews for publishers such as 'Penguin, 'Random House', and more to analyze a pattern of why these customers think this book from this publisher is most helpful. This would help the publishing company to follow the business strategies of 'Penguin, 'Random House', and more to ensure higher book sales.

Discussion

Application and Value for Organizations:

By leveraging sentiment analysis of Amazon book reviews, our book recommendation system offers an approach to catering to reader preferences while profiting from publishing profitable books. This system promises enhanced personalization in book recommendations, which is pivotal for increasing reader engagement, satisfaction, and, consequently, book sales and circulation. It provides value for both the business (publishing companies) and consumers (readers) by giving insights into market trends and reader preferences.

Generalizability of Insights:

The Top 10 books, publishers, categories, and authors by total score or total number of helpful review insights derived from our analysis are adaptable across different demographics. This versatility ensures that the system remains relevant and effective in diverse reading communities and changing market dynamics.

Critical Analysis of Findings:

While our system provides innovative recommendations based on sentiment analysis, we have to acknowledge potential biases inherent in user-generated content. The effectiveness of the system is contingent upon the quality of data and its representativeness of the broader reader base.

SWOT Analysis:

Strengths of our system include its data-driven approach and the personalized user experience it offers. However, it faces weaknesses like potential biases in user-generated reviews and reliance on external data sources. Opportunities lie in expanding into new markets and integrating AI technologies for enhanced analysis. The system must navigate threats such as rapidly evolving technology, competitive landscapes, and data privacy concerns.

Exhibits (4 pages) * refer to slides for bigger images

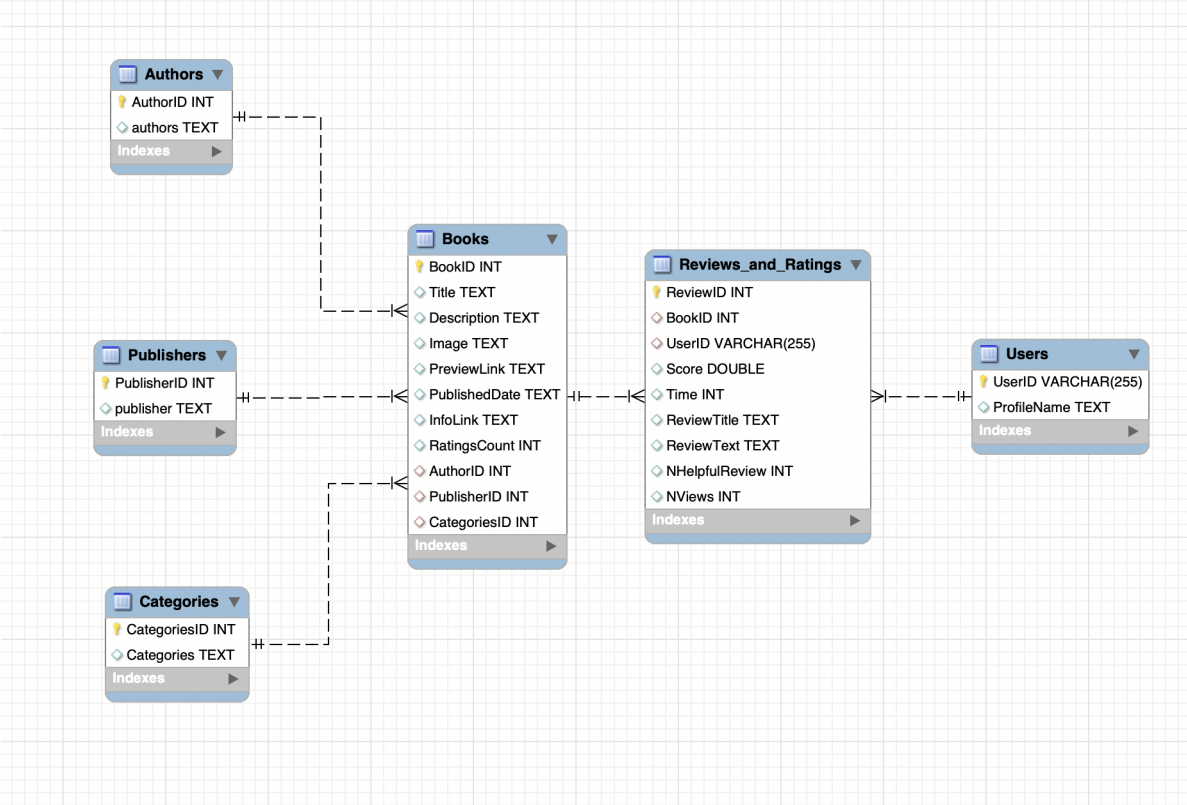
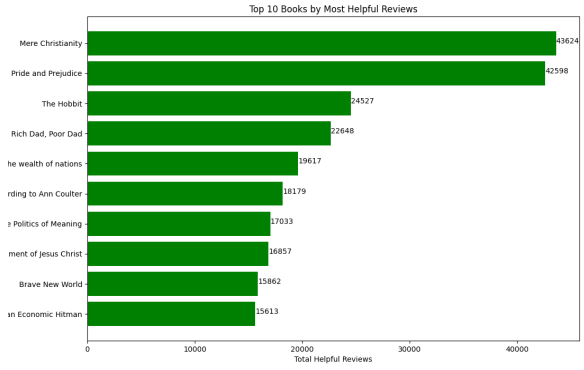


Exhibit 1. ER Model of Amazon Books Review Data

Title	TotalHelpfulReviews
Mere Christianity	43624
Pride and Prejudice	42598
The Hobbit	24527
Rich Dad, Poor Dad	22648
An inquiry into t...	19617
How to Talk to a ...	18179
Liberal Fascism: ...	17033
THE BOOK OF MORMO...	16857
Brave New World	15862
Confessions of an...	15613



Title	NHelpfulReview
Through the Stone...	194.0
The Lazarus Pit	189.0
Quines Son Los Do...	174.11764705882354
Zondervan NIV Stu...	145.0
Deck planner: 25 ...	132.0
A practical Sansk...	123.0
The Losers Club: ...	112.875
How To Grow Organ...	110.0
School Nurse's Su...	102.0
Good People Beget...	101.45454545454545

Title	Score
The Hobbit	85040.0
Pride and Prejudice	81125.0
Mere Christianity	24826.0
Great Expectations	23557.0
Brave New World	21573.0
The Great Gatsby	17844.0
Harry Potter and ...	17490.0
The Hobbit There ...	17177.0
The Hobbitt, or t...	17098.0
The Hobbit or The...	16862.0

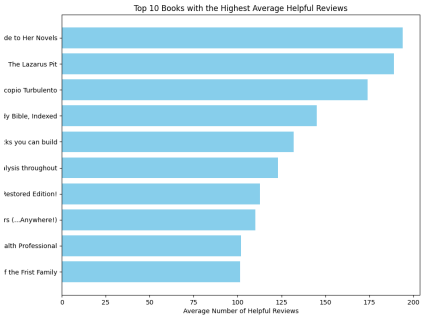
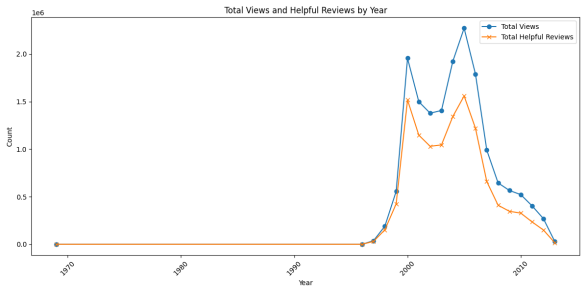


Exhibit 2. Top 10 Books by the number and highest average number of Helpful Reviews, and Top 10 Books with the Highest Total Score



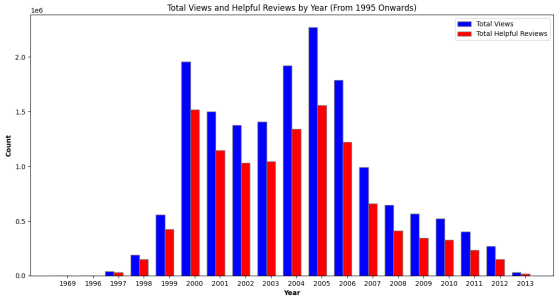


Exhibit 3. Total Reviews & Helpful Reviews from 1970-2013

categories	NHelpfulReview
['Fiction']	1880384
['Religion']	517201
['History']	412754
['Biography & Aut...	391334
['Juvenile Fiction']	315930
['Business & Econ...	291577
['Political Scien...	168173
['Health & Fitness']	163818
['Cooking']	159376
['Body, Mind & Sp...	146234

Exhibit 4. Top 10 Categories by Total Number of Helpful Reviews

authors	NHelpfulReview
['C. S. Lewis']	67169
['Jane Austen']	65720
['J. R. R. Tolkien']	51364
['Adam Smith']	39546
['Charles Dickens']	33593
['John Steinbeck']	32584
['James Joyce']	24696
['Kurt Vonnegut']	24610
['Robert A. Heinl...]	23904
['Sharon L. Lecht...]	22648

authors	Score
['J. R. R. Tolkien']	133057.0
['Jane Austen']	115277.0
['Charles Dickens']	63929.0
['John Steinbeck']	53607.0
['C. S. Lewis']	45201.0
['Kurt Vonnegut']	42719.0
['Harper Lee']	38172.0
['John Ronald Reu...]	36154.0
['F. Scott Fitzge...]	34797.0
['Lewis Carroll']	31745.0

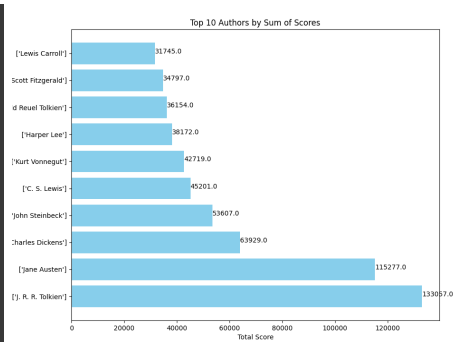


Exhibit 5. Top 10 Authors by Total Score, Top 10 Authors by Total Number of Helpful Reviews

Publisher	RatingsCount
Penguin	77862
Random House	48596
Simon and Schuster	48477
Harper Collins	47901
Vintage	32985
Knopf Books for Y...	26910
Macmillan	16524
Hachette UK	15145
Vintage Canada	14111
Bantam	13876

Publisher	NHelpfulReview
Penguin	389211
Simon and Schuster	374548
Harper Collins	258886
Vintage	153702
Random House	132079
John Wiley & Sons	106292
Courier Corporation	103254
Macmillan	99950
Penguin UK	93012
Hachette UK	77359

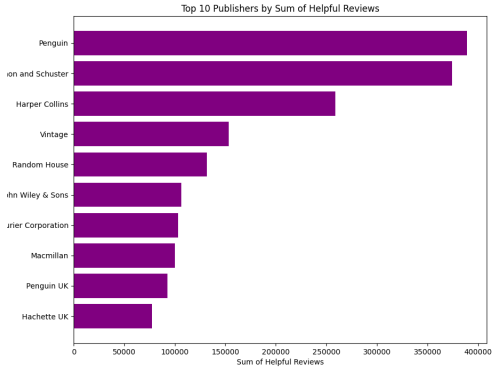


Exhibit 6. Top 10 Publishers by Total Number of Helpful Reviews, Top 10 Publishers by Total Number of Ratings