Optimising Online Bookstore Management:
Data-Driven Strategies for Inventory and Marketing
Group W17G7

Ravon Chew
1243550
ravonc@student.unimelb.edu.au

Reuben Mattam
1271217
rmattam@student.unimelb.edu.au

Viane Dorthea Tiwa
1413279
vtiwa@student.unimelb.edu.au

Executive Summary

This report delves into the dataset from an online bookstore to uncover strategic insights that will guide inventory selection and enhance the sales. Utilising a combination of advanced data processing and sophisticated machine learning techniques, the analysis identifies pivotal factors influencing book rating and user engagement. The datasets include detailed information on 18,185 book titles, profiles of 48,299 users, and over 200,000 user reviews spanning from 1920 to 2005. This comprehensive analysis should instill confidence in the bookstore manager about the strategic direction of the inventory selection.

Significant findings for this report reveal positive sentiment for specific types of books, particularly among younger adult demographics (20-40 years). These books also gain more favorable reviews, underscoring their popularity and potential profitability. The analysis further highlights a trend of increased interaction with books published after 2000, suggesting that newer publications engage customers more effectively, likely due to enhanced marketing strategies and greater online visibility.

The insights obtained advocate for a strategic focus on stocking books that receive positive sentiment and leveraging modern marketing techniques to boost visibility for those who reside in English-speaking country. Additionally, the enhanced understanding of customer preferences for a certain demographic characteristic can facilitate more accurate book recommendations, which can improve user satisfaction and foster loyalty, ultimately driving increased revenue for the bookstore.

1. Introduction

The shift towards digital platforms has markedly transformed how books are sold and consumed in the evolving retail landscape. With the existence of online bookstores, a new paradigm has emerged where data-driven insights have become central to optimising business strategies. This report focuses on leveraging comprehensive datasets from an online bookstore to discern trends and patterns that can inform strategic decisions for the manager to enhance inventory management and customer engagement. The objective is to analyse various facets of user interactions and book popularity to determine which books should be stocked and promoted, thus maximising sales and customer satisfaction.

The online bookstore, with an extensive catalog of 18,185 titles and a diverse user base of 48,299, presents a unique opportunity to explore how different demographics interact with various genres and authors. The dataset spans books published between 1920 to 2005, encompassing a wide range of types, from classic literature to contemporary bestsellers. This time frame also allows for a historical analysis of trends over nearly a century, providing insights into the evolution of reader preferences and publishing patterns.

The primary datasets employed in this analysis include:

1. *BX-Books.csv*: This dataset provides fundamental details about each book, such as the International Standard Book Number (ISBN), title, author, year of publication, and publisher. These attributes are crucial for identifying trends in genre popularity, author success rates, and the impact of publication periods based on user's review rating.

2. *BX-Users.csv*: This dataset contains anonymised information about the bookstore's users and includes data points such as user ID, city, state, country, and age. Analysing this information helps understand the demographic distribution of the bookstore's buyers and their reading preferences, which is instrumental in tailoring marketing and stocking strategies.

3. *BX-Ratings.csv*: This dataset captures book user reviews, providing further information into customer satisfaction and preferences. It includes the user ID, book ISBN, and the rating given. Analysing these ratings is vital for assessing the popularity of books and understanding which features contribute to higher customer ratings and, further, the potential sales.

This analysis explores vast data resources and actionable business strategies through data processing techniques and machine learning models. By dissecting user ratings, demographic information, and book attributes, this report seeks to construct a detailed picture of the bookstore's market dynamics. The insights derived from this approach are expected to guide decisions regarding which books are likely to be successful in the future and how to effectively engage different segments of the bookstore's audience.

Moreover, the comprehensive nature of this analysis allows for the exploration of more profound questions, such as the influence of socioeconomic factors on reading habits, the correlation between books' digital presence and their physical sales, and the potential for predictive models to forecast future trends based on past and current data.

In conclusion, the integration of these datasets provides a robust platform for uncovering the underlying patterns that drive book sales and user engagement. Thus, the research question this report aims to answer is "**What type of books tend to be highly rated by users, and how is this influenced by their demographic**?". This report will offer a snapshot of current trends and give corresponding recommendations for the bookstore manager. By aligning inventory and marketing strategies with data-driven insights, bookstore managers can not only meet but anticipate customer needs, securing a competitive edge in the rapidly changing retail environment of books.

2. Methodology

The methodology deployed in this report is structured to efficiently process and analyse the online bookstore datasets to gain strategic insights to inform bookstore management. This section outlines the systematic approach utilised for data preparation, analysis, and interpretation, focusing on advanced techniques and tools.

2.1 Data Preprocessing

To ensure the reliability and accuracy of the analysis, comprehensive data preprocessing steps were implemented, including:

- Data Cleaning: Incorrect values were identified and removed, correcting entries to prevent any skewing of results due to erroneous data. This step is crucial in ensuring that any analysis or decision-making based on the data is reliable and accurate. For instance, the action taken for each data set is detailed as follows:
  - *BX-Books.csv*: Removing extra quotation marks in the book's title, replacing invalid titles with the correct ones, removing improper entries via regular expression. Improper book titles were replaced with correct titles through API requests utilising the ISBN that were hence stored to a separate csv file, tester.csv.
  - *BX-Users.csv*: Removing users with missing information such as age and country, removing any additional quotation marks from the country name, converting age data to an integer format, and removing any users whose age was found unreasonable, specifically those who were less than five years old or older than 100 years old.
  - *BX-Ratings.csv*: This book dataset covers removing quotation marks from book's title, correcting an invalid book's title, and removing books that only have one rating.

  Furthermore, a new dataset is created to allow deeper analysis using all the information from the *BX-Books, BX-Users,* and *BX-Ratings* dataset.
  - *Merge-dataset.csv*: The aforementioned datasets were merged on common keys ('userID' and 'ISBN'), to consolidate user profiles with their book preferences and ratings for a holistic view. This dataset was then cleaned by filtering out books with only one rating, authors with less 3 books and publisher with less than 5 books, to ensure significance of rating results and minimize noise.

- Feature Engineering: New data columns were derived from existing features in order to enhance model legibility and insight depth. For instance, 'Year-Of-Publication' and combined with the total number of reviews to produce 'Reviews-Per-Year' to facilitate the analysis of corresponding trends over time. Furthermore 'English-Speaking-Ratio' was created to provide a continuous value for nominal data.
- Outlier Treatment: The dataset was further restricted to only including the IQR (Interquartile Range) to identify and treat outliers in continuous variables like ratings that could form a bias during data analysis. Furthermore, outliers such as 'English-Speaking-Ratio' being 0 or 1 were utilized to isolate trends that were skewed by that data.
- Data Integration: Critical to the approach was the integration of user demographic data with their corresponding book ratings and book details. This was accomplished through:
    - Aggregation: Summarising detailed data into insightful metrics, such as average ratings per book and total ratings per user, to facilitate deeper analysis of user engagement and book popularity.

2.2 Analytical Techniques

With the data prepared, a blend of statistical and machine-learning techniques was applied to analyse the data and extract relevant insights:

- Statistical Analysis: As a baseline for evaluation, statistical methods such as Bayesian averages and Wilsons Lower Bound were used to explore tendencies and dispersions within the data. Those statistical methods were then used to produce a popularity ranking list by considering the relations between review count, frequency, and value.
- Machine Learning: Various techniques were implemented to understand and forecast user behavior and corresponding book success. This included:
    - Regression Analysis: When trends were formed in data, this was used to fit the observed trends to a quantifiable model for predicting continuous developments.
    - For classification models, the Sentiment Analysis is used to classify books into categories whether they receive a positive, neutral, or negative sentiment.
    - The Nearest Neighbor Method, as an unsupervised method, is used to find a pattern and thus give further recommendations of five books to purchase, which are based on the rating given by the user.

2.3 Tools and Software

The analysis was supported by a suite of data analytics tools:

- Python: Utilised for its robust libraries including Pandas for data manipulation, Scikit-learn for machine learning, and Matplotlib for data visualization.
- SPSS: Used for advanced statistical testing and analysis, providing a comprehensive platform for the application of statistical theories and tests.

3. Data Exploration and Analysis

This section delves into the comprehensive data exploration and analysis conducted on the datasets provided by an online bookstore. The aim was to uncover underlying patterns, trends, and relationships within the data, including book details, user demographics, and ratings. Through a combination of descriptive statistics, visualizations, and inferential analysis, this section provides insights into various aspects of the bookstore's operations and customer interactions.

3.1 Descriptive Statistics

Initially, the analysis began with descriptive statistics to provide a foundational understanding of the data characteristics of each dataset as follows:

- Books Dataset: The dataset includes 18,185 books, whereas 17,868 entries are used further for analysis. The books are published in the range of 1920 to 2005. The highest number of publication year in Figure.1 was in 2000 to 2010, indicating a relatively modern collection of books, as well as the beginning of the digital transformation era in publishing.

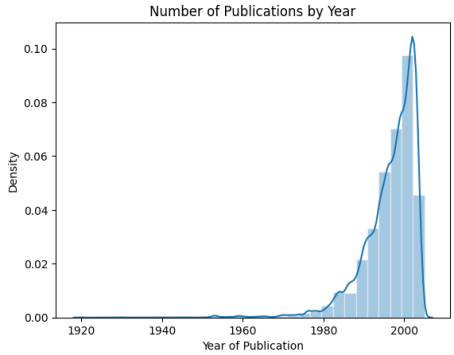

Figure 1: Histogram of Books Published Each Year

- Users Dataset: The user base consists of 48,299 individuals, whereas 28.503 entries are used further for analysis. Figure 2 depicts the age distribution of users, which is concentrated within the range of 30-40 years old with 30.92 percent, indicating a user base primarily composed of young adults. The distribution also is right-skewed, suggesting a gradual decrease in user frequency with increasing age.

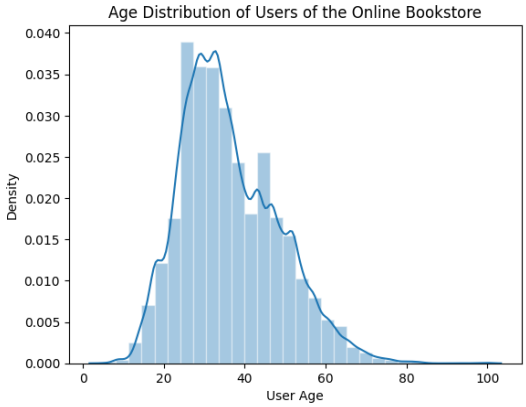

Figure 2: Histogram of User Age Distribution

- Ratings Dataset: Over 200,000 ratings were provided, whereas 136,548 ratings are used further for analysis., with an average rating of 7.75 out of 5. This suggests a generally positive reception of books by the users. The rating distribution was slightly left-skewed, indicating that lower ratings were less frequent.

3.2 Further Descriptive Statistics

Additionally, the demographic background of the userbase was also explored to provide a holistic view of the factors underlying book purchase trends and popularity. The ratio of users from English-speaking countries compared to those from non-English-speaking countries was created and compared with average user age to produce the following density plot.

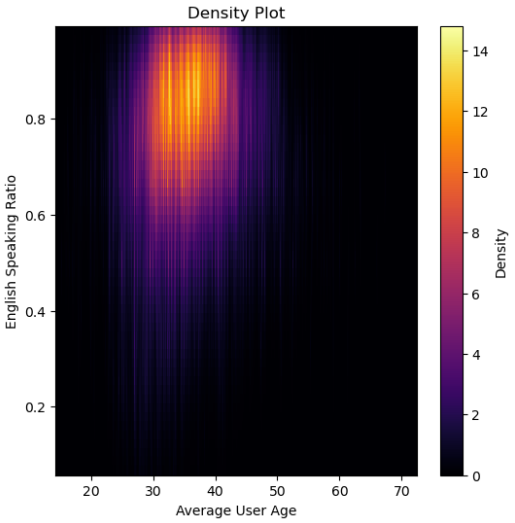

Figure 3: Density Plot of the Average Age and Ratio of English-Speakers of Users

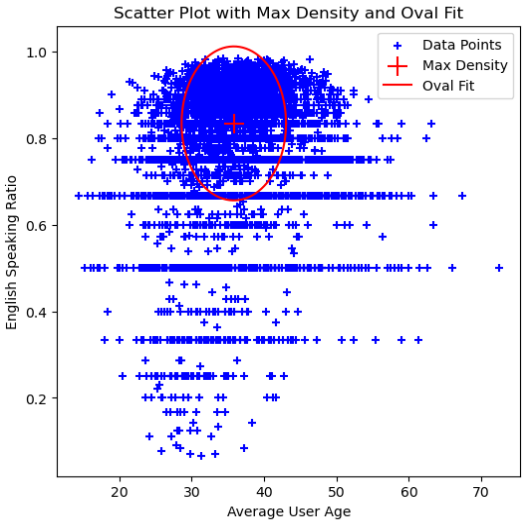

Figure 4: Scatterplot of the Average Age and Ratio of English-speakers of Users

As circled in Figure 4, an English-speaking ratio between 0.628 and 0.984 represents the region of highest density indicating the greatest significance in exploring trends among the user demographic. Comparing the number of ratings of books with an English-speaking ratio within this margin against those outside of this margin produces Figure 5.

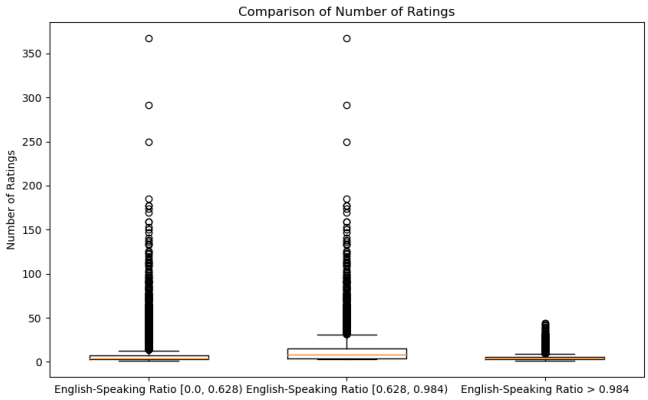

Figure 5: Boxplot of Books' English-Speaking Ratio against their Total Number of Reviews

The IQR of the middle plot representing books within the highest user density is (3.0, 31.0), with outliers representing books that are more popular than expected. By restricting both the number of reviews and the English-speaking ratio to the margin of significance, a discrete exponential scatter plot emerges (Figure 6).

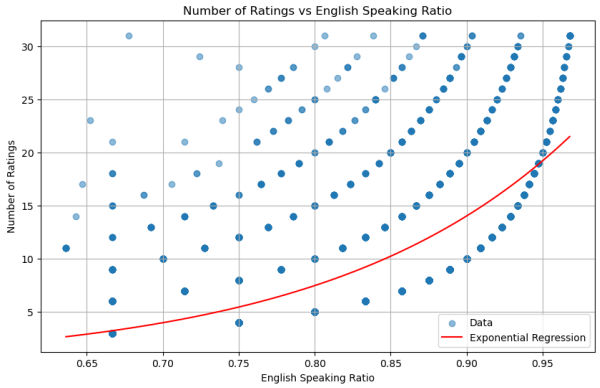

Figure 6: Scatterplot of Users' Average Book Rating against Total Number of Reviews

Which returns the parameters (0.04856743, 6.29536343) after the application of exponential regression, with an adjusted R-squared score of 0.50. This suggests a significant exponential correlation between the ratio of English-speakers of a book and the number of ratings received.

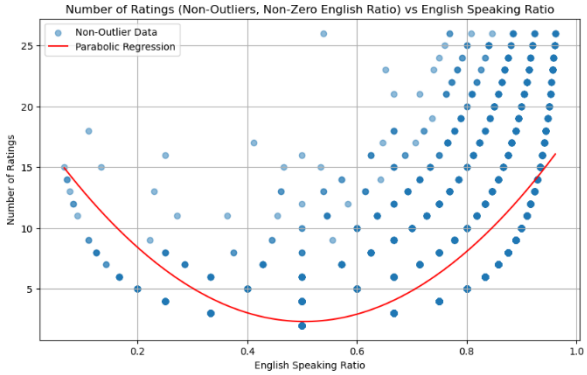

Figure 7: Scatterplot of the Ratio of English-Speakers to Number of Ratings of Books without outliers

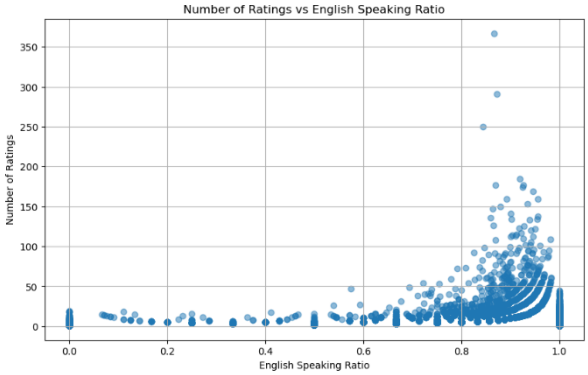

Figure 8: Scatterplot of the Ratio of English-Speakers to Number of Ratings of Books with outliers

When also considering books with a lower English-speaking ratio, the resulting plot (Figure 7) can be fitted to a parabolic curve with regression parameters (65.86438412, -66.44506919, 19.0730352) and an Adjusted R-squared score of 0.49. In combination with Figure 8, books tend to be more popular with an English-speaking ratio tending towards 0 or 1, but not if their user demographic is solely English-speaking.

3.3 Analysis of Popularity

Further analysis focused on user engagement, which is assessed by examining the number of ratings per user and the corresponding number of rating reviews. A scatter plot of these two variables (Figure 9) indicated a positive correlation, where users who rated more books tended to give higher overall ratings. This suggests that more engaged users are likely to be more satisfied with their reading choices, or possibly that more satisfied users are more engaged in rating activities.

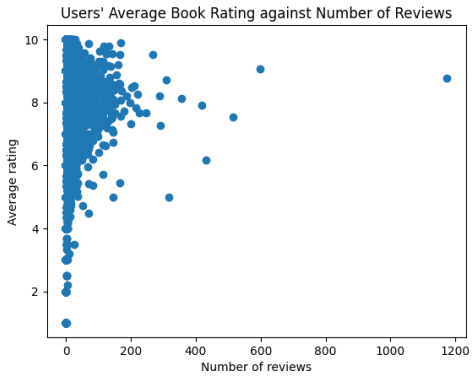

Figure 9: Scatterplot of Users' Average Book Rating against Total Number of Reviews

3.4 Correlation Analysis

Correlation analysis was performed to identify relationships between various factors:

- Rating and year of publication: The scatter plot (Figure 10) illustrates the relationship between the WLB-adjusted average book ratings and their respective years of publication. The data spans from the 1960s through the early 2000s, revealing several notable trends worth to be further discussed.

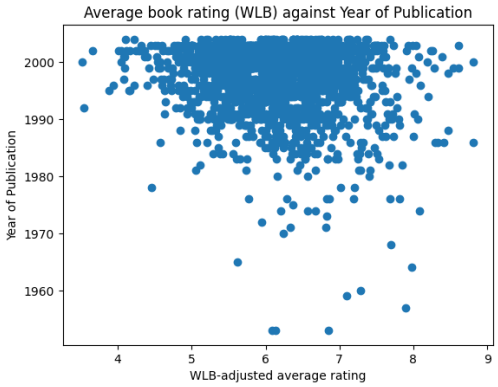

Figure 10: Scatterplot of Users' Average Book Rating against Total Number of Reviews per Year

- Rating and number of reviews: Figure 11 provides the relationship between the adjusted average book ratings and the number of reviews received per year. This scatterplot highlights several key aspects of reader engagement and its impact on perceived book quality. Notably, there is a dense cluster of data points with a high concentration of books receiving fewer than 5 reviews per year, which typically exhibit a wide range of ratings from approximately 4 to 9. As the number of reviews per year increases, there appears to be a trend toward higher average ratings, with books receiving more than 5 reviews per year generally maintaining ratings above 6. In addition, a significant observation is the presence of fewer books with a high number of reviews per year, yet these books tend to sustain higher average ratings, which could suggest that books with greater visibility and reader engagement are more favorably received. However, there are outliers, such as books with relatively few reviews per year but exceptionally high ratings, indicating that lesser-reviewed books can still achieve high perceived quality.

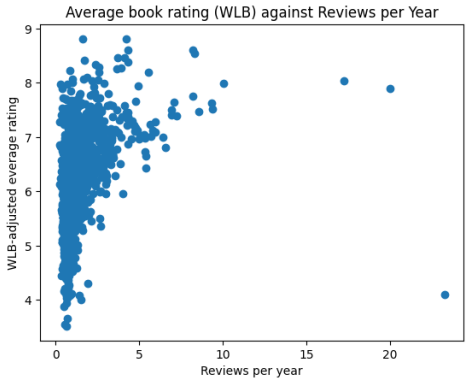

Figure 11: Scatterplot of Users' Average Book Rating against Total Number of Reviews per Year

4. Result

The analysis conducted in this report utilised machine learning methods to extract deep insights from the data, aiming to address the critical questions posed in the introduction regarding optimal stock management and effective customer engagement strategies. This section summarizes the key results derived from these models, focusing on their implications for decision-making within the online bookstore context.

1. Sentiment Analysis

   1.1. Objective: To identify the most popular book based on features such as the highest number of reviews, user country, publication year, and author popularity.

   1.2. Model Used: Sentiment Analysis – Via TextBlob and scikit-learn

   1.3. Key Findings:

   Analysis of book titles' sentiments reveals insights into popularity, particularly regarding preferences across age ranges. Figure 12 suggests how indicates that older readers tend to favor books with heightened sentiment polarity, be it positive or negative, over those with a neutral tone. Therefore, suggesting a preference for narrative variability. Furthermore, as the sentiment

score increases, reflecting stronger emotional resonance within titles, there is a correlation between increased ratings. Except in cases where the sentiment score is 0. This highlights a connection between emotional resonance and book popularity.

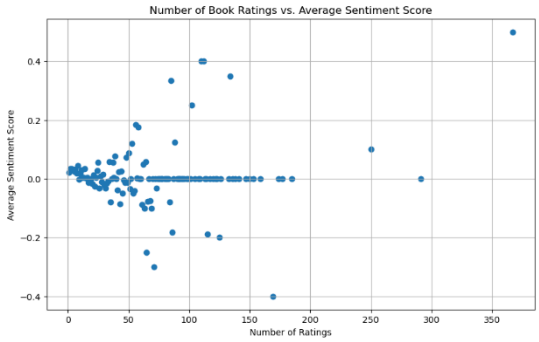

Figure 12: Scatterplot of Number of Book Rating against Average Sentiment Score

Furthermore, there is a strong correlation between the distribution of age ranges and sentiment scores. As shown in Figure 13, the density map shows how sentient score variance increases as we approach this age range.

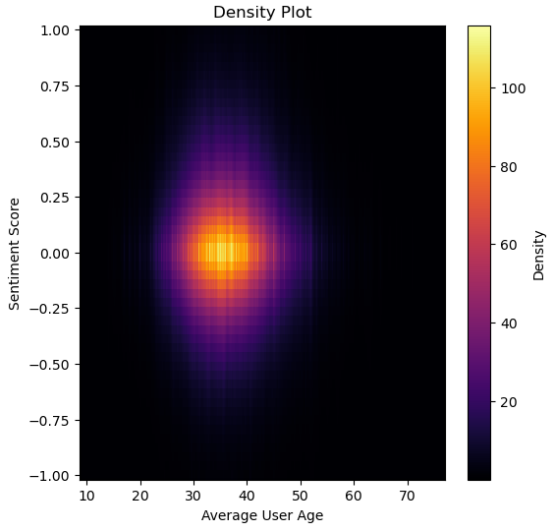

Figure 13: Density plot of sentiment score against average user age

It is especially relevant to note that this region of density in terms of average user age aligns with the density distribution found when comparing average user age and English speaking ratio. Thus suggesting, indirectly, as user age approaches this point, both sentiment score and English speaking level increase.

2. Nearest Neighbor

   2.1. Objective: To identify distinct segments within the user base to tailor marketing strategies more effectively.

   2.2. Model Used: Nearest neighbor method.

   2.3. Key Findings:

   By calculating the total rating count, further book recommendations are identified among the vast collection of books. Popular books, which received more ratings (above the threshold), were thus identified as books that would be the preference of the user which interested in books with similar ratings. These findings displayed different preferences and engagement levels, providing a basis for targeted marketing and stock selection strategies. Figure 14 provides five book recommendations for a particular book. The figure showcases the performance of the nearest neighbor, with identified five closest neighbors to be further recommended to the user who purchases the book with the title *To Kill a Mockingbird*.

```
for i in range(0, len(distances.flatten())):
    if i == 0:
        print('Recommendations for {0}:\n'.format(english_country_user_rating_pivot.index[query_index]))
    else:
        print('{0}: {1}, with distance of {2}:'.format(i, english_country_user_rating_pivot.index[indices.flatten()[i]], distances.flatten()[i]))
```

```
Recommendations for To Kill a Mockingbird:

1: Bridget Jones's Diary, with distance of 0.7922409190562208:
2: The Da Vinci Code, with distance of 0.8157933245815823:
3: The Lovely Bones: A Novel, with distance of 0.8236139520752827:
4: The Secret Life of Bees, with distance of 0.8348861442612808:
5: Life of Pi, with distance of 0.8439310921271841:
```

Figure 14. Code and example of book recommendations result

2.4. Implications of Results

The insights from these models can significantly enhance decision-making as follows:

2.4.1. Inventory Management: Knowing which books are likely to receive higher ratings allows the bookstore to prioritise these books in stock decisions.

2.4.2. Marketing Strategies: Understanding user segments enables personalised marketing, such as recommending new books with similar ratings or promotional discounts on associated books in certain countries, such as those in English Speaking countries.

5. Discussion and Interpretation

This section delves into a detailed interpretation of the findings obtained from the comprehensive data analysis of an online bookstore. Each insight is not only discussed in the context of its direct impact on the bookstore's operations but also in terms of broader trends in consumer behavior, technological influences, and market dynamics. The analysis brings forth a combination of expected trends and surprising patterns, each providing unique strategic value.

1. Significance Book Ratings on Publication Year

1.1. Finding: Books published around 2000 received a general consistency in book quality or rating criteria over time significantly higher ratings.

1.2. Interpretation:

1.2.1. This trend is reflective of the condition whereas the more recent publications, particularly those from the 1990s onwards, tend to cluster towards the higher end of the rating spectrum, which could indicate an improvement in book quality, changes in publishing standards, or possibly a shift in reader expectations and review practices over the years.

1.2.2. There is a visible concentration of ratings between the 6 to 8 range across all decades. Notably, books are sparse with very high ratings (above 8.5) before the 1980s, which might reflect the evolving nature of literary criticism or changes in the types of books being published.

1.2.3. The dispersion of points also becomes slightly denser towards the later years, suggesting an increase in the volume of books published or possibly in the number of reviews they receive, which could influence average ratings. This graph serves as a valuable visual tool for understanding how literary quality, as perceived by readers or critics, has potentially evolved alongside the publishing industry over the latter half of the 20th century and into the 21st century.

1.2.4. Strategic Implication: Focusing inventory and promotional efforts on the books published can capitalise on their current popularity, ensuring that the bookstore remains relevant and appealing to its most engaged customer segments.

2. Demographic Preferences and Their Impact on Book Selection

2.1. Finding:

By categorizing age values into distinct bins representing age ranges - children, young adults, middle-aged, and older adults - we uncover the role age plays in sentiment score analysis, as demonstrated in Sentiment Analysis (Figure 12). Notably, specific age demographics exhibit preferences for book titles.

Additionally, grouping English speaking countries including the United Kingdom, the USA, Canada, Australia, New Zealand, and Ireland allows for the derivation of an "English speaking ratio," enabling comparison of how country variability in book ratings influences popularity. As illustrated in FIgure 6, an exponential correlation exists between book popularity and the English-speaking ratio, particularly evident when focusing on the region of highest density concerning English speaking ratio versus average age per book.

This region, delineated by areas of density with ratings higher than average, provides insights into the most popular books while concurrently comparing country demographics and age.

2.2. Interpretation:

2.2.1. Reading preferences evolve with age, influenced by changes in interests, lifestyle, and even cognitive engagement needs. Younger readers might seek escapism and entertainment, which modern fiction genres provide, while older readers may prefer the intellectual stimulation or nostalgia offered by non-fiction and classics.

2.2.2. Marketing strategies that do not consider these preferences may fail to engage effectively with all customer segments, particularly those who live in English-speakers countries.

2.2.3. Unexpected Insight: Interestingly, the English-speaking ratio and number of reviews tend to group and clump at specific points, creating individual exponential distributions, suggesting that though have a slow start, as the ratio of English speaking increases, the number of book reviews also will increase significantly, reflecting its user engagement. Further analysis is needed on these individual exponential lines and their own adjusted R-squared scores.

2.2.4. Strategic Implication: Tailoring marketing and stock based on demographic data, which is the location of the user can enhance customer satisfaction.

3. User Engagement Levels and Their Correlation with Satisfaction

3.1. Finding: Higher engagement levels, measured by the number of reviews posted, correlated with higher overall ratings from users.

3.2. Interpretation:

3.2.1. The more engaged users are likely to be more satisfied with their reading choices, or possibly more satisfied users are more engaged in rating activities. Engaged users are likely more invested in the community aspects of the bookstore, such as participating in reviews, which can enhance their satisfaction as they feel part of a reader community.

3.2.2. Alternatively, satisfied users may be more inclined to engage by leaving reviews, creating a positive feedback loop.

3.2.3. Strategic Implication: Enhancing user engagement through community-building features like book clubs, review rewards, and interactive recommendations can foster a more active and satisfied customer base.

The insights gleaned from this analysis provide a multi-dimensional understanding of the operational and strategic dynamics of the online bookstore. The bookstore can optimise its current operations and strategically position itself for future development and transformation in a rapidly evolving market by linking detailed data analysis with user behavior and market trends. The findings underscore the importance of a data-driven approach in retail, highlighting how nuanced insights can lead to significant business advancements and customer satisfaction improvements.

6. Limitations and improvement opportunities

Despite the comprehensive insights gained from this analysis, several limitations must be acknowledged, which could impact the applicability and precision of the findings:

- Data Completeness: The datasets used may not encompass all relevant variables that influence book sales and customer satisfaction, such as detailed customer purchase history, external economic

factors, or comprehensive metadata on books (e.g., awards won). The absence of these variables might result in an incomplete analysis of factors driving book popularity.

- Temporal Dynamics: The analysis largely treats the data as static, not accounting for potential changes in customer behavior or market conditions over time. Seasonal trends and economic cycles, which could significantly affect book sales and ratings, were not considered.
- Potential Biases: The user data are self-reported and could contain inherent biases. For instance, users who actively participate in rating books might not be representative of the broader customer base. This selection bias can skew the results towards more engaged, possibly more positive experiences.
- Geographic Representation: The data may disproportionately represent certain geographic areas, particularly where the bookstore has a stronger presence. This could limit the generalizability of the findings to global markets.

Recommendations:

To enhance the robustness and applicability of future analyses, additional data sources should be integrated to fill gaps in the current dataset. Implementing time series analysis would help understand the temporal effects on sales and ratings. Efforts to mitigate bias through stratified sampling or post-stratification in the analysis phase could also provide more accurate reflections of the entire customer base. Finally, expanding the geographic diversity of the data would help generalise the findings across different markets, ensuring that the strategies developed are effective globally.

7. Conclusion

This report has thoroughly explored and analysed the extensive datasets from an online bookstore, unveiling significant insights into book sales, user preferences, and engagement trends. The findings have illuminated vital patterns, such as the pronounced preference for specific genres like Science Fiction and Fantasy among younger demographics and the impact of publication years on book ratings. The demographic-based segmentation and association analysis have also highlighted potential opportunities for targeted marketing and strategic inventory management.

Applying the machine learning method has provided a deep understanding of customer behavior and preferences, enhancing operational strategies. Specifically, the predictive models have shown how various book features influence ratings, while clustering has revealed distinct user groups with unique buying behaviors.

The strategic recommendations derived from this analysis, including focusing inventory on popular genres and tailoring marketing efforts to specific customer segments, have the potential to improve customer satisfaction and profitability significantly. By implementing these data-driven strategies, the bookstore can optimise its current operations and position itself firmly for future growth and adaptation in the competitive retail landscape.

In conclusion, this data-centric approach offers a blueprint for leveraging analytics to drive business decisions, underscoring the critical role of comprehensive data analysis in transforming insights into actionable strategies and tangible business outcomes.

References

Book-Crossing: User review ratings. (2021). Retrieved from https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset

IBM. (n.d.). IBM SPSS 28 Software. Retrieved from https://www.ibm.com/analytics/spss-statistics-software

Python Software Foundation. (n.d.). Python Language Reference. Retrieved from https://www.python.org