Final Project

Amazon Books

30 June 2024

Group: 2
Authors: Arda Cem Çakmak (2836118)
Haojia Lu (2761879)
Sinemis Toktaş (2837025)
Isabelle de Beijer (2741258)

Course Name: X_400645: Project Big Data
Period/Year: P6/2024
Instructor: Dr. Alessandro Zocca

A report submitted for the course Project Big Data

Table of Contents

1. Introduction

Amazon is an American multinational company founded by Jeff Bezos in 1994 (*Amazon Books*, n.d.). Being one of the largest e-commerce companies in the world, Amazon offers a wide variety of products which include books. Having established a subsidiary named Amazon Books, Amazon claims to be "the Earth's Biggest Bookstore" due to its largest selection of titles (*Amazon Books*, n.d.). Amazon Books, being the main retailer of books and other book related products, also has a membership system that allows its users to rate the books that they are buying. Having established that, our dataset is a collection of books, ratings, as well as the users of Amazon Books platform (see Dataset Description section for details). Using this data, we aim to explore, uncover, and understand the trends in books, authors and publishers as well as the behaviors of the users based on their age and location. Specifically, we aim to give an answer to the below listed research questions within each subsection of the Exploratory Data Analysis section (see Section 4).

4.1. Best and Worst Books, Authors, and Publishers for All and for Different Age Groups

- What are the top 10 best and worst rated books of all time?
- What are the top 10 best and worst rated authors of all time?
- What are the top 5 best and worst rated publishers?
- Which authors and books have the most ratings?
- What are the best and worst rated books per age group?
- What are the best and worst rated authors per age group?
- Which book do the users who do not like the best rated book, like instead?

4.2. The Effect of Location and Age on User Ratings

- Could the age of the user be affecting how a user rates?
- Could the location of the user be affecting how a user rates?
- Could the location of a user indicate if a user has a tendency to be an extremist rater?
- Could the age of a user indicate if a user has a tendency to be an extremist rater?

4.3. Gender Bias

- Is there a gender bias towards female authors against male authors?

4.4. Sentiment Analysis

- Is there a correlation between the sentiment of the title of a book and its ratings?

4.5. Prediction and Recommendation Models

- Is it possible to implement a user age predicting model based on our data?
- Is it possible to implement a user country predicting model based on our data?
- Is it possible to create a user recommendation system based on our data?

2. Dataset Description

Our dataset consists of three ".csv" files, namely, books.csv, ratings.csv, and users.csv retrieved from Kaggle (*Books Dataset*, n.d.). Detailed explanation of each file is as listed below.

1. **books.csv** | 77.79 MB | 271380 rows and 8 columns
   This file contains 8 columns which from left to right contain a unique International Standard Book Number (ISBN) value of the book, the title of the book, the name of the author of the book, the year of publication, the name of the publisher of the book, a link redirecting to the small-sized cover image of the book, a link redirecting to the medium-sized cover image of the book, and a link redirecting to the large-sized cover image of the book. Each row of this file is designed to hold the information of a unique book, thus, no duplicates are expected within this file.

2. **ratings.csv** | 30.68 MB | 1149781 rows and 3 columns
   This file contains 3 columns which from left to right contain a unique User Identity Number (User-ID) who rated a book, the unique International Standard Book Number (ISBN) value of the book that is rated, and a rating value of the book by the user ranging from 0 to 10 where 0 is considered to be an implicit rating and values from 1 to 10 are considered to be and explicit rating. The rows of this file may contain duplicate values that account for the ratings given by the same user on different time stamps as the file itself holds no information about the time that the rating was given.

3. **users.csv** | 12.28 MB | 278860 rows and 3 columns
   This file contains 3 columns which from left to right contain a unique User Identity Number (User-ID), the location of the user in a format of comma separated values of state, city, country, and the age value of the user. Each row of this file is designed to hold the information of a unique user, thus, no duplicates are expected within this file.

The above mentioned three files of our dataset can be seen in the figure below which contains exemplary rows of the files to visualize the structure of each file (see Figure 1).

Figure 1: Dataset files from left to right, ratings.csv, books.csv, and users.csv also containing exemplary rows.

3. Data Cleaning

Upon inspecting the dataset retrieved from Kaggle (*Books Dataset*, n.d.), we have observed various erroneous values in each file. Having observed so, each file has been cleaned of the erroneous values as explained below.

1. **books.csv** | 0.2% of this file has been cleaned
   3 rows that were parsed incorrectly were corrected. Invalid values from the *ISBN* column were filtered out according to the standards of ISBN (Pearce, 2021). The datatype of the column *Year-Of-Publication* has been changed to integer. 4 rows with missing values were dropped since it was an insignificant amount compared to the whole dataset. The columns *Book-Title*, *Book-Author*, and *Publisher* were cleaned from unwanted characters and capitalized.

2. **ratings.csv** | 66.64% of this file has been cleaned.
   38% of the column *Book-Rating* were implicit ratings which are not directly provided but inferred from user behavior. These ratings were filtered out. Ratings that were referencing non-existent books in the *books* dataset were filtered out.

3. **users.csv** | 77.03% of this file has been cleaned.
   Upon further looking into the datasets, many impossible age values were found. Values outside of the age range of (6 - 122) were replaced with NaN values. Since the average age at which a child learns to read is between 6 and 7 years old, 6 was chosen as the lower limit. It was assumed during the analysis that they could read, in order for the user to make a valid rating. 122 was chosen as the upper limit since it was the age of the longest-lived person on record (*List of the Verified Oldest People*, n.d.). After the filtering of age values, 40% of the *users* dataset had NaN age values. These values have been replaced with the median to increase robustness to the outliers in the age distribution. The values from the column *Location* have been parsed into the new columns of *City*, *State*, and *Country*. During this process, rows without the sufficient information were dropped. The invalid string values of columns *City*, *State*, and *Country* were replaced with NaN values. Rows with missing *City* and *Country* values were dropped, however the empty values of the column *State* were replaced with the *Country* value with the assumption that they were stateless countries. Users that did not have any ratings in the *ratings* dataset were filtered out.

4. Exploratory Data Analysis

4.1. Best & Worst Authors, Publishers, and Books for All and for Different Age Groups

4.1.1. Methodology

In this analysis, we first grouped the *ratings* by the column *ISBN* to calculate the number of ratings for each book and as a result we can see the distribution of rainting counts per book from the Figure 2 below.

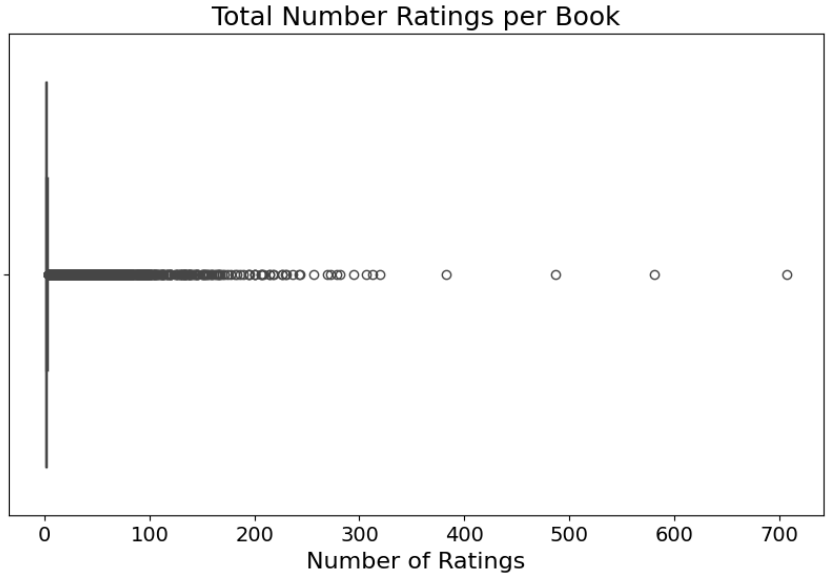

Figure 2: Distribution of Rating Counts per Book, a box plot where x-axis showing the number of ratings.

As it can be seen from the Figure 2 above, a minimum rating count threshold is necessary to ensure statistical significance as many ratings are smaller than 40. Rating counts this low should be filtered out to reduce the noise of low numbered ratings and increase reliability by including a larger sample size in the analysis. After filtering them out, we chose the min rating count threshold between Q2-Q3 or Q3-Q4 (depending on the question). This approach ensured that the rankings for the top and bottom 10 authors, books, and publishers were based on more substantial user feedback, providing a clearer view of user preferences and book performance in the dataset. Then, we merged the *ratings* dataframe with *books* dataframe using the 'ISBN' as a key to combine detailed book information with user ratings. We grouped the merged dataset by *Book-Author*, *Publisher*, *ISBN*, and age groups respectively for each question and then calculated the average rating and rating count for each book. After that, we used the rating count values to filter based on the minimum rating count threshold.

- To determine which authors, publishers, and books had the best/worst ratings, we identified the top and bottom entries in the sorted merged dataset.
- To determine which authors and books had the most ratings, we analyzed the rating count values of the merged dataset and identified the top entries.
- For evaluating the best and worst rated books and authors per age group, the users dataset was merged with the ratings and books datasets at the initial step. Ages were split into groups as such: 6-18, 19-30, 31-40, 41-50, 51-60, 60+. After the common analysis steps described before, we reached our results by identifying top and bottom performers per demographic.

- To identify alternative book preferences among users who disliked the highest-rated book, we used a bad rating threshold of 5. Then once we identified the users, we performed similar analysis steps described above and compared other ratings of these users to find the books they rated highly, revealing their different user preferences. During this analysis, we did not use a minimum rating count threshold as our sample size was much smaller.

4.1.2. Results and Discussion

Figure 3 represents the top and bottom-rated authors by average user ratings. Maud Hart Lovelace and Art Spiegelman top the list with an impressive 9.29 rating who are known for their work in children's literature and graphic novels, respectively. Dr. Seuss maintains a strong presence with a rating of 9.19 together with Bill Watterson. On the other end, Rich Shapero stands out with a low rating of 4.39, reflecting mixed reception to his experimental works like "Wild Animus." Similarly, Barbara Cartland, known for her romance novels, has a rating of 5.04, indicating varying reader opinions within the romance genre. Our results show a wide range of appreciation for different authors.

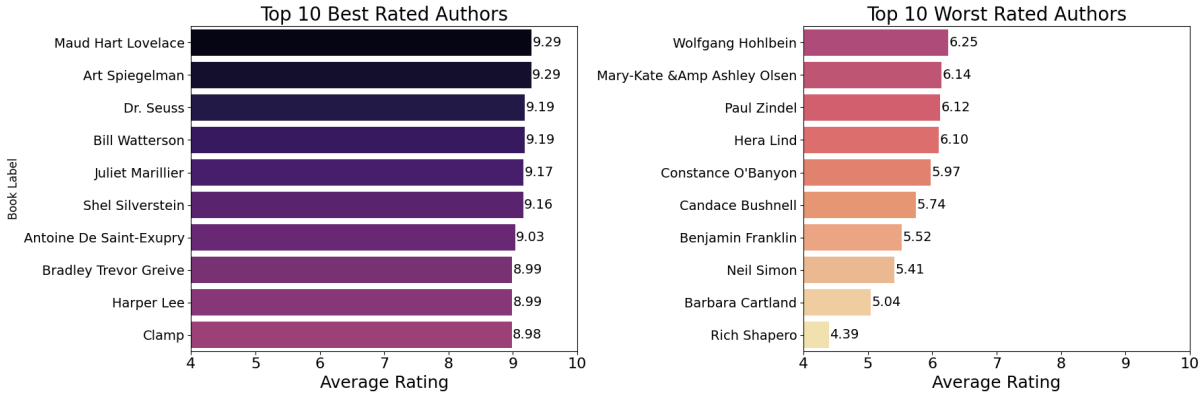

Figure 3: Best (left) and Worst (right) Rated Authors for All Ages, bar plots where x-axis showing the average ratings and y-axis showing the authors.

The top 5 best and worst rated publishers are shown in Figure 4 below. The ratings of the top publishers lie close to each other, near a 8.9 average. Carlsen Verlag GmbH has published the German Harry Potter and Twilight, hence its popularity with these famous titles. Arthur A. Levine Books is the US publisher for the Harry Potter book series. The worst rated publishers score between a 4.0 and 6.5. Worst publisher is Too Far, the publisher of Wild Animus by Rich Shapero, which is the only author they publish from, as Rich Shapero is the founder of this publishing company, (*Rich Shapero*, 2023).

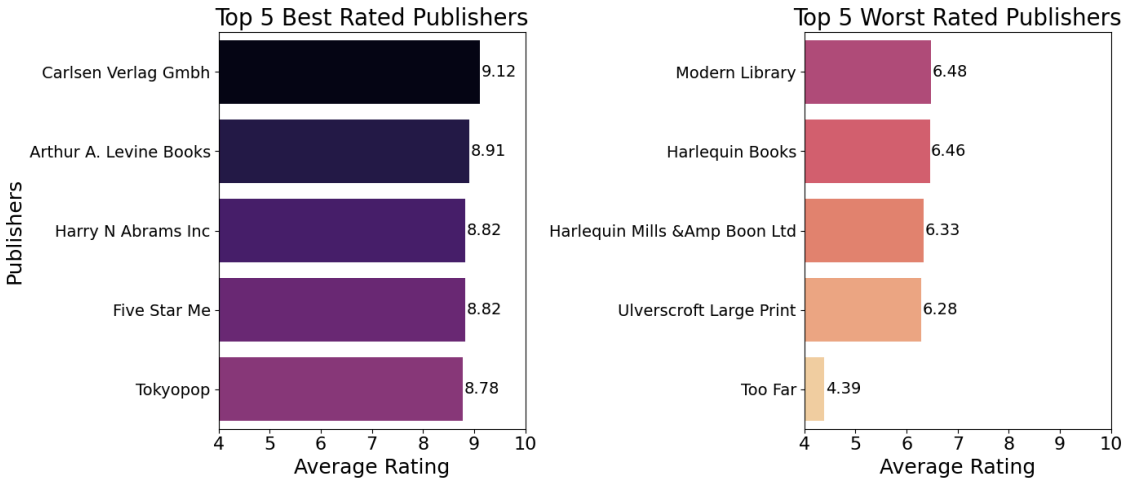

Figure 4: Top 5 Best (left) and Worst (right) Publishers, bar plots where x-axis showing the average rating and y-axis showing the publisher.

The results below showed the dominant book series of the top 10 best rated books were the Harry Potter and the Lord of the Rings series (see Figure 5), which was not surprising considering their popularity. The results of the worst rated books showed much more variety compared to top 10 best books. In the end, Return of the King was the best and Wild Animus was the worst rated book of our dataset.

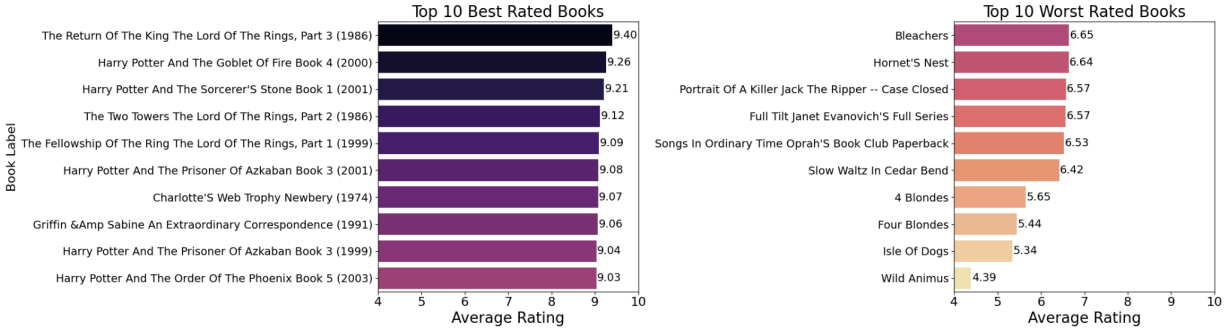

Figure 5: Top 10 Best (left) and Worst (right) Rater Books for All Ages, bar plots where x-axis showing the average rating and y-axis showing the book.

After analysis, we found 2 users that disliked the best rated book. Among the ratings of these users there were 11 books with an average rating score of 10. The titles of these books can be seen in the word cloud below. The results here showed that there may be a preference for more whimsical and less intense fantasy series like the Oz series for the users who disliked the most popular book (see Figure 6).

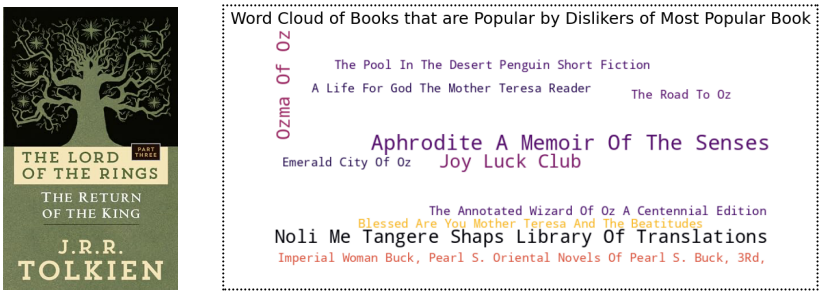

Figure 6: The Best Book (left) and The Books that are Liked by the Users Who Dislike The Best Book (right).

The results of our best and worst rated books per age group analysis showed that each age group had different favorite books with similar average rating scores, while the dominance of Lord of the Rings series was still noticeable. Whereas, Wild Animus appears to be consistently low-rated across all age groups, showcasing its widespread dissatisfaction among readers (see Figure 7).

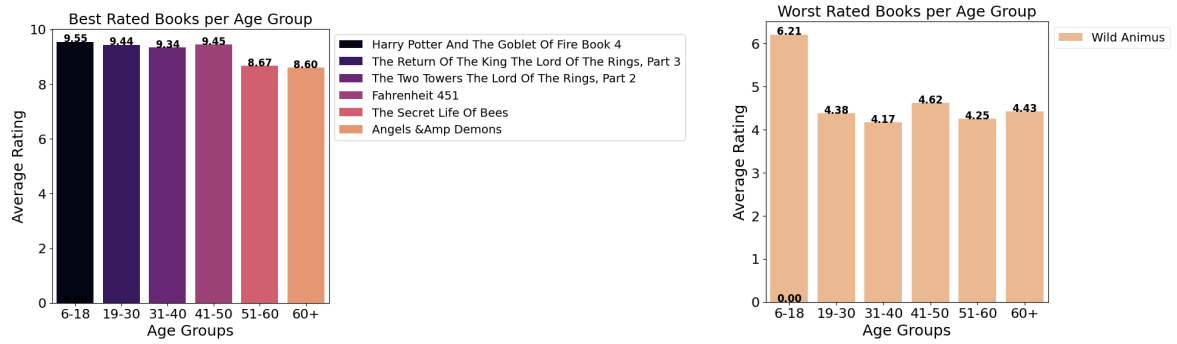

Figure 7: Best (left) and Worst (right) Rated Books per Age Group, bar plots where x-axis showing the age groups and y-axis showing the average rating.

The results of our best rated authors per age group analysis was very similar to the results of the best rated books per age group analysis since each age group again had different favorite authors with similar average rating scores (see Figure 8). While Rich Shapero, who is no one other than the author of Wild Animus, appears to again be a low-rated choice across most age groups, further demonstrating its common unpopularity among readers. We also see new author names for the least liked authors of the youngest and oldest age groups, maybe indicating a generational preference difference.

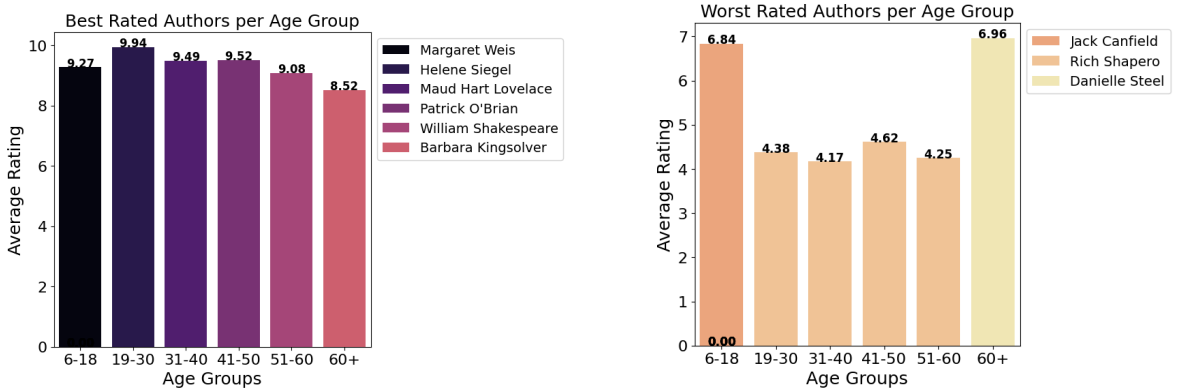

Figure 8: Best (left) and Worst (right) Rated Authors per Age Group, bar plots where x-axis showing the age groups and y-axis showing the average rating.

4.2. The Effect of Location and Age on User Ratings

4.2.1. Methodology

To investigate the effect of location and age on user ratings, the data was merged from the users, ratings, and books datasets. Ages were categorized into the same bins of 6-18, 19-30, 31-40, 41-50, 51-60, 61-122, and locations were analyzed by country. The average rating is then calculated per age group and per country. For both age and location, a Kruskal-Wallis test was performed to determine if there were significant differences in user ratings across groups. This non-parametric test was chosen because it does not assume a normal distribution of the data, making it suitable for comparing multiple independent groups with potentially skewed distributions. This approach provided insights into whether user demographics influenced their rating behaviors significantly.

Another question that was raised during the analysis of location and age on the ratings of the users was whether the location of a user or the group that the user is in have an effect on how extremely they rate a book. To investigate this question we have categorized the ratings into extreme and moderate where the extreme ratings are defined as those with values of 1 or 10 and the moderate ratings are those with values between 2 and 9. We have then categorized the raters into their countries and age groups and calculated their extreme ratings to moderate ratings ratio. For better generalizability of the findings for the whole country we have set a threshold that each country should contain at least 1000 ratings with also 100 ratings per category. To test if there is a correlation between both the country of origin and being an extremist rater, and the age group and being an extremist rater, we have created a contingency table and conducted a chi square test per correlation to be tested.

4.2.2. Results and Discussion

The analysis aimed to determine if age influences user ratings, utilizing a Kruskal-Wallis test on Figure 9, depicting average ratings across different age groups. The test yielded a non-significant result (H-statistic = 5.0, p-value = 0.416), indicating no statistically significant differences in ratings among age groups. This finding suggests that age alone may not be a determining factor in how users rate books, implying other variables or factors could play a more significant role in shaping user preferences and ratings.

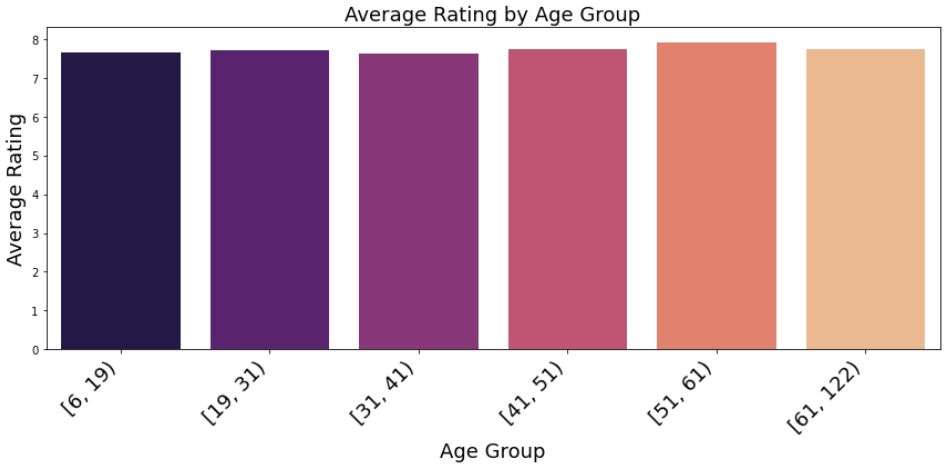

Figure 9: Average Rating per Age Group, a bar plot where x-axis showing the age groups and y-axis showing the average rating.

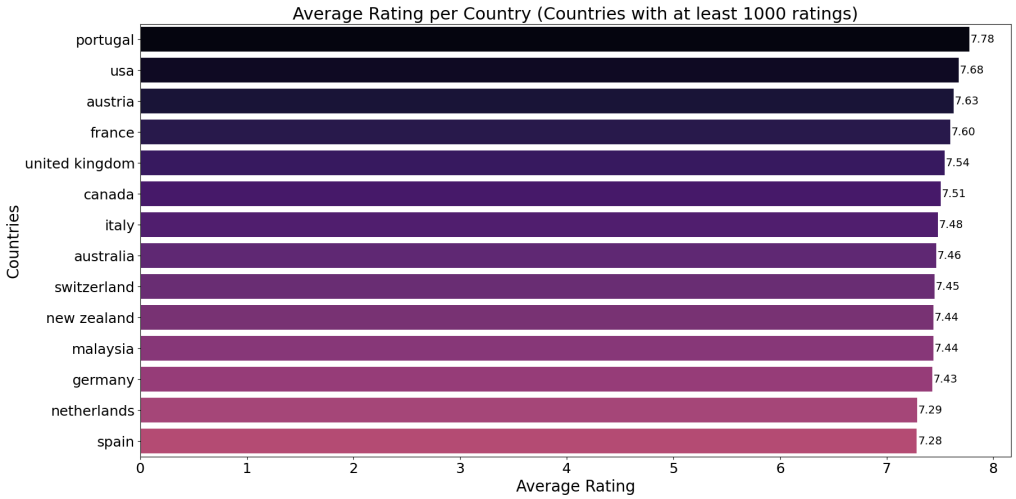

Figure 10: Average Ratings per Country, a horizontal histogram where x-axis showing the average ratings and y-axis showing the countries.

The investigation explored whether users' geographic locations influence their ratings, supported by Figure 10 displaying average ratings across different countries. A Kruskal-Wallis test was performed, yielding a highly significant result (H-statistic = 1178.31, p-value < 0.001), indicating substantial variability in average ratings between countries. Portugal emerges with the highest average rating of approx. 7.8, while Spain records the lowest average rating near 7.3, among countries with at least 1000 ratings. This disparity underscores the influence of geographical context on user perceptions and evaluations of books, suggesting regional factors may play a role in shaping rating behaviors.

The extreme to moderate rating ratio per age group can be seen as below with the most extremist age group being 6-18 with the ratio of 0.32 (see Figure 11). The chi squared test results with the value of 719.12 and with a p value of 3.60×10^{-51}, indicating that there is a significant relationship between the age group and the rating type ($p < 0.05$).

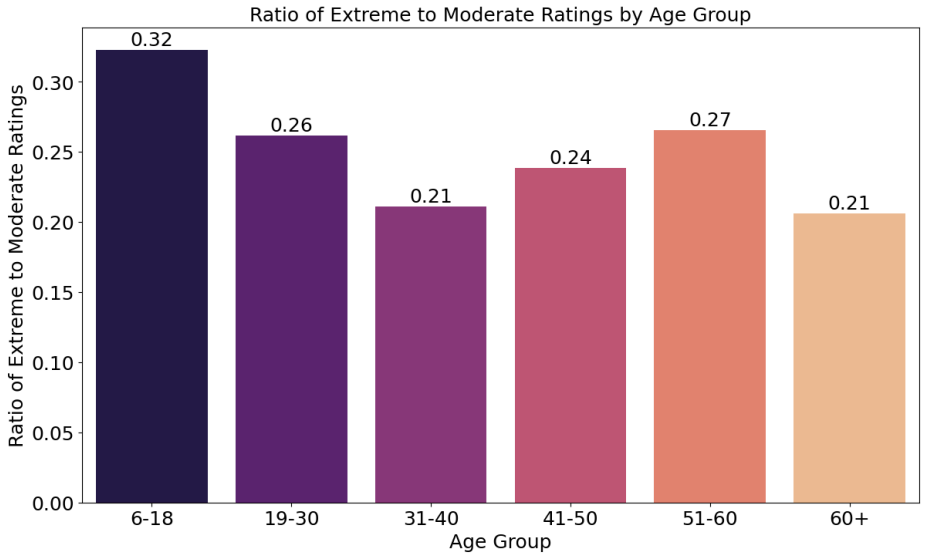

Figure 11: Ratio of Extreme to Moderate Ratings per Age Group, a bar plot where x-axis showing the age groups and y-axis showing the ratio of extreme to moderate ratings.

The extreme to moderate rating ratio per country can be seen as below with the most extremist country being Austria with the ratio of 0.37 (see Figure 11). The chi squared test results with the value

of 1510.98 and with a p value of 0.000, indicating that there is a significant relationship between the country and the rating type ($p < 0.05$).

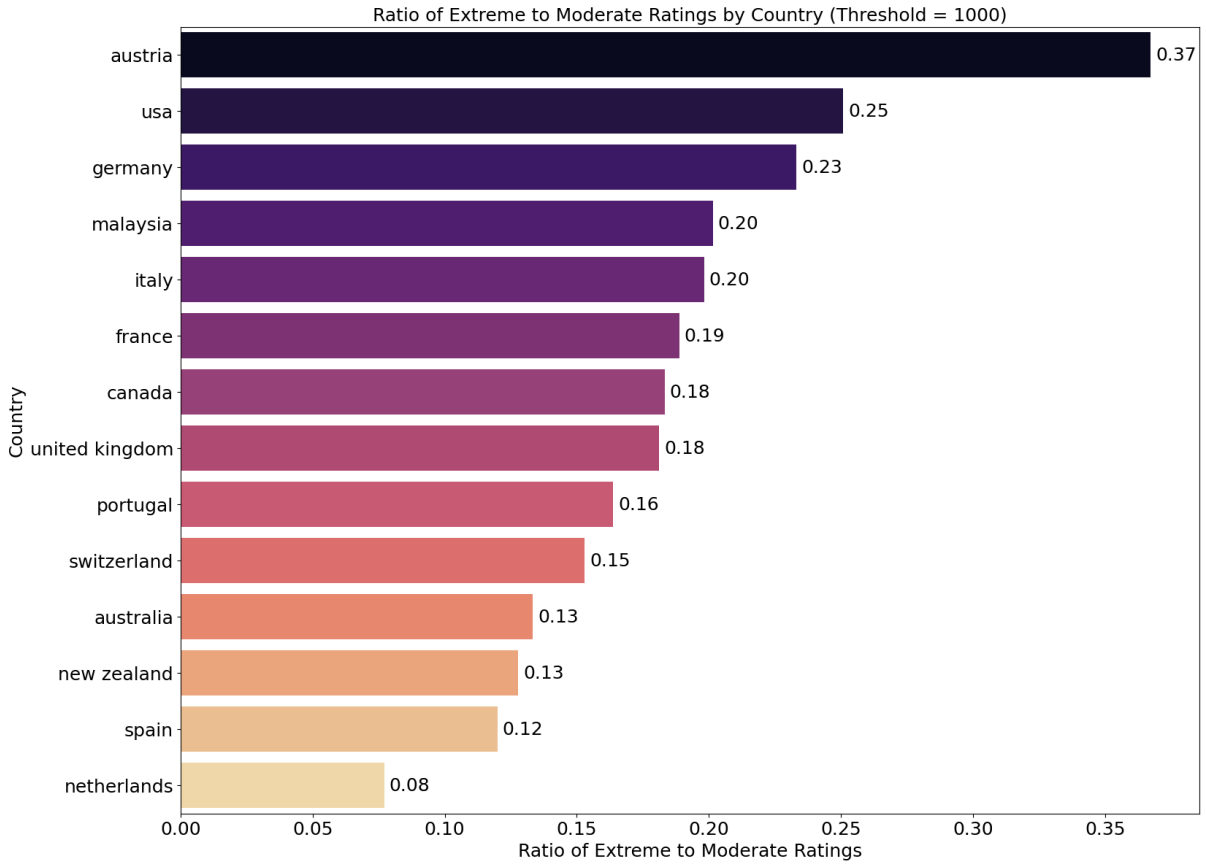

Figure 12: Ratio of Extreme to Moderate Ratings per Country, a horizontal bar plot where x-axis showing the ratio of extreme to moderate ratings and y-axis showing the countries.

The youngest age group having the users with highest probability of being extremist raters might be related to the development of emotional regulation towards the adulthood as it can also be seen from the figure (see Figure 12), as the age group progresses, their ratio of extreme to moderate rating decreases until the age group reaches an onset of 51, an indicator of adults having better critical viewpoints emotional regulation while the teenagers and elderly tend to be quickly-tempered, leading them to extreme rating. In terms of the country, cultural practices and the daily language used by the country could be the primary driver of why some countries rank the highest in the extreme to moderate ratio, however, to exactly know underlying reason, one must conduct a deeper analysis encompassing other variables related to the cultural practices and the daily life within those countries.

4.3. Gender Bias

4.3.1. Methodology

One of the trivial questions that arose upon seeing such a big data of author ratings was whether it is possible to categorize them into genders to see if there is a significant difference between ratings of female authors against male authors to observe whether there exists a gender bias against female authors, possibly leading them to have lower ratings. As our dataset did not contain any gender information of the authors, we had to consult to an external tool, and in doing so, we have utilized the *gender_guesser.detector* package (Pérez, 2016). *gender_guesser.detector* is a package equipped with a *Detector* class that has a method named *get_gender* that takes a string as an input, indicating a name of an individual, and outputs a gender based on the first name of the individual. It can result in six possible outputs, listed as, female, male, mostly_female, mostly_male, andy, and unknown where andy refers to an androgynous name and unknown refers to a name that is not found in the database of the package (Pérez, 2016). Having parsed all the names of the authors, each name has been assigned with a gender by utilizing the get_gender method, and afterwards, to have the most accurate results, only the female and male authors have been subject to the further analysis (the author gender distribution can be seen in Figure 13). Having done so, the rating distribution per rating for each gender has then been retrieved and plotted signifying how many ratings each gender gets per rating category from 1 to 10 (see Figure 14). One-Sided Mann-Whitney U Test has then been conducted on these distributions to see whether there is a significant difference between the rating distributions of female and male authors. To gain insight of the effect size of the difference, a Rank-Biserial Correlation has been conducted on these distributions. Having also seen that male authors are in abundance compared to female authors, the distributions are then normalized by dividing the total number of ratings to the total number authors of the respective genders (see Figure 15). Another One-Sided Mann-Whitney U Test has then been conducted on these normalized distributions to see whether there is a significant difference between the normalized rating distributions of female and male authors. To conclude the results, a Welch's t-test has been performed on the mean ratings of female and male authors.

4.3.2. Results and Discussion

The gender distribution that *gender_guesser.detector* package results in based on the names of the authors is as below (see Figure 13). It can be observed that there are 30504 female authors and 47529 male authors. The gap between the female authors and male authors can be seen as 17025. This gap indicates that male authors are in abundance compared to the female authors, a possible candidate for a contributor to the idea of gender bias for female authors.

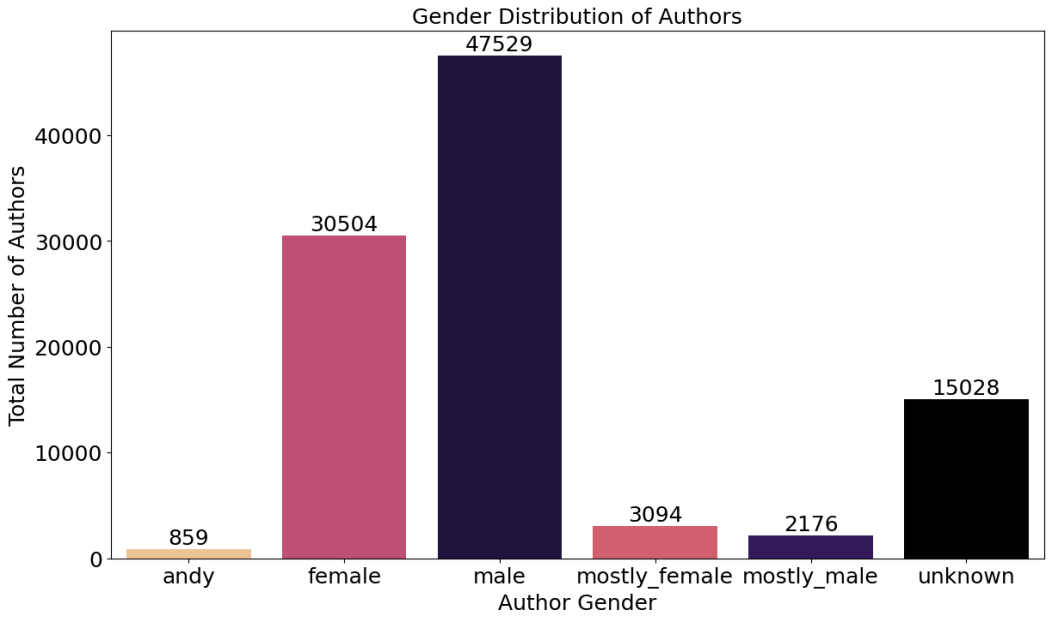

Figure 13: Gender Distribution of Authors, a bar plot where x-axis showing the gender and y-axis showing the total number of authors per gender.

To explore better if there exists a gender bias against female authors, the rating distribution of both male authors and female authors in comparison can be seen as below (see Figure 14). Here, it can be seen that the ratings of male authors surpass that of female authors for all possible explicit rating values from 1 to 10, and in fact, One-Sided Mann-Whitney U-Test conducted on these distributions results in a U value of 12319899979.5 and a p value of 0.000. This test indicates that there is, in fact, a significant difference between the rating distribution of female authors against male authors ($p < 0.05$), and further suggests that there might be a gender bias against female authors. Upon computing the effect size of this difference through Rank-Biserial Correlation, it has been found out to be -0.022, suggesting that female authors are slightly rated lower than male authors.

Figure 14: Rating Distribution of Authors per Gender, a bar plot where x-axis showing the possible rating values and y-axis showing the total number of rating entries per gender.

Considering the abundance of male authors compared to female authors and also considering the low effect size value, it was appropriate to normalize the distributions of the ratings by dividing the total number of ratings per gender and rating value to the total number of authors per gender to obtain the normalized distribution as below (see Figure 15). Here it is revealed that although male authors possess a higher probability to get higher ratings (see rating values between 8 to 10), they also possess a higher probability to get lower ratings (see rating values between 1 to 4), suggesting that when normalized, there might not be significant difference between the ratings of female and male authors. Another One-Sided Mann-Whitney U-Test, in fact, results in a U value of 47.0 and a p value of 0.425, suggesting that there is no significant difference between the normalized rating distributions of male and female authors as we fail to reject the null hypothesis (there are no differences between female author rating distribution and male author rating distribution) due to p value being higher than 0.05.

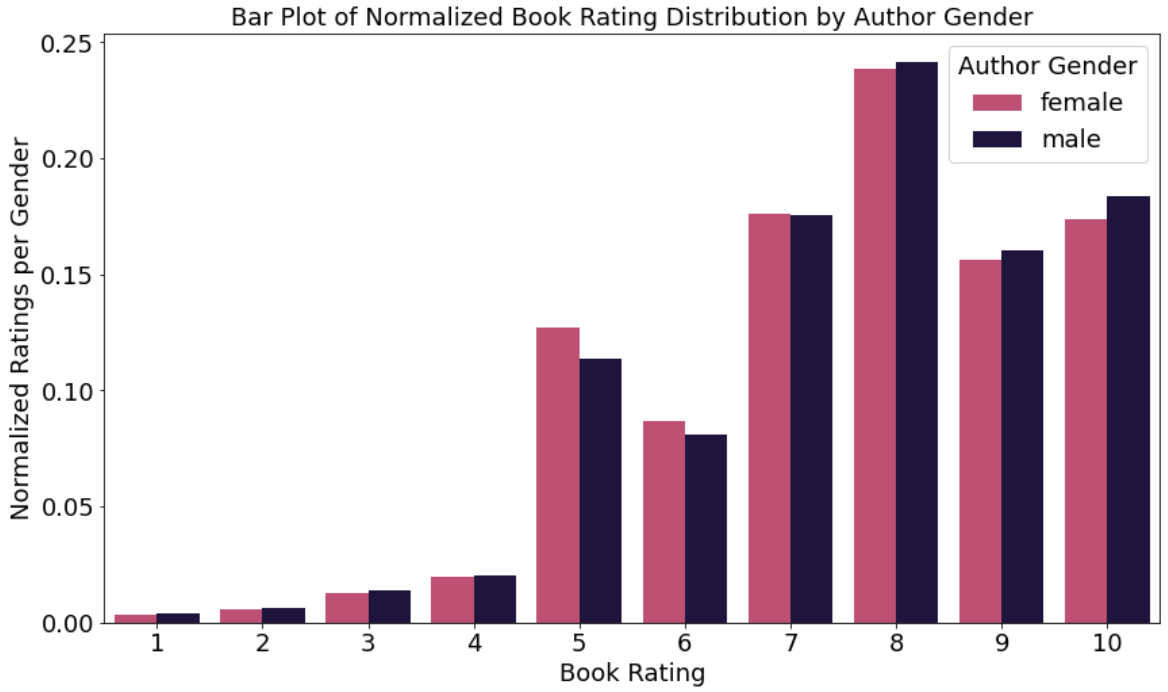

Figure 15: Normalized Rating Distribution of Authors per Gender, a bar plot where x-axis showing the rating values and y-axis showing the total number of rating entries per gender.

Considering the discrepancy amongst the Mann-Whitney U-Tests when the distributions are normalized and non-normalized, it was appropriate to see if there is significant difference between the mean ratings of female and male authors. As neither of the distributions are normally distributed (as both distributions fail the Shapiro-Wilk Test of normality), Welch's t-test was the appropriate method to see the difference between the mean ratings. The results indicate that the mean rating of male authors is 7.63 whereas it is 7.57 for female authors with a gap between them being 0.06. Welch's t-test confirms that this gap is significant with a t value of 9.363 and a p value of 0.00. This end result allows us to conclude that not only the male authors are in abundance against female authors (with a difference of 17025) but also there is a significant gender bias towards female authors against male authors leading them to have lower ratings although the difference is considerably small (with a difference of 0.06).

4.4. Sentiment Analysis

4.4.1. Methodology

One of the questions that arose upon seeing the titles of the books was whether there is a correlation between the sentiment of the title and the rating that a book gets, and if so, would a book with a positive or a negative title get higher ratings by appealing to the emotions of the rater? As our dataset only contained the title of the book, we needed an external source to extract the sentiment of the titles, and for that, we have outsourced Valence Aware Dictionary and Sentiment Reasoner (VADER) tool (Hutto & Gilbert, 2014). VADER is a "lexicon and rule-based sentiment analysis tool" that contains a class named *SentimentIntensityAnalyzer* that is equipped with a method named *polarity_scores* which takes a string as an input and outputs the polarity score by summing the valence scores for each word in the lexicon that sums up to a value ranging from -1 to 1, where -1 indicates the most negative sentiment and 1 indicates the most positive sentiment (Hutto & Gilbert, 2014). Having imported VADER, we have retrieved the explicit ratings and the book titles being rated, and processed them through polarity_scores to retrieve their scores. Having categorized the books into positive, negative, and neutral sentiment according to the documentation (Hutto & Gilbert, 2014), we have created a heatmap showing how many books there are per category for each rating value (see Figure 16). To test if there is a correlation between the title sentiment and the rating, we have computed the Spearman correlation coefficient to obtain the final result.

4.4.2. Results and Discussion

The heatmap of the total book numbers categorized by their sentiment and their ratings is as below (see Figure 16). It can be seen that each rating has the most abundant book in the neutral category, indicating that there is a tendency to choose book titles with a neutral sentiment over positive or negative sentiments. To explore how much of this tendency correlates with the ratings of the books, the Spearman Correlation coefficient has been computed resulting in a value of 0.0092 with p value being 0.0000000144. It can be seen from these values that although there is a significant correlation between the title sentiment and the rating of a book ($p < 0.05$), the magnitude of the correlation coefficient is negligibly low (note 0.0092).

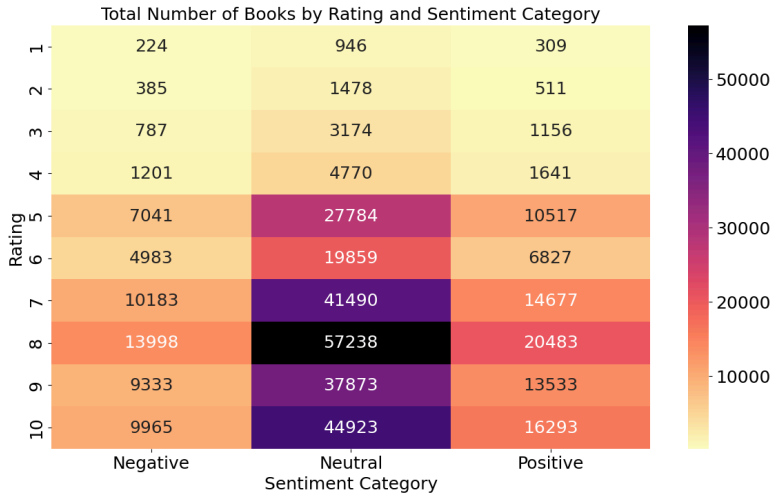

Figure 16: Total Number of Books per Rating Value and Sentiment Category, a heatmap where x-axis showing the sentiment category and y-axis showing the unique rating values, each cell indicates the total amount of books.

4.5. Prediction and Recommendation Models

4.5.1. Methodology

In order to predict the ages of users accurately, it is necessary to include an adequate amount of features and testing data. So we have only included users with 30 or more ratings in the training and testing datasets. By inspecting the dataset, it is clear that there are more than 3000 users who rate more than 100 books, so the dataset is large enough for training and testing. This ensures that we train and test the algorithm on a dataset that only includes users with enough information. Features that we use to predict the age include the age and the top 5 books that a user likes the most and least as well as the location of the user. Also, we aim to predict the country of origin of the users with their preferred books and other information. In fact, the datasets contain many inconsistent names of countries, for example, the United States of America can be written as US, U.S, USA, America, New York, etc. Since it is difficult to unify the names of countries here, we just keep the origin names from the user dataset. To predict the age of users, we use the multilayer perceptron as the model here. The MLP algorithm works on non-linear prediction problems and can deal with large input data. The structure of the MLP used in the task includes 4 layers, including 1 input layer, 2 hidden layers, and 1 output layer. We train the model with the training dataset for 10 times and plot the mean absolute error (mae) over runs of training, and the mae at the end of training is also printed. For predicting the countries, a MLP with the 1 input layer, 2 hidden layers and 1 output layer is used.

A recommendation system is designed to suggest books to users based on their past preferences. A collaborative filtering approach using Singular Value Decomposition (SVD) is used here. First, a subset of users who have rated at least 2 books with high ratings (>7) is created. This subset ensures the recommendation model is trained on users who have provided sufficient data for accurate predictions. Next, the recommendation model is trained using SVD to decompose the ratings matrix into latent factors representing user preferences and book characteristics. This decomposition captures patterns and similarities in user preferences, needed to make personalized recommendations. During training, the SVD algorithm computes three matrices: U (user-to-concept), Sigma (singular values), and VT (concept-to-item), which together represent the underlying structure of user-book interactions. These matrices are leveraged to estimate how users would rate books they haven't yet rated based on their historical ratings and similarities with other users. Finally, the recommendation model is tested by randomly selecting users from the subset and generating personalized book recommendations.

4.5.2. Results and Discussion

We adapt the dataset, define the MLP, and train the MLP with the adapted dataset for multiple epochs. From the figure below, it is clear that the model can achieve a stable mean absolute error after 2 epochs (see Figure 17). Although the training error still decreases after that, the validation error stays stable. This shows that the trained model overfits as the validation error stays high. At the end of training, the training is about 3 and the validation error is about 9, showing that the overfitting is very severe. The final testing error of this experiment is about 9.34, which is better than the model that is trained only on average rating and number of ratings. This result shows that, although our prediction is not accurate enough, it can approximate the age with the preferred books of a given user.

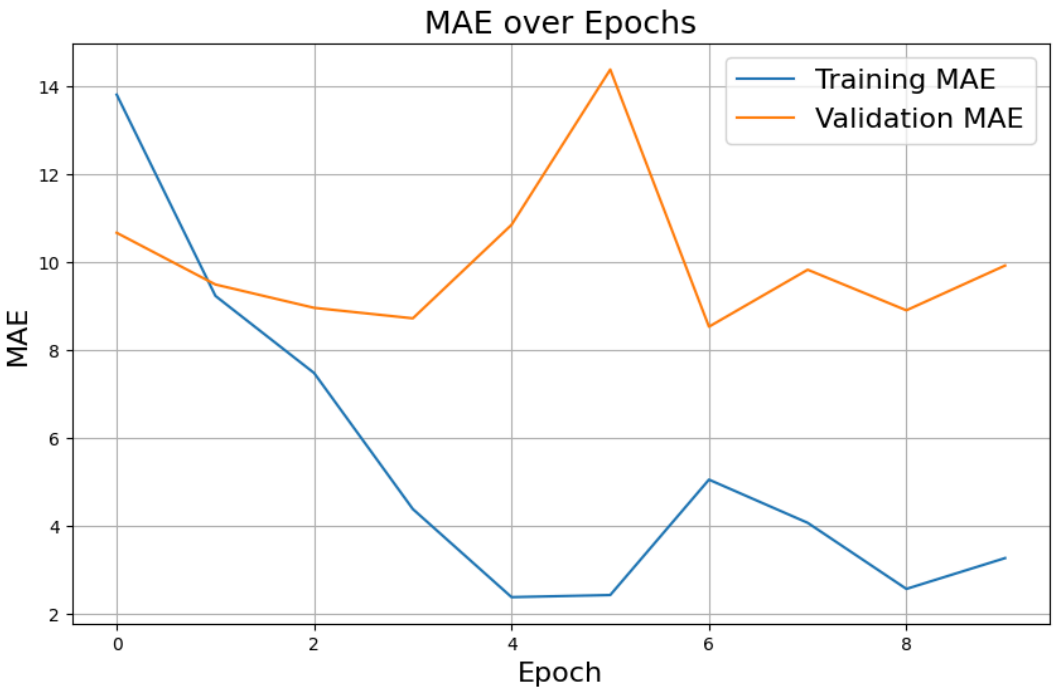

Figure 17: The Mean Absolute Error of the Model over Epochs of Training.

It seems that the prediction of age is not as accurate as we expect in this task. The reason why this happens is that, first, the prediction of age is a very complicated issue, and the connection between age and reading preferences may not be strong enough to support a very accurate prediction of age. In particular, the curse of feature space is the most important reason that makes this task difficult. There are more than 270,000 books within the dataset, and people from the same age group only share a very small intersection of preferred books. To deal with categorical features, one-hot encoding is used, making the feature space very large. For example, if there are 20,000 preferred books in the training dataset, 20,000 new features are required to represent the list of preferred books of each user. This also holds for preferred publishers and countries of origin. So, a simple multilayer perceptron is not sufficient to achieve this in our case. A possible solution is to analyze the age group of books separately, then use this information to analyze a user from his or her book lists. Another factor that limits the performance of our model training is the limitation of the size of the computer memory. During the training of the model, more than 2 GB of memory space is utilized, and due to the size of our devices' memories, we cannot build a larger MLP model that contains more neurons and hidden layers. In practice, a rule of thumb for the number of hidden layers is about 2 / 3 of the input layer plus the size of the output layer. But in this case, the size of the input layer is higher than 50,000, while the size of the hidden layer is only 4096. If the machine supports, defining a MLP of a suitable size may improve the result. In contrast, the prediction of countries works better. On the training dataset, the accuracy of prediction keeps stable after training for 3 epochs and is about 60%. The final testing accuracy of the prediction is also about 75% (see Figure 18). Although this accuracy is not as high as expected, it shows that the model can predict a person's country from his or her reading preference more accurately than just guessing randomly. Considering the class imbalance in this task, an accuracy of 75% can be considered as a good result. As predicting the country of a person is also a very complicated task, we believe this accuracy is acceptable for this task. Similarly, including more features, building a larger MLP network, or using other powerful algorithms like evolutionary algorithms may be a resort to the issue of low accuracy.

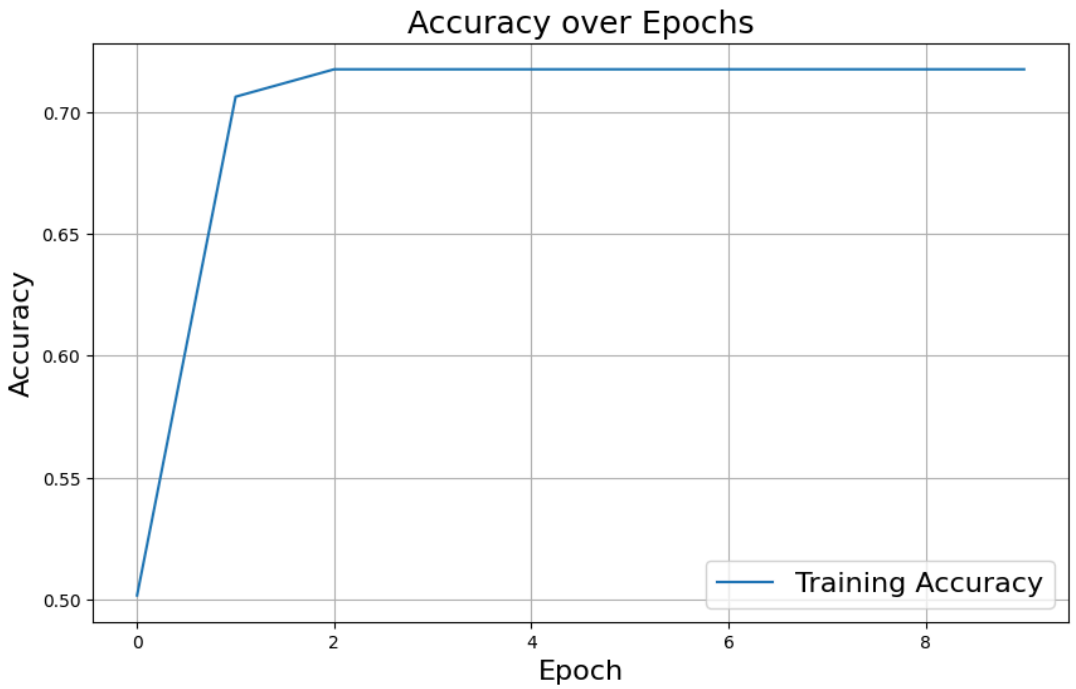

Figure 18: The Training Accuracy over Epochs of Training.

Due to the intensive computations required by the Singular Value Decomposition (SVD) approach used in the recommendation system, significant computational resources are necessary. SVD involves matrix factorization and iterative operations on large datasets here, demanding substantial memory and processing power. As a result, the model encounters continual kernel crashes during execution due to memory constraints or processing limitations. These crashes prevent the completion of model training and hinder the generation of reliable recommendations. Thus, despite its theoretical promise, practical challenges have prevented the model from being tested and implemented. Future research could explore the recommendation system by searching for simplified models and alternative algorithms with lower computational demands. Adjusting parameters and testing on diverse computing platforms might also alleviate the current issue of kernel crashes. Due to time constraints these possible solutions have not been examined.

5. Conclusion

In this analysis of Amazon Books' dataset, we explored various perspectives of user behavior, book preferences, and demographic influences on their ratings. Our research into the best and worst rated books, authors, and publishers showed notable trends and disparities across different age groups. For instance, all-time favorites like the "Harry Potter" and "Lord of the Rings" series emerged as top performers, while experimental works such as "Wild Animus" consistently received lower ratings across all demographics. These findings underscored the diverse preferences among readers and showed the importance of considering user feedback when it comes to decision making for publishers.

Looking at the demographic factors, we found that while age did not significantly influence rating behaviors, geographic location did play a role. Users from different countries demonstrated varying average ratings, suggesting regional factors shape their perceptions of literary works. Furthermore, our analysis indicated a correlation between younger age groups and higher tendencies towards extreme ratings, underscoring potential differences in emotional regulation across demographics.

Gender biases in author ratings were also inspected, revealing a landscape where female authors tended to receive slightly lower ratings compared to their male colleagues. This disparity, albeit small, raised questions about the underlying factors contributing to such biases (in literary evaluations). Additionally, sentiment analysis of book titles suggested a weak correlation between title sentiment and user ratings, highlighting the complex interplay of emotional appeal and actual content in influencing reader perceptions.

Finally, our prediction and recommendation models provided insights into the potential for predicting user demographics and offering personalized book recommendations based on historical preferences. While these models showed promise, particularly in predicting user countries of origin, challenges such as feature complexity and model overfitting revealed the need for further refinement.

In conclusion, this analysis not only sheds light on the diverse landscape of reader preferences and behaviors on Amazon Books but also offers the potential for leveraging data-driven insights to enhance user experiences and inform strategic decisions in the e-commerce and publishing industries. Future research could explore additional variables and refine methodologies to uncover more complex insights into literary preferences and user engagement in online book platforms.

6. References

Amazon Books. (n.d.). Amazon.com. Retrieved June 30, 2024, from https://www.amazon.com/books-used-books-textbooks/b?ie=UTF8&node=283155

Books Dataset. (n.d.). Kaggle. Retrieved June 30, 2024, from https://www.kaggle.com/datasets/saurabhbagchi/books-dataset

Pearce, B. (2021, August 31). *Anatomy of a 10-digit ISBN*. ISBN Information. Retrieved June 30, 2024, from https://isbn-information.com/the-10-digit-isbn.html

List of the verified oldest people. (n.d.). Wikipedia. Retrieved June 30, 2024, from https://en.wikipedia.org/wiki/List_of_the_verified_oldest_people

Pérez, I. S. (2016, December 5). *gender-guesser 0.4.0.* PyPI. Retrieved June 30, 2024, from https://pypi.org/project/gender-guesser/

Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. GitHub. Retrieved June 30, 2024, from https://github.com/cjhutto/vaderSentiment