Finding correlations between User Age and User Rating

Group W06G4

Cathy Wu	**Piripi Martin**	**Lukas Krunic**
1453041	1462615	1460182
cawu1@student.unimelb.edu.au	piripim@student.unimelb.edu.au	lkrunic@student.unimelb.edu.au

Executive Summary

This report explores the relationship between user age and book ratings to inform recommendation strategies for online bookstores towards their buyers. Employing data preprocessing techniques such as scaling, encoding, discretizing, dimensionality reduction and text processing, alongside machine learning methods such as K-Nearest Neighbors and Linear Regression, we found correlations between user ages and their respective ratings for books. We found that there was a discernible trend of increasing preference for older books among older age groups, plateauing at seniors, where there was a slight decline in preference. This suggests the importance of targeted marketing strategies, advocating for older books to be promoted to adults and middle-aged buyers, and more modern books to be pushed towards children, teenagers, and seniors.

Introduction

The central question addressed in this analysis is how user age influences book preferences and ratings. Leveraging data from an online bookstore, which includes user demographics, book details, and book ratings by users, this report explores the relationship between these variables. By employing data preprocessing techniques alongside machine learning algorithms such as K-Nearest Neighbors and Linear Regression, we seek to find trends between user age and ratings that can guide online bookstores in tailoring their books to a diverse customer base, ultimately improving sales.

Methodology

Data Preprocessing

1. Handling missing data, format errors and range errors

In our handling of missing values, we decided to exclude rows with missing values from our calculations in order to maintain accuracy in the data. The primary reason for excluding these rows was to prevent the introduction of biases into our analysis. In order to handle format errors, we remove non-alphanumeric values from our target columns to ensure that all the data that we're using is in uniform format. Using the lambda function and regex, we replace all matches of non-alphanumeric values with an empty string, so that values such as ' 26" ' and "The Last Command (Star Wars: The Thrawn Trilogy, Vol. 3)" are changed to "26" and "The Last Command Star Wars The Thrawn Trilogy Vol 3" respectively. Moving on to handling range errors, we removed all publication years of 0 and limited the maximum year to 2024. Furthermore, we assumed that a reasonable age to read books would be between 5 and 100, and thus we set our age range of users to be between these two values.

2. Text Processing

We then applied text processing techniques to our data for the purpose of improving our machine learning models, which rely on text-based classification. First, we applied case-folding to the data in our target column. Then, we tokenized our text so that we can remove stop words. Following this, we then applied lemmatization to reduce words to their lemma. We decided to use lemmatization as opposed to other techniques such as stemming to ensure that the resulting words are valid and meaningful, i.e., the title "Beauty Fades, Dumb Is Forever: The Making of a Happy Woman" lemmatizes to "beauty fade

dumb forever making happy woman", which preserves the original meaning of the title whilst also allowing for more accurate machine learning classification. Furthermore, we found that there were a few words, which, due to data input inconsistency, would mislead the machine learning algorithms. As such, words like "novel", "book", and "paperback" were removed from the book titles using regex. In this way, a book like "Visions of Sugar Plums: A Stephanie Plum Holiday Novel" would not be classed as similar to "Kiss of the Night (A Dark-Hunter Novel)", thus improving the precision of our machine learning algorithms.

3. Text Vectorization

We used TF-IDF vectorizer to process book titles, transforming the processed text of book titles into numerical representations. This process was necessary for machine learning, particularly the K-Nearest Neighbor algorithm. This is because K-NN is a distanced-based algorithm, and thus can only be calculated between numerical data points. We chose to vectorize our text using TF-IDF over other methods, such as encoding, due to its ability to capture the nuances of the title since it gives higher weights to words that are more frequently occurring.

Machine Learning Techniques

1. K-Nearest Neighbor

We used the K-NN (K-Nearest Neighbor) machine learning algorithm to create a model to predict the rating that users will give books based on their age, the book's title and the book's publishing date. We used it for its simplicity, along with the finite value of the dataset it was trained on. For a relatively larger dataset, other models should be considered.

We used a relatively small k value of 3, to ensure flexible decision boundaries and to prevent overfitting due to ignoring smaller potential patterns in the data.

In the process of implementing the K-NN algorithm, we first had to transform the data that we input into it. For the "Year-of-Publication" and "User-Age" inputs, we had to reshape the values in each input into two dimensional arrays using the *.reshape()* method from the NumPy library. For the "Book-Title" data, we used the method of vectorization using the *TfidfVectorizer()* method from the scikit-learn library. This method transformed the incompatible string types into numerical format, specifically, into the form of vectors, which could then be inputted into the K-NN algorithm. The next step was to categorize the book ratings into low, medium, and high (low being ratings between 1-3, medium being 4-7 and high being 8-10), and save them into a new column named "Rating Category". We then used *hstack* to combine these three data sets and fit the model on the training data as the x train, with Rating Category as the y train.

2. Linear Regression

We used the linear regression model to investigate the relationship between age and preferred recency of publishing date. We used this model to find a tangible value that we could associate with preference. In our case, this value is the Coefficient of "year of publication", where a higher value signifies a preference towards newer books, and a lower value signifies a preference towards older books. This interparabilty was important in determining a suitable model. Additionally, Linear Regressions simplicity and efficiency were advantageous to us.

In the process of implementing the Linear Regression algorithm, we first had to reshape the data in the "year of publication" data. We then fit the model to the publishing date and rating values using the *linearregression()* method from the scikit-learn library. We then recorded the regression model coefficient of the year of publication. We repeated this for ratings from the age ranges of children to senior, recording the corresponding coefficients, attempting to observe a relationship between age and

susceptibility to publishing date in relation to book rating. We plotted the predicted rating for each data point, and plotted the regression line on the same graph.

Data Exploration & Analysis

Finding correlations in data

To address our research question of how the age of a user affects their rating of a book, we first decided to examine the different attributes of book that are provided to us, being title, author, year of publication and publisher. In particular, we focused on the specific attributes of year of publication and book title, as we believed these would yield the highest degree of correlation. This is because we assumed that younger users would be more likely to read books published more recently, compared to older readers. Further, we also assumed that similar titles would correspond to similar genres of books, which would then appeal to different ages.

In our exploration of the correlations between user and book attributes, and their corresponding ratings, we decided to group both user age and ratings into buckets. For user ages, we broke them into 5 separate buckets: child (5-12); teen (13-19); adult (20-39); middle-aged (40-59); senior (60+). For ratings, we broke them into 3 separate buckets: low (1-3); medium (4-7); high (8-10).

1. **Correlation between age, date of publication and rating**

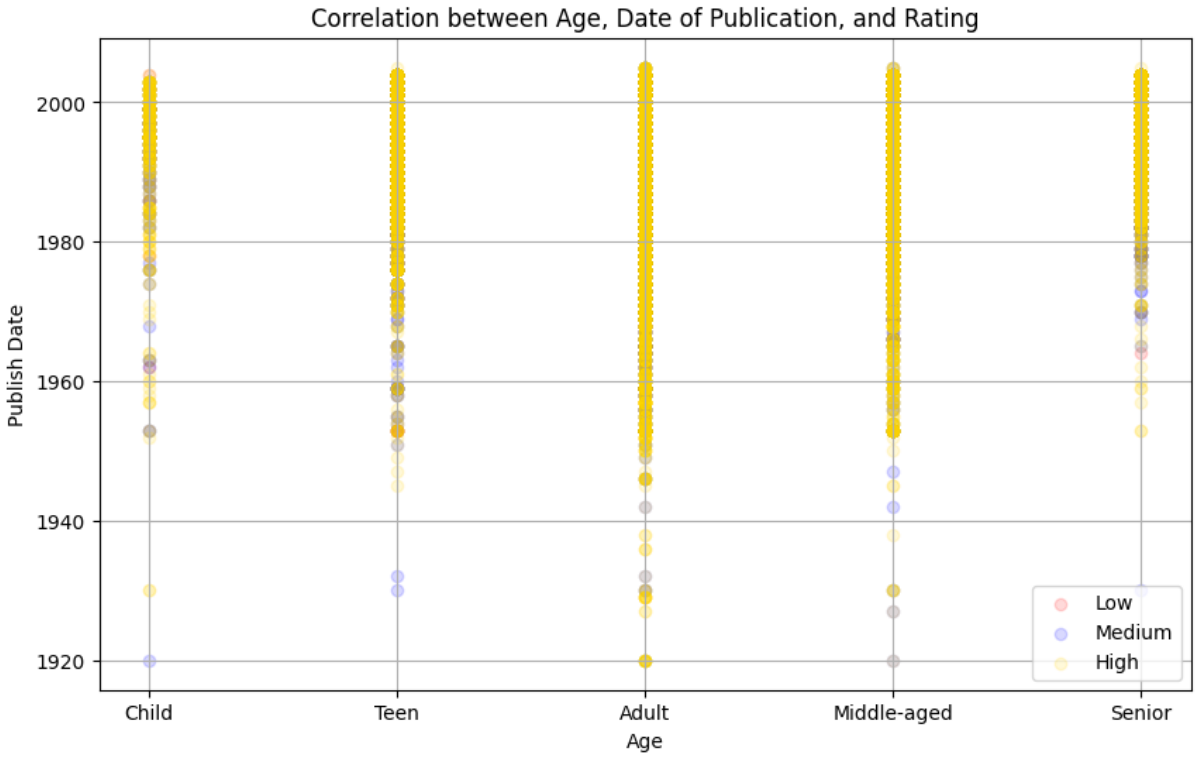


Figure 1: Graph depicting the relationship between how users in different age buckets rate books, depending on their year of publication.

This scatter plot shows the different age groups and the relation between these groups and the publish date. It exhibits the books each age group is reading, and the rating that people are giving these books. From the graph, we can clearly see that those in the child age range are very rarely reading books that are published before 1960, which is very similar to those in the senior age bracket. Conversely, those in the adult and middle-aged age brackets are reading a lot more books from the years before 1960 and even some before 1940, while teens still read some books from before 1960 and not as many before 1940. We can conclude from the graph that for children, more modern books are being read rather than an older style of writing, while for senior readers, they are potentially reading newer books since

they have already read older books and are now limited to more recently published books in order to read what they're interested in. Other reasons could include a lack of reading skills in both of these age groups and thus need to read a more modern style of writing. As for the other 3 age groups, there is a much wider spread of books read, especially in the adult age group. This shows that adults expand their reading skills and can read and enjoy all styles and ages of writing. Teenagers are still more interested in reading more modern books as they are yet to expand their reading taste, and middle-aged people are now narrowing down the types of books they want to read.

2. Correlation between age, book title and rating

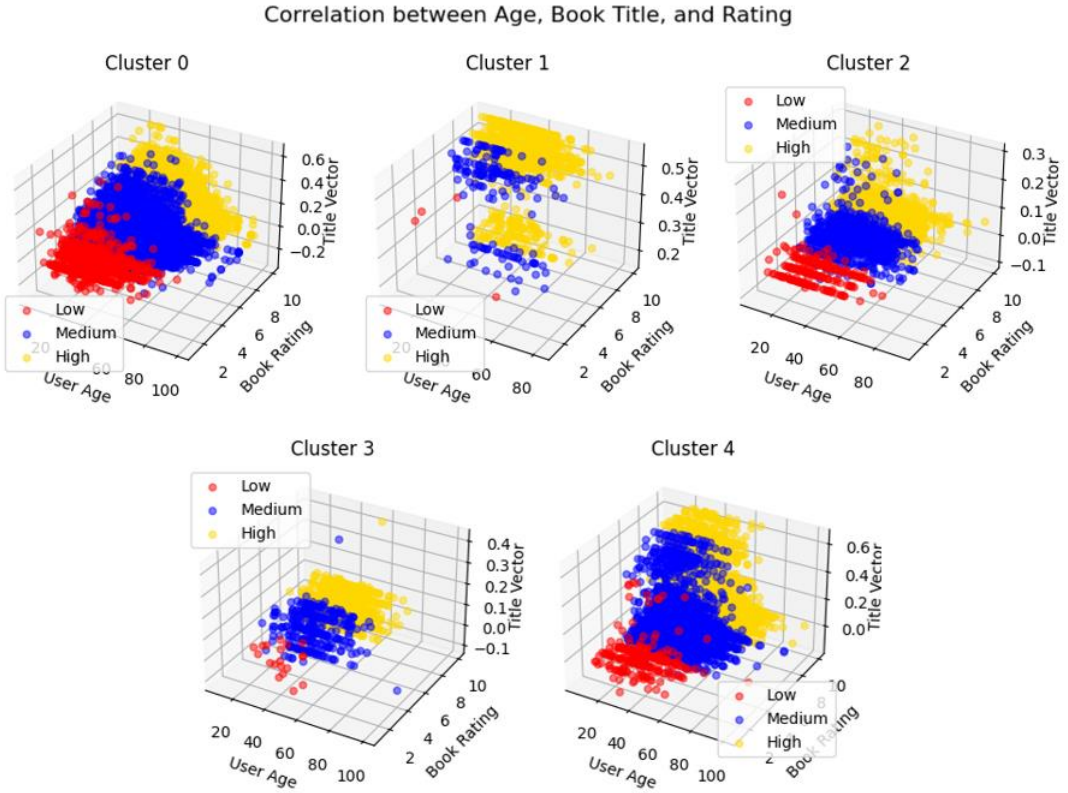


Figure 2: Graphs depicting the relationship between how users in different age buckets rate books, depending on their title.

These scatter plots show the relationship between user age, book title and rating. The axis "Title Vector" represents book titles upon TF-IDF vectorization and SVD truncation. The book titles are grouped into different clusters based on KMeans clustering with five clusters, which are displayed below.

Table 1: Five randomly sampled book titles allocated to the five clusters.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
red fox	science harry potter magic really work	bridget jones guide life	secret way	love trouble story black woman
boy tale childhood	j k rowling wizard behind harry potter	proust change life	secret history	dark river heart
wielding red sword incarnation immortality	trouble harry	secret diary anne boleyn	secret life bee	diet new america food choice affect health happiness future life earth

moon sun	harry potter sorcerer stone harry potter	shrub short happy political life george w bush	night life	marry man dance
catfantastic	harry potter et la chambre de secret	life time michael k	wild life	late dont start create second life forty

From randomly sampling distinct book titles in each cluster, we can evidently see that some clusters are more related than others. Reading from these sampled titles, we can visibly see that Cluster 1 represents a group of titles, mainly relating to the Harry Potter series. Applying this information to Figure 2, we can see that users of all ages mainly rate these books in the medium to high ranges, suggesting that they are popular for all demographics. Likewise, Cluster 2 seems to have few low ratings. From the titles in this cluster, it seems that a lot of them include "life", and have names, which suggests that this cluster could potentially encompass the genre of autobiography. The plots are more concentrated in the adult to middle-aged age category, suggesting that they are the main demographic for these books. Looking at Cluster 0, the titles seem to indicate books in the fiction region, and words like "boy tale", "catfantastic" and "sword…immortality" seem to be quite juvenile. This corresponds to a more left-aligned Figure 2, suggesting these books are being read by a younger age demographic. There also seems to be a large portion of users rating books in this category lower, which could suggest that the quality of these books are not as high. This aligns with our theory that these books are mainly targeted towards a younger audience, and could therefore not be of the greatest writing ability.

However, we have also noted that there doesn't seem to be extremely clear trends in ratings between age demographics. This could potentially be due to the decision to only choose 5 clusters, which does not accurately represent the large number of genres of books available. Nevertheless, this could also be due to the fact that book titles are not a strong determining factor for different age demographics in their ratings of books.

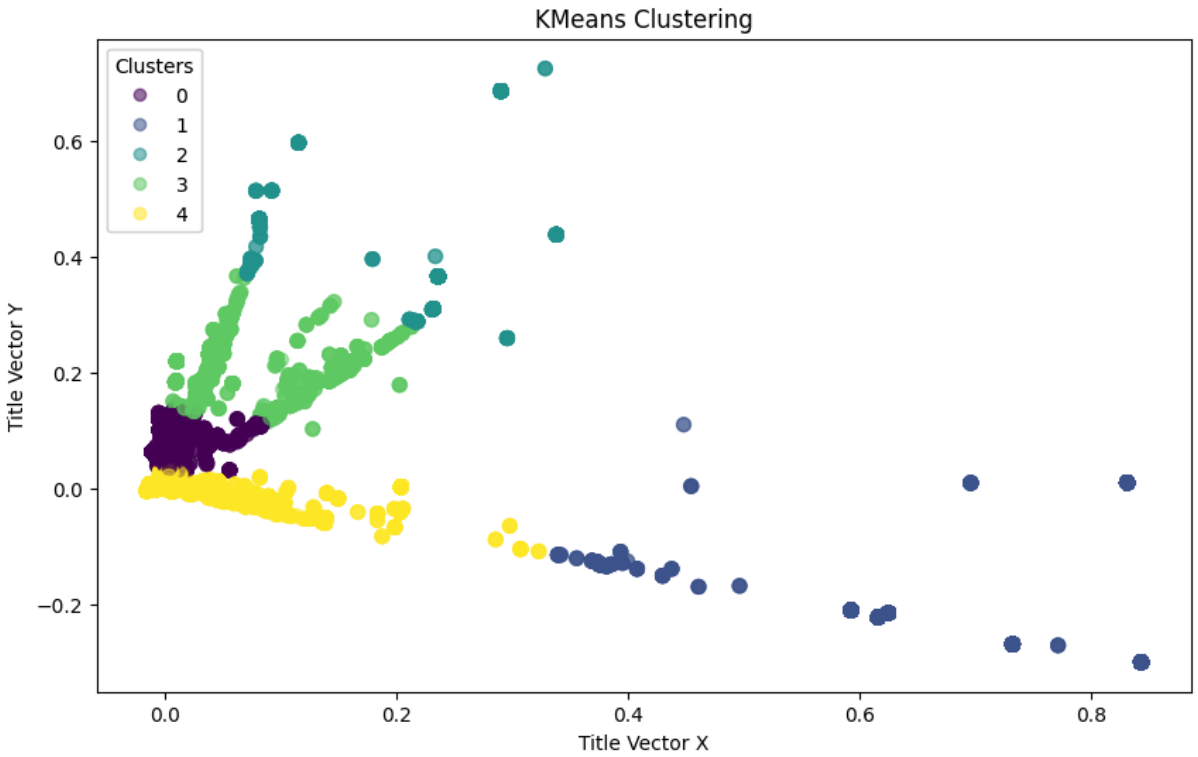


Figure 3: Scatter plot representing the KMeans clustering of book titles.

This is further supported by Figure 3, which shows the different cluster groups. There don't seem to be many cluster centroids, but instead, the data distribution is quite linear. This correlation along one dimension potentially be due to the application of SVD, but it could also be due to a lack of relationship between age, book title and rating.

Results, Discussion & Interpretation

K-Nearest Neighbor Algorithm

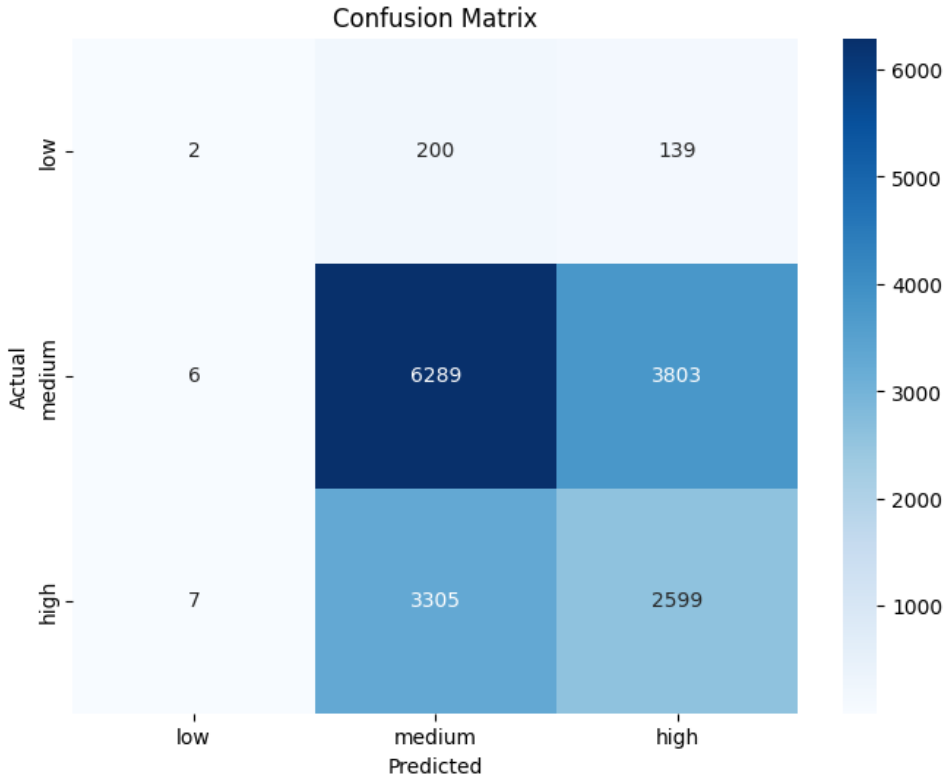
The first machine learning algorithm that we used was the K-Nearest Neighbor algorithm, which we used to try and find the relationship between user age, book publication date, book title and book rating. We based our choice of K-NN inputs on the data analysis conducted above, specifically operating under the assumption that correlations between age, book title and rating are higher than what was indicated.

We segmented our analysis of the K-NN into separate steps: accuracy evaluation and confusion matrix analysis.

1. **Accuracy Evaluation**
   The K-NN model produced an accuracy score of 0.5437308868501529, which suggests that the model is performing slightly better than random chance. This suggests that the model is indeed learning some patterns from the data, but is not particularly reliable for making predictions.

2. **Confusion matrix analysis**
   In order to understand the distribution of correct and incorrect predictions across our different rating categories of low, medium and high, we decided to graph a confusion matrix.

Figure 3:



Confusion matrix resulting from the KNN which predicts ratings based on user age, book title and book publish year.

From this matrix, we can see a clear distribution disparity in rating buckets – there are comparatively fewer low ratings compared to medium and high. As such, we have a very low True Positive value for low ratings, but this also corresponds to low False Positive values for low ratings. Regardless, it seems that our algorithm is predicting medium to high ratings for books that should in reality be lower. This

could potentially be because the algorithm is biased towards predicting medium to high ratings, as it has been trained on a dataset where medium and high ratings are more prevalent. Hence, the algorithm is struggling to accurately predict low ratings due to the limited representation. Comparatively, our algorithm is much better at predicting medium and high ratings, as our True Positive values for these are comparitively higher.

Furthermore, we calculated both the micro and the macro F1 scores of our confusion matrix, which are listed below:

Micro average F1 score: 0.5437308868501529
Macro average F1 score: 0.35366447813139046

Understandably, our micro F1 score corresponds to the accuracy score generated by our KNN matrix. However, our low macro F1 score indicates and further supports the idea that our model has a poorer precision in generating ratings for low book ratings.

Given the low degree of accuracy in predicting user ratings by considering user age, book title and date of publication, we have concluded that perhaps the TF-IDF vectorization of book title does not have as high of an impact on book ratings as we originally thought. As such, we have decided to exclude book titles from our analysis and draw upon the seemingly linear relation between user age and publishing date of book,

Linear Regression

We used Linear regression to try and find links between User Age and publishing date of the book they are reviewing, specifically the rating they give. In particular, we used this model to find a tangible value that we could associate with preference. In our case, this value is the coefficient of "year of publication", where a higher value signifies a preference towards newer books, and a lower value signifies a preference towards older books. Using these values, we plotted the book rating vs age of publication for the reviews of each age group, finding a regression equation for each, with each review representing a data point, and the predicted rating.

In comparison to our K-NN algorithm which predicted book ratings into 3 distinct buckets, we took advantage of the linear regression model's ability to predict continuous values. In this way, we hope to predict the exact rating of books based on the attributes user age and publishing date.

We segmented our analysis of the Linear Regression algorithm into separate steps: accuracy evaluation and comparison of regression lines for different age groups.

1. **Accuracy Evaluation**

The mean squared error (MSE) of our linear regression model was 3.8489, and as such the root mean squared error (RMSE) is roughly 1.961, meaning that on average the predictions of the model are off by about 1.961 units compared to the real values. Given that a model which randomly chooses ratings has a RMSE of roughly 2.6, this suggests that our model does indeed capture some meaningful patterns in the data. However, whilst this improvement in accuracy by about 0.64 units indicates that our linear regression model is more effective in predicting book ratings compared to random guessing, this accuracy is still not ideal. An explanation for this is that perhaps user preferences for different genres, authors etc. are not included in the model, and that a linear relationship between age, date of publication and ratings does not completely encompass the complexities involved in a user's rating of books.

2. **Results**

In order to visualize the relationship between publication year and ratings by different age groups, we selected two large markets for books (kids and adults), to see how their rating trends compare.

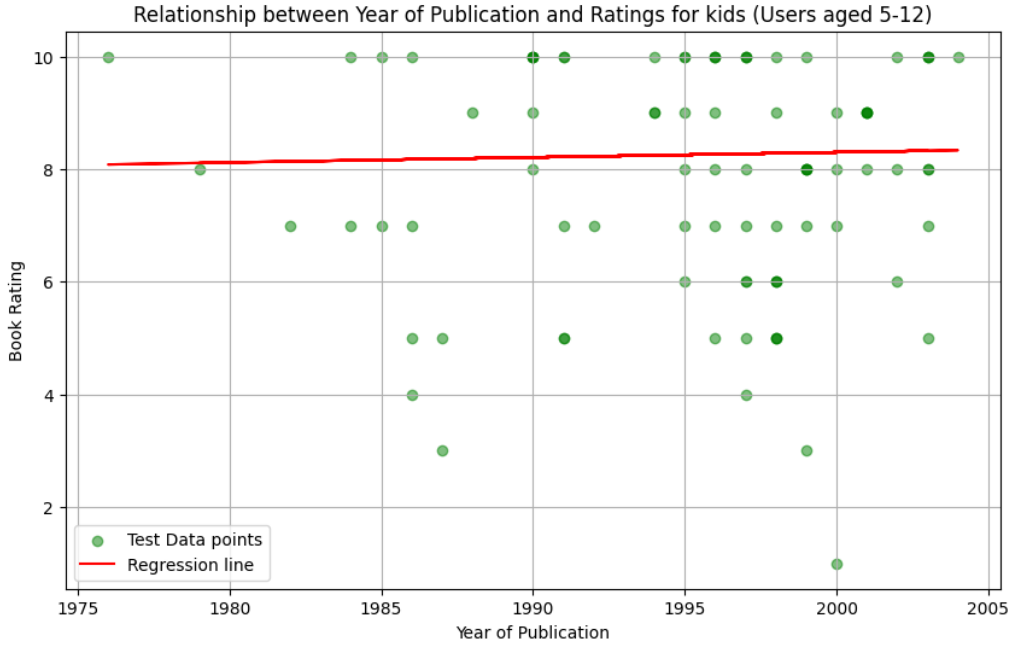


Figure 4: Scatter plot of test data points for children, with the regression line plotted.

From observation of Figure 4, the regression line has a positive gradient, with the predicted book rating increasing from older to newer books. This suggests that children prefer books with a more recent date of publication.

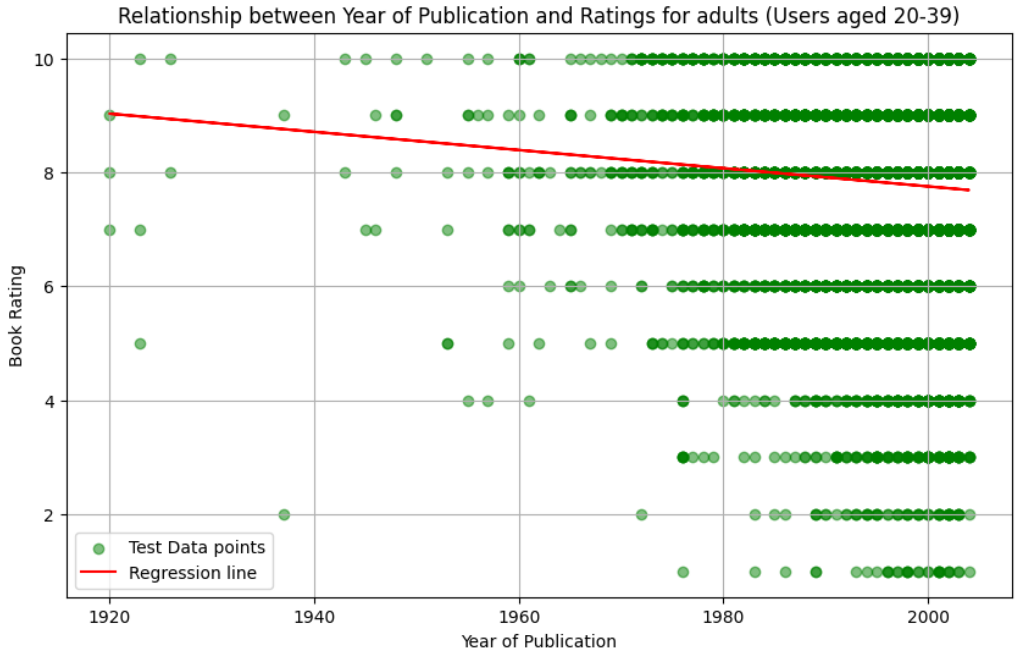


Figure 5: Scatter plot of test data points for adults, with the regression line plotted.

The regression line in Figure 5 has a negative gradient, with the predicted book rating decreasing from older to newer books. This suggests that the adult age group prefers books published earlier.

Indeed, upon plotting the linear regression equations for each individual age group, we can see some clear trends.

Table 2: Linear Regression equation for each age group

Age group	Linear Regression equation
Child	$\text{Rating} = 0.00905357 \times (\text{Year of Publication}) - 9.81$
Teen	$\text{Rating} = 0.00159106 \times (\text{Year of Publication}) + 4.60$
Adult	$\text{Rating} = -0.01595608 \times (\text{Year of Publication}) + 39.67$
Middle Aged	$\text{Rating} = -0.02083377 \times (\text{Year of Publication}) + 49.46$
Senior	$\text{Rating} = -0.00823979 \times (\text{Year of Publication}) + 24.13$

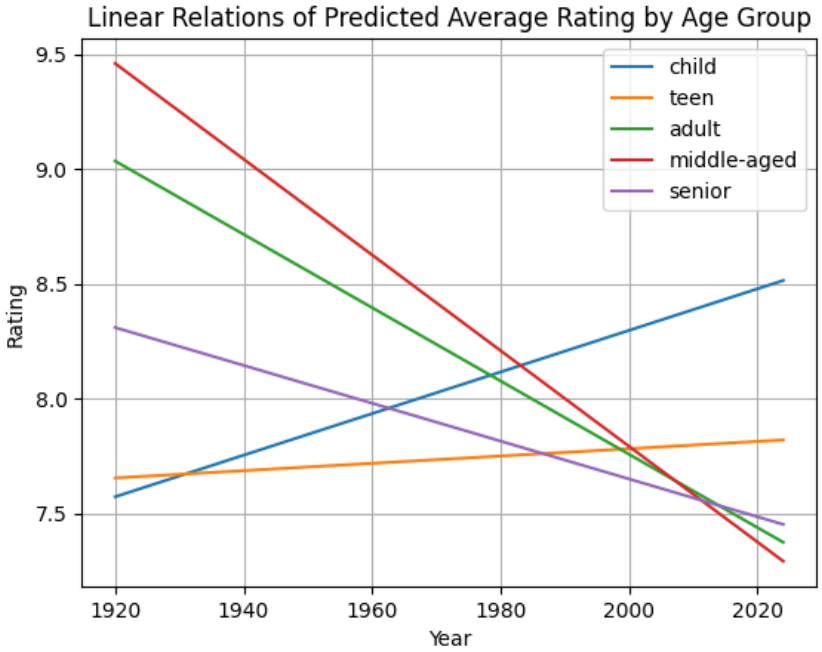


Figure 6: Date of publication vs predicted rating for each age group.

From the regression equations found from each age group seen in Table 2 and plotted in Figure 6, we make the observation that the more positive the absolute value of the coefficient, the stronger the relationship is between rating and publish year, either positively or negatively. For instance, it is observed that whilst children and teens will both rate newer books higher, children rate these books higher to a greater extent (0.00905357 with children vs. 0.00159106 with teens). Observing the rating trends in the older generation, we can see that adults, middle-aged readers and seniors have a declining interest in books the more recent they are. However, even within these demographics, we can see that middle-aged people have the strongest preference towards older books, indicated by the lowest year coefficient of -0.02083377. This could be attributed to the fact that as age increases, ability to read more complex and archaic texts increase also. Younger demographics may lack the reading level required to understand and appreciate older books, leading to a lower liking of them. Furthermore, older books are regarded as “classics” which could lead to increased receptiveness from an older demographic of readers.

There does seem to be an outlier in this linear increase of the coefficient for publishing date in the equation, being the senior category. Whilst they do still enjoy older books to newer books, they enjoy them to a lesser extent than their adult and middle-aged counterparts. We believe this could be due to the fact that reading capabilities lessen as people age, and hence seniors prefer to read more modern books. Another possible explanation is that they are recommended books by the younger people in their lives, such as their grandchildren, and hence prefer to select books that are published more recently.

In terms of making predictions, from these regression lines, we can predict that a book published in 1920 will receive a rating higher than 9 from adults and those who are middle-aged, whereas children and teens would rate them around 7.6. Comparatively, a book published in 2020 would receive a rating of less than 7 from those older than 20, whereas a child would rate it around 8.5.

We can also note from our graphs that there is a clear preference for users to rate highly. As we can see, the predicted ratings for books, regardless of publishing date, has a mean above 7. This suggests that users hesitate to rate books too harshly.

From the relationship between age, book rating and publishing date observed, it seems beneficial to recommend younger users of the online store newer books, while recommending older users of the store older books. One way to implement this is to customize recommendation pages of the website for each age group, showing mostly new books to children and teenagers, and more old books to adult, middle aged and senior customers.

Limitation and Improvement Opportunities

There exist a few limitations and improvement opportunities in our method and analysis. For starters, we decided to exclude all rows with missing values. When considering how to replace missing data, we considered both mean/median/mode imputation and proportionally filling in missing values by generating random variables based on ratios derived from existing data. However, we thought that this could skew the rating distributions as it would be inaccurate to assign user ratings to artificially generated ages. Instead, we could have potentially applied our machine learning algorithms of predicting rating of users, input these values back into the missing values in the dataset, and then reapplied the ML models. Another way we could've processed our data better was to restrict the age demographics even more – considering that a bulk of our rating data for ages seemingly laid between 10 and 80, restricting it to these ages could've allowed for better data visibility.

In our text processing, we only chose a select few misleading words to remove. For a more comprehensive handling of this data, we could've generated a list of the most commonly appearing words in the titles, and removed all of the words that we deemed misleading, such as titles which contained publisher names etc.

Furthermore, in our analysis of data, we only considered the key categories of age, book rating, book titles and publishing date. For a more comprehensive understanding of the dataset, we could have incorporated analysis on attributes such as publishers and authors. Whilst we surmised through our K-NN algorithm that book titles may not have been as great of an indicator on how different age groups rated books, we could have included publishers or authors into our parameters for both K-NN and linear regression, which would potentially have improved our F1-score and MSE.

Conclusion

In conclusion, our report delved into the relationships between user age, book attributes, and ratings, aiming to uncover patterns that could create recommendation systems and improve sales for online book

platforms. Through data preprocessing and analysis, we found links between the age of users and their preferences for certain book attributes. Our machine learning models, including K-Nearest Neighbors and Linear Regression, helped us predict user ratings for different types of books, showing us distinct trends in user ratings based on age, with younger readers having a preference for newer publications and older readers for older publications. By implementing age-specific recommendations, online bookstores can improve customer satisfaction and drive sales.