An analysis on the types of books rated highly by various ages and locations

Group W{12}G{4}

Aditya Sarkar
1354041
assarkar@student.unimelb.edu.au

Amelia Coccia
1271992
acoccia@student.unimelb.edu.au

Will O'Keefe
1461010
wtokeefe@student.unimelb.edu.au

Executive Summary

This report explores the preferences of book genres among users of various ages and locations based on the analysis of a dataset containing books, ratings and user demographics with the objective to assist the online bookstore in deciding which books to buy for future best sales based on what the largest age cohort and locations rate the highest. For the analysis, we used data pre-processing techniques such as text stripping, data cleaning, binning, mapping, data merging, lemmatization and TF-IDF before using clustering and descriptive analytics for the Kruskal Wallis test and Wilcoxon rank sum tests. From this, the key findings were that across the age cohorts, children's books and fiction were the most popular and therefore we recommend the online book store buy fiction and children's books for future best sales.

Introduction

The purpose of the report is to analyse the book ratings of users from different demographics to assist the online book store in deciding which books to buy for future best sales. The research question that guides the analysis asks what specific genres of books are preferred by individuals across geographical regions and specific age cohorts, in particular; child, teen, young adult, middle age and senior. By answering this question and relating it to the provided user information, we will be able to decide which genres of books to buy - books in the genres rated highly by the largest age cohorts and regions.

The analysis draws on a dataset of books, users and ratings found in the csv files; BX-Books.csv, BX-Users.csv and BX-Ratings.csv respectively.

BX-Books.csv contains a unique ISBN, the books title, the author, the year of publication and the publisher for 18,185 books. Each ISBN is unique and in the 10 digit format, with books originally assigned a 13 digit ISBN converted to the 10 digit conversion, signified by an X in the 10th position. The errors in the files include some encoding errors, resulting in strange text like "QuerschÃƒ?Ã, Â¼sse" (found in title from line 101 of BX-Books.csv) the files are encoded in UTF-8, but the error would be caused by the original title being copied from a different encoding format. This error was mostly found in the titles and authors. Additionally, the same book can be found multiple times, published by different publishers, in different languages or years or different forms (paperback, etc.). Bias in the data occurs due the majority of the books are written in english, meaning we expect that users will also be from english speaking countries and our insights outside of those countries won't necessarily be useful.

BX-Users.csv contains a User-ID, the users city, state, country and age. Some of the data is inputted with spelling mistakes, randomised or empty which suggests that the data was user generated (other than User-ID), some users may have chosen not to share their data. There are also unexpected characters and punctuation that will need to be cleaned before analysis can occur. We will use the data from users who've voluntarily shared all of their data. The majority of users are from English speaking countries so we can expect a bias towards English books being more commonly rated. We expect a bias towards english books, with non-english books being underrepresented in number of resumes

BX-Ratings.csv contains 204,164 reviews and captures the User-ID of the reviewer, the ISBN of the book and the book rating. The same User-ID can be seen rating multiple books, which could be useful for comparing books to each other. As users are more likely to go out of their way to positively recommend a book than the inverse, we expect a bias towards a higher average rating for all books. On a typical 1-10 scale, 5 is average, but we expect the real average to be higher. As a result of this bias, we will adjust our perspective on what a “high” rating is accordingly.

Methodology

The first stage of pre-processing was to look at the provided data and identify what techniques must be used to prepare the data for analysis. A simple first step was to use a text pre-processing technique to strip all punctuation, leading and trailing whitespaces from BX-Books.csv and BX-Users.csv. For the purpose of our analysis, punctuation and whitespaces would only hinder our ability to compare the data. The removal of punctuation does not cause the book titles to lose meaning, and all other cases of punctuation occurring are also not vital to the analysis so its removal is appropriate.

Mapping was used heavily to convert the unique user-imputed data in BX-Users.csv to be standardised. Users' cities and countries information would be mapped to a list of all countries and cities found in worldcities.csv. Mapping was also used to correct invalid data such as empty or unexpected data to a standardised “Error” string, which would signal invalid data and allow for it to be easily ignored later in the analysis. This was an appropriate pre-processing technique as it left the data standardized and allowed for comparative analysis between users from the same locations.

It was also decided that binning would be used to group the ages of the users into cohorts; Child, Teen, Young Adult, Middle Age and Senior, so that meaningful patterns in what different age cohorts like and dislike would be able to be found when analysing the data. Binning is an appropriate pre-processing technique because it allows for better recommendations for new users as we will be able to find books that people in their age range enjoy, which would be a larger and therefore more accurate list of books than the books enjoyed by users the exact age of the new user. It also reduces the impact of outliers on the data

To easily find the user information for each book rating, data merging was used to add the user information from Clean-BX-Users.1.csv to the BX-Ratings.csv to create Clean-BX-Ratings-UserInfo.1.csv. Data merging was an appropriate pre-processing technique to use as the merged data allows for ratings to be categorised according to specific user information, allowing for smaller samples of ratings that have specific shared user information in common to be collected for the purpose of analysis.

Lemmatization was used as a text processing technique in the data pre-processing stage on the book titles in BX-Books.csv. This created a new column in the Clean-BX-Books-Valid.1.csv file containing a list of each lemma in the title of the book for each book. The list of lemmas highlighted the key words in the title of each book and was later used in term frequency inverse document frequency (TF-IDF) to increase the accuracy of the clusters generated. This was because similar words such as ‘drive’ and ‘driving’ with related meanings would be counted together for the term frequency. Lemmatization was an appropriate pre-processing technique to use as it was vital in preparing the book titles for TF-IDF.

Once the book titles were preprocessed into lemmatised lists of words, the scikit-learn library could be used to convert the strings into TF-IDF vectors. This created a vector for every title in the cleaned dataset that was approximately seventeen thousand dimensions (the number of unique words in all titles in the data-set). The value of each dimension represented the TF-IDF score of the accompanying lemmatised word which is a value between zero and one that indicates the importance of that word in the title. Since most titles were less than ten words, the vast majority of the seventeen thousand dimensions in each vector were zero.

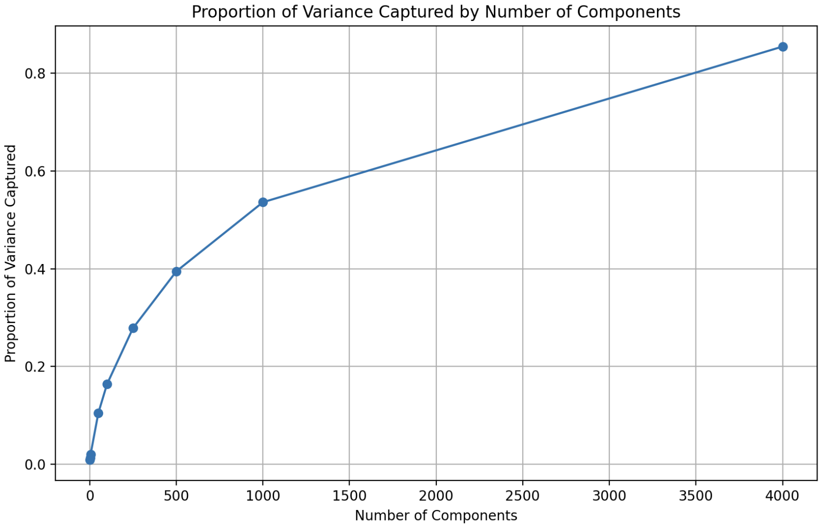


Figure 1: The amount of information retained in the TF-IDF vectors as the number of Pricipal Components (dimensions) decreases, measured by explained variance.

Thus it was appropriate to use dimensionality reduction techniques to improve the information efficiency of the vectors. This was done for 2 main reasons. The first is the 'Curse of Dimensionality' which means that in high dimensional spaces, euclidian distances between 2 vectors break down and are no longer the best distance measure, this is particularly evident when the number of information conveying dimensions in the high dimensional vector is low. The key issue is that as dimensions, or features are added, the volume of the search space becomes so large that it is difficult to prevent the vectors from becoming sparse, i.e. the majority of the vector is zero as seen in the TF-IDF vector. (Udacity, 2015) Secondly, because larger dimension vectors are harder to generalise, reducing the dimensionality of the TF-IDF vectors serve to increase computational efficiency without significantly affecting the final product (the quality of book title clusters). This was because the information could be conveyed in fewer dimensions with minimal loss as the majority of the dimensions (words in titles) convey limited information about the meaning or contents of the title. The amount of information retained as the dimensions (principal components) were reduced through principal component analysis can be seen in Figure 1. 4000 principal components were chosen as it was a substantial reduction from the 17000 initial dimensions, but still retained more than 80% of the explained variance.

With this new, more informationally efficient, lower dimension group of TF-IDF vectors, k means clustering was used to sort the books into similar groups based on the titles. In order to maximise the 'goodness of fit' the number of clusters must be chosen such that the error metric is minimised while also ensuring the clusters are not spread so thin that they are meaningless. Additionally, a larger number of clusters caused the clustering to be more computationally intensive. Thus, the elbow method was used to determine the correct number of clusters.

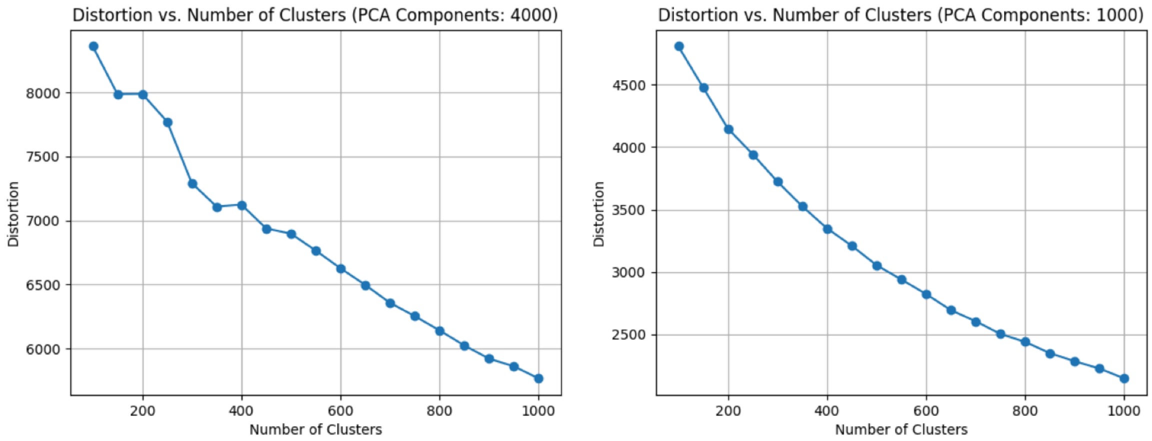


Figure 2: Distortion (Squared Sum of Errors) from cluster centroid to members as number of clusters increase for both 4000 and 1000 principal components.

Ultimately, the elbow method is relatively subjective, but a cluster count of 500 was chosen. The output of this step however was simply a list of 500 clusters of various books, there was no simple way to classify what each cluster was or how it was clustered without manually looking at the clusters. To solve this problem, a web scraper was built to randomly sample books in each cluster and obtain the genre of the book from GoodReads. This was used to classify the clusters and remove uninformative clusters. There were 13 final overarching genres used to classify each cluster

After completing clustering, it was noted that Clean-BX-Books-Valid.1.csv contained duplicate books that were only different as they were published by different publishers and/or at different times. This was causing those books to appear more frequently in the clusters generated, making it appear in the data as though there were more unique books in that cluster than in reality. It was presumed that users would be unlikely to buy the same book from a different publisher, and therefore the duplicate books were removed in a new csv file, Clean-BX-Books-Valid.2.csv. The pre-processing technique of data cleaning here was highly appropriate as the deduplication of books meant that when clustering was repeated, the clusters generated were more accurate to the list of books available.

The Goodreads web scraper returned nearly 100 genres for each book and so it was also decided to use the same code for the title clustering to cluster the books based on the genres from Goodreads in order to compare the results.

To begin analysing the clusters anomalous entries marked with 'Error' in the pre-processing stage were removed for the analysis and the cluster dataframe was transformed to a long format using the 'melt()' function, normalising the data. It was important to have a record of every rating, user-details and their respective cluster and genre in a unified csv file for observing correlations.

Observations of the data using statistical measurements and manipulation were applied under varying conditions. Aggregate functions and group-by techniques were used to proportionally analyse the data. This was important to gain an understanding of a demographic's relative behaviours as percentages were used to provide a relative measure of preferences amongst cohorts. As well as this, overlap in the books rated by each cohort were analysed using these techniques, providing insight on the likeness of demographics.

Descriptive analytics were used to calculate highest rated books among demographics and identify trends across age and region. These observations are turned into actionable insights using mean calculations, ordering and ranking.

Continuing, each cluster was associated with a unified genre. The csv containing the clusters are in long form with each cluster containing a series of books containing genres in the form of [‘genre1’,’genre2’]. Each genre entry is parsed as a list and then "exploded" to create individual rows per genre within each cluster. This restructuring allows for the calculation of genre frequency within each cluster. The frequencies are then tabulated, and for each cluster, the genre with the highest count is determined as the most popular. This aggregated genre data is merged back into a dataset containing average book ratings per cluster, providing genre specificity. This process effectively links book ratings with the most prevalent genres per cluster, clarifying understanding of genre popularity within clusters.

Clusters are then ranked over the cohorts based on the classification that a positive review is a rating of 8 or higher, returning a mean rank. A weighted score is created for each group by adding the average rating to half the count of positive ratings, prioritising groups with higher frequencies of positive feedback, aiming to minimise the effect of outlier positive reviews. This analysis identifies which book genres resonate most with specific demographic groups, based on their collective ratings.

The Kruskal Wallis test was used to prove that there is a statistically relevant margin of difference in tastes in genres between demographics on the basis of age cohort and regions. It is used to assess the distribution of genre preferences across groups. A p-value of less than 0.05 suggests that the median genre preferences significantly differs among the tested demographics. For this analysis, the 6 most popular genres from our dataset were identified through analysis of the top rating clusters as an identifying genre has been established for each cluster through ranking. These genres were fantasy, fiction, childrens, classics, comics and non-fiction. Working within the subset of books, lists of genres reviewed by each cohort are compiled and applied to the Kruskal Wallis test using the scipy library.

For deeper analysis, the ratings dataset was split into subsets by the previously binned ages. Child, teen, young adult etc. Each subset was then further split based on the genre of the book being rated. For each age group, each genre was compared with every other genre using Wilcoxon rank sum tests, which is a statistical comparison of location between 2 ordinal datasets. This test is particularly useful because it does not rely on a normally distributed underlying population and makes fewer assumptions in general, thus allowing the analysis to be more robust. Through this test, it could be determined whether there was a preference for one genre over another (based on ratings), or whether there was no statistically significant preference.

Using this information, genres were assigned scores similar to a tournament where if a genre was rated higher than another, it was awarded the absolute value of the test statistic in points while the lower rated genre was given 0. This is because the test statistic is a measure of the distance between the locations of the samples but also takes into account the size of the samples. Two genres with very large samples and a relatively large difference in location will have a large statistic whereas two genres with the same difference in location and much smaller samples will have a smaller statistic. Thus, the scoring of the genres also takes into account the confidence and the magnitude of the preference. If the genres had no statistically significant difference in ratings, both genres would also be awarded 0 points. Ultimately, this determined which genres were most highly rated for each age group and thus which were most appealing.

Data Exploration and Analysis

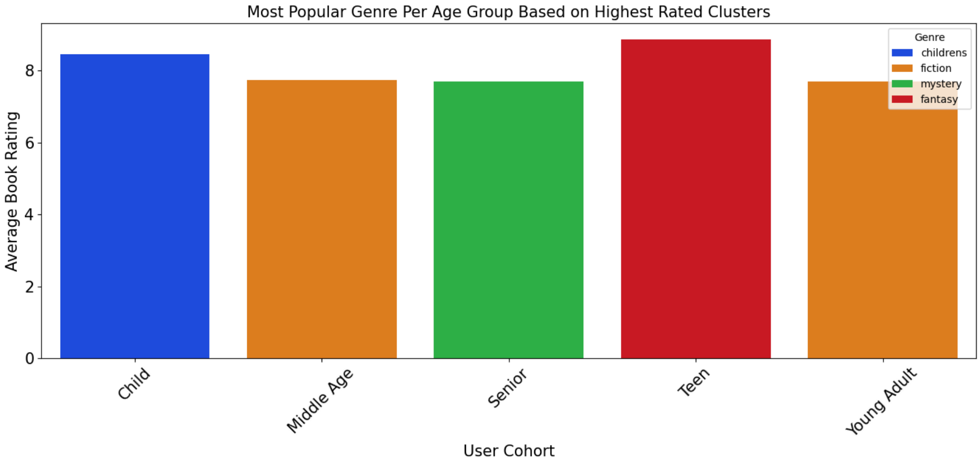


Figure 3: Aggregate determination of top genres per cohort with weighting.

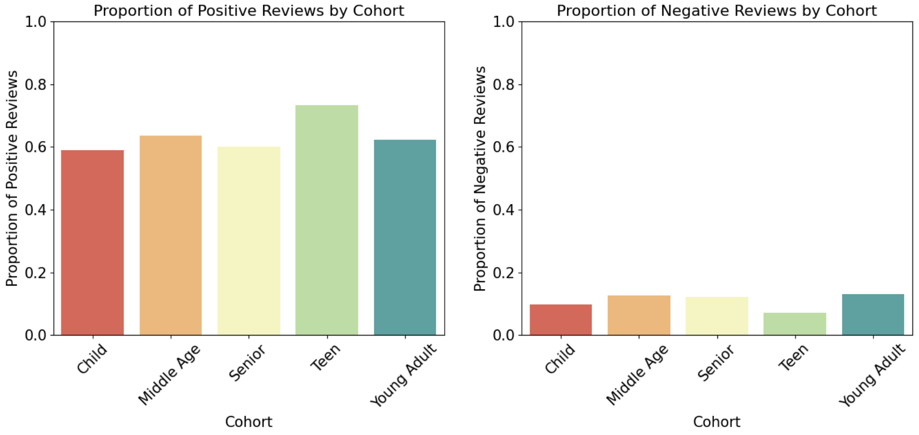


Figure 4: Proportion of positive and negative reviews by cohort.

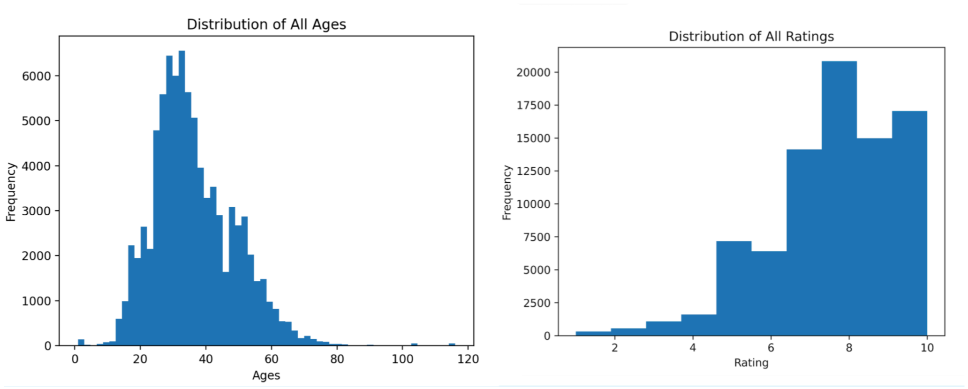


Figure 5: Proportion of ages and ratings.

Table 1: Descriptive statistics of age and ratings.

Data	Mean	Median	Standard Deviation	Coefficient of Variation
Age	36.27	34.0	12.58	0.347

Ratings	7.8	8.0	1.8	0.231

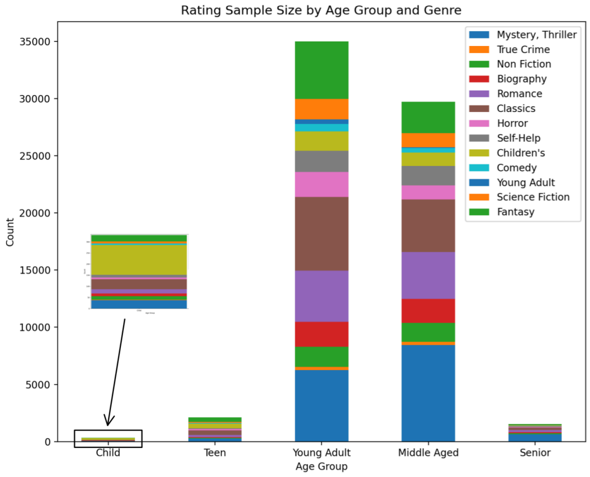


Figure 6: Cleaned sample size of ratings by age group and genre.

Table 2: Genre preferences by age group. Scores of genres are in brackets.

Preferences	**Child (6-12)**	**Teen (13-18)**	**Young Adult (18-35**	**Middle Age (36-64)**	**Senior (65+)**	**Total**
1	Children's (19.7)	Fantasy (70.38)	Children's (223.11)	Children's (168.93)	Classics (13.24)	Children's (140.52)
2	Biography (4.38)	Comedy (64.23)	Comedy (176.19)	Comedy (99.29)	Science Fiction (12.5)	Comedy (103.71)
3	-	Romance (21.27)	Fantasy (153.88)	Classics (69.12)	Mystery, Thriller (12.3)	Fantasy (91.81)
4	-	Children's (20.94)	Classics (80.28)	Fantasy (58.44)	Biography (11.46)	Classics (55.64)
5	-	Biography (17.18)	Biography (53.89)	Biography (32.71)	Non Fiction (10.22)	Biography (32.6)
6	-	Science Fiction (10.36)	Science Fiction (45.4)	Self-Help (30.08)	Fantasy (10.12)	Self-Help (26.74)
7	-	Mystery, Thriller (5.75)	Self-Help (44.16)	Non Fiction (28.77)	Self-Help (9.55)	Science Fiction (26.06)
8	-	Classics (5.34)	Non Fiction (40.97)	Horror (26.21)	Children's (8.78)	Non Fiction (23.89)
9	-	-	Mystery, Thriller (6.8)	Science Fiction (24.79)	Romance (0)	Horror (9.81)
10	-	-	Horror (5.0)	True Crime (16.66)	-	Mystery, Thriller (7.09)

11	-	-	Romance(0)	Mystery, Thriller (6.77)	-	True Crime (2.67)
12	-	-	-	Romance (6.14)	-	Romance (2.18)

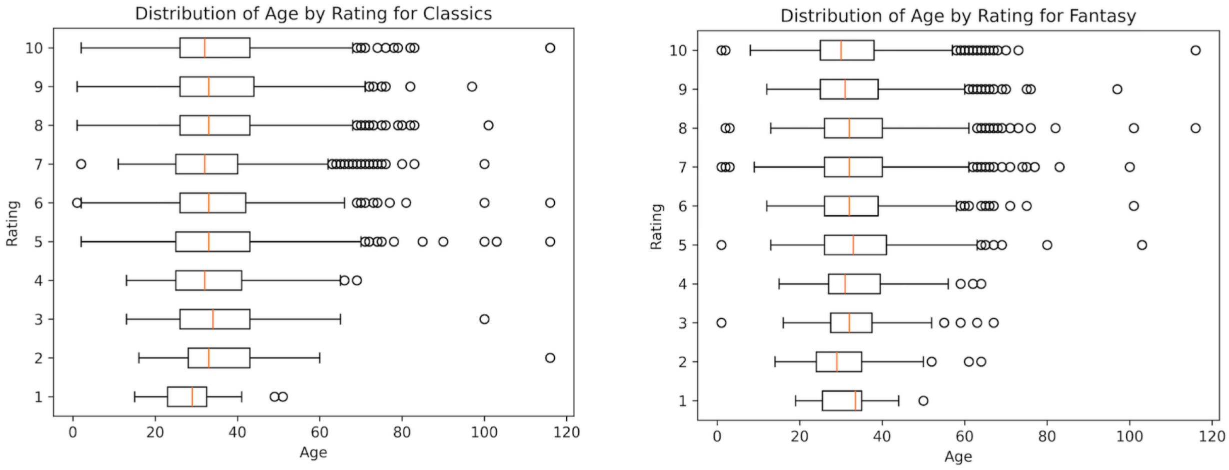


Figure 7: Cleaned sample size of ratings by age group and genre.

Results

The Kruskal-Wallis test conducted on cohorts and regions yielded very small p-values, thus providing strong evidence to reject the null hypothesis, which posits no difference in genre preferences across demographics.

Table 3: Kruskal-Wallice test results on cohort and region for genre

Cohort	p = 8.91 * e^-28
Region	p = 1.98 * e^-21

The test assessed two primary components: statistical differences in genre preferences and identification of the cohorts that are most or least distinct in their tastes, i.e. the books they're rating.

Applying the Kruskal-Wallis to evaluate the variation in book ratings among the demographics revealed that Young Adults exhibited the most variation in their ratings, while Teenagers showed the least. In terms of regions, Asia demonstrated the greatest diversity in book ratings, whereas Africa showed the least.

Test on which groups have the most overlap in the books that they're ratings:

Table 4: Cohort overlaps in books rated, with the percentage each demographic1 overlaps with demographic2 in the total variation of books rated.

Overlap in books rated	**Demographics by Region and Cohort**	**Overlaps**	**Percentage overlap demographic1**	**Percentage overlap demographic2**
Most	Young Adult and Middle Age	13092	85.26% of Young Adult overlaps with Middle Aged	86.03% of Middle Aged overlaps with Young Adult
	Americas and Europe	4449	27.17% of Americas overlaps with Europe	79.63% of Europe overlaps with Americas
Least	Child and Senior	118	19.38% of Child overlaps with Senior	5.29% of Senior overlaps with Child
	Oceania and Africa	17	0.99% of Oceania overlaps with Africa	28.81% of Africa overlaps with Oceania

The repeated use of Wilcoxon tests yielded a more granular perspective of the specific tastes of each age group, ranking each genre by preference. Below is a list of key insights that will be further explored in the discussion section:

- Various genre preferences seem to be mildly related with age, for example, the preference level of classics increase with the seniority of age groups.
- Teens displayed a strong preference for fiction genres.
- Romance was read frequently but was relatively strongly disliked
- Children's books were broadly appealing

Discussion and Interpretation

Diversity of reading habits was assessed using the Kruskal-Wallis test **(See Table 3)**, which evaluates whether the median ratings across different demographics significantly differ. This non-parametric method ranks data regardless of the group size, which means it does not inherently adjust for the number of ratings from each demographic. Consequently, demographics with fewer ratings might be underrepresented, potentially skewing the understanding of diversity in reading habits. Nevertheless, the test offers valuable insights into the characteristics of the database as a useful comparative tool. To more accurately reflect all demographics, especially those with fewer ratings, statistical approaches involving weighting may be considered.

Young Adults are noted as the most diverse cohort in terms of books rated. This diversity could be attributed to the wide range of interests and life stages within this group, from late teens to those in their late twenties, possibly incorporating both educational texts and leisure reading. Teens are described as the least diverse, which might suggest a more uniform taste within this group or a focused interest in certain popular genres such as fantasy.

Asia is noted as the most diverse region. This could reflect the vast range of cultural influences and languages that contribute to diverse literary tastes and the availability of a wide variety of genres. Africa, identified as the least diverse, could indicate more uniform preferences or limitations in the availability of a broad range of books.

Figure 4 presents the distribution of positive and negative book reviews across various age cohorts, with ratings above 8 classified as positive and those below 5 as negative. Despite the broader negative range, the analysis reveals a predominant preference for positive reviews, with all demographics exhibiting more than 60% in positive feedback.

This observation suggests a positivity bias among users, where individuals are predisposed to record and share positive experiences rather than negative ones online. Notably, the Young Adult cohort displays the highest percentages of both positive and negative reviews. This lends to the fact that young adults are the largest subset in the data **(See Figure 5)**, and are reviewing the most diverse range of books as found through the Kruskal-Wallis test.

This pattern indicates that Young Adults are not only consuming a wide variety of books but are also most actively critiquing them. This suggests that this demographic's tastes are varied and their reviews are relatively discerning, making them a critical audience for a bookstore. Focusing on the preferences and detailed feedback of Young Adults could enable curation that's better tailored to meet the expectations and interests of a key demographic, thereby potentially increasing sales and expanding market reach.

Adults and Middle-Aged cohorts show a high number of overlapping book ratings **(See Table 4)**, suggesting similar literary tastes and preferences. This could be due to the implication that the proximity in age between the Adult and Middle-Aged cohorts contributes to similar life stages, experiences, and interests. This closeness in age likely influences their literary choices, leading to a higher overlap in book ratings.

This finding is contrasted with the clear distinction in preferences observed between widely spaced age groups like children and seniors. The low overlap between these demographics indicates very distinct preferences. Children’s books target the younger demographic, whereas seniors prefer more classics and science fiction, reflecting a significant generational gap in book preferences.

The significant overlap in book ratings between the Americas and Europe could suggest a cultural or economic similarity that leads to comparable reading habits and preferences. The minimal overlap between Oceania and Africa could be due to many factors, taking note of the bias towards English titles, and the low subset of users from Africa in the dataset.

Figure 1 displays the preferences of different age cohorts with respect to book genres. Each colour represents a different genre, illustrating the highest-rated genre within each cohort based on aggregated book ratings:

The highest rated genre for ‘Children’ is 'children's', represented in blue. This outcome is predictable, given the targeted nature of this genre towards this group.

Both Young Adults and Middle-Aged cohorts show a rating of the ‘fiction’ cluster the highest. This observation is particularly significant as it reinforces the findings of **Table 4** which identifies that Young Adult and Middle Aged groups are the most similar demographics in terms of books rated. We know from **Figure 4** that a rated book is over 60% likely to be rated positively for all age groups, hence it can be expected that an overlap in books rated alludes to a high overlap in books rated highly.

Young Adults and Middle-Aged readers are likely inclined towards fiction due to the broad appeal of the genre which encompasses a variety of subgenres that cater to a wide range of interests. The blurring of traditional age-specific preferences may also reflect broader societal shifts where significant life milestones are being approached differently, leading to extended engagement with youth-oriented cultural products into the early Middle Age, which we have classified to begin at 36 years old. This outcome supports the results of the Wilcoxon test (**See Table 2)** which states that the two demographics top rated genres are ‘children’s’, thus falling within the fiction category. Such a trend suggests that marketing strategies could be streamlined for these demographics.

It is observable in most bookstores the presence of a ‘Tik-Tok table’, which is a targeted, central display in the store catering for these demographics, typically featuring ‘young-adult’ books that fall within the fiction category and have garnered popularity online. This is an excellent method not only in aiding in a seamless shopping experience for those hoping to find popular books they have encountered in online spaces, but as a recommendation system for people within this demographic.

Seniors recorded a preference for 'mystery' suggesting a favour for genres that offer mental stimulation and intricate plotting, which can provide cognitive benefits and maintain engagement. This analysis aligns with previously observed trends in other data visualisations **(See Figure 6)**, confirming consistency across different measures of preference.

In the context of this report, the aggregate determination of genres is produced to provide a preliminary look at preferences within the dataset, deriving a metric that reflects both the sentiment of ratings and quantity of positive feedback among data points. The results differ from that of the Wilcoxon test, but generally can be considered to sharing the same sentiment

The primary focus on the Wilcoxon test is assessing whether the distributions of two populations are identical without assuming them to follow the normal distribution. As the test uses ranks, it is less sensitive to outliers than mean calculations. Clusters with a few very high ratings can disproportionately affect the mean in the weighted scoring system but will have less impact in the Wilcoxon test. It involves ranking all the observations from both groups together, then analysing these ranks to determine if one group tends to have higher or lower values than the other. In comparison to the weighted means approach, which is a representational method of data aggregation that includes weights to adjust the importance or influence of certain observations within a dataset.

The ranking of preferences using repeated Wilcoxon tests provided a more granular view of the preferences of each age group as well as the population as a whole. From **Table 2** it appears as though some genre preferences are related with age. For example classics are least preferred by Teens but increasingly preferred as seniority of age groups increases, with Seniors preferring classics over all other genres. Similarly, Fantasy is very popular with teens bu.t declines in preference with the seniority of age-groups. However, when analysing the distributions of ages across ratings (**Figure 7**) there is no statistically significant relationship between the rating of books of these genres. This is likely due to the Wilcoxon analysis being an analysis within age-groups and whereas the correlation between age and rating is a analysis across age groups. For example, Seniors prefer classics over other genres, but in absolute terms, do they prefer them more than other age groups? It is possible that on average, Seniors rate all books lower, thus making it seem as if there is no correlation between age and genre preference, and this is exactly what is seen in the data. While teens have an average rating of 8.00, seniors, on average, rate books 0.41 points lower at 7.59 points. This is meaningful as it is a substantial portion of the standard deviation.

Table 5: Mean rating by age group.

	Child	Teen	Young Adult	Middle Age	Senior
Mean Rating	7.94	8	7.8	7.81	7.59

Romance was read frequently but was relatively strongly disliked. Romance did not appear in any age group's preferences apart from teens, and in the preferences of the entire sample, appeared last (**See Table 2**). This is despite Romance being the third most popular genre read across the sample (**See Figure 6 and Appendix A**). Romance having a score of 0 for all age groups other than teen simply means that every Wilcoxon test with romance for that age group suggested either no statistically significant preference or a preference for the other compared genre. The conclusion here is twofold, either readers believe they will enjoy romance more than they actually do when selecting romance and chick-lit novels, thus the disconnect between how many people are rating them and how highly they are rated; or users are reviewing these books because they are bad at a higher rate than other genres. It is important to remember that reviews are biased toward extreme views, rather than the average opinion.

Finally, one atypical trend was that Children's books were very broadly appealing, in fact, this genre was the highest preference for both Middle Age and Young Adult, the two largest age groups in the sample (**See Figure 2**). This is surprising as it would be expected that older audiences are less likely to like children's books, Children's books were also not a very frequently read and reviewed genre and thus it could be said to be a relatively niche genre.

There are a number of factors that can be hypothesised to lead to this observation. First, Young Adult and Middle Age age groups are the most likely age groups to have young children. Thus they may not be rating the book for themselves but rather for their children, or perhaps more abstractly, their perception of the act of reading this book to/with their child. There are two features in the data which support this theory, firstly, Teens and Seniors, the groups which are least likely to have young children, rate Children's books relatively low at fourth preference and eighth preference, compared to the first preference of Young Adult and Middle Age groups. The second is that the strength of preference decreases between Young Adult and Middle Age age groups, dropping from 223.11 to 168.93 points. This would make sense as Middle Aged readers, up to 64 years old, are marginally less likely to have young children compared to Young Adults.

Another reason for the preference of children's books could be childhood nostalgia which causes readers to rate highly. Some anecdotal evidence for the previous two hypotheses can be found in Appendix B. Finally, Children's books are designed to be unproblematic, apolitical and educational and as such are uniquely positioned to avoid negative reviews. Recalling that reviews are by nature biased towards extremes, and thus, if there are no harsh, negative opinions to be had, there are no negative reviews, leaving only particularly positive reviews.

Limitations and Improvement

Limitations in the data pre-processing:

Due to the encoding error causing strange characters to occur, all books containing these characters were removed from the books dataset. Many of these books were non-English books and so the data would become more biassed towards English books. An improvement would be to search online for the book using the ISBN and replace the broken title with the correctly encoded one that is returned from the search.

There were a few books in the dataset that were identical, except they were published by different publishers or in different years. For the purpose of our analysis we did not use either the year of publication or the publisher so when removing duplicate books, only the first occurrence was kept, irrespective of when or who published it. Therefore, further analysis on the publishers would be invalid as they were chosen essentially at random.

The data pre-processing cut the input data for books and ratings, BX-Books.csv from roughly 18,000 to 10,000 books and ratings from 200,000 to 120,000. While the pre-processed data was enough to gather useful insights from, improvements to accuracy could easily be made by cutting less data out of the input data.

Ultimately, the book clustering based on the scraped genres was more relied on for analytics. Although the title based clustering provided insights, the quality of clustering was limited. For example, out of 500, just 1 cluster had 4000 out of 10000 titles. Furthermore, clusters seemed to be based on particular key words in titles which at times meant that books that were completely unrelated were clustered together. Many titles were also too short to be useful ithrough this method of natural language processing. The fundamental issue was that the amount of information that can be extracted about the contents of a book from the title is limited. By contrast, the books clustered by genres were much more useful and insightful. Another limitation was that although 500 clustered seemed to minimise the distortion, it was far too many clusters for meaningful analysis as clusters were too niche and often had too few members to perform statistical analyses. For this reason, the amount of clusters were later reduced to just 20 which were later further condensed down to 13 genres. Once clustered into 20 clusters, the largest cluster had only approximately 1000 books, suggesting a more even split.

Ultimately, the elbow method is a subjective method to determine the number of clusters and it is important to prioritise the purpose and the needs of the exploratory data analysis.

Conclusion

Our analysis of the data substantiated the research question that asks what specific genres are preferred across cohorts regarding age and region, revealing that genre preferences vary significantly across different age groups and regional demographics.

We utilised a diverse array of methodologies to determine genre preferences and correlated these preferences with specific demographic groups. These methods involved data pre-processing techniques such as text stripping, data cleaning, binning, mapping, data merging, lemmatization and TF-IDF before using clustering and descriptive analytics for the Kruskal Wallis test and Wilcoxon rank sum tests. The consistency of genre preferences between cohorts across different analytical methods, such as the Kruskal-Wallis, the Wilcoxon and distribution analysis, enhances the reliability of the findings.

Through our analysis, we determined distinct genre preferences among various cohorts which can be used to recommend books to specific groups. While younger audiences preferred a variety of fiction genres such as fantasy and Comedy, older age groups preferred classics and, to an extent, non-fiction genres such as biography. Romance was preferred only by teens. The ‘Children's’ category in particular should be considered by bookshop owners as it is unlikely to deviate through trends like other genres as it is designed to appeal to an age group, rather than specific sensibilities.

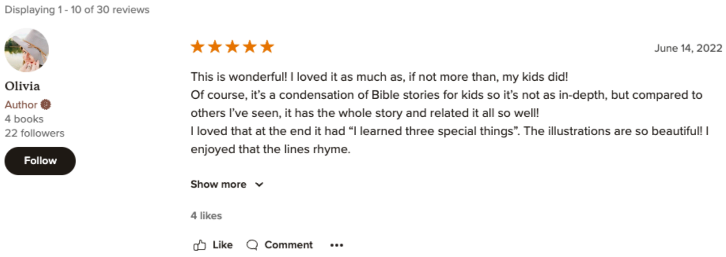
These insights allow us to make informed recommendations to bookshop owners as through analysis of the data, we can provide actionable recommendations to bookstores on inventory curation, emphasising genres like Children's and fiction due to their broad and enduring appeal, while recommending genres like Fantasy and Comedy to younger audiences due to their specific tastes.

Appendix A

Preference	Child (0-12)	Teen (13-18)	Young Adult (18-35)	Middle Aged (36-64)	Senior (65+)
Mystery, Thriller	38	260	6260	8455	630
True Crime	2	2	264	259	13
Non Fiction	16	61	1750	1666	83
Biography	12	80	2194	2102	86
Romance	19	185	4472	4112	190
Classics	45	403	6471	4573	230
Horror	8	100	2177	1253	42
Self-Help	12	77	1865	1699	74
Children's	134	406	1693	1164	33

Comedy	7	46	641	395	18
Young Adult	0	10	402	77	0
Science Fiction	9	84	1776	1227	49
Fantasy	29	404	5025	2731	88

Appendix B



Displaying 1 - 10 of 30 reviews

Olivia

Author

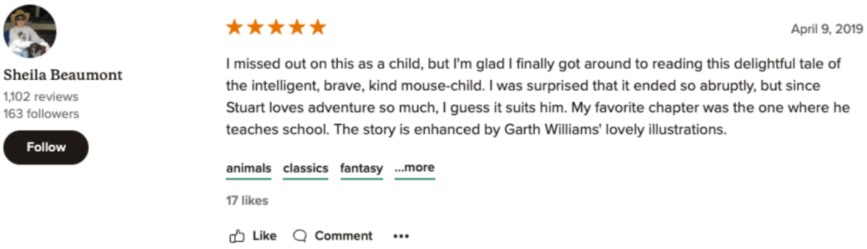
4 books

22 followers

★★★★★ June 14, 2022

This is wonderful! I loved it as much as, if not more than, my kids did!
Of course, it's a condensation of Bible stories for kids so it's not as in-depth, but compared to others I've seen, it has the whole story and related it all so well!
I loved that at the end it had "I learned three special things". The illustrations are so beautiful! I enjoyed that the lines rhyme.

4 likes



Sheila Beaumont

1,102 reviews

163 followers

★★★★★ April 9, 2019

I missed out on this as a child, but I'm glad I finally got around to reading this delightful tale of the intelligent, brave, kind mouse-child. I was surprised that it ended so abruptly, but since Stuart loves adventure so much, I guess it suits him. My favorite chapter was the one where he teaches school. The story is enhanced by Garth Williams' lovely illustrations.

animals classics fantasy ...more

17 likes

References

DataTab. (n.d.). *Kruskal-Wallis test*. Retrieved 2024-05-01, from https://datatab.net/tutorial/kruskal-wallis-test

Goodreads (n.d.) *Genres* Retrieved 2024-05-03 from www.https://goodreads.com

SimpleMaps. (2024-03-19). *World cities data*. Retrieved May 3, 2024, from https://simplemaps.com/data/world-cities

Udacity (2015-02-23). "Curse of Dimensionality - Georgia Tech - Machine Learning". Retrieved 2022-06-29.