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

*Presented at **Stage III: Oral Presentation***
as part of Assignment 2
Element of Data Processing

W17G7

Ravon Chew
Reuben Mattam
Viane Dorthea Tiwa

Melbourne, 16 May 2024

OUTLINE

1

Introduction

THE IMPORTANCE OF DATA IN DECISION MAKING

> *With the existence of online bookstores, a new paradigm has emerged where data-driven insights have become central to optimising business strategies.*

GOAL

Leveraging comprehensive datasets from an online bookstore to discern **trends and pattern**

inform strategic decisions for **the bookstore manager**

Enhance inventory management and customer engagement to improve the bookstore's sales

Online Bookstore Data

BX-Books.csv

Catalog of 18,185 book titles

This dataset provides details about each book, with the International Standard Book Number (ISBN), title, author, year of publication, and publisher.

BX-Users.csv

Diverse user base of 48,299 users

This dataset contains information about the bookstore's users, including user ID, city, state, country, and age.

BX-Ratings.csv

204,164 user's reviews

This dataset captures reviews from book users, including user ID, ISBN, and the rating given by the user.

RESEARCH QUESTION AND EXPECTED OUTCOME

Research Questions

The research question formulates in this report:

(1) What types of books tend to be highly rated by users?

(2) How is this influenced by their demographic?

Expected Outcome

This report offers:

A snapshot of current trends in popular books.

Give recommendations to the bookstore's manager for aligning inventory.

Characteristics of the targeted user.

Bookstore's manager can tailor marketing strategies to meet customer needs.

2

Methods, Techniques, and Tools

PRE-PROCESSING

Data Cleaning

- ✓ **Removing improper entries** – users, book details, ratings – through **Regular Expression**
- ✓ **Data Augmentation**, using provided ISBNs to replace improper book titles through an **API request***.
- ✓ **Outlier removal**.
  - Ages above 99
  - Years of Publication before 1750.
  - Books with only 1 Rating
  - Authors with only 3 Books
  - Publishers with only 5 Books
- ✓ **Feature Engineering**. Creating continuous data from nominal data of countries.
- ✓ **Data Integration**. Aggregation of average rating per book and total rating per user.

***storing the data into a separate *.csv* file**

ANALYTICAL TECHNIQUES

STATISTICAL METHODS

Bayesian averages and **Wilson's Lower Bound**

To explore tendencies and dispersions within the data and produce a popularity ranking list

- ✓ **Density maps**, **Scatterplots**, and **Histograms.**

MACHINE LEARNING

Classification Models

Sentiment Analysis and Subjectivity Analysis of Book Titles.

Using Vader sentiment analysis, to obtain a probability score per book title on the likelihood of the sentiment is positive, negative, or neutral. **Using Textblob**, to provide the corresponding overall sentiment score and to calculate subjectivity scores.

Regression Model

Linear and non-linear regressions are calculated to find correlations in the plotted dataset.

Nearest Neighbour Method

To find the book's neighbour and give **further recommendations** of books to purchase based on the rating given by the user.

3

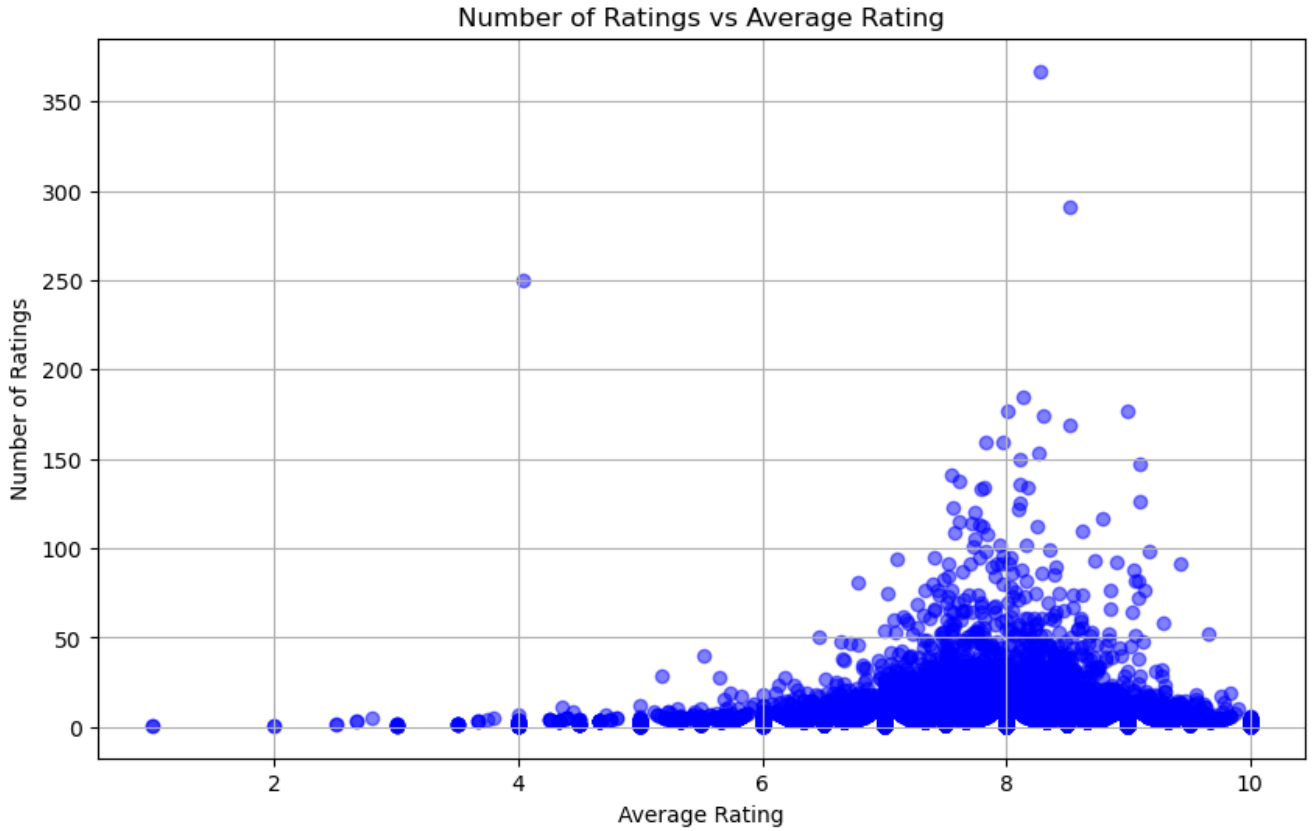
Result

The first and most obvious question to ask is....

Does Higher Average Book Ratings Correlate to a Larger Readership Base ?

No, it doesn’t.

Graphing the number of ratings per book vs their average rating clearly shows this is false



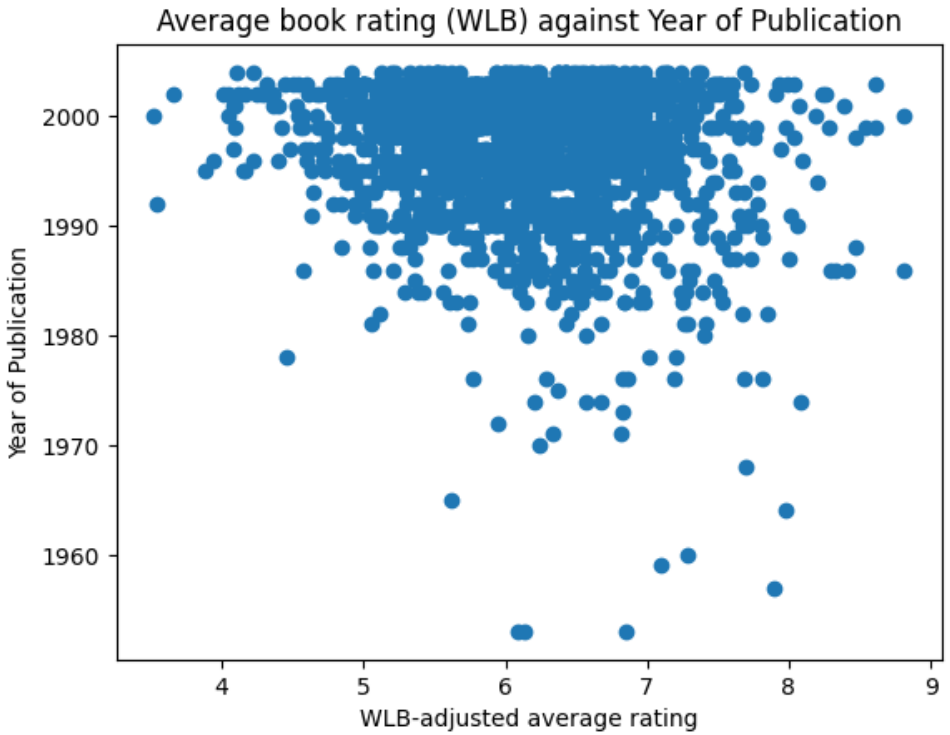
So why does the number of ratings clump at an average book rating of around 8 ?

Which leads us to the question of...

What factors cause readers to clump around this point.’

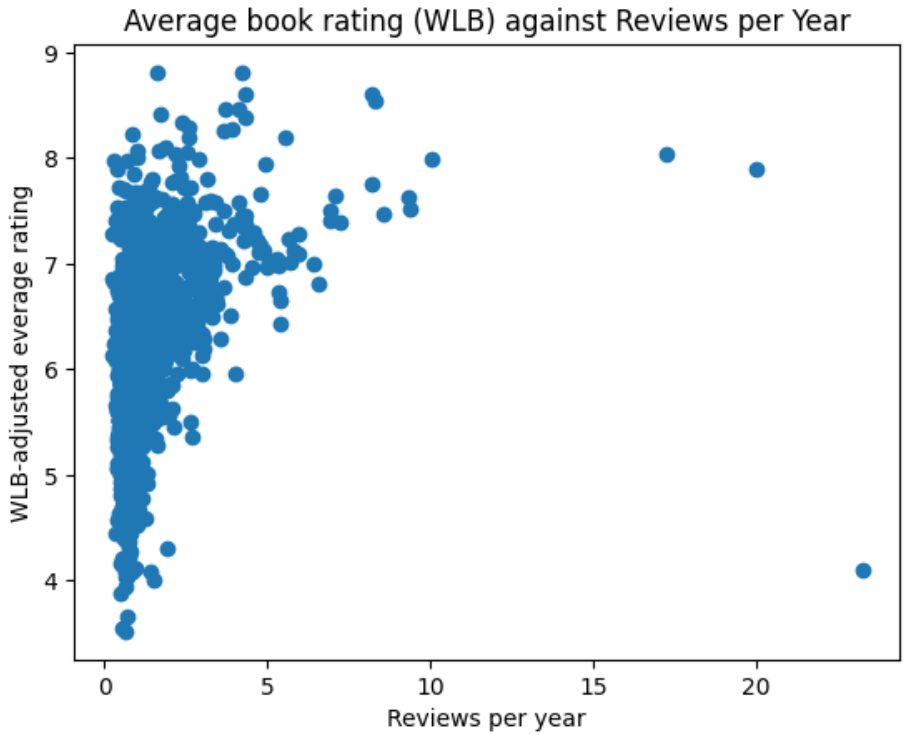
And, thus we ask,

Correlation Analysis



Using WLB, revealing correlation between average rating and year of publication.

Using WLB, revealing correlation between average rating and total number of review



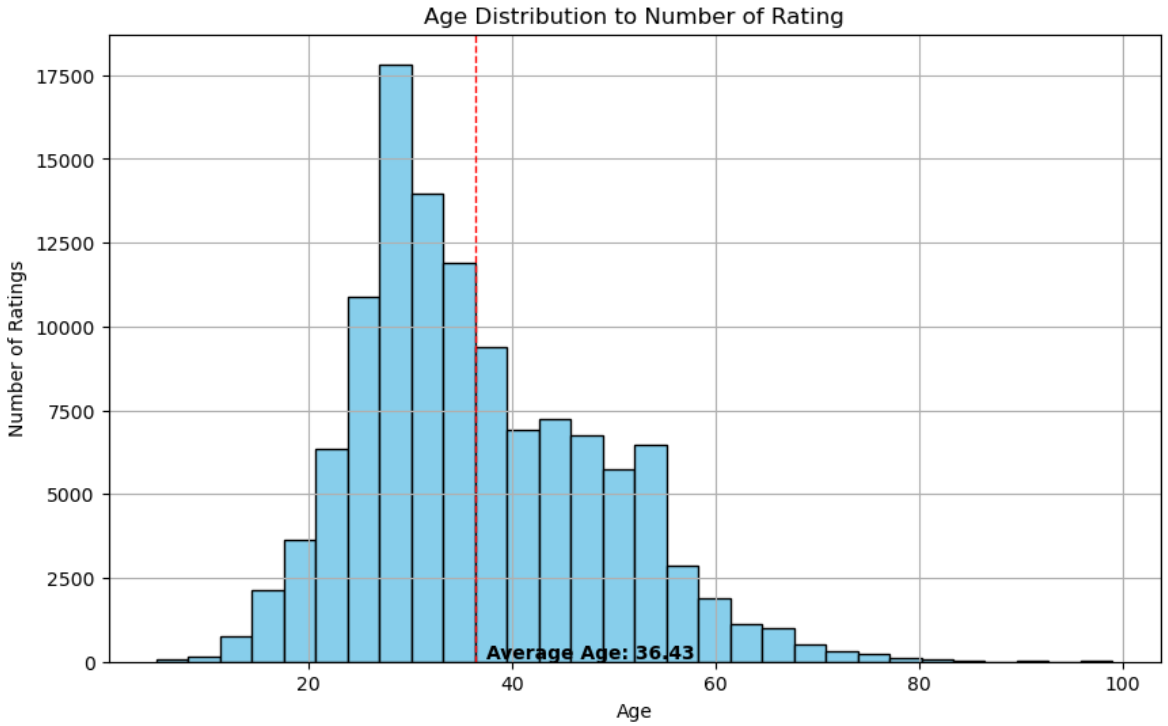
Clumping of data around the average rating

Therefore, what role does

Readership Demographics

play in influencing book popularity?

Age



Age demographic and distribution play a major role in the popularity of a book. Something that will heavily explored later on throughout this presentation.

It is important to note however, the specific age demographic of this tends to favour a middle aged demographic, thus skewing the data in favour of their definition of 'popular'. As seen on the left.

The other demographic explicit in the dataset is location.

Before addressing location, its necessary to examine an

Assumption from the data

Answer how does location impact popularity?

The Assumption

From the context of the dataset,

- We presumed that books deemed more popular would likely possess a substantial non-English speaking audience.
- Consequently, at a certain ratio of English to non-English ratings, we should see a peak of average book popularity.
- Thus, as readership tends towards a 0% and 100% English speaking level, a corresponding decline in popularity is expected.

From this assumption, we feature engineered our own continuous dataset

English Speaking Ratio

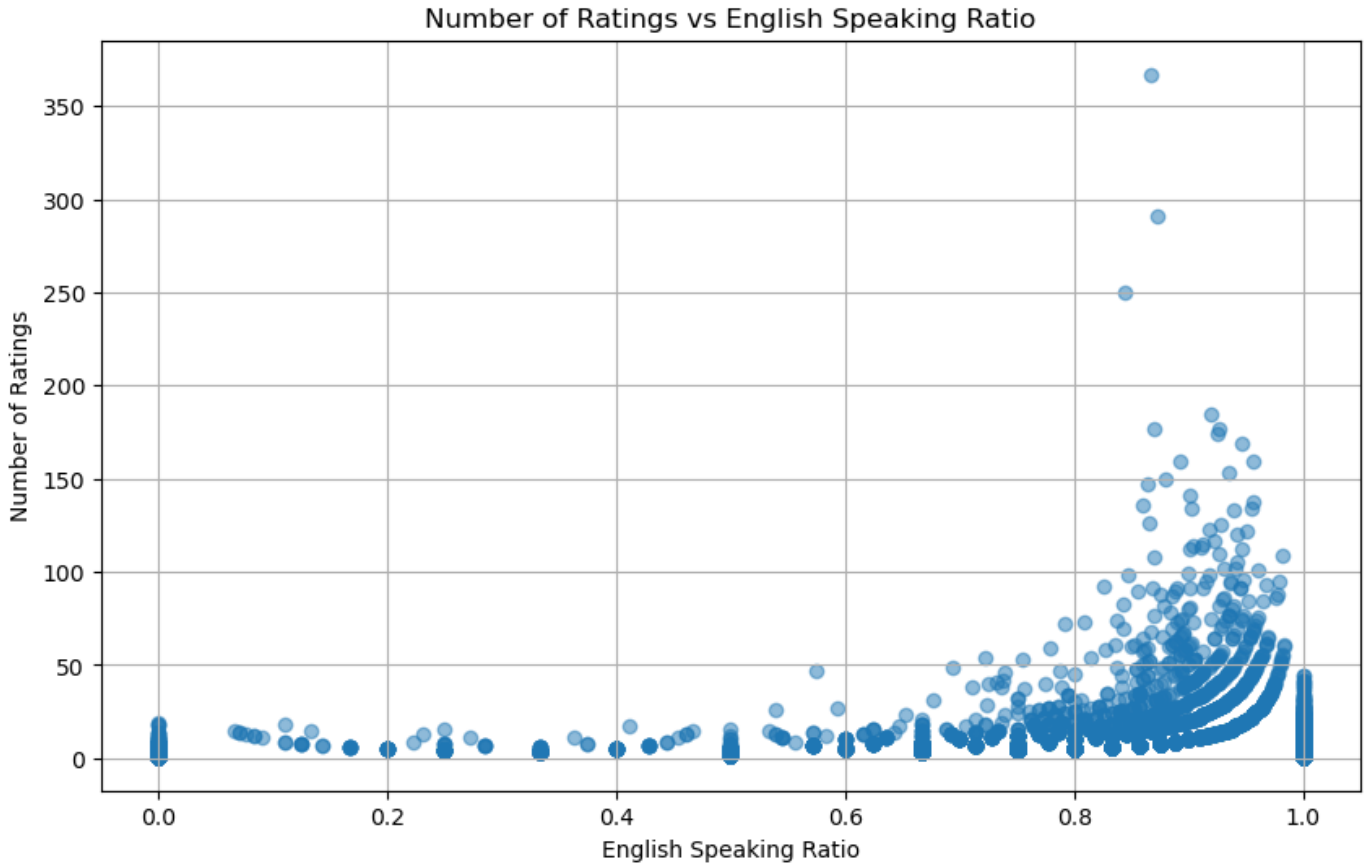
English speaking ratio is defined as the ratio between:

- The number of people whose ratings are from 'usa', 'canada', 'australia', 'ireland', 'new zealand', 'united kingdom'.
- Its noted that there was no deviations suggesting that countries were set options and henceallowing us to create this ratio.

- This score was divided by the total number of reviews.
- Instead of dividing by the total number of reviews with a specified country, we opted for this approach under the assumption that the ratio of missing countries across a substantial dataset would balance out, thus averting any potential skewing in the results of our regression analysis.

- This returned an 'English-Speaking-Ratio' score between 0 to 1 for every book in our dataset.

Validating the Assumption



The natural progression hence is graphing the number of ratings vs English speaking ratio, testing to see if the assumption made was right.

Visually, the correlation is evident in reinforcing the assertions claimed by the assumptions.

Clumping of data around the 0 and 1 part heavily skewed data however.

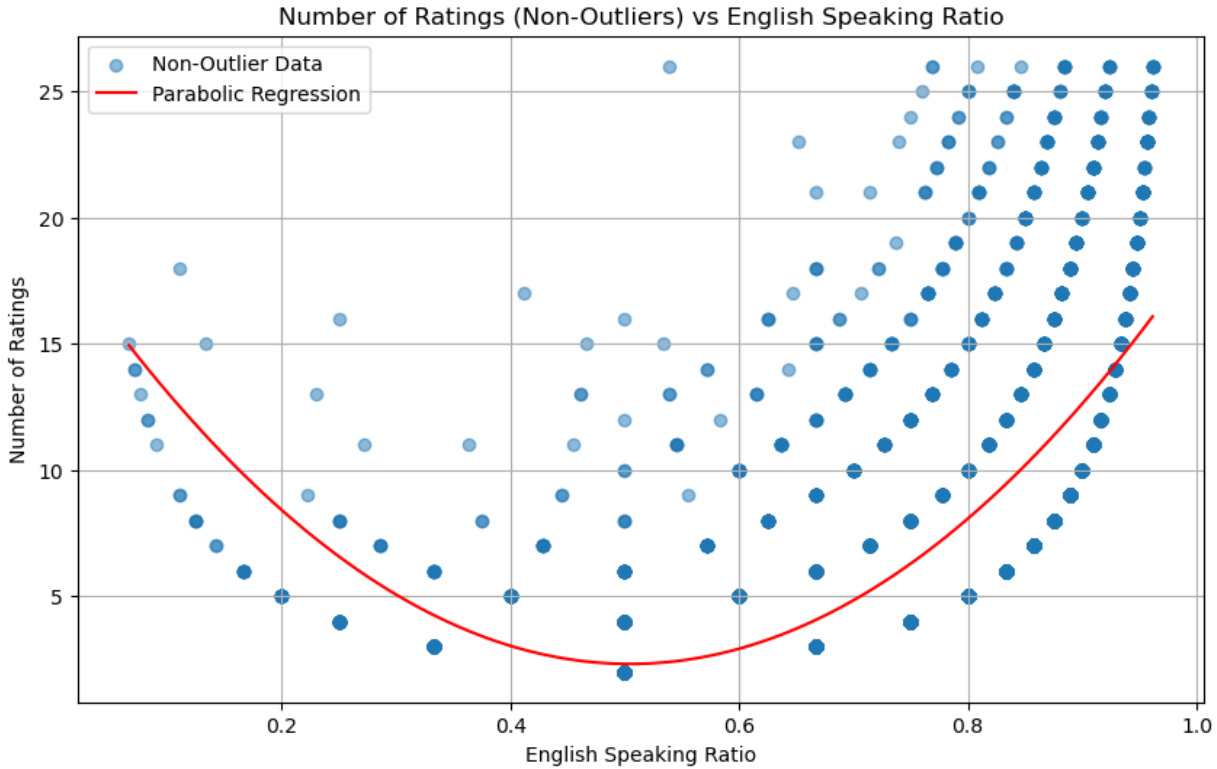
Outliers

As previously stated, books with an English-speaking ratio of 0 or 1 are additionally classified as outliers within the dataset.

This categorization stems from the clustering effect observed at these points, which heavily skew the data in a non-meaningful manner. Moreover, these data points lack statistical significance as the focal point of popularity we are examining does not fall within their range, nor is it influenced by the grouping observed at these points.

This brings us to the average case, ignoring all outliers.

The Average Case




```
Parabolic regression parameters (a, b, c): [ 65.86438412 -66.44506919  19.07303652]
Adjusted R-squared score: 0.4879424074881056
```

Restricting number of ratings between its interquartile and removing outliers from English Speaking ratio provides us a sample of the average books popularity.

Visually, an parabolic correlation is explicit from the scatterplot as points of 'grouping' form parabolic functions on top of each other.

Utilizing a parabolic regression, we return an adjusted R-Squared score of 0.488. An adjusted R-Square score is a measure of variance in the data that penalizes points that don't improve the model. A score close to 1 means less variance.

Limitations

The grouping of overlapped of points is a reoccurring feature, stemming from the 'continuous' data of English Speaking Ratio. This overlaps forms as when compared to the discrete number of ratings, it can only take on a certain number of values as English Speaking Ratio is a ratio of the total number of reviews.

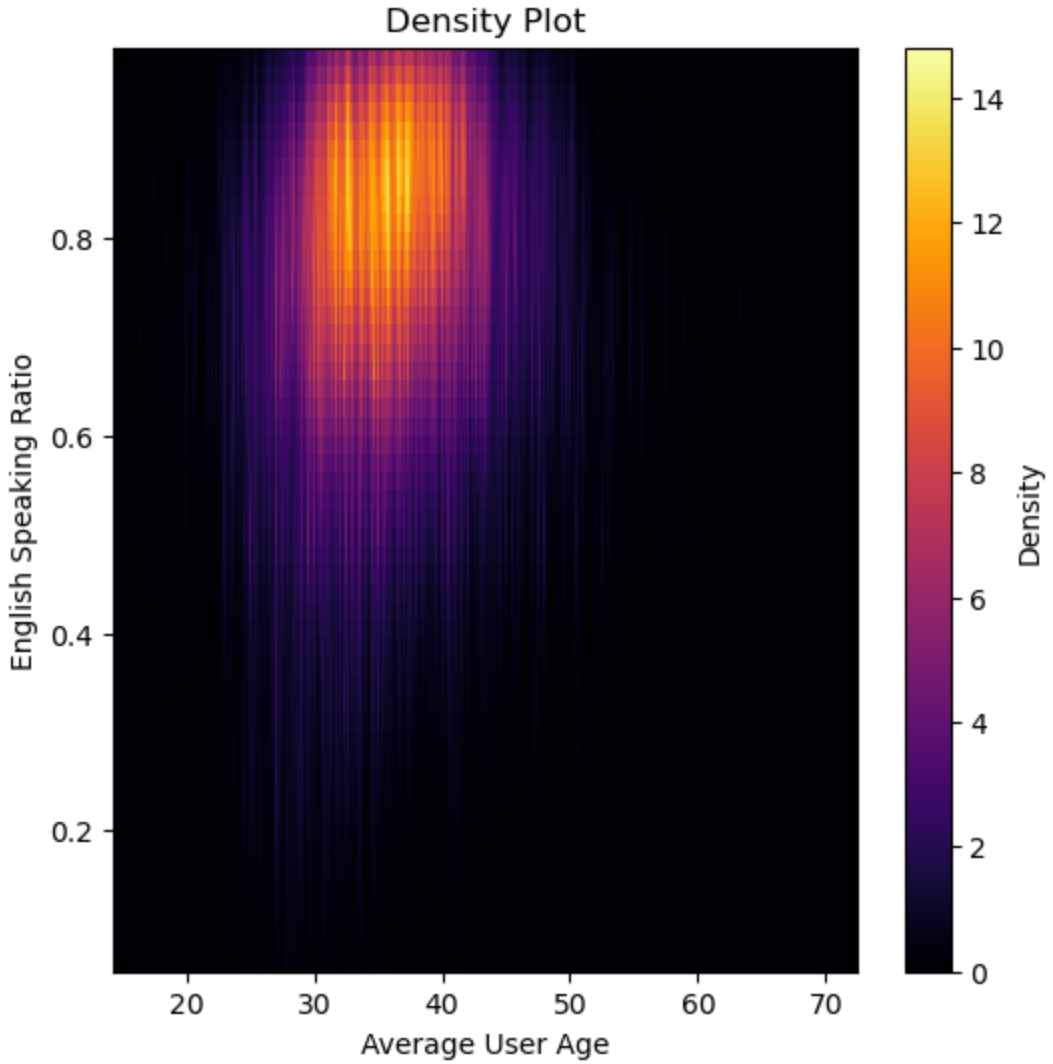
Consequently, the presence of heightened variance is an expected outcome hence impacting the adjusted r-squared score, creating higher levels of variance and resulting in suboptimal suitability for the data. Despite the correlation coefficient of 0.488 falling short of 1, the underlying nature of the data suggests this is a strong correlation.

This will point will be revisited again throughout the course of presentation.

Confluence of Demographics

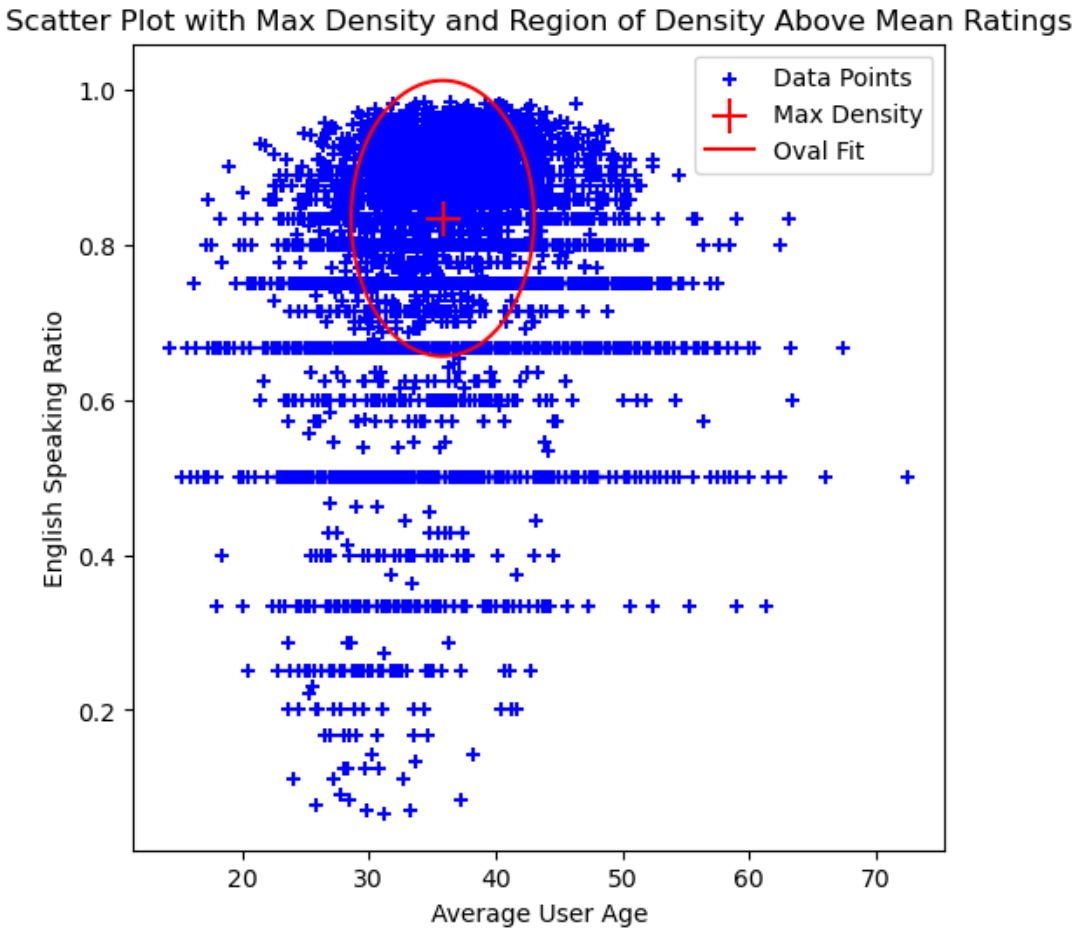
By plotting English speaking ratio and Average User Age across a density plot, we are able to visualize how the 3 main variables of our focus all interplay together.

By utilizing density to represent the number of ratings, the graph delineates the area where the highest ratings occur in relation to age and English-speaking ratio.



This is the basis of our upcoming analysis.

Locating the Region of Interest



Each point of density represents the number of ratings per that intersection of English Speaking Ratio and Average User Age. Hence, by setting the density threshold to the mean number of ratings, we find the region of significant density – circled in the red.

From this we have an age range of:

- 28.994 to 43.416

And a English Speaking Ratio of:

- 0.628 to 0.984

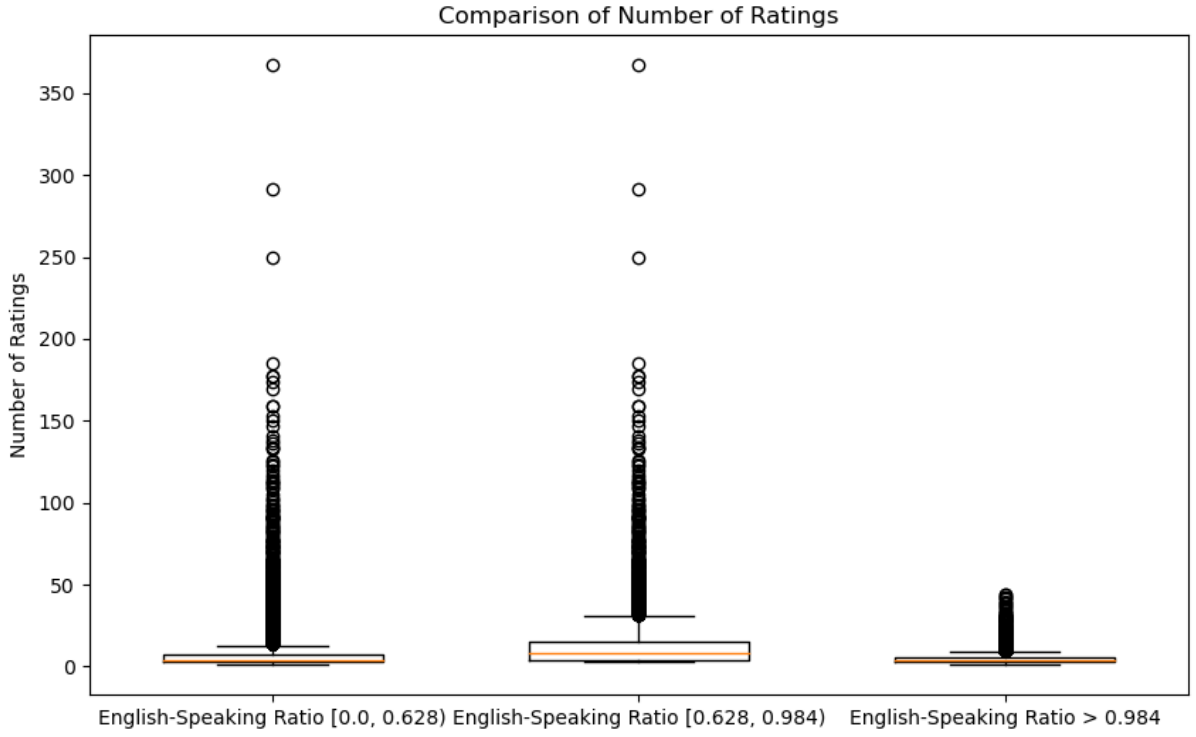
Focussing on the region of interest,

Behavior of Outliers

Before describing the behavior of the focal region, understanding the behavior of outliers is important to understanding how books are rated outside of non-outlier cases.

Beyond the average context, we see even in our region of interest, average outlier values – hence average outlier number of ratings – and significantly higher outside of the IQR of said region.

This observation suggests that even outside the confines of the 'average case', the region of interesting is still the most popular.




```
Percentage of outliers for first boxplot: 10.09%
Average value of outliers for first boxplot: 29.39
Lower Quartile for first boxplot: 1.0
Upper Quartile for first boxplot: 13.0
```

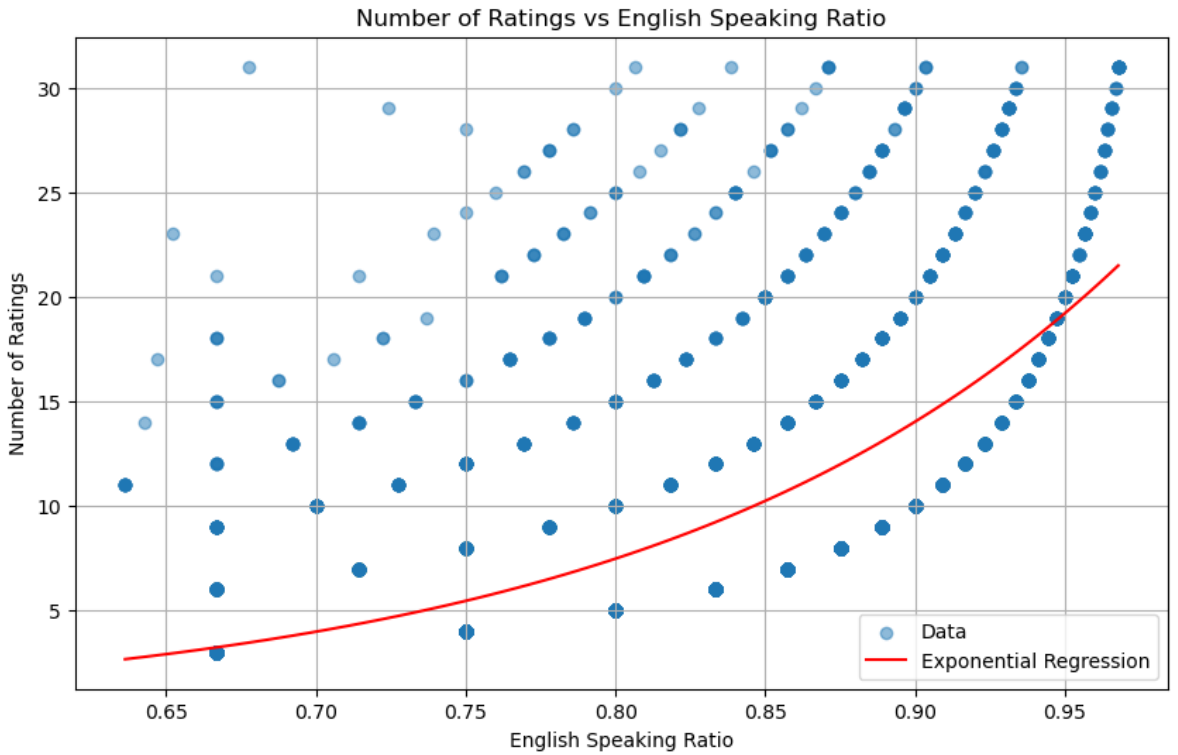
```
Percentage of outliers for second boxplot: 8.81%
Average value of outliers for second boxplot: 59.33
Lower Quartile for second boxplot: 3.0
Upper Quartile for second boxplot: 31.0
```

```
Percentage of outliers for third boxplot: 7.78%
Average value of outliers for third boxplot: 14.68
Lower Quartile for third boxplot: 1.0
Upper Quartile for third boxplot: 9.0
```

Targeting the Region of Interesting

By defining our region of interest as the intersection between the highest density of English speaking readership and the number of ratings within its interquartile range for this region, we observe an exponential relationship between the 2 variables.

Utalising an exponential model, we are able to verify this, obtaining an adjusted R-squared score of 0.500. As previously mentioned, given the nature of the dataset, this score is anticipated and indicates a correlation consistent with our visual observations.




```
Exponential regression parameters (a, b): [0.04856743 6.29536243]
Adjusted R-squared score: 0.49550417144637193
```

How does this target varying demographics ?

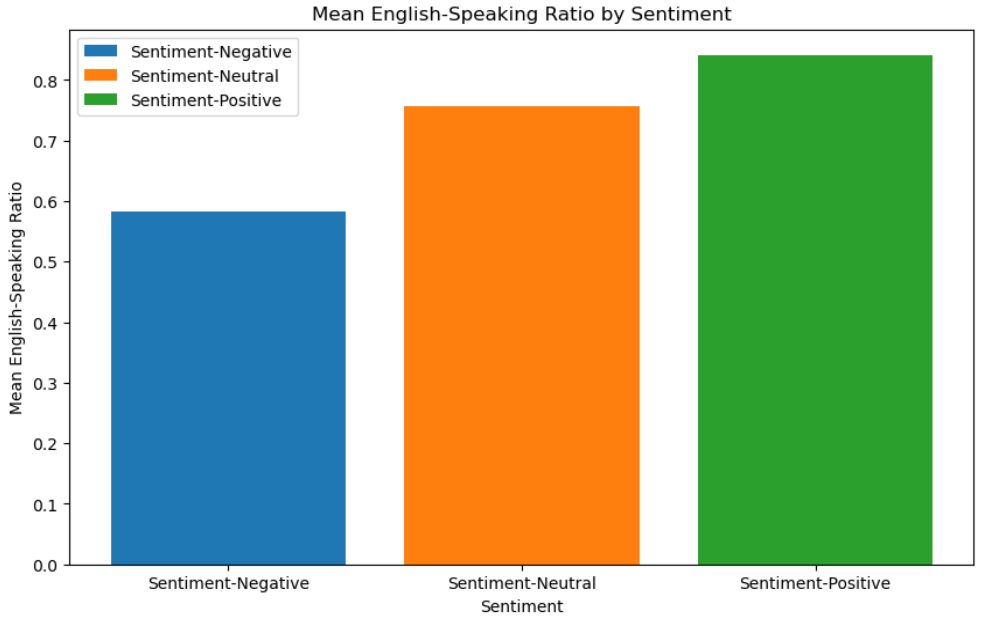
We explore this through...

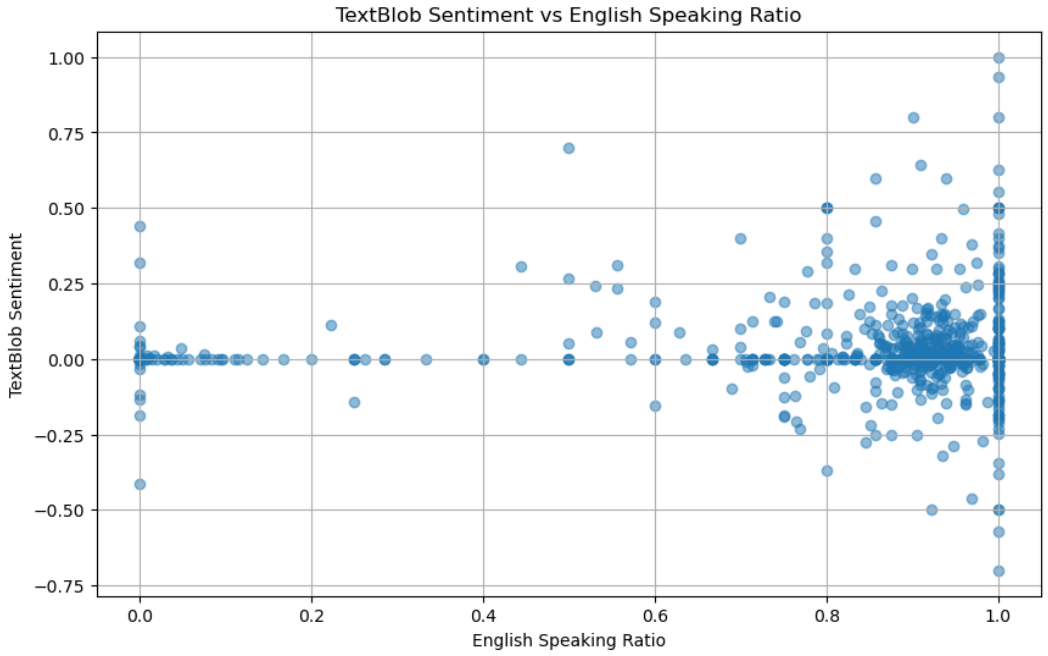
Sentiment Analysis

Sentiment Analysis

Through the use of VaderSentiment Analysis, we can obtain a probability score that indicates the level of positivity, negativity, or neutrality conveyed by the title. Plotting these probabilities illustrates a trend: there is typically a preference for more positively-toned book titles among readers high English speaking readership, and vice versa.

Additionally, utalising TextBlob analysis provides a score ranging from –1, for negativity, 0, for neutrality, and 1 for positivity. The scatterplot indicates how approach 0 and 1 english speaking ratios reflect in a want for books with greater variance in tone.

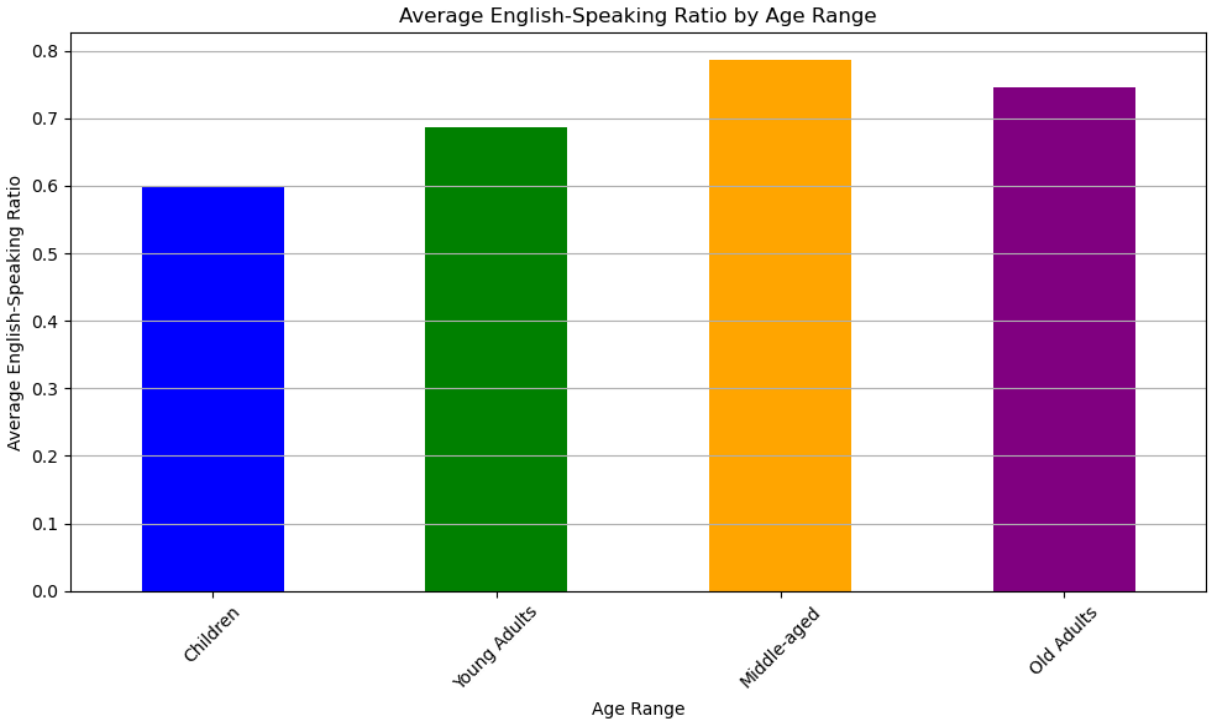




Age Ranges

Another aspect to consider is the age range of our target audience. By categorising individuals into Children, Young Adults, Middle Aged and Old Adults, we more comprehensive understanding of the interpay between variables.

As suggested on the graph and logically, as you increase in age range, the probability of you reading a book with a higher English-speaking ratio increase.



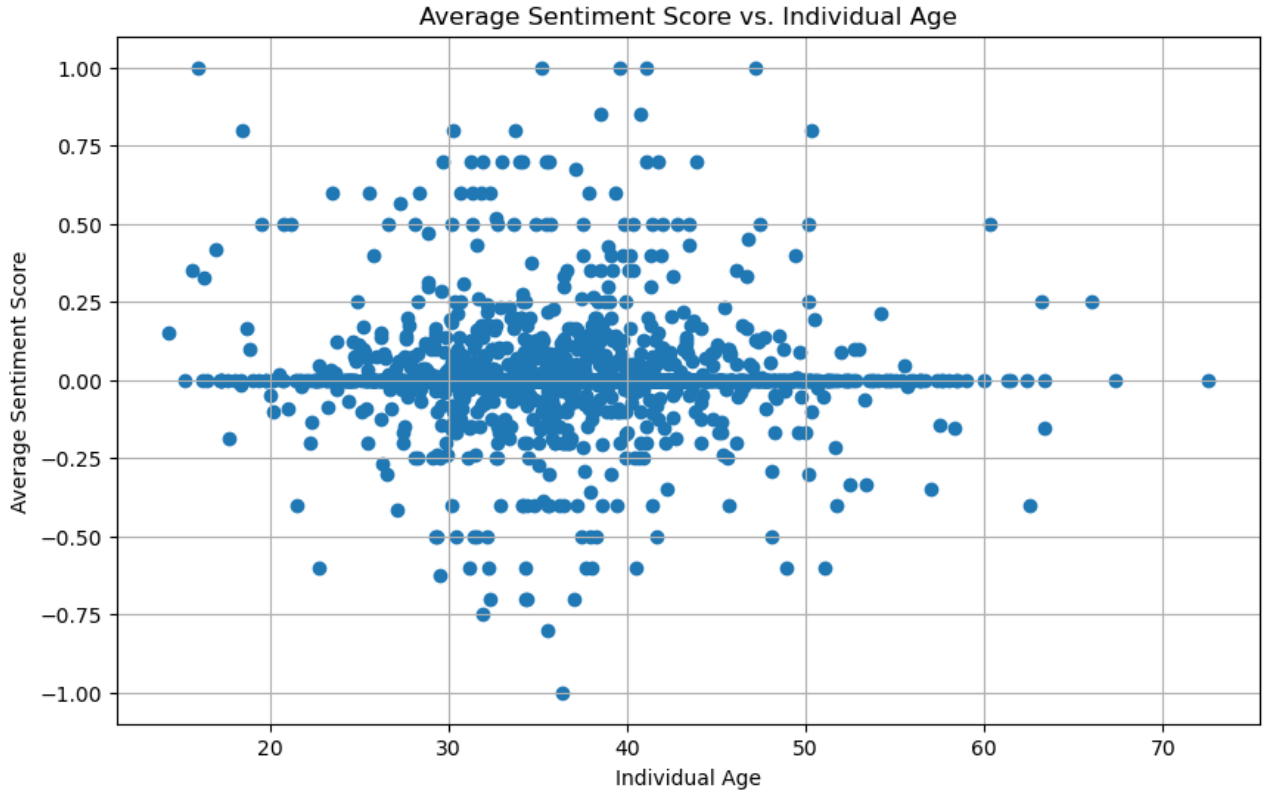
How does Age impact Sentiment?

Age vs Sentiment Score

The interest in variance in sentiment score, and hence the variance in tonality of the book title is heavily influenced further by the age of the individual.

As the users age classifies them into the Middle Aged, between 31 and 45, their interest in books with more variance increases.

Thus while tracking popularity overall, the mean sentiment score that correlates with the target region would reflect in the most popular option, this idealistic demographic does not reflect our bookstores demographic most likely.



Which leads us to...

How do we target popularity based on...

The Bookstore Demographic

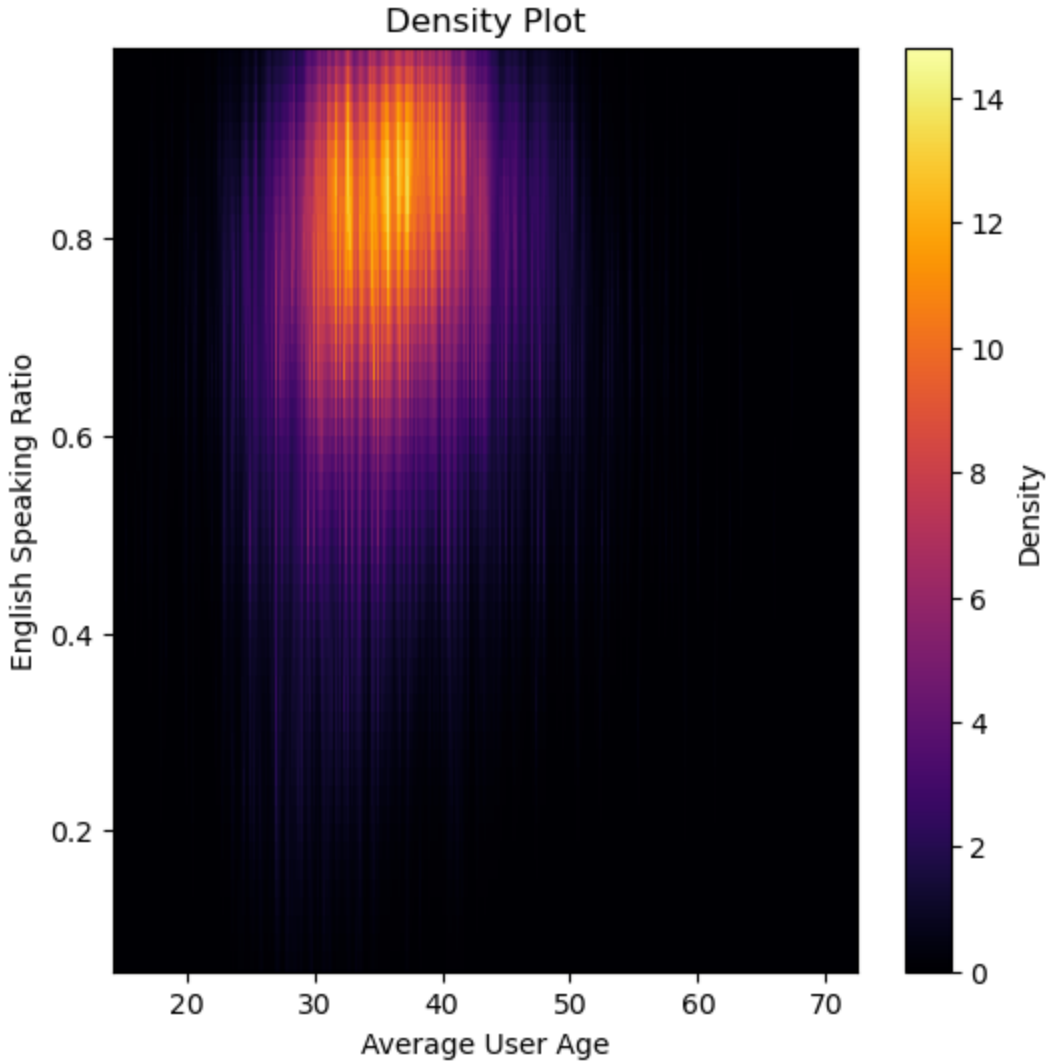
and what defines those demographics?

Age and Location

Given an age range and/or location demographic, we can tailor the recommendations using density plots.

To determine user age, we either utilise the provided age range demographic, or calculate the average age of a given location.

Subsequently, we can obtain the English speaking ratio by identifying the peak density point at that age value/range, thus allowing us to provide recommendations accordingly.



Nearest Neighbour Method

By calculating the total rating count, book recommendations are identified **among the vast collection** of books.

- ✓ Popular books, which received more ratings (above the threshold), were identified as books that would be the preference of the user which interested in books with similar ratings.

Five closest neighbors are being identified as the further recommendations book to the user who purchases the book with the title "To Kill a Mockingbird".

```
Recommendations for To Kill a Mockingbird:

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

Example of book recommendation

4

Findings and In-Depth Interpretation

FINDINGS AND RECOMMENDATIONS

Main Finding

Books published around 2000, consistently received **higher ratings**.

This could indicate improved book quality, changes in publishing standards, or possibly a shift in reader expectations, and review practices over the years.

Other findings

- ✓ Visible concentration of ratings between the 6 to 8 range across all decades. Notably, books are sparse with very high ratings (above 8.5) before the 1980s.
- ✓ The dispersion of points becomes slightly denser towards the later years, suggesting an increase in the volume of books published or possibly in the number of reviews they receive, which could influence average ratings.

Recommendation

The bookstore's manager should focus the inventory and promotional efforts **on the books published in recent years**. This ensuring that the bookstore collection remains relevant and appealing to its customers.

Read More

FINDINGS AND RECOMMENDATIONS

Demographic Preferences and Their Impact on Book Selection

Findings

- ✓ Age (children, young adults, middle-aged, and older adults) have a significant role in **Sentiment Score Analysis**, whereas specific age demographics exhibit preferences for book titles.
- ✓ An exponential correlation exists between book popularity and the English-speaking ratio*, particularly evident when focusing on the region of highest density concerning English speaking ratio versus average age per book.

*The United Kingdom, the USA, Canada, Australia, New Zealand, and Ireland

More Insights

- ✓ Reading preferences evolve with age, which influenced by changes in interests, lifestyle, and even cognitive engagement needs.

Younger readers might seek escapism and entertainment, which modern fiction genres provide.**Older readers** may prefer the intellectual stimulation or nostalgia offered by non-fiction and classics.

- ✓ Interestingly, the English-speaking ratio and number of reviews tend to group and clump at specific points.

This is suggesting that though have a slow start, as the English-speaking ratio increases, the number of book reviews also will increase significantly, reflecting **its user engagement**.

Recommendation

The bookstore's manager should focus the marketing based on user's demographic data, specifically age and location. For users who lived in the English-speaking countries, book inventory and marketing strategies should focus for young adults preferences.

FINDINGS AND RECOMMENDATIONS

User Engagement Levels and Their Satisfaction

Main Finding

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

The more engaged users are likely to be **more satisfied with their reading choices**.

The more satisfied users are **more engaged in rating activities**.

Creating a Positive Feedback Loop

Recommendation

The bookstore's manager should also **enhance user engagement** through **community-building features,** such as promoting book clubs, provide review rewards, and interactive recommendations.
This can foster a more active and satisfied customer base.

5

Limitations and Improvement Opportunities

LIMITATION AND IMPROVEMENT OPPORTUNITIES

Limitation

Data Completeness
The datasets used may not encompass all relevant variables that influence book sales and customer satisfaction, such as detailed customer purchase history, external economic factors, or comprehensive metadata on books (e.g., awards won). The absence of these variables might result in an incomplete analysis of factors driving book popularity.

Temporal Dynamics
The analysis largely treats the data as static, not accounting for potential changes in customer behavior or market conditions over time. Seasonal trends and economic cycles, which could significantly affect book sales and ratings, were not considered.

Potential Biases
The user data are self-reported and could contain inherent biases. For instance, users who actively participate in rating books might not be representative of the broader customer base. This selection bias can skew the results towards more engaged, possibly more positive experiences.

Geographic Representation
The data may disproportionately represent certain geographic areas, particularly where the bookstore has a stronger presence. This could limit the generalizability of the findings to global markets.

Improvement

- To enhance the robustness of future analyses, additional data sources should be integrated to fill gaps in the current dataset (30% data is missing).
- Implementing time series analysis would help understand the temporal effects on sales and ratings.
- Expanding the geographic diversity of the data would help generalise the findings across different markets.

Photo by vnwayne fan on Unsplash