A Book Story

Data Analysis Project
by Lainey Odette

The Project Overview

Considerations

This project explored a dataset of **readers**, **books**, and their **ratings** of books read.

The dataset was narrowed to examine books **published between 1980 - 2000** to analyze trends and patterns for the purpose of insight discovery.

Goals

The goal of this project was to develop insight to help **predict how a book will be rated by readers** and discover factors that support a prediction.

Approach

Data **cleaning** and **descriptive**, **exploratory**, and **visual analyses** were conducted with Python in a Jupyter notebook. The analyses and insights derived drove the hypotheses explored.

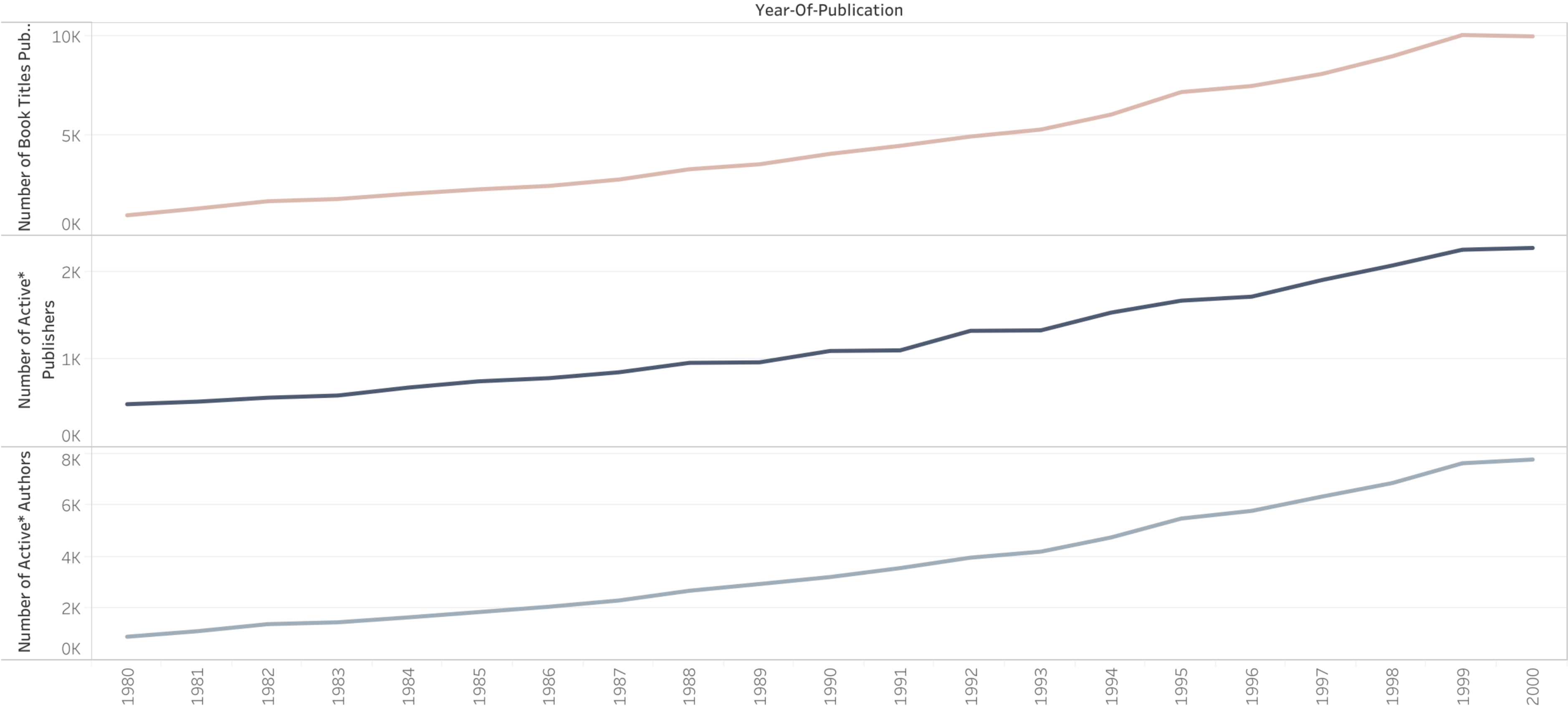
Final insights, visuals, and story were assembled in Tableau.

The Publishing Landscape

The publishing landscape **grew dramatically** from 1980 through 2000.

Overall:
- more books
- more authors
- more publishers

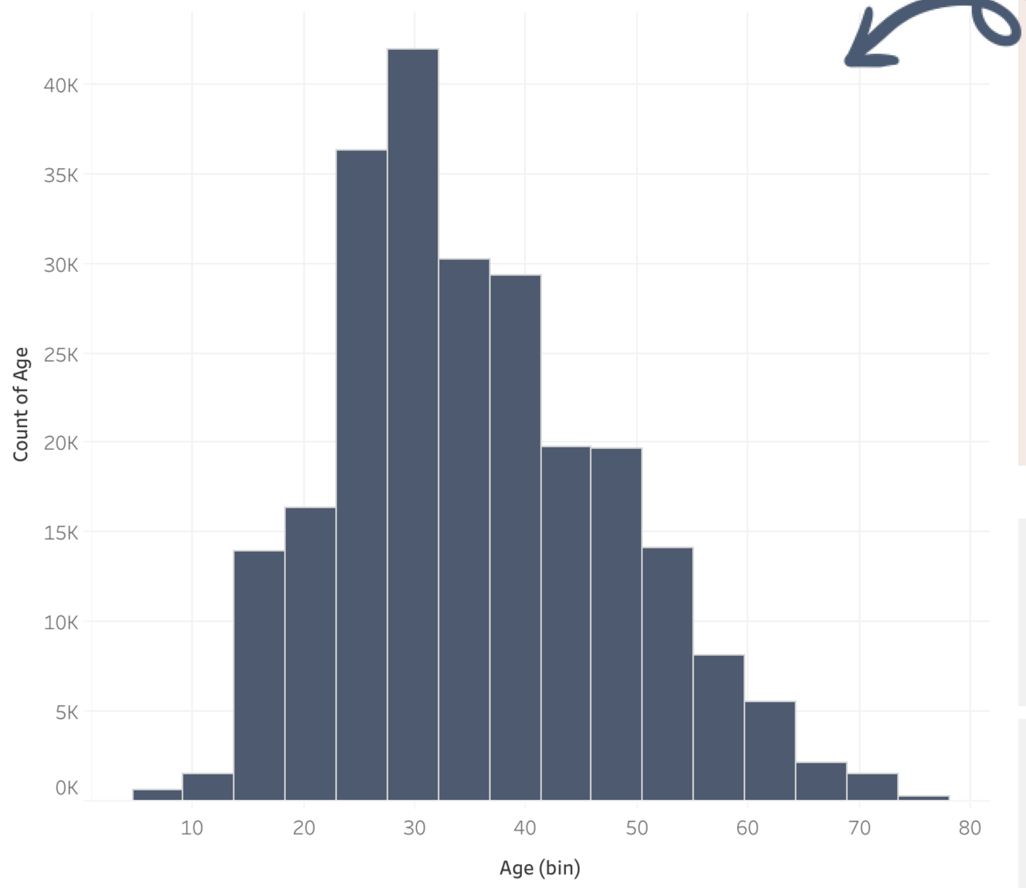
Publishing Landscape (1980-2000)



*Active defined as book published in that year

The Reader Landscape

Number of Readers by Age



Reader Age Range

Readers used in the analysis vary in age with the majority in their mid-20's to late-30's.

Reader Location

Readers span a wide range of countries across the globe with the majority in the U.S.

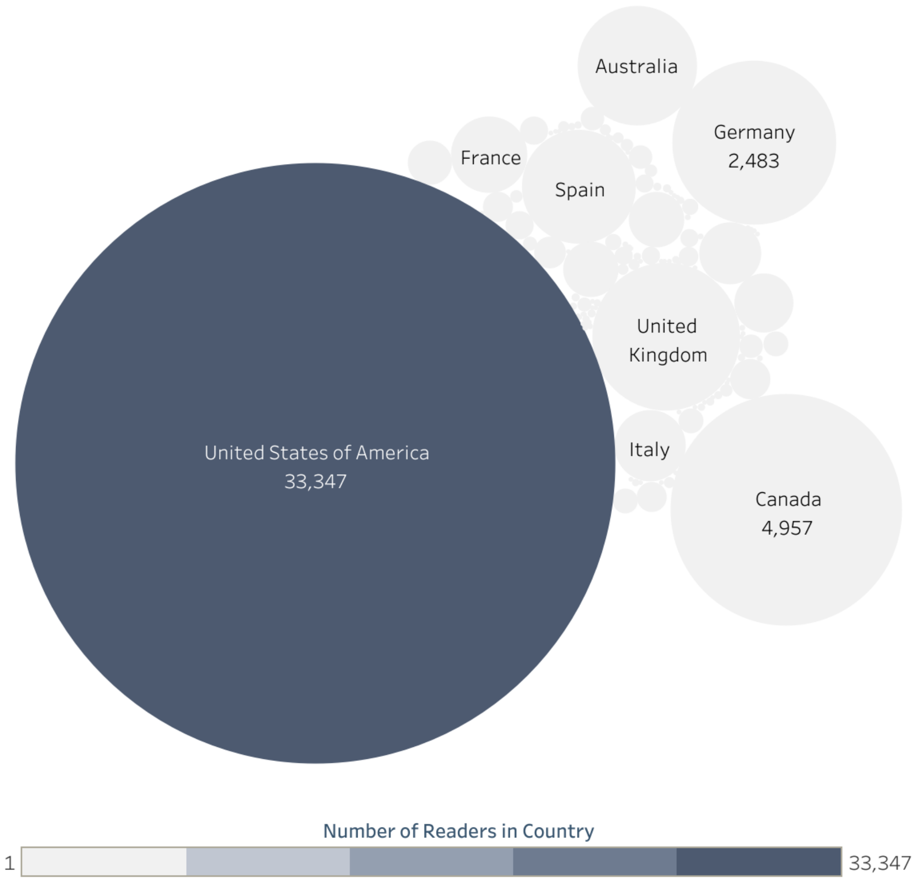
Number of Readers

49,162

Number of Countries

113

Number of Readers by Country



The Big Question

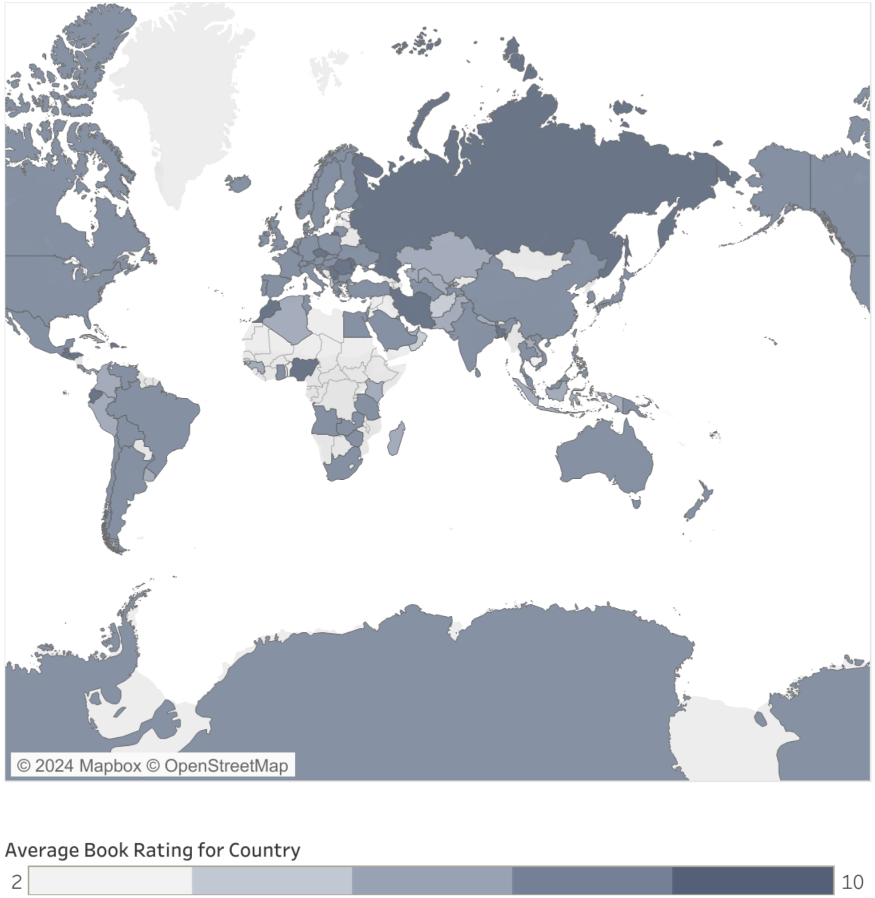
Is there a factor that can help predict book ratings?

Hypothesis 1:
Reader location affects book rating.

Hypothesis 2:
Reader age affects book rating.

Does a reader’s location help predict book rating?

Average Book Rating by Country


© 2024 Mapbox © OpenStreetMap
Average Book Rating for Country
2
10

Spatial Analysis

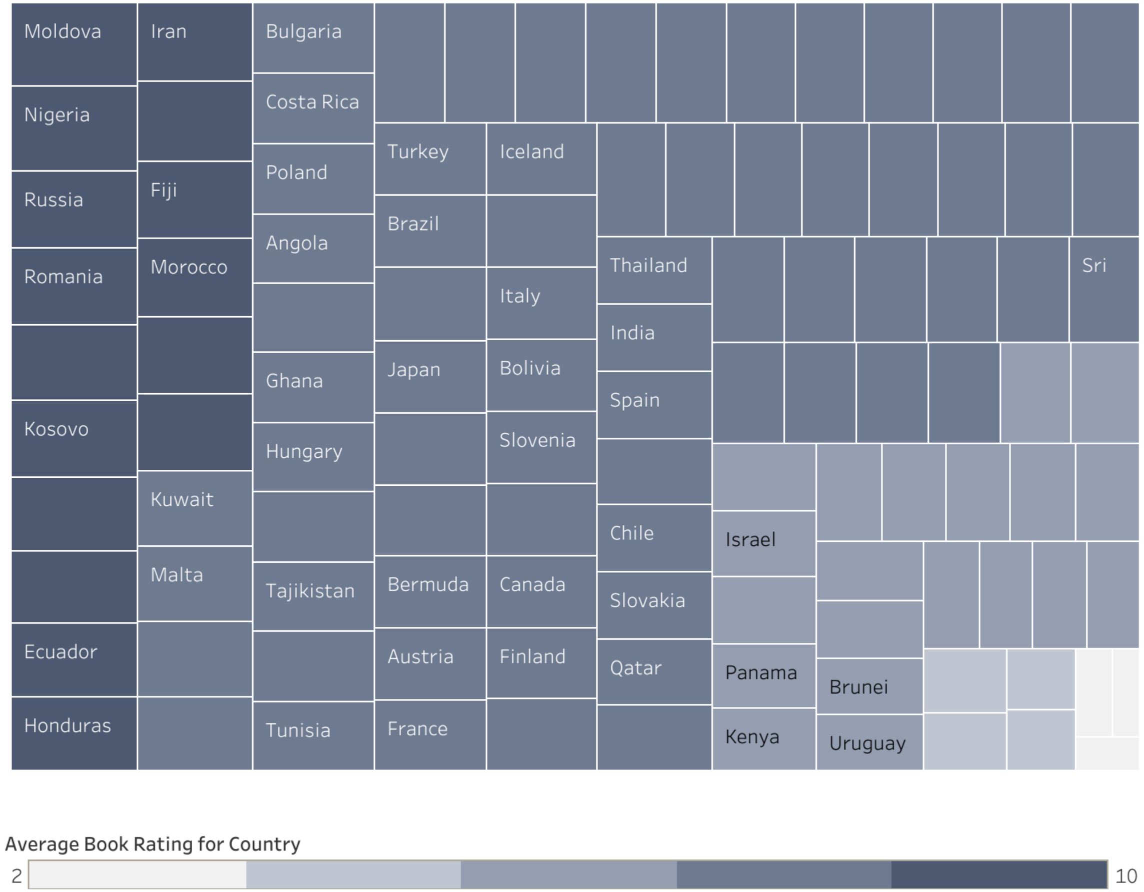
Mapping analysis did not reveal any clear pattern to indicate a reader’s country is a fair predictive factor in book rating.

Treemap Analysis

Additional visual analysis confirms that country is not a clear predictive factor.

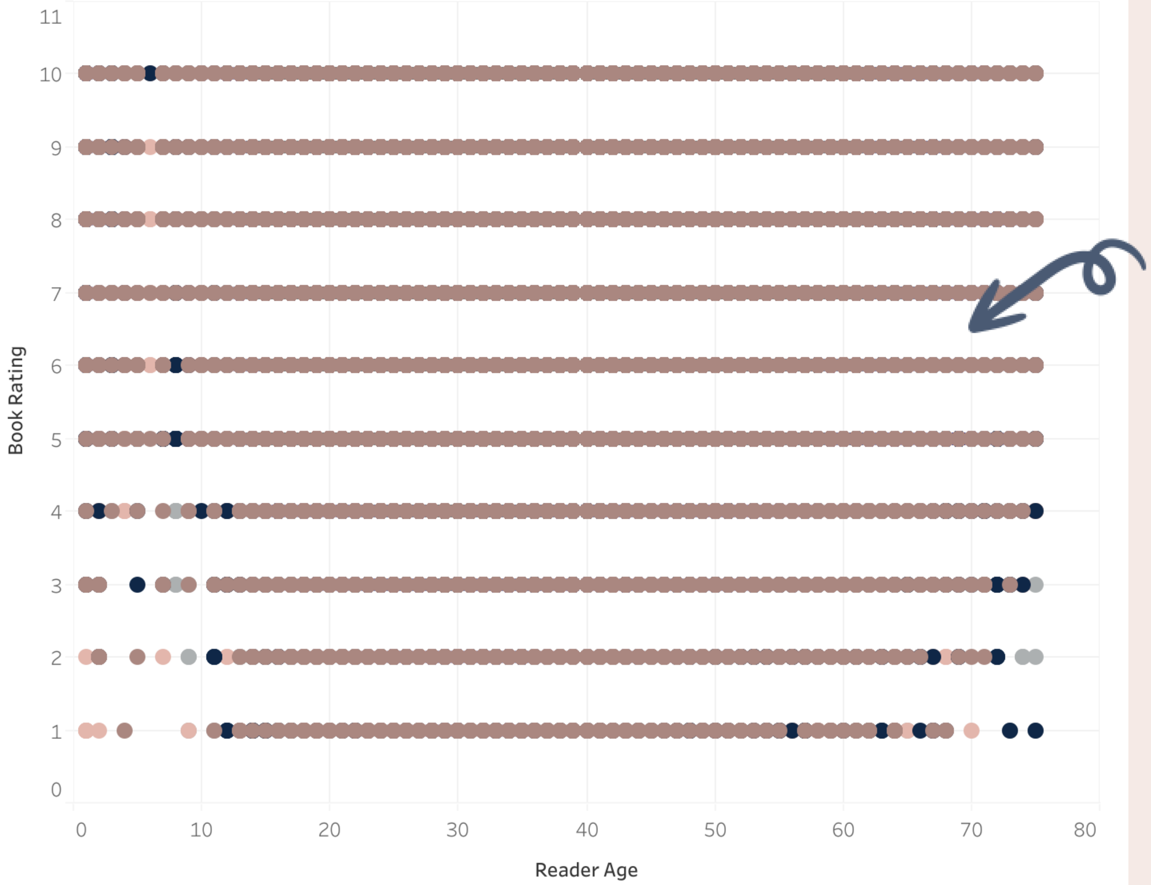
Logic: the most frequent rating for a book is ”8.” The same is true for the majority of countries, indicating country is not a factor.

Average Book Rating by Country



Does a reader’s age help predict book rating?

Scatterplot Cluster Analysis (Reader Age v Book Rating)



Scatterplot Analysis

The pattern with scatterplot alone hinted that younger and older readers were less likely to provide a high or low rating.

Scatterplot Cluster Analysis

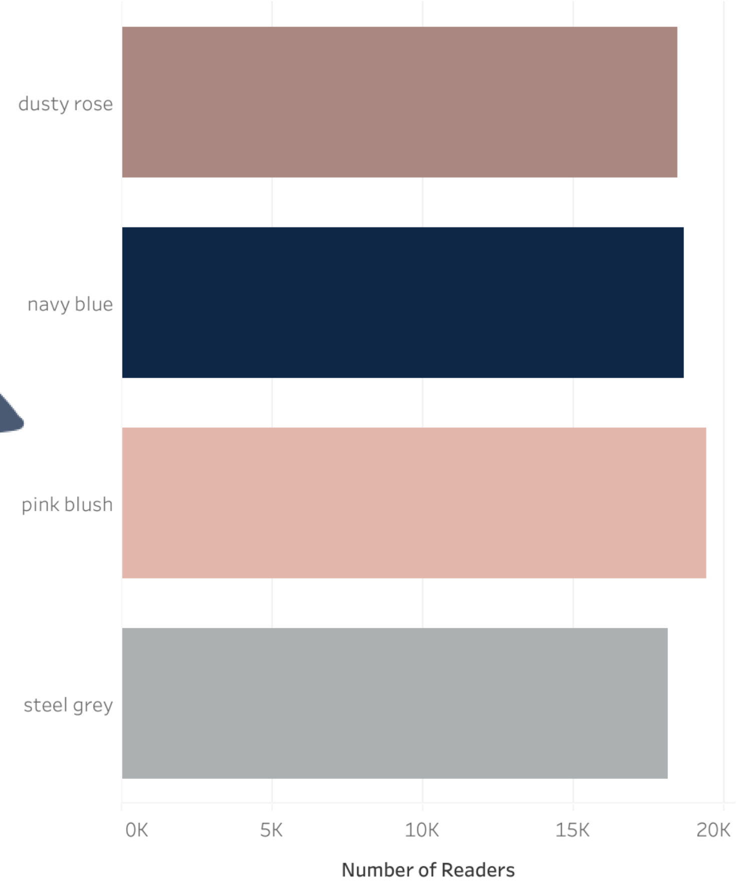
Paired with a cluster analysis, that potential pattern was debunked.

As each cluster is essentially equal in size, the gap in the clusters is more indicative of a lack of readers at the age versus a lack of rating present for that age.

Regression Analysis

Was completed but found not suited for this project. Results n..

Number of Readers by Cluster



Conclusion

Reader location alone is not a clear indicator of book rating.

Reader age alone is not a clear indicator of book rating.

Additional analysis and variables needed to uncover predictive factors.

Next Steps

Additional Data

Secure **additional data** for further analysis.

- genre detail on books
- more recent publications (for more data points)
- more recent rating information (for more data points)

Further Analysis

Calculate **new variables** - such as number of books published by author - to deepen analysis.

Use new and additional data to **continue analysis**.

Appendix

Datasets

Book, User, and Ratings Dataset downloaded from Kaggle
https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset?rvi=1

Collection Methodology: Collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. Contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

Book Title, Book Author, Year of Publication, Publisher pulled from Amazon Web Services.

License: CC0: Public Domain

Countries GEO JSON Files downloaded from Kaggle
https://www.kaggle.com/datasets/ktochylin/world-countries/data

Limitations

- Author only lists first author vs all authors of a work.

- Due to the large number of missing age values, the missing values were imputed with random data based on the make-up of the present age values.

- A large number of the "ratings" values were "0" and therefore removed to prevent the data from skewing analysis results.

- The location information was not consistent and contained city, state/region, and country information all in one column. These were separated to isolate country but some lines did not provide enough context to determine the correct country so some lines were removed altogether for the geographical analysis.

- Only two numerical variables were included in dataset so analysis such as correlation and regression was limited.

- Scrub from Book Crossing was completed in 2004, which is early in the internet days. No telling how complete the information actually is, as it is unclear where Book Crossing got their book information from. Also user information was not uniform as there was clearly not a lot of effort in collection to extract accurate/consistent information. Also, simply by the fact that this is around the inception of the internet, the users who have provided ratings are possibly skewed to more tech-savvy or other bias in that this was early on in the usage of online tools. Many opportunities for bias surrounding all data in set.

Appendix, cont.

Image Credits

People on books graphic on Cover by graitofendyn on Vecteezy
https://www.vecteezy.com/free-vector/reading-book

Woman Reading Book graphic on Overview and Appendix slides by doraclub on Vecteezy
https://www.vecteezy.com/free-vector/reading-book

Woman Leaning on Question Mark graphic on Big Question slide by frehanvect on Vecteezy
https://www.vecteezy.com/free-vector/question-mark

Woman Pointing graphic on Conclusion slide by gstudiomimagen on Vecteezy (combined with art from Question Mark graphic)
https://www.vecteezy.com/free-vector/woman-pointing

Pile of Books graphic on Next Steps and Appendix pages modified from original "People on Books" graphic by graitofendyn on Vecteezy
https://www.vecteezy.com/free-vector/reading-book

Arrows throughout storyboard created by Peri Priatna on Vecteezy
https://www.vecteezy.com/free-vector/arrow