Readership and Gender in 2000s America

Jenna Poliansky
School of Computing and Data Science

***Abstract*—This study aims to explore the trends in readership and book genres in the U.S. and investigate their relationship with gender. By using various data on book sales, genre trends and demographic reading data on gender, the research explores the connection between gender and reading over time. Through data visualization and analysis, the study explores the reasoning behind the gender gap in readership. Notably, this study reveals the driving sociocultural and technological forces behind this gap.**

Keywords*—*Print Book, E-book, Tiktok, #BookTok, Genre

I. Introduction

American readership trends have changed throughout the years but never more significantly than in the 21st century.

II. Datasets

A. Source of dataset

The first dataset, sourced from Statista, tracks the number of print books sold from 2004-2024. The second, also sourced from Statista, shows the number of E-Books sold from 2010-2020. These were both compiled by a media researcher who looks at the fast-paced, ever-changing media landscape and how it impacts traditional formats. In this study, these datasets were used to look at the relationship between these formats and readers.

The third dataset from Statista looks at another technological impact on readership: TikTok. The social media app TikTok, created in 2016, has grown in usership immensely. One specific community on TikTok using the hashtag #BookTok, has also grown in users. This dataset looks at how many Americans, per state, are reading more because of this hashtag.

The fourth dataset called "Best Books Ever", sourced from Kaggle, is a comprehensive list of popular books published between 1916-2017. By using a stacked surface area plot to graph the genre of these books, the study looks at how the trend of genres has evolved over time. Lastly, the fifth dataset from Statista looks at the gender comparison of readers categorized by genre. By using this dataset alongside the fourth one, a correlation between gender and genre can be seen.

B. Character of the datasets

The print book and e-book datasets each have two variables; year and number of books sold. These datasets were combined to be used for a model by merging each of their number of books sold columns and using a range of years included in both sets.

TABLE 1. Total Books Sold data

Field Name	Description	Format
Year	A year books were sold	YYYY
Print books sold	The number of print books sold per year	Integer
E-books Sold	The number of e-books sold per year	Integer
Total books sold	The sum of print and e-books sold per year	Integer

The third dataset contains a column of states and a column of the percentage each state is reading more because of TikTok. The fourth dataset had numerous, comprehensive variables for over 50,000 books. However, not all were necessary for this study. To look at the trend of genres, the variables used were "bookId", "genres" and "firstPublishDate". The "bookId" variable was used to clean the data for each book.

TABLE 2. Best Books Ever relevant data

Field Name	Description	Format
bookId	The numeric ID of the book followed by its name	XXX.name
genres	The genre(s) the book is a part of as a list	['genre1', 'genre2'..., genreX]
firstPublishDate	The date the book was first published	Date (MM/DD/YYYY)

To track the trend of genres per year, each genre a book was a part of had to be counted. Data cleaning was performed to go through each book's list of genres and strip out each genre, putting it in its own list. Then, a count was established for each genre. Additionally, any rows of data that were empty or duplicated (same book and year) were removed. The year was extracted from firstPublishDate as well.

The fifth dataset has three variables; an unnamed column referring to genre, a column for male readers and a column for female readers, both in percentages. The male and female reader percentages refer to the percentage of each gender reading books in the specified genre.

III. Methodology

A. Data Visualization

In order to conduct data visualization, data cleaning and analysis were performed using Python in Visual Studio Code using Jupyter Notebook. Libraries include numpy, matplotlib, pandas, seaborn and ast. The aim of this step was to provide clear visualizations that create insights into the topic of this study.

Tools and Techniques:

- Visual Studio Code: VS Code was used to establish a Python environment for the Jupyter Notebook used for data cleaning and visualization.
- numpy: Used for numerical operations, specifically on arrays and matrices.
- pandas: Allows for ease of data cleaning and manipulation.
- matplotlib: Used to plot data in various formats in order to visually assess trends and distributions.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

- seaborn: Another library used for graphs and has a higher-level interface and design aspects.
- ast: A library that works with Abstract Syntax Tree of Python source code. For this project, it was used to evaluate strings contained in a list for one of the models.

A. *Scatter Plot Models*

For the two datasets on books sales, and the combined set made for this study, scatter plots were used to display the trends. Because the data is discontinuous and graphed over time, a scatter plot was the best model.

B. *Bar Chart Models*

A bar chart model was used for data that was organized categorically.

C. *Stacked Surface Area Models*

To track the trend of the popularity of genres over time, a stacked surface area model was used.

Advantages: A stacked surface area plot makes the cumulative trend of genres more clear over time.

Disadvantages:

IV. Results

A. *Books Sold Over Time*

Three scatter plots were made to show the number of books sold over time. The first one shows print books sold from 2004-2024. The second tracks e-books sold from 2010-2020. The purpose of these is to show the impact of technology on readership and the medium readers use.

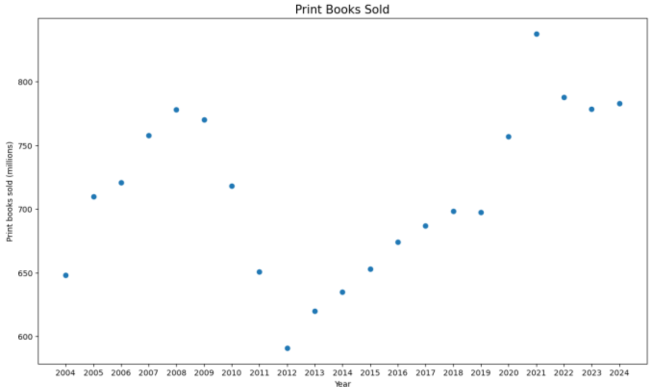

Fig 1. The number of print books sold in the U.S. from 2004-2024

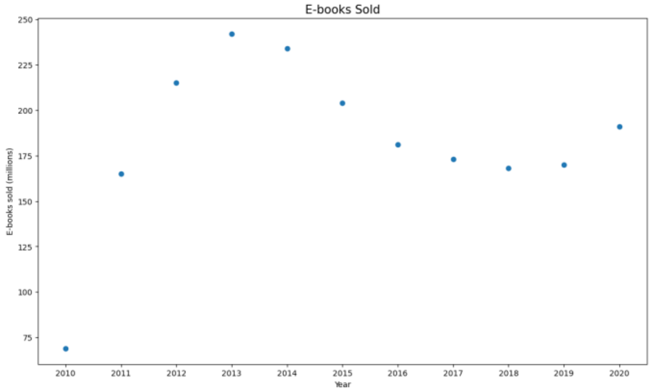

Fig 2. The number of e-books sold in the U.S. from 2010-2020

The third scatter plot combines print and e-books to display the overall trend of books sold in the U.S. While varied, the trend shows an overall increase over time.

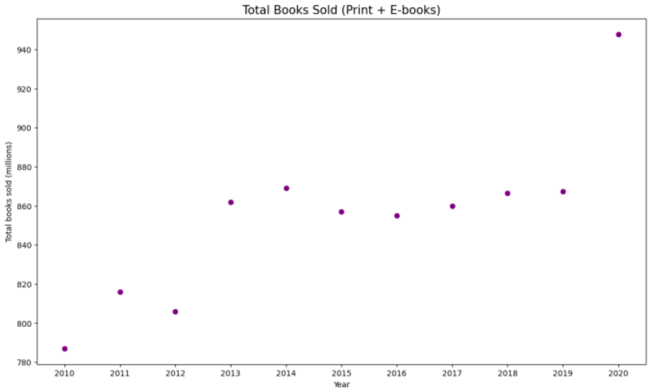

Fig 3. The total number of books sold in the U.S. from 2010-2020

B. *TikTok Users Reading More Because #BookTok*

For this model, the data was sorted categorically (per state). A bar chart was used to show the percentage of TikTok users per state reading more because of #BookTok.

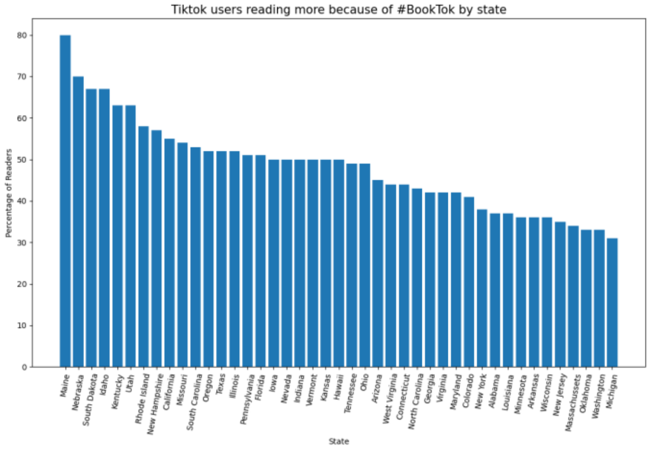

Fig 4. The percentage of TikTok users per state

C. *Genre Popularity Over Time*

The following model is a stacked surface area plot the shows the count of genres of the most popular books published from 1916-2017. The genres displayed are any genre that were in the top 5 genres in a given year.

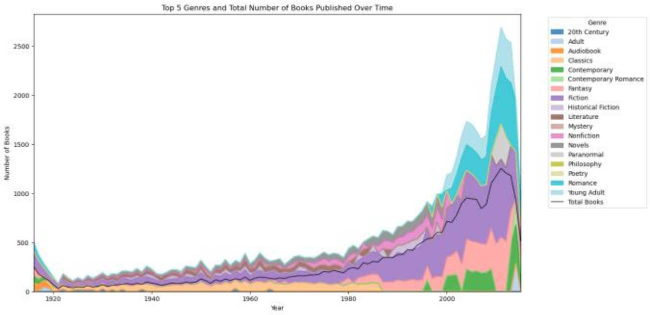

Fig 5. The popularity of the top 5 genres of any given year over time

D. *Genre Popularity in the 2000s*

The following model uses the same dataset as the previous but only displays the trend of top 5 genres that were in the top 5 of any year post 1999.

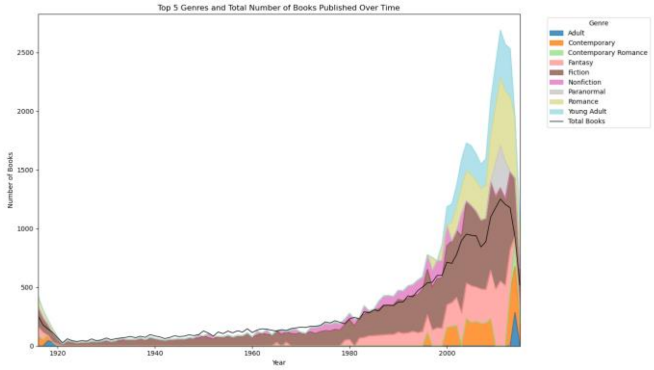

Fig 6. The top 5 genres only included in the 2000s

E. Genre Popularity by Gender

The following model displays the percentage of readers per genre categorized by male and female readers. This data was taken from 2015, and research shows that this trend has persisted. A bar chart was used in order to display each genre and then the response variable categorized by gender.

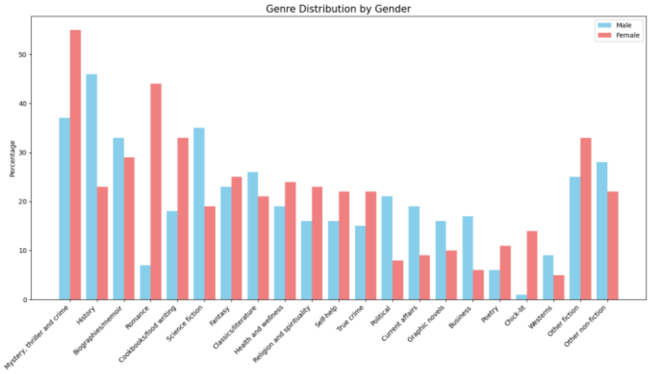

Fig 7. The percentage of respondents categorized by gender per genre

V. Discussion

VI. Conclusion

Acknowledgment

References

[1] A. Watson, "U.S. e-book unit sales 2017,"*Statista,* 2025. https://www.statista.com/statistics/426799/e-book-unit-sales-usa/

[2] A. Watson, "U.S. print book sales 2020," *Statista*, Jan. 18, 2024. https://www.statista.com/statistics/422595/print-book-sales-usa/

[3] "#BookTok impact on reading books in the U.S. by state 2023," *Statista*. https://www.statista.com/statistics/1398645/united-states-booktok/

[4] "Leading book genres in the U.S. by gender 2015," *Statista*. https://www.statista.com/statistics/470748/favorite-book-genres-gender-usa/

[5] Pooria Mostafapoor, "Best Books Ever Dataset," *Kaggle.com*, 2024. https://www.kaggle.com/datasets/pooriamst/best-books-ever-dataset?resource=download