Report I

Data Visualisation: IESM 343

Influence of People's Age on Reading Frequency

Professor: Gurgen Hovakimyan
E-mail: ghovakimyan@aua.am

Team:	**E-Mail:**
Arman Manukyan	arman_manukyan@edu.aua.am
Mariam Manukyan	mariam_manukyan@edu.aua.am

I. Introduction

As people grow older, their preferences and behavior naturally evolve from the books they read to the way they shop. But how exactly does age influence everyday decisions? Can we observe meaningful patterns just by looking at reading habits or purchase data? These were the questions we wanted to explore through this project.

To do so, we combined data visualization techniques with a real-world dataset that reflects both cultural interests and consumer behavior. Our goal was to design an interactive dashboard that not only reveal hidden trends but also help make sense of them.

Purpose and Objectives

This project came from our curiosity about how age affects everyday behavior, not just what people read, but also how they spend their money. We wanted to explore how habits and preferences change as people grow older, and whether we could find patterns by looking at both global and local data. Instead of studying just one type of behavior, we focused on two different parts of life: reading habits and shopping activity.

To do this, we created an R Shiny dashboard. This is built using cleaned data from Kaggle and focused on reading preferences across different genres, countries, and age groups. We were especially interested in seeing whether younger people prefer Fiction more, and how genre diversity changes with age. The dashboard included regression analysis, genre distributions, and country comparisons to show how people's reading choices develop over time.

By comparing the reading and shopping behaviors side by side, we hoped to get a better overall picture of how age shapes people's decisions, both culturally and economically.

Our goal wasn't just to make nice visualizations, but to create tools that actually help people explore and understand the data. With the help of statistical summaries and interactive plots, we wanted to make these patterns easier to see, explain, and learn from.

Overview of the Dataset

We used **International Readers'** data sourced from Kaggle: Book Readers Dataset. It includes information on 3524 readers (after cleaning), enabling a comprehensive examination of reading patterns and genre engagement across demographics. After cleaning, it allowed us to study how different age groups engage with genres, how reading habits shift across countries, and whether Fiction remains popular as readers grow older.

This dataset provided both cultural and behavioural context, helping us understand how personal habits evolve with age, whether it's about what people choose to read or how they spend money. We also considered the real-world relevance of this data, especially for publishers, marketers, and analysts looking to serve age-targeted audiences better.

Reading Dataset (Kaggle)

A global dataset with over 5000 records originally, filtered and cleaned down to relevant fields such as age, genres, publisher, and country. It allowed us to examine age distributions by genre, track Fiction popularity, and compare geographic reading trends.

After cleaning up the data, this dataset had 3524 entries, each containing a user's age, genre preferences, and reported reading frequency. The steps we followed were:

- Removed unrealistic ages (below 10 or above 80)

- Focused on the most common genres, especially Fiction and Nonfiction
- Converted frequency responses to numeric scales for comparison
- Created scatter plots of reading frequency by age, colored by genre
- Added regression lines to track genre preference changes
- Built bar plots grouped by age ranges (15-25, 26-35, etc.)
- Added histograms and density plots to show the shape of frequency distributions
- Used skewness and kurtosis to assess distribution characteristics per genre

Highlighted genre shifts across age with line plots

II. Methodology

Explanation of Data Processing Steps

We performed all data cleaning and transformation steps using R, mainly with the dplyr, tidyr, and lubridate packages. These were the main actions:

- For the reading dataset (df_newcleaned.csv), we first removed invalid entries such as rows with missing age, unrealistic ages (under 10 or over 80), and empty genre fields. We used separate_rows() to split users who reported multiple genres into individual rows, which allowed us to calculate frequency and age statistics per genre.
- We used stringr and countrycode to clean country names and convert them into ISO3 format for mapping.
- Age was treated as a numeric variable throughout, and reading frequency responses (like “often” or “rarely”) were converted into numerical values for comparison.

- We computed skewness and kurtosis (using the moments package) for both age and revenue variables to understand distribution shapes and support our histogram interpretations.
- We grouped and summarised the dataset using group_by() and summarise() functions to calculate count, mean, median, and spread values per category.

The processed dataset was then used in the Shiny dashboard as reactive data frames, allowing user filtering and dynamic visual updates.

Data Cleaning and Processing

We used dplyr, tidyr, and lubridate in R for transforming the dataset.

Cleaning steps included:

- Removing null or unrealistic values (e.g., ages < 10 or > 80)
- Splitting multiple genres per reader using separate_rows()
- Converting timestamps to extract Hour, Weekday, and Month
- Grouping revenue per age, gender, and product
- Transforming reading frequency into numeric scales for analysis
- Mapping country names to ISO3 codes for geographic visualizations

Dataset was saved as df_newcleaned.csv and reloaded dynamically in the dashboard.

Justification for Visualization Choices and Techniques Used

Our visual design choices were driven by both interpretability and user experience.

- We used **histograms** and **density plots** to show distribution of age and revenue because they're intuitive for visualizing skewness and concentration.
- **Bar plots** were ideal for categorical comparisons (e.g., top genres, top products) and were paired with color encoding (such as average price) to add another variable without overloading the viewer.
- **Boxplots** allowed us to compare variation in age or revenue across categories (e.g., genres, genders, age groups), and highlighted outliers clearly.
- **Scatter plots with regression lines** helped us demonstrate trends.For example, the age-based Fiction reading trend was visualized with both a linear and a logistic model to check fit.
- **Treemaps** were used to visually represent genre volume in a compact, area-based form, which works well when there are many categories of unequal size.
- **Facet plots** grouped genre preferences by age range and made age comparisons across genres much easier.
- **Line plots** were used for monthly and weekly revenue trends to show seasonality and peak patterns over time.
- For geographic insights, we used **choropleth maps** to show reader count, average reader age, and genre diversity by country, taking advantage of spatial layout to tell regional stories.
- We also included a **summary statistics panel** and embedded data tables for transparency and further exploration.

We structured the dashboard with tabs and filters to allow users to explore findings interactively and at their own pace, avoiding visual overload while maintaining flexibility.

Visualization Tools and Design Choices

We used the following R packages:
- ggplot2 for all core plots
- plotly to add interactivity
- treemapify for genre area visualization
- rworldmap and countrycode for mapping country-level stats
- shiny, shinyWidgets, and shinythemes for dashboard UI
- moments to calculate skewness and kurtosis of distributions

We chose a tab-based layout in our dashboard to guide users through exploration. Each plot was labeled and grouped logically (e.g., temporal patterns, genre-age relationships, demographic breakdowns). We also added download buttons and filters so users could interact with the data in real-time.

III. Findings and Insights

Justification for Visualization Choices and Techniques Used

When selecting our visualizations, we tried to strike a balance between clarity and depth. Since our dataset dealt with human behavior over time and across categories, we needed visuals that could show variation, trends, and group differences clearly, without overwhelming the viewer. Each plot type was chosen to highlight a specific relationship or distribution in the data.

For example, we used histograms and density plots to examine how values like age or revenue were distributed. These are intuitive tools for identifying skewness and outliers, and helped us understand whether our assumptions about “typical” users were accurate. When we wanted to compare spending or reading behavior across categories (like genre

or gender), we used grouped bar charts and boxplots, these made differences more visible and were easier to interpret than tables or raw numbers.

Scatter plots with regression lines were used when we wanted to explore relationships between continuous variables, like how Fiction reading changes with age. In some cases, we applied logistic regression or LOESS smoothing to help illustrate these trends more clearly.

We also included map-based visualizations to show geographic variation, which added a spatial dimension to the reading dashboard. Treemaps were useful for summarizing categorical distributions (like genre popularity) when space was limited, and they helped quickly identify which items dominate.

Overall, we tried to build a dashboard that allowed users to explore freely, using tabs, filters, and interactive components where appropriate, but also structured enough to tell a clear story through each visual section. Every visualization was chosen based on what kind of question we were answering and how best to show the answer visually.

Interpretation of Visualizations and Key Takeaways

Across the dashboard, several clear behavioral patterns emerged from the visualizations.

The reading dashboard showed that Fiction is most popular among people aged 15 to 40. After that, its popularity gradually declines. We saw this clearly in the scatter plots, where a negative regression slope highlighted the trend. Nonfiction had a more gradual increase, becoming more popular from the early 30s onward.

Bar plots grouped by age showed that the 26–35 group had the highest average reading frequency. Histograms and density plots revealed that Fiction reading was right-skewed, indicating a concentration of frequent readers in younger groups. Nonfiction had a more balanced distribution, with lower skewness and less variation.

Skewness and kurtosis values supported these observations:

- Fiction had higher positive skew and moderate kurtosis
- Nonfiction had near-zero skew and lower kurtosis

These patterns suggest that Fiction appeals more to a specific age segment, while Nonfiction is read more evenly across ages. It also hinted that younger readers are more likely to engage intensely with one genre.

Summing up, from the **reading dataset**, we learned that:

- **Fiction is most popular among younger users**, especially those between 15 and 40. This was seen in both the bar plots and regression trend lines, which showed Fiction declining steadily with age.
- In contrast, **Nonfiction and Biography genres increase in popularity with age**, which was especially visible in the boxplots and faceted genre-age histograms.
- **Treemaps and bar charts** highlighted that just a few genres dominate the overall volume, even though some lesser-known genres have concentrated age groups.
- **Map plots** revealed strong geographic variation. For instance, the United States had the highest Fiction reader count, but Iran and Canada had notably younger average readers. Genre diversity also varied, with Western countries having broader genre engagement.

These insights confirm that user behavior is heavily influenced by age, not just in what people choose to read, but also when and how they spend money.

Business or Real-World Implications of the Results

The patterns we discovered have several practical applications, especially for industries that rely on understanding user behavior, such as publishing, retail, and digital marketing.

In the case of the global reading data, **publishers and content platforms** could use these insights to better target their audiences. For example, since Fiction is clearly more popular among younger users, companies could tailor campaigns to that age group using more youth-oriented channels like Instagram or TikTok. Meanwhile, Nonfiction genres, which are more common among readers 30 and older, could be promoted through professional networks or newsletter formats.

Libraries and cultural institutions could also benefit by designing age-specific reading programs. Knowing which genres attract certain age groups can help them organize events, curate recommended reading lists, or adjust funding for collections.

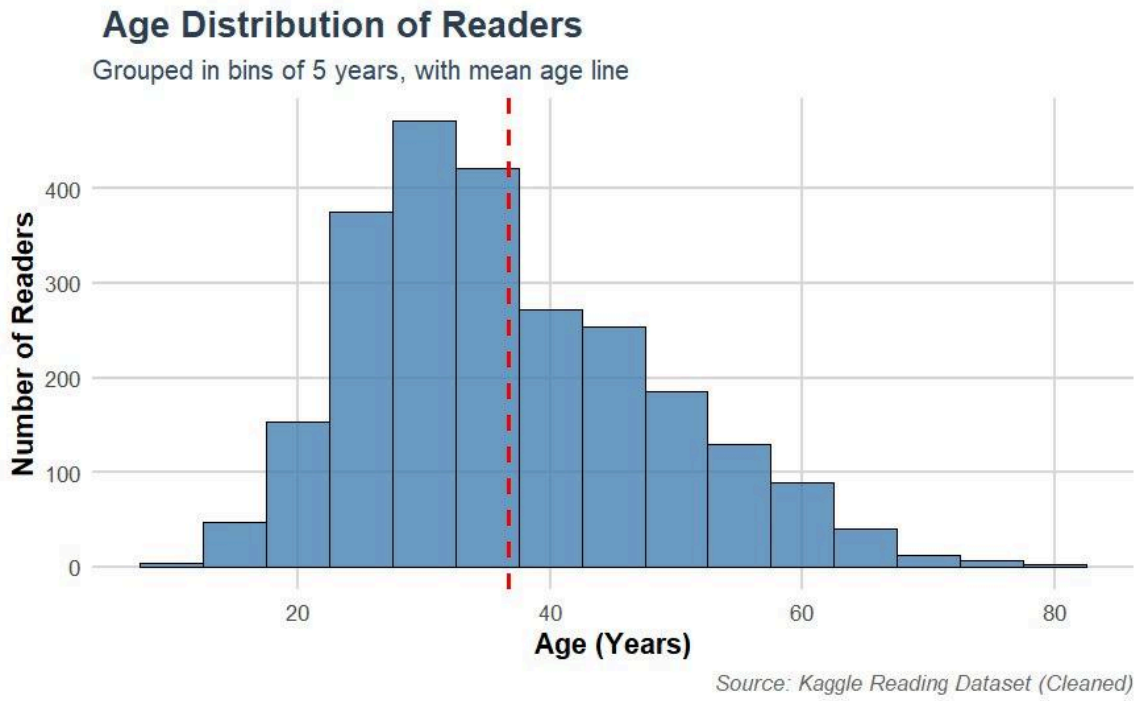
Product pricing strategies can also be informed by understanding which age groups spend more and on what types of items. **Marketing teams** could use this data to customize offerings by age range, pushing higher-end items to older segments or promoting more frequently purchased items to younger shoppers.

Finally, for any business with a digital presence, an **interactive dashboard like the one we built could be adapted for internal analytics**, giving non-technical users easy access to customer insights that are usually buried in raw spreadsheets.

IV. Visuals in the Paper

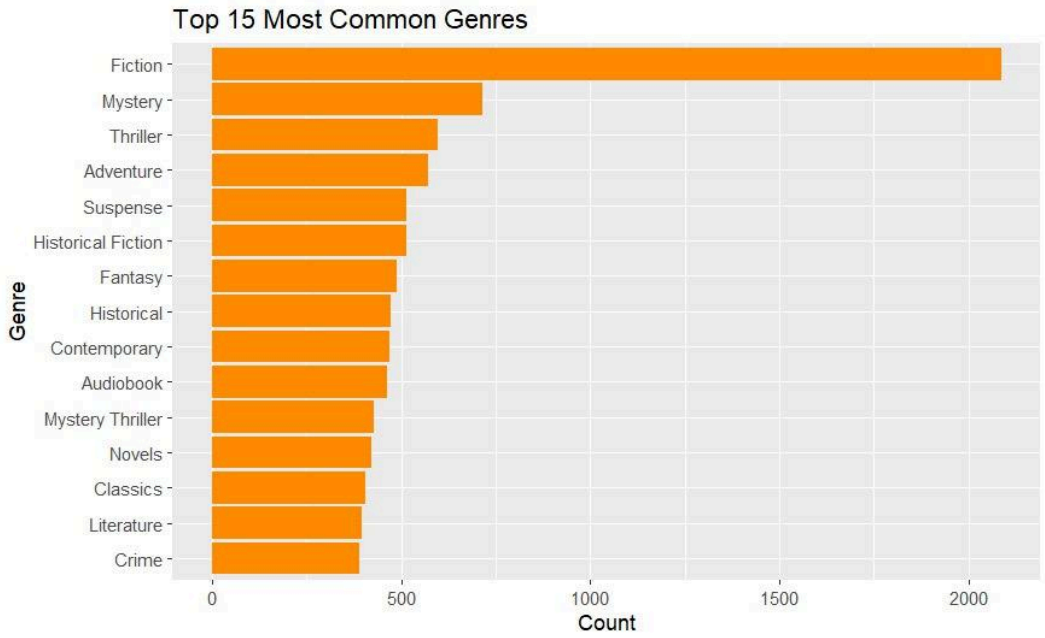
Each figure in our report directly reflects the code and outputs from the dashboard. Visuals were made using ggplot2 and customized for clarity.

Figure 1: Age Distribution of Readers



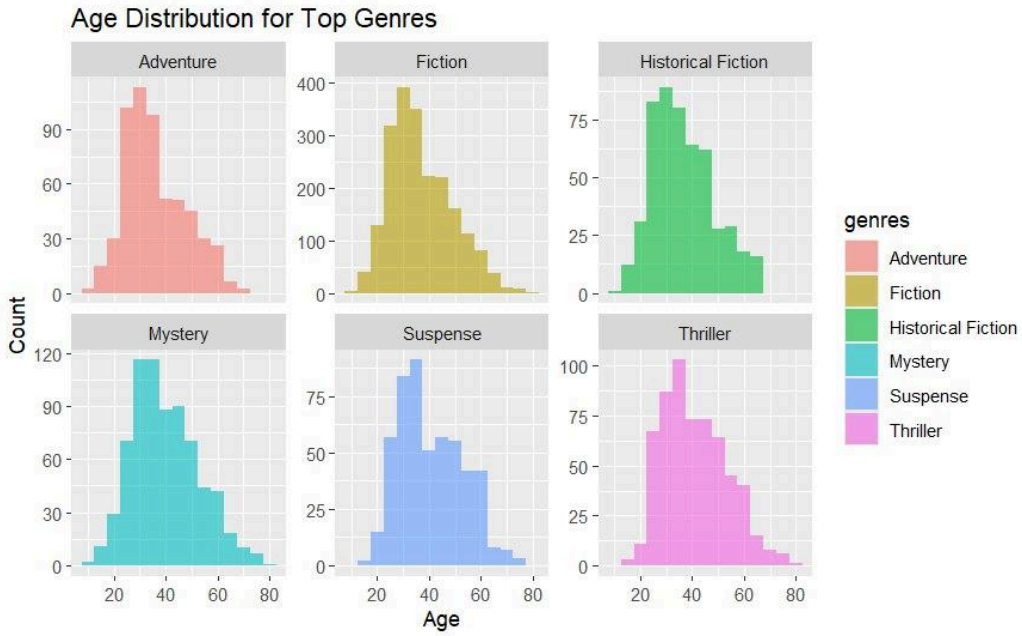
This histogram shows the number of readers across different age groups, grouped in 5-year bins. The red dashed line represents the mean age of the dataset. The plot reveals that most readers are in their late 20s to mid-30s, with a gradual decline in older age groups.

Figure 2: Top 15 Most Common Genres



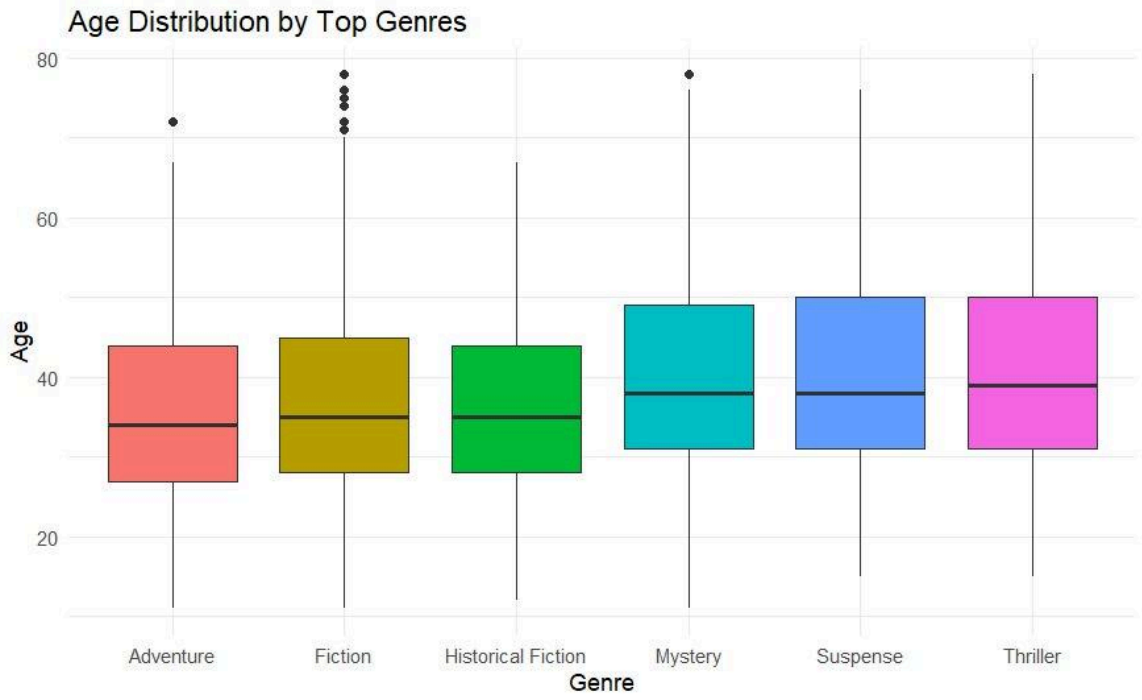
This horizontal bar chart ranks the 15 most frequently read genres based on user data. Fiction overwhelmingly leads, followed by Mystery and Thriller. The graph highlights a strong preference for narrative-driven genres, suggesting that most readers gravitate toward immersive and suspenseful content.

Figure 3: Age Distribution for Top Genres



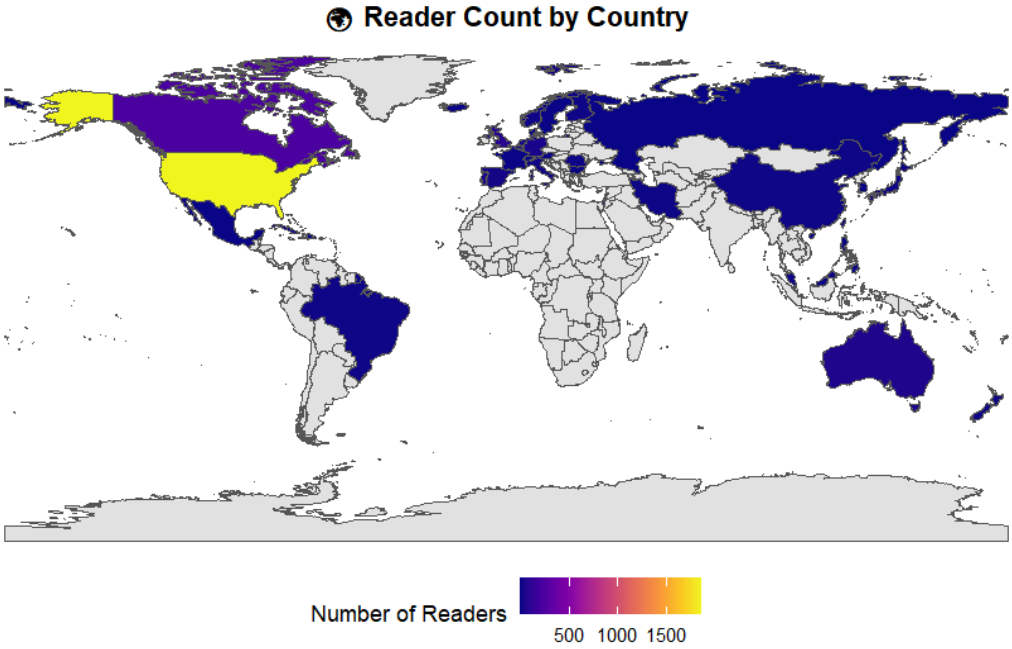
This set of histograms displays how age groups are distributed across six of the most popular genres: Adventure, Fiction, Historical Fiction, Mystery, Suspense, and Thriller. Each subplot reveals genre-specific age trends: Adventure and Fiction skew slightly younger, while genres like Suspense and Thriller show broader age appeal, especially among middle-aged readers.

Figure 4: Age Distribution by Top Genres



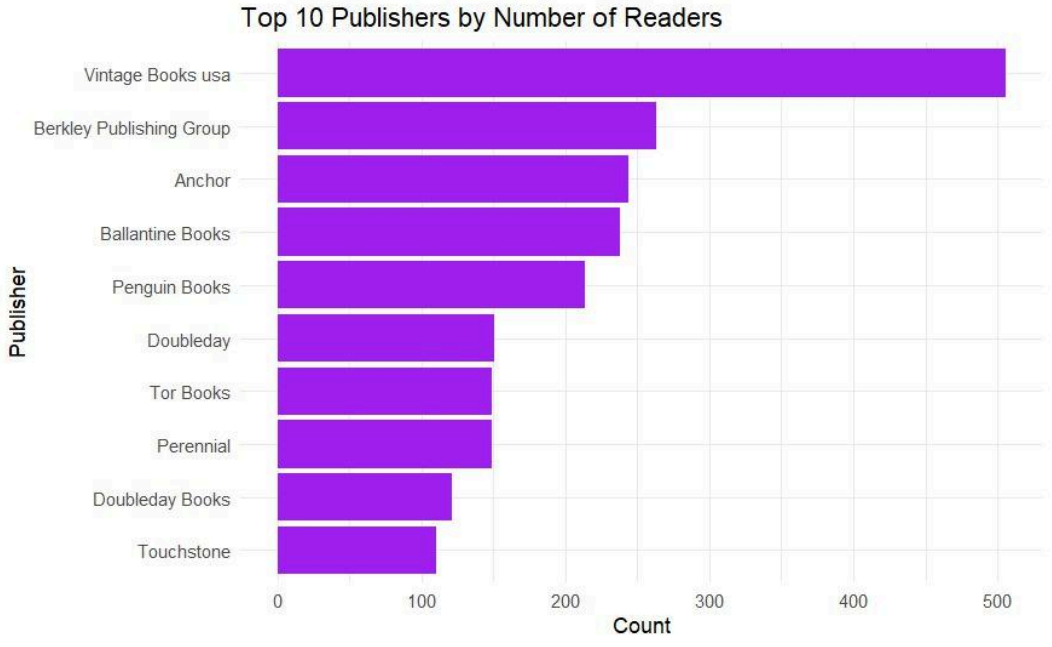
This boxplot illustrates the spread of reader ages for six leading genres. Fiction, Adventure, and Historical Fiction tend to attract slightly younger audiences, while genres like Mystery, Suspense, and Thriller show a wider age range, with median ages closer to the late 30s and greater variability in readership age.

Figure 5: Reader Count by Country



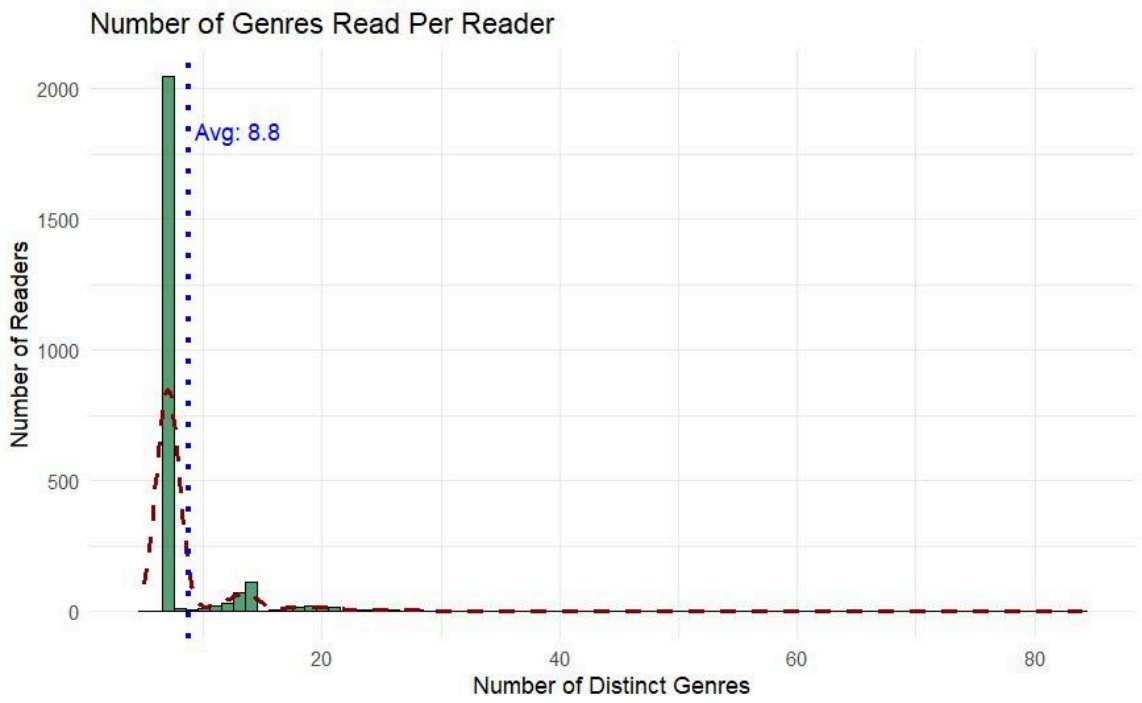
This choropleth map shows the geographic distribution of readers by country. The United States, highlighted in yellow, has the highest number of readers, followed by Canada and Australia. Most countries fall within the lower reader count range, as indicated by darker shades of blue.

Figure 6: Top 10 Publishers by Number of Readers



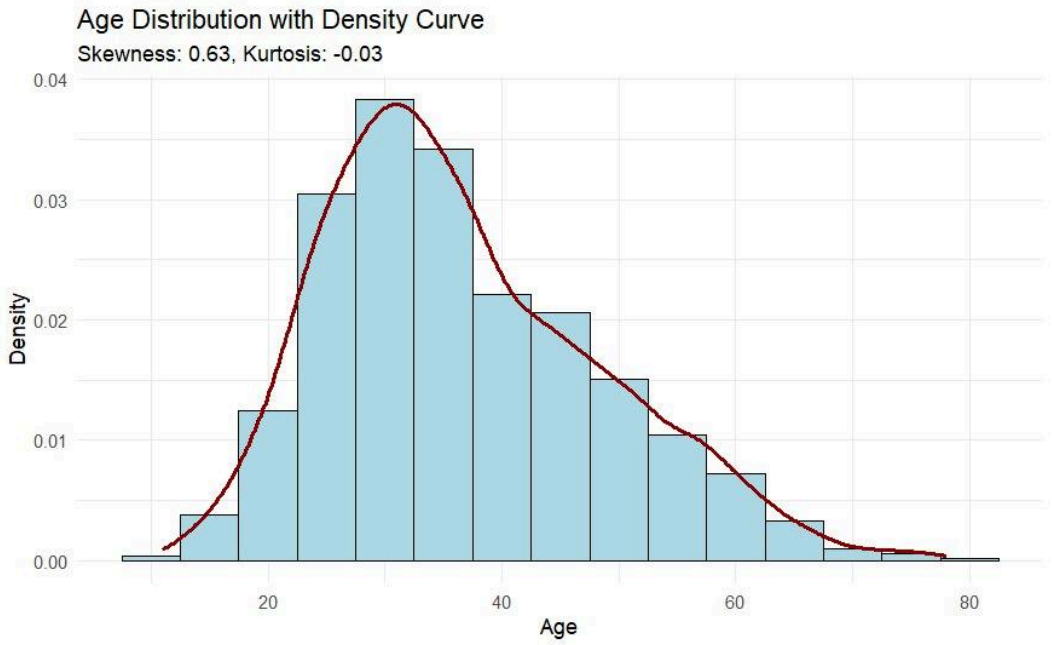
This horizontal bar chart displays the publishers with the highest number of readers in the dataset. Vintage Books USA leads by a significant margin, followed by Berkley Publishing Group and Anchor. This figure highlights the dominance of a few major publishers in attracting reader interest.

Figure 7: Number of Genres Read Per Reader



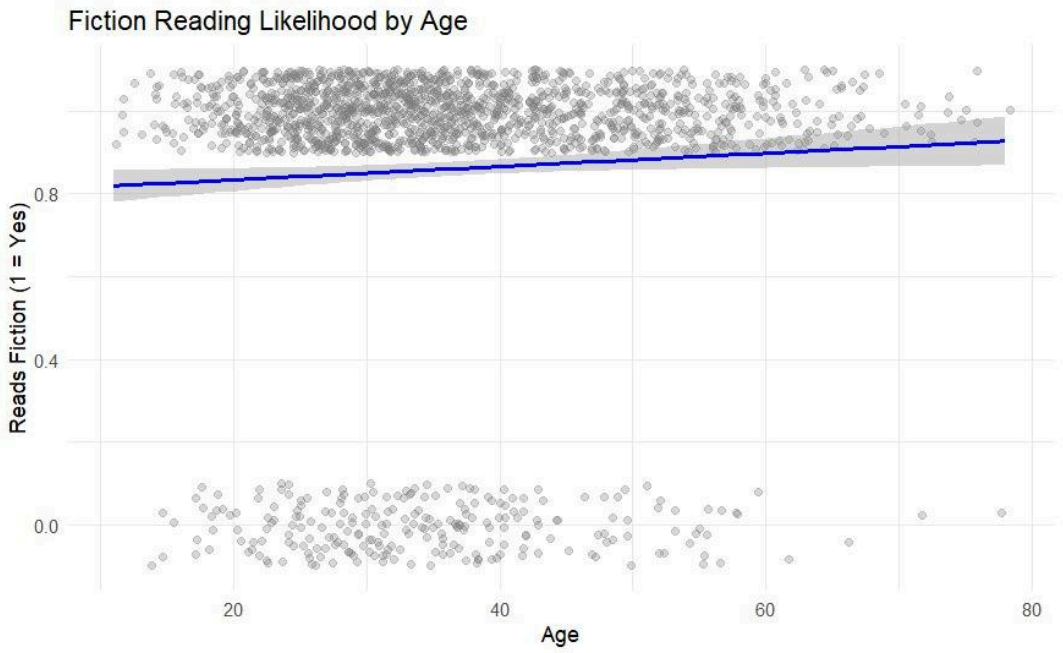
This histogram shows how many distinct genres each reader explores, with most users concentrated at the lower end of the scale. The vertical dotted line indicates the average number of genres read per user (8.8). The distribution is highly right-skewed, meaning most readers focus on a few genres, while a smaller number engage with a broad variety.

Figure 8: Age Distribution with Density Curve



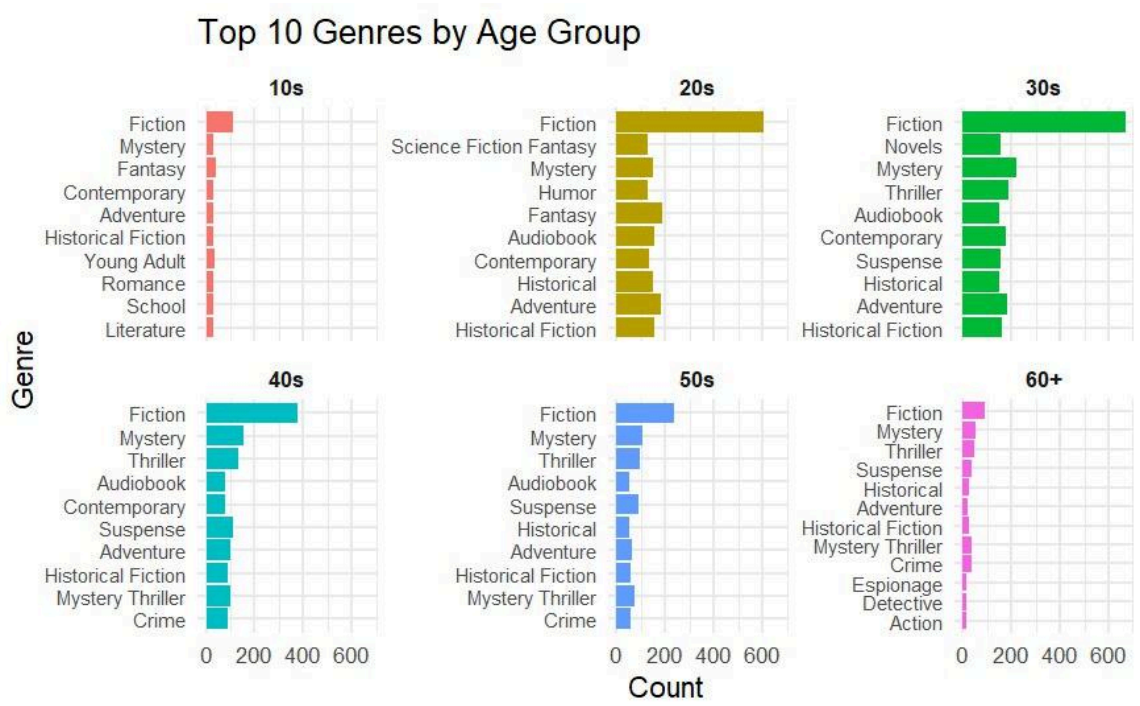
This histogram displays the distribution of reader ages with a smoothed density curve overlaid. The age distribution is moderately right-skewed (Skewness = 0.63), indicating that most readers are younger, with a longer tail toward older ages. The kurtosis value (−0.03) suggests a shape close to normal, though slightly flatter at the peak.

Figure 9: Fiction Reading Likelihood by Age



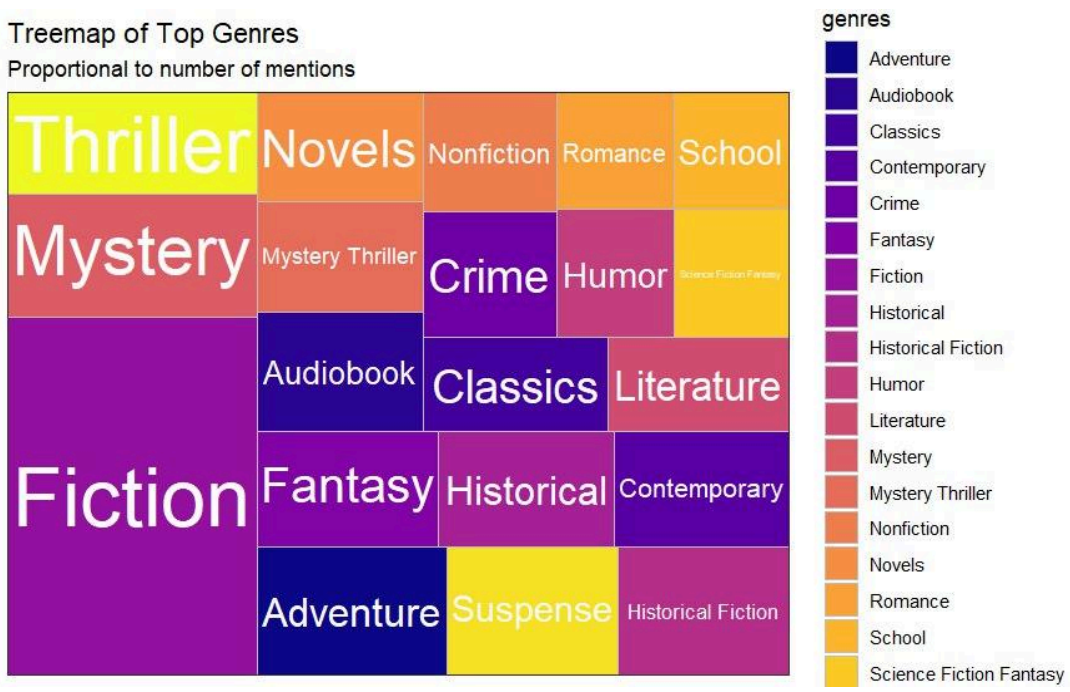
This scatter plot with a logistic regression line models the probability of a user reading fiction based on their age. The graph shows a slight upward trend, indicating that older readers are marginally more likely to read fiction than younger readers. The shaded area represents the confidence interval of the regression.

Figure 10: Top 10 Genres by Age Group



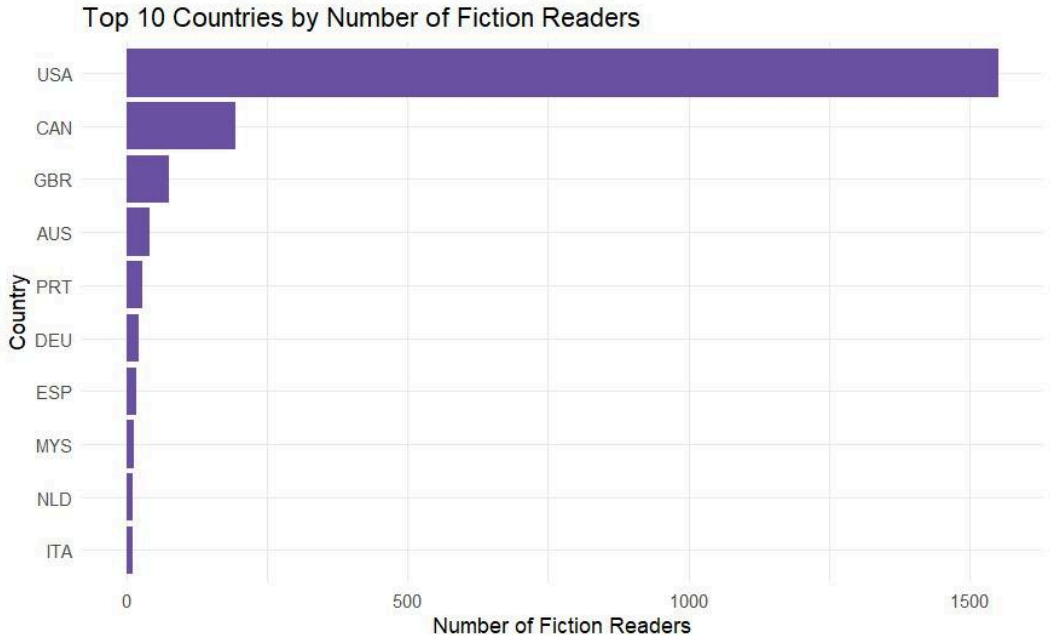
This panel chart displays the ten most read genres within each age group, from teenagers to those aged 60 and older. Each subplot reveals distinct preferences by age: younger groups lean toward genres like Fantasy, Young Adult, and Adventure, while older readers show stronger interest in Historical Fiction, Suspense, and Crime. The consistent popularity of Fiction across all age groups is notable.

Figure 11: Treemap of Top Genres



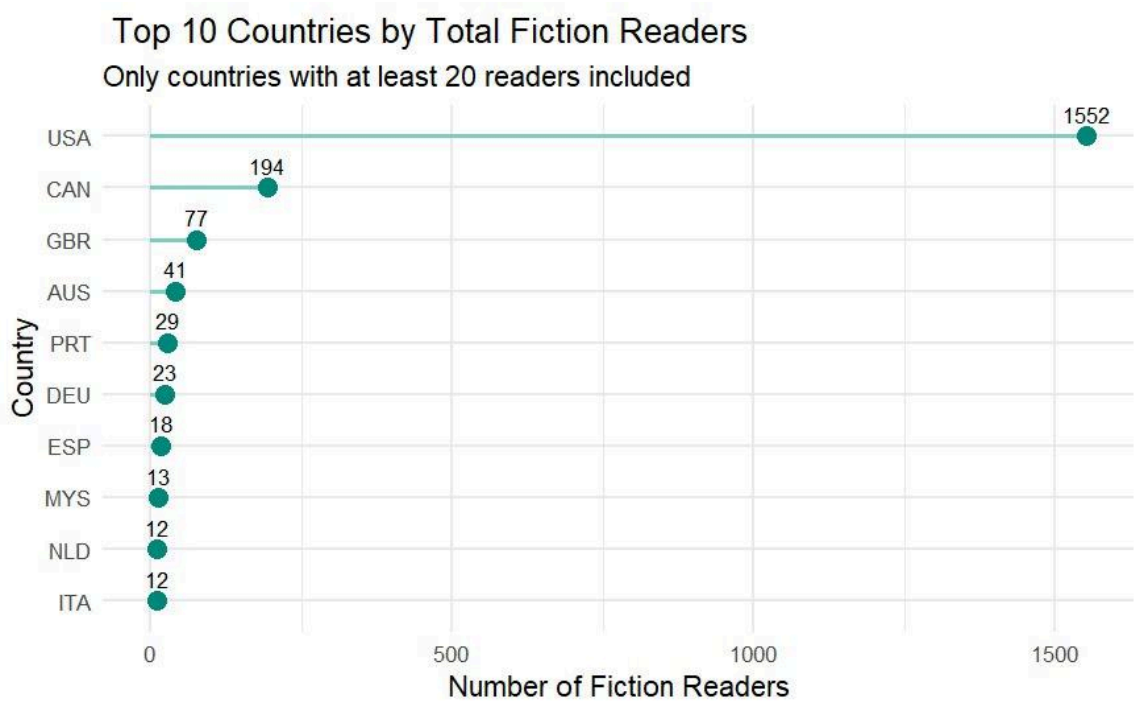
This treemap visualizes the popularity of book genres based on the number of mentions. Larger boxes represent genres with higher readership, highlighting Fiction, Mystery, and Thriller as the most common. The color-coded layout helps differentiate genres while emphasizing their relative frequency in the dataset.

Figure 12: Top 10 Countries by Number of Fiction Readers



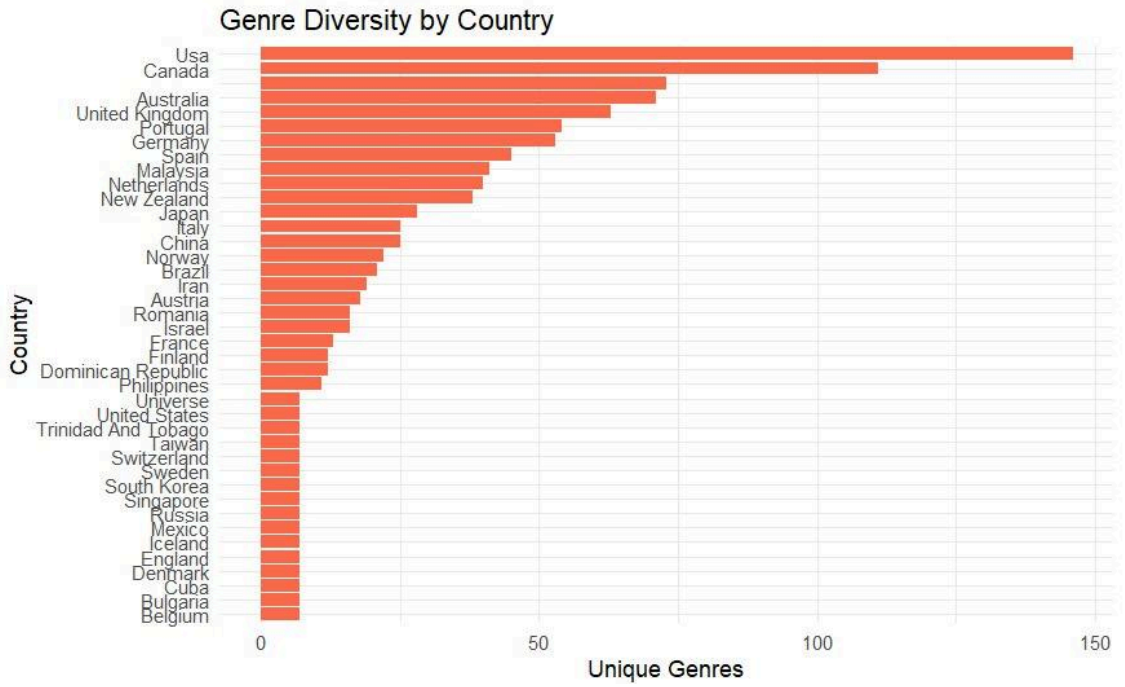
This horizontal bar chart shows the top 10 countries with the highest number of fiction readers. The United States leads by a significant margin, followed by Canada and the United Kingdom. The visualization highlights regional concentration and dominance in fiction readership, indicating cultural or market factors influencing genre preference.

Figure 13: Top 10 Countries by Total Fiction Readers



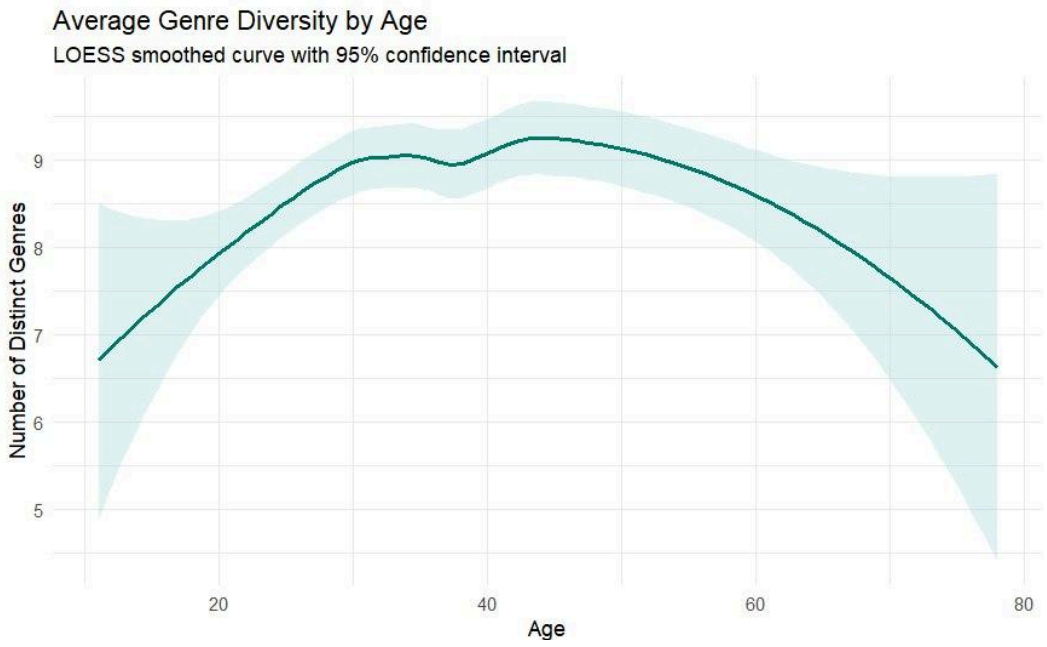
This dot plot highlights the top 10 countries with the most fiction readers, restricted to those with at least 20 readers. The United States far exceeds others with 1,552 readers, followed by Canada (194) and the UK (77). This detailed labeling emphasizes the scale difference among countries and the global spread of fiction readership.

Figure 14: Genre Diversity by Country



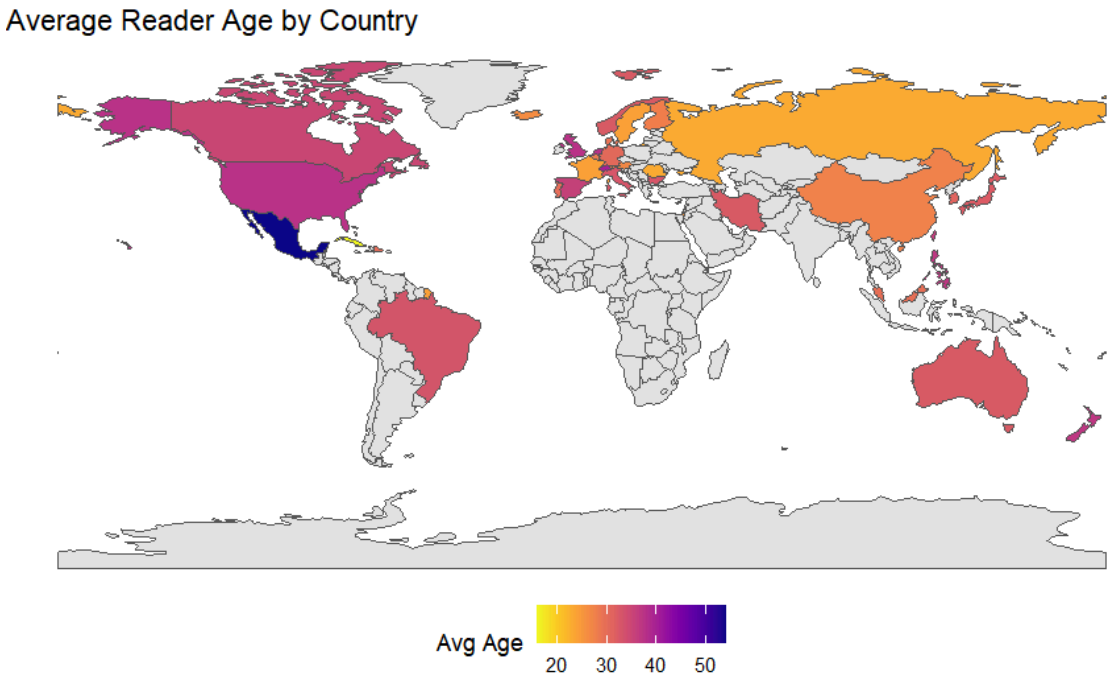
This horizontal bar chart displays the number of unique genres read in each country, indicating literary diversity. The United States leads with nearly 150 unique genres, followed by Canada, Australia, and the United Kingdom. This suggests broader reading preferences and possibly more access to varied content in these countries.

Figure 15: Average Genre Diversity by Age



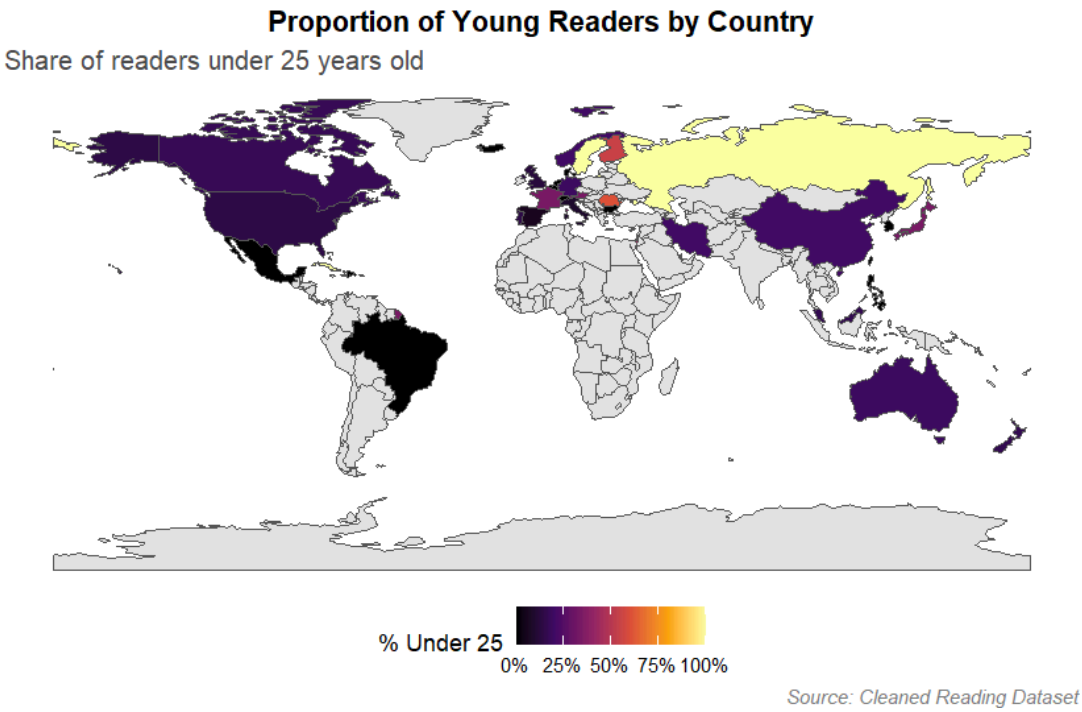
This line chart visualizes the average number of distinct genres read across different age groups using a LOESS smoothed curve. Genre diversity increases steadily from early teens through the 30s and 40s, peaking around age 40. It then gradually declines, suggesting younger and middle-aged adults tend to explore more varied literary genres compared to older readers. The shaded area represents the 95% confidence interval.

Figure 16: Average Reader Age by Country



This choropleth map displays the average self-reported age of readers by country. Lighter shades (yellow) represent younger average ages, while darker shades (purple) indicate older populations. For instance, Mexico and Southeast Asia have younger readers on average, while countries like Russia and some parts of Europe have older average readerships. This geographic trend helps identify regions where reading is more popular among specific age groups.

Figure 17: Proportion of Young Readers by Country



This choropleth map illustrates the share of readers under 25 years old in each country. Darker shades (black and purple) represent countries with a higher proportion of young readers, while lighter shades (yellow) show a lower percentage. Countries in Latin America and parts of Asia have a large youth readership, in contrast to regions like Russia and parts of Eastern Europe, where the proportion is relatively low. This figure helps highlight where younger populations are more engaged in reading.

Conclusion

This project allowed us to explore how age influences two very different types of behavior: reading and shopping. While Fiction reading clearly declines with age, spending behavior stays more active for longer, especially in the 25 to 45 age range. The dataset showed that age is a powerful variable in shaping interests and decisions.

On the technical side, we learned how to prepare real-world data for an interactive dashboard, how to apply both descriptive statistics and visualization techniques, and how to design layouts that communicate insights effectively. Working with an international data gave us perspective on behavioral trends that are shared across countries.

The dashboard we created are not only tools for exploration but also show how visual storytelling can make data easier to understand. This project helped us connect technical skills with real human behavior in a way that was both analytical and creative.

References

- **Kaggle Reading Dataset. (n.d.).** Retrieved from **https://www.kaggle.com/datasets/dk123891/books-dataset-goodreadsmay-2024**
- Chang, W. (2018). *R Graphics Cookbook: Practical Recipes for Visualizing Data*. O'Reilly Media.
- Wickham, H. (2021). *Mastering Shiny*. Retrieved from https://mastering-shiny.org