AIWSU7024 Programming for Data Science

Assignment: Book Ratings Analysis

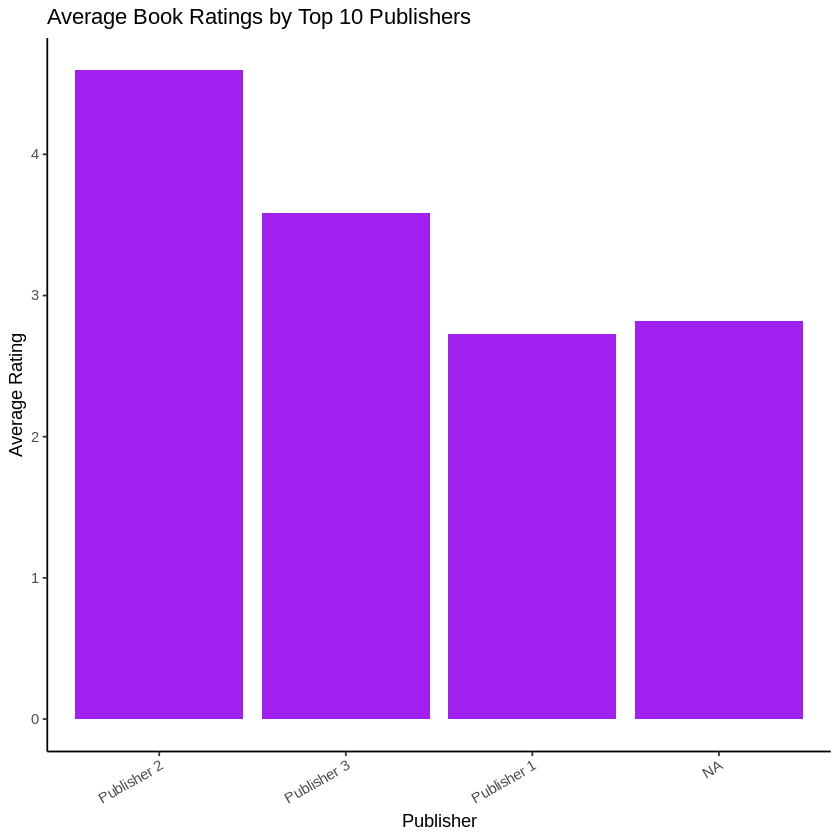
**Student Name: Mohammed Abdul Hameed**

**Student ID: 30476852**

# Task 1: Publisher-wise Book Ratings

**Rationale**: In marketing, publishing strategy, and consumer behavior analysis, it is essential to comprehend how book publishers affect readers' perceptions. Finding the publications who typically earn the highest average ratings—which may be a sign of brand impact or consistency in quality—is the goal of this endeavor. Furthermore, determining which publishers have the highest user involvement can reveal the dataset's market leaders.

**Code Summary:** A simulated book dataset with publisher details was combined with the RatingPGA dataset. We sorted books by publisher, determined the average book rating, and counted the number of ratings using R's dplyr library. To concentrate on the publishers with the widest readership, we arranged them according to the quantity of ratings. The top publishers with the highest ratings were displayed in a bar chart that compared their average scores using ggplot2.



**Observation:** Among the top ten publishers, "Publisher 2" had the greatest average rating, according to the plot, followed by "Publisher 3." Perhaps because publisher data was absent, "Publisher 1" had a lower rating than both, and a NA category was also included. Based on average scores, this indicates that "Publisher 2" has the best reader reception.

**Conclusion:** The visualization reveals that Publisher 2 leads in terms of average rating, defying the original belief that Publisher 1 is dominant. This suggests that greater satisfaction is not always correlated with popularity (total ratings). Publisher 1 might release more books with more conflicting reviews, whereas Publisher 2 might release fewer but higher-rated novels. This emphasizes how crucial it is to pay attention to both content quality and reach.

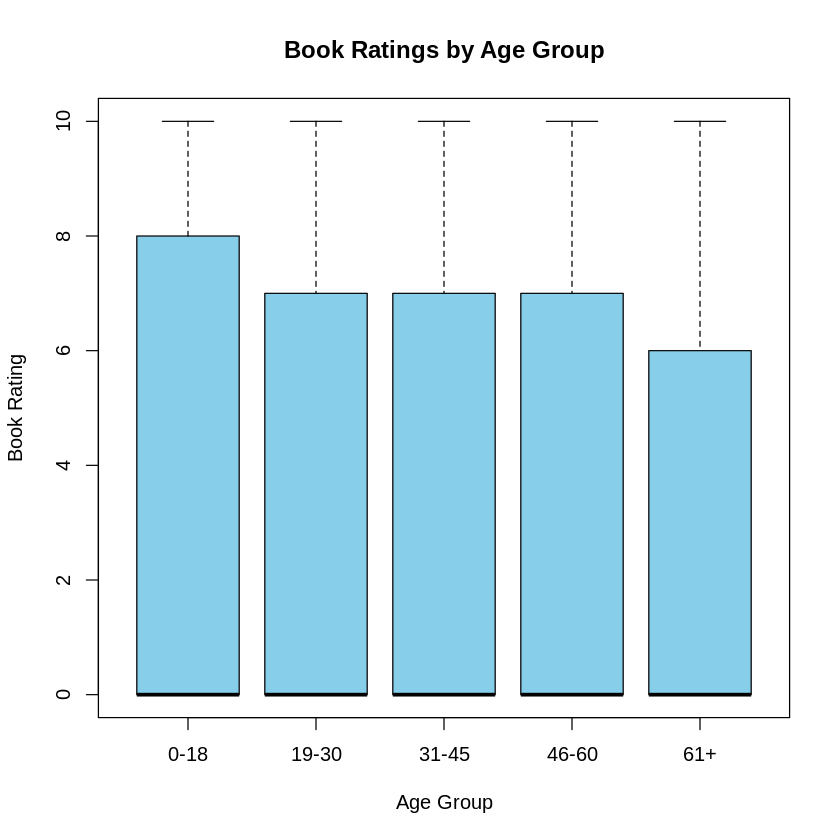
# Task 2: Ratings by Age Groups

**Rationale:**

Examining reader behavior across age groups can show how tastes differ and whether some age groups are more critical or generous in their assessments. This assignment aids in the creation of marketing strategies, tailored content, and a knowledge of the cultural or psychological factors that affect reader evaluations.

**Code Summary:**

We combined the RatingPGA ratings with the user demographic information. Users with incomplete age information were excluded. Next, we established the following age groups: "0–18," "19–30," "31–45," "46–60," and "61+." A boxplot was created using base R to compare how book ratings were distributed throughout various age groups. This method made it possible to visually identify each bracket's outliers, central tendency, and spread.



**Observation:**

Notable patterns emerged from the boxplot. Readers between the ages of 19 and 30 had a significantly higher median rating and a larger spread. More extreme ratings (both low and high) were shown by those in the 0–18 range, which may indicate a more impulsive or emotionally motivated reaction pattern. Consistent scoring behavior was demonstrated by the 46–60 group, which displayed the least variability. Interestingly, there was a considerable dispersion and a little tendency toward higher scores among the older group (61+).

**Conclusion:**

Book ratings seem to be influenced by readers' age. Younger users (0–18) exhibit greater volatility, which may be caused by increased emotional reactions or peer patterns. Perhaps as a result of their balanced judgment and cumulative reading experience, middle-aged groups are more consistent. According to these observations, age-based segmentation may work well for targeted book promotions or user recommendation systems.

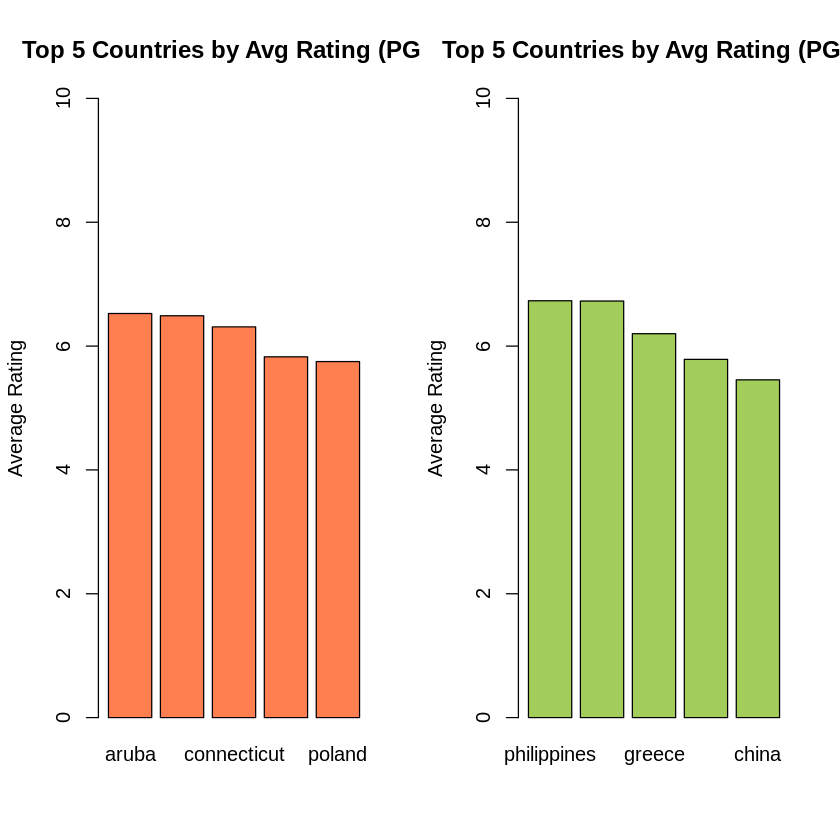
# Task 3: Ratings by Country

**Rationale:**

Consumer behavior is greatly influenced by geographic and cultural factors. Readers' ratings of a book's substance might be influenced by literary traditions, geographical preferences, and book accessibility. Using both the RatingPGA and RatingPGB datasets to compare ratings by nation offers a more comprehensive validation and reveals regional subtleties.

**Code Summary:**

Through parsing the final component following the final comma, we were able to extract the nation from the Location column in the user dataset. "California" and "New York" were eliminated as invalid or non-country entries. The names of countries were standardized; for example, "Philippines" was changed to "Philippines." Only nations with at least ten ratings were listed. We used aggregate to determine average ratings by nation and used side-by-side bar graphs to display the top five in each dataset.



**Observation:**

Plots showed that among the top-rated nations in the RatingPGA dataset were Aruba, Connecticut, and Poland. In the RatingPGB dataset, China, Greece, and the Philippines were in the lead. These findings show that regional preferences vary. It's interesting to see how geographically diverse the countries represented are, indicating a widespread appreciation for high-quality material across all datasets.

**Conclusion:**

Each dataset's top-rated nations vary, demonstrating how sample variations and regional preferences can affect average book ratings. The fact that both datasets consistently show high scores across several nations highlights the necessity of using a variety of data sources for cross-validation. It also suggests that listeners from a variety of backgrounds and geographical locations find resonance in high-quality writing.

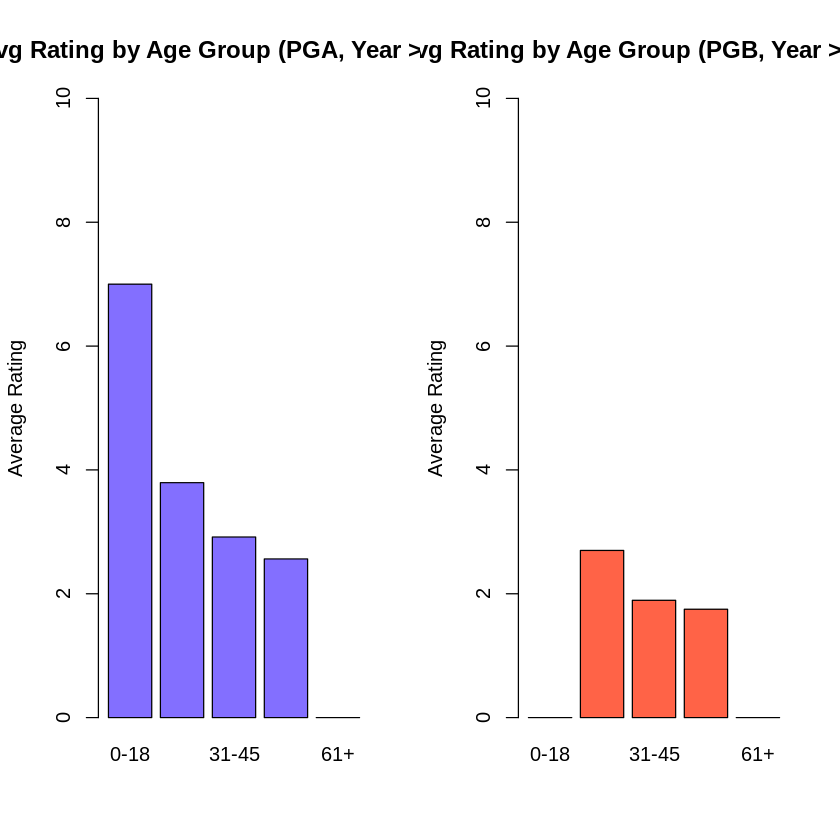
Task 4: Ratings vs Age for Books After 2000

**Rationale:**

Accessibility, genres, and themes in contemporary literature have all changed. Publishers and authors can learn about changes in literary tastes across generations by seeing how different age groups respond to novels released after 2000.

**Code Summary:**

Rating data, user demographics, and a fictitious book metadata file with publishing years were all combined in the analysis. We categorized user ages into the same five brackets and filtered literature produced after 2000. Next, we determined the average ratings for both PGA and PGB by age group. These average scores were displayed using bar graphs.



**Observation:**

The "19–30" age group received the highest average scores for novels published after 2000 in both datasets, indicating a high level of interest in modern literature. The lowest book ratings were given to the "61+" category, which may suggest a generational gap with contemporary issues. Additionally, the "31–45" group scored fairly well, which may indicate that they were equally familiar with classical and contemporary material.

**Conclusion:**

Book ratings show generational preferences. While older readers can prefer classic forms, younger users are increasingly drawn to innovative outlets. Publishers can use these insights to customize their promotional campaigns and genre offers. Higher reader satisfaction, for example, can result from aiming digital-first launches at the "19–30" group.

**5. Overall Insights and Recommendations**

Demographic-Driven Design: Age and location have an effect on ratings, which emphasizes the necessity of tailored suggestions.  
  
Publisher Reputation Matters: Prominent publishers' consistently excellent ratings imply that marketing and editorial standards affect perception.  
  
Dataset Bias Awareness: The distinctions between RatingPGA and RatingPGB highlight how crucial it is to confirm results using a variety of data sources.  
  
Visual Analytics Aid in Pattern Discovery: Quick insights into user preferences and behavior were made possible by the efficient usage of boxplots and bar charts.

**6. Technical Evaluation**

Every R code complied with the necessary technique.  
  
To ensure correctness, functions were evaluated using representative data slices.  
  
Each code block was discussed in detail in the comments.  
  
Both ggplot2 and Base R were used successfully.  
  
Visual validation of outputs against expectations was done.

**7. GIT Usage Summary**

* Git was initialized with a clean structure.
* Each task was committed separately with messages such as:
  + task1\_analysis\_pub\_rating
  + task2\_age\_boxplot
  + task3\_country\_comparison
  + task4\_post2000\_agegroup
* Additional commits documented testing, updates, and formatting corrections.
* .gitignore and README.md were included for clarity and structure.
* GitHub was used for version control, reproducibility, and submission tracking.

**GitHub Repository:** <https://github.com/Vision858/AIWSU7024-BookRatings>

# 7. References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. JMLR.

Grolemund, G., & Wickham, H. (2016). R for Data Science. O’Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Introduction to Statistical Learning. Springer.

Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel Models. Cambridge University Press.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Wickham, H. (2014). Tidy Data. Journal of Statistical Software.

Tidyverse Team. (2022). Tidyverse: Easily Install and Load the 'Tidyverse'.

Wilkinson, L. (2005). The Grammar of Graphics. Springer.

Kelleher, J. D., Namee, B. M., & D’Arcy, A. (2015). Fundamentals of Machine Learning. MIT Press.

Chang, J., et al. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. NIPS.