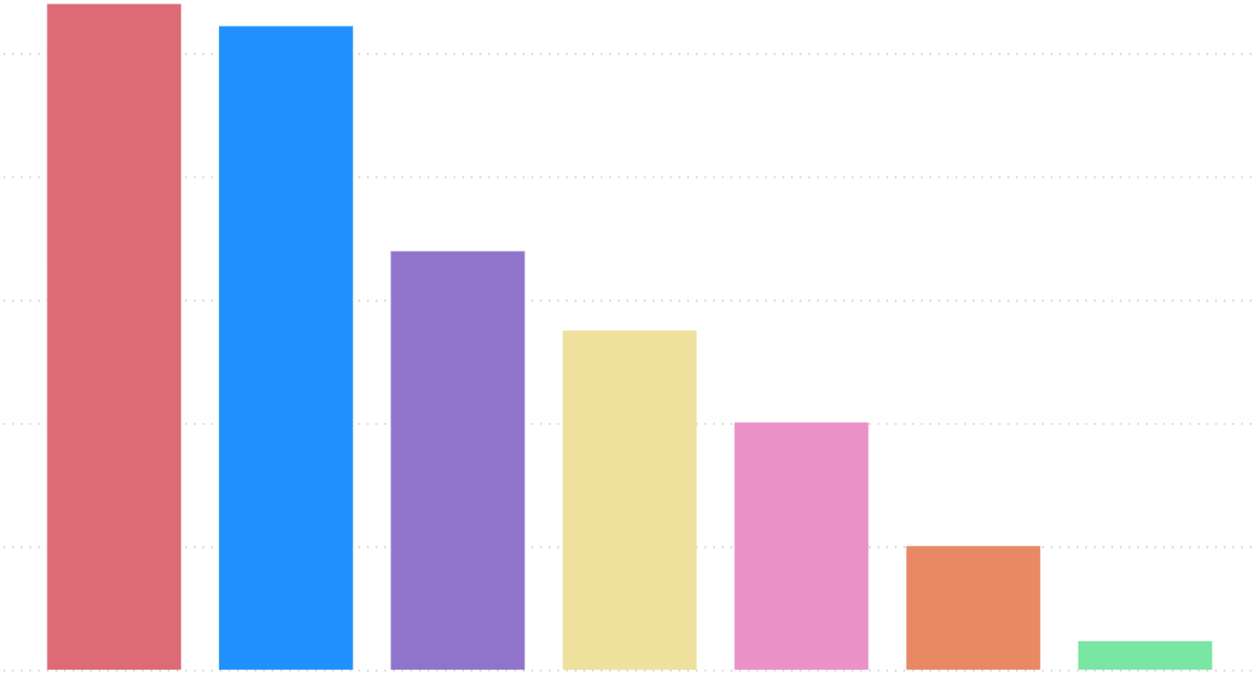
Rachel UWIMANA

23684

Book Reading Analysis

Big Data Final Exam Report



3rd JULY 2025

Abstract

The average respondent (user in my sample data) is around 49 years old, and a typical respondent reads around 4 books a year.

Female respondents read significantly more books on average than male respondents.

The relationship between age and books read is not strong.

Print books are the most popular reading format, followed by ebooks and then audiobooks.

Laptops and tablets are the most commonly owned devices in our sample.

PART 1: PROBLEM DEFINITION & PLANNING

Problem Statement

The central problem this project addresses is to quantify and analyze generational shifts in reading frequency and their implications for public engagement with literature. The analysis aims to determine if there has been a decline in the average number of books read per year, with a specific focus on how younger generations, particularly those who were Gen Z in 2016, compare to older generations.

The main objectives of this analysis are:

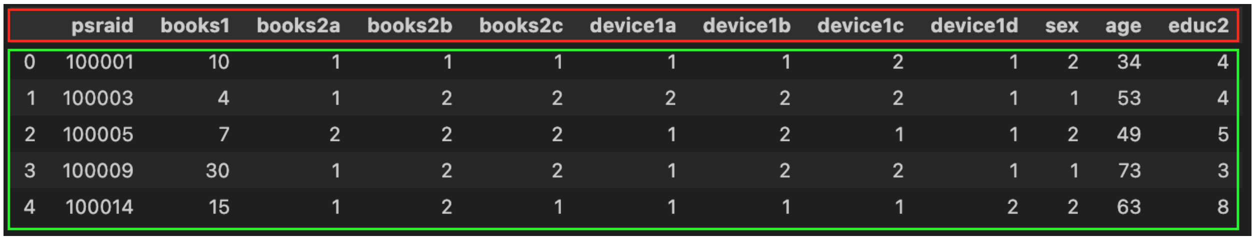
1. To determine and compare the average number of books read annually across different age groups and analyze this data to identify any a decline in reading frequency over time.
2. To identify how generational preferences for different book formats, such as print, e-books, and audiobooks, may correlate with changes in reading frequency.
3. To present these findings through clear and compelling visualizations to provide actionable insights into generational reading trends.

PART 2: PYTHON ANALYTICS TASKS

How do demographic factors (such as age, sex, and education) and technology usage (including the ownership of devices like e-readers) influence the number of books read and the format in which they are read (print, audio, or e-book) among the survey respondents?

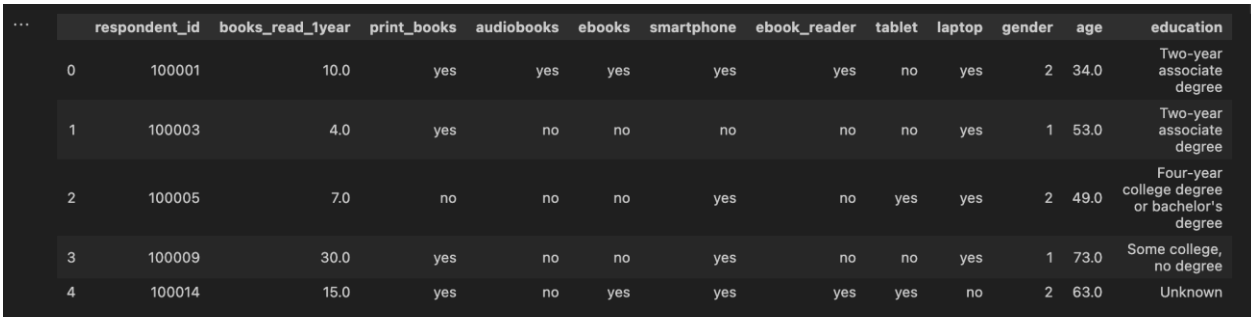
1. **Feature Selection:** We chose the columns relevant to our problem.
2. **Column Renaming:** We gave those columns meaningful names.
3. **Handling Coded Values:** I cleaned all the columns with coded values, including:
   - The binary 'Yes/No' columns, handling tricky data type issue.
   - The books_read_1year and age columns, correctly converting 98 and 99 to NaN.
   - The education column, I handled the unknown codes (7 and 8) by creating a new 'Unknown' category, a responsible and ethical choice to minimize data loss as it encompassed 284 values.

From this: coded columns and values with hidden meaning:



	psraid	books1	books2a	books2b	books2c	device1a	device1b	device1c	device1d	sex	age	educ2
0	100001	10	1	1	1	1	1	2	1	2	34	4
1	100003	4	1	2	2	2	2	2	1	1	53	4
2	100005	7	2	2	2	1	2	1	1	2	49	5
3	100009	30	1	2	2	1	2	2	1	1	73	3
4	100014	15	1	2	1	1	1	1	2	2	63	8

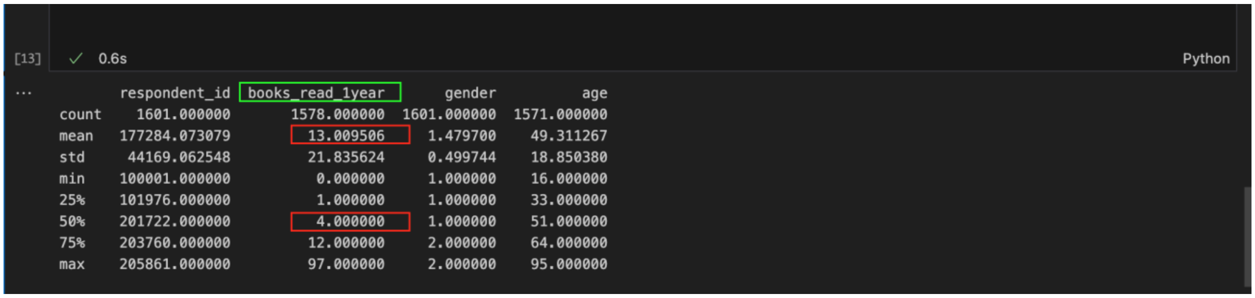
To This: Meaningful values and columns with reference from the dataset questionnaire.



	respondent_id	books_read_1year	print_books	audiobooks	ebooks	smartphone	ebook_reader	tablet	laptop	gender	age	education
0	100001	10.0	yes	yes	yes	yes	yes	no	yes	2	34.0	Two-year associate degree
1	100003	4.0	yes	no	no	no	no	no	yes	1	53.0	Two-year associate degree
2	100005	7.0	no	no	no	yes	no	yes	yes	2	49.0	Four-year college degree or bachelor's degree
3	100009	30.0	yes	no	no	yes	no	no	yes	1	73.0	Some college, no degree
4	100014	15.0	yes	no	yes	yes	yes	yes	no	2	63.0	Unknown

Exploratory Data Analysis

Individual key columns analysis: age, books read in 1 year and education



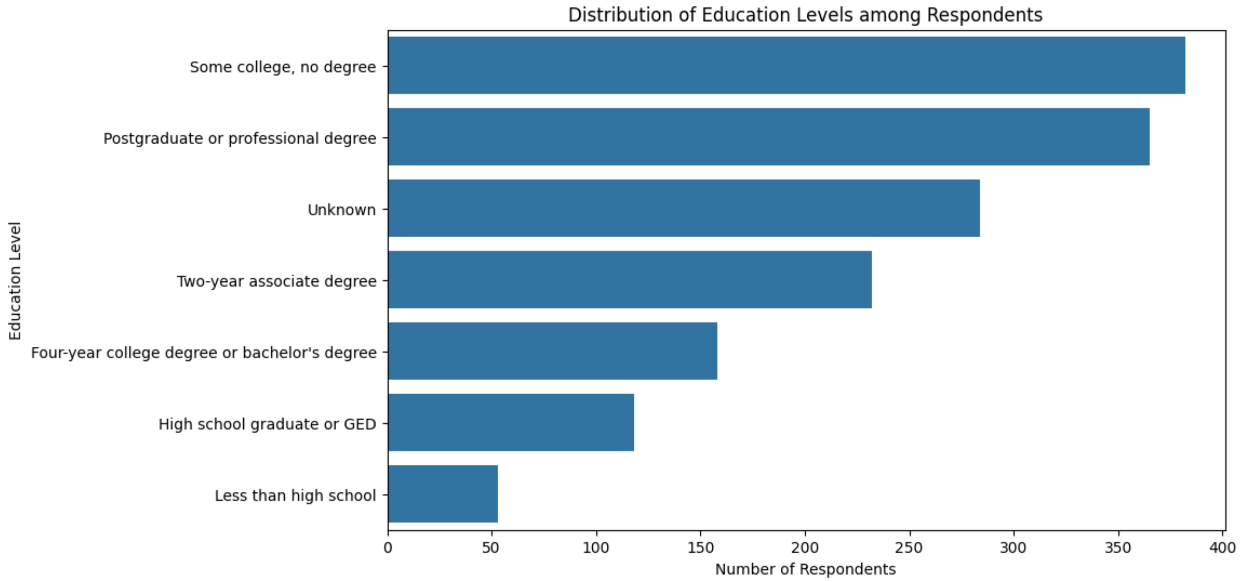
```
...        respondent_id  books_read_1year       gender          age
    count    1601.000000       1578.000000  1601.000000  1571.000000
    mean   177284.073079         13.009506     1.479700    49.311267
    std     44169.062548         21.835624     0.499744    18.850380
    min    100001.000000          0.000000     1.000000    16.000000
    25%    101976.000000          1.000000     1.000000    33.000000
    50%    201722.000000          4.000000     1.000000    51.000000
    75%    203760.000000         12.000000     2.000000    64.000000
    max    205861.000000         97.000000     2.000000    95.000000
```

Analysis of books_read_1year column

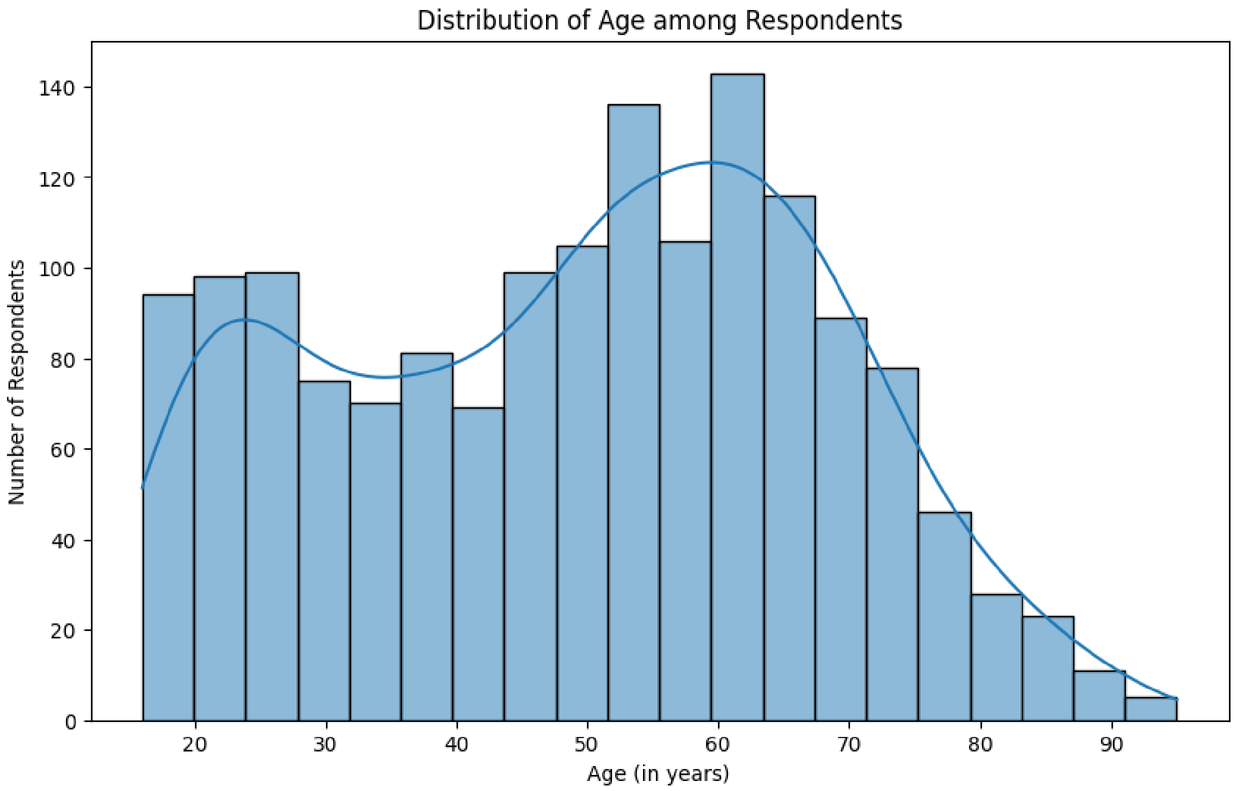
- The mean is about **13 books**, and the median is **4**. This difference suggests that the data might be skewed, with a few respondents reading a very large number of books, pulling the average up.
- The min is 0, and the max is 97, which is the questionnaire’s code for "97 or more."

Analysis of age column

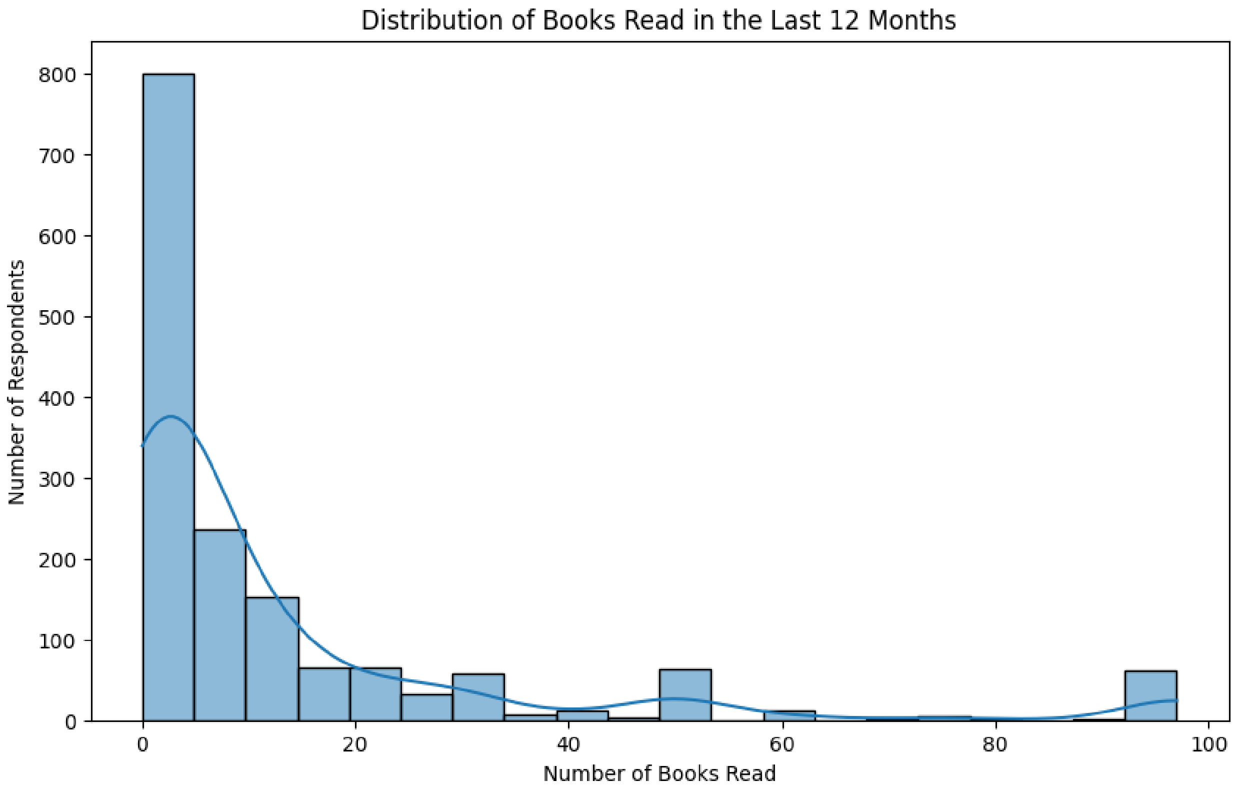
- The count of 1571 is very close to the total of 1601, which confirms that thecleaning step successfully converted the 98 and 99 codes for refused and no value into NaN values, as describe() excludes null values.
- The mean age is around **49 years old**, with a standard deviation of about 18.8 years.
- The age range is from 16 to 95.
- The median (50%) age is **51**, meaning half of the respondents are 51 or younger.



- **Most Common Education Levels:** The most common education levels among the respondents are "Some college, no degree" and "Two-year associate degree." This gives us a clear picture of the educational background of our sample.
- **Least Common Education Levels:** The "Less than high school" category has the fewest respondents, followed by "High school graduate or GED."
- **The "Unknown" Category:** From here I can tell my decision to create an "Unknown" category was a good one. It's a significant portion of the data, and by keeping it, I have an understanding that some people not in these categories were respondents and can contribute in other analyses.

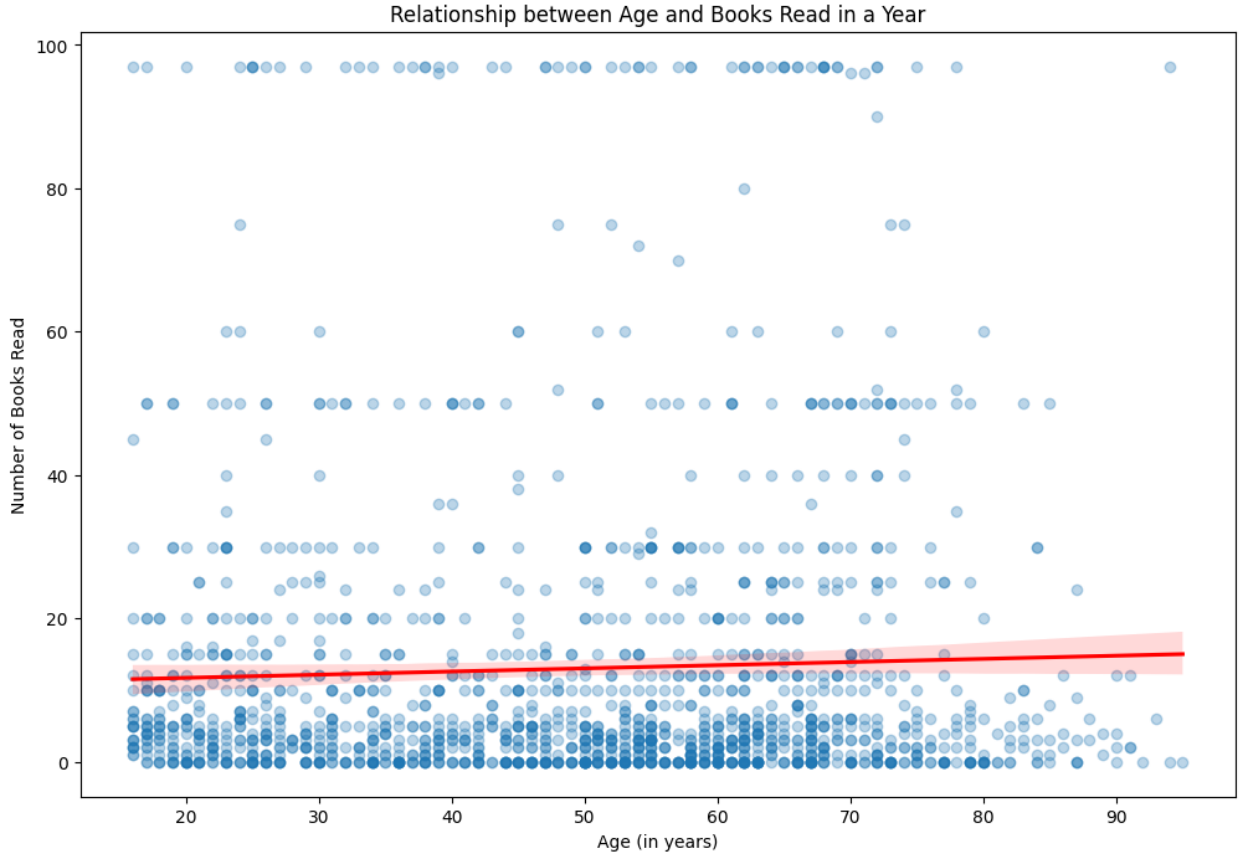


- **Bimodal Distribution:** The distribution is not a smooth bell curve. It appears to have two main peaks, or modes, one for respondents in their 20s and another for those in their 40s to mid-50s. This suggests our sample may be composed of two distinct age groups.
- **Skewness:** The distribution has a long tail to the right, indicating that there are a number of older respondents, which is consistent with the describe() output where the maximum age was 95.
- **Overall Spread:** The majority of respondents seem to be concentrated between the ages of 20 and 60.



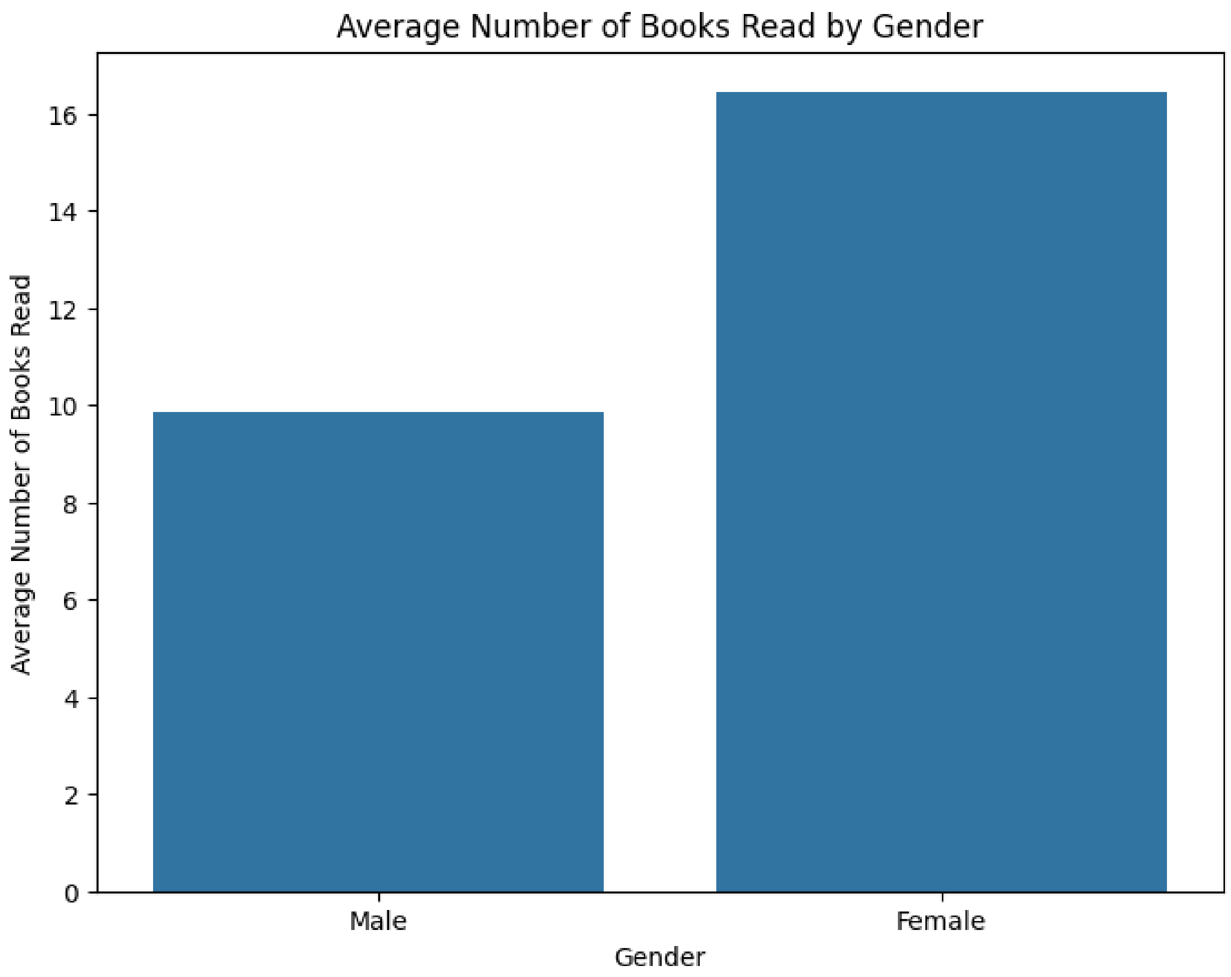
- **Extreme Skewness:** The distribution is heavily skewed to the right. The vast majority of respondents read a small number of books, which is why the histogram is so tall on the left side.
- **Outliers:** The long tail of the distribution, which extends all the way to 97 books, confirms our earlier observation that a small number of people read a very large number of books. This is what pulled the mean (13) away from the median (4).
- **Clear Picture of Reading Habits:** This plot gives us a much more accurate picture of the reading habits of the average person in our sample than the mean alone.

Relationship analysis between all columns.

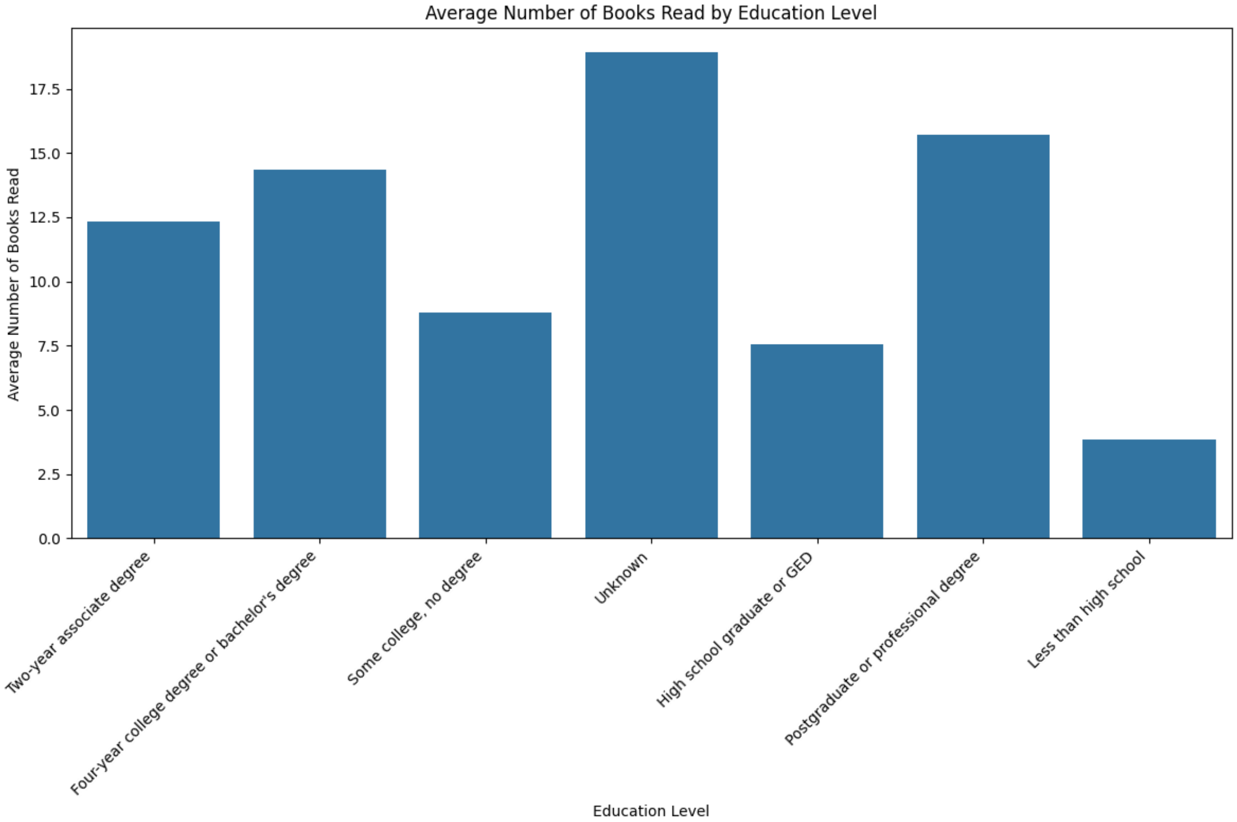


- **No Strong Correlation:** The red line of best fit is almost perfectly flat, with a slight upward trend. This indicates that there is **no strong linear relationship** between a person's age and the number of books they read in a year. The number of books read doesn't seem to increase or decrease significantly with age.
- **High Variability:** The data points are widely scattered all over the graph, especially in the lower half. This confirms that there's a huge amount of variability in reading habits regardless of age.
- **Outliers:** You can clearly see the outliers in the data, with a few respondents reading a very high number of books across all age groups.

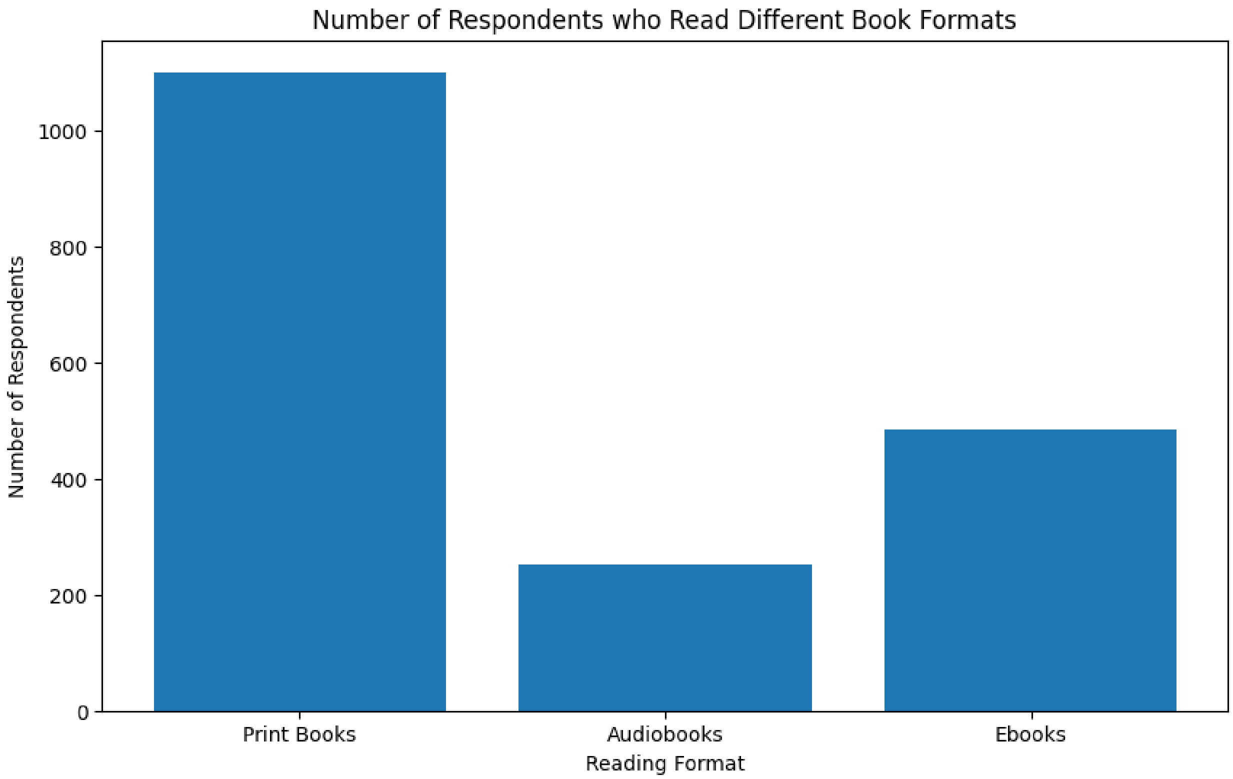
This means I might need to look at other variables to explain differences in reading habits.



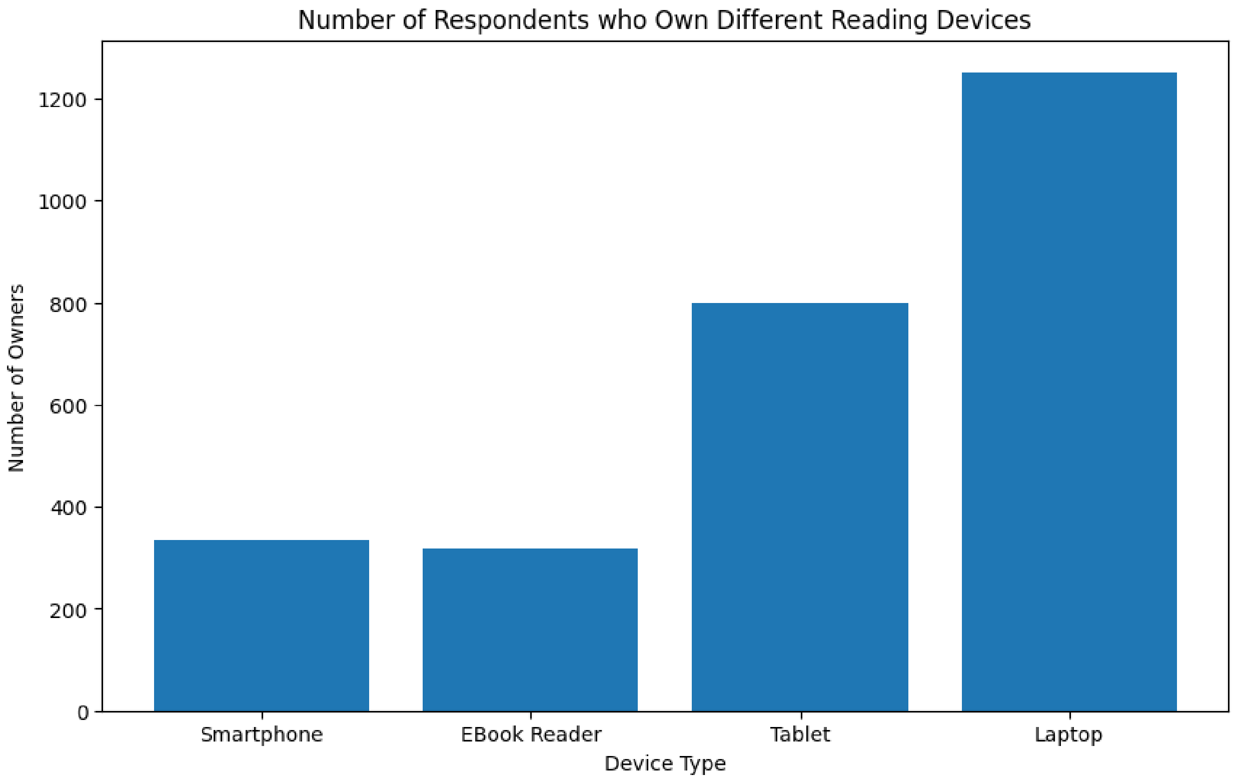
- **Significant Difference:** There is a clear difference in the average number of books read between male and female respondents.
- **Female Readers:** On average, female respondents read a significantly higher number of books per year (approximately 16 books).
- **Male Readers:** Male respondents read an average of approximately 10 books per year.



- **The "Unknown" Category:** The highest average number of books read is in the **"Unknown"** category. This is a surprising finding which validates my earlier concern about dropping this data. The respondents whose education levels were not specified are, on average, the most avid readers in this sample.
- **Postgraduate Readers:** Respondents with a "Postgraduate or professional degree" also read a high number of books on average.
- **Lowest Readers:** The lowest average number of books read is in the "Less than high school" category.
- **Complex Relationship:** The plot shows that the relationship between education level and books read is not a simple linear trend.

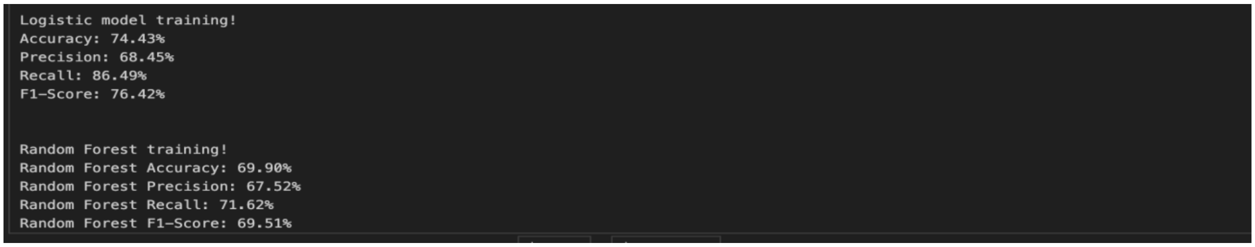


- **Print Books are the highest:** A very large majority of the respondents read hard copy books, with over 1000 people saying "Yes."
- **Ebooks are the Second Choice:** Ebooks are the next most popular format, with about 500 people reading them.
- **Audiobooks have the Lowest Readership:** Audiobooks are the least common format among the three.



- **Laptops are Most Common:** Laptops are by far the most commonly owned device among the respondents, with well over 1200 owners.
- **Tablets are Second:** Tablets are the second most owned device, with around 800 owners.
- **Ebook Readers and Smartphones:** Ebook readers and smartphones are the least owned devices. It's interesting to note that the number of respondents who own a smartphone is quite low in our sample, which is an unexpected but important finding.

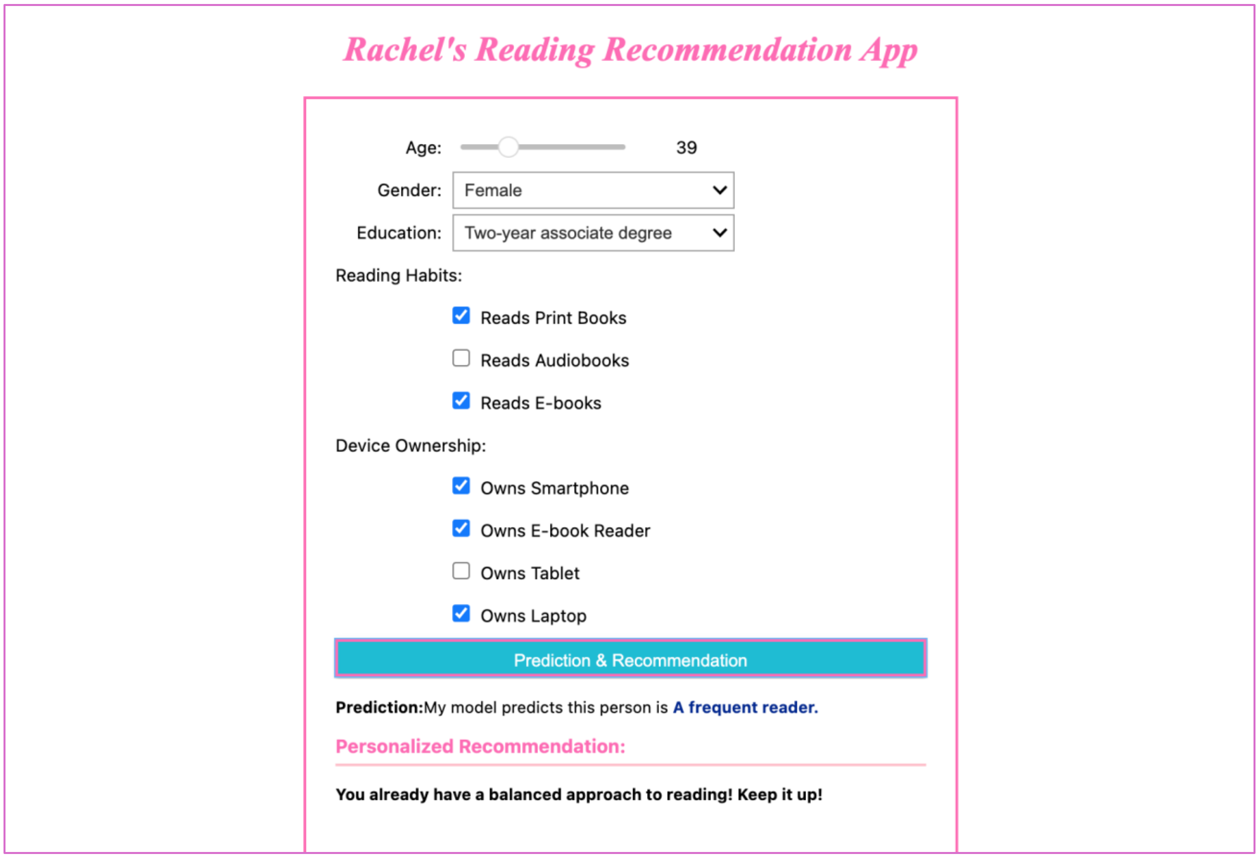
Logistic Regression/Random Forest Classification Model



```
Logistic model training!
Accuracy: 74.43%
Precision: 68.45%
Recall: 86.49%
F1-Score: 76.42%


Random Forest training!
Random Forest Accuracy: 69.90%
Random Forest Precision: 67.52%
Random Forest Recall: 71.62%
Random Forest F1-Score: 69.51%
```

I chose the Logistic Regression model because it had higher accuracy than the Random Forest one, to build my small prediction and recommendation app as an added innovation to the project.



Rachel's Reading Recommendation App

Age: 39

Gender: Female

Education: Two-year associate degree

Reading Habits:

☑ Reads Print Books

□ Reads Audiobooks

☑ Reads E-books

Device Ownership:

☑ Owns Smartphone

☑ Owns E-book Reader

□ Owns Tablet

☑ Owns Laptop

Prediction: My model predicts this person is **A frequent reader.**

Personalized Recommendation:

You already have a balanced approach to reading! Keep it up!

Part 3: Power BI Analysis

This can be found documented in the read me and the .PBIX files can be found in the same reports folder as you found this document.