Data Source

Summary

How much do Americans read? This dataset is the result of a survey conducted by Pew Research Center in 2021. According to their website, "Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world. We conduct public opinion polling, demographic research, content analysis and other data-driven social science research." This survey was conducted over the phone from random calls and includes 1502 people across the United States. The survey questions also focused on how much people use the internet and related technology questions, but for this project I'm only interested in the questions related to books and reading.

Relevance

I chose to do this project because I have always loved reading books, and I'm curious to know how much other people read and whether there are characteristics that correlate with how much a person reads.

Data Profile

Data Cleaning & Consistency Checks

I started by removing some of the columns about other of the survey questions that don't pertain to my project. Then I renamed the remaining columns so that they are more clear. Next, I changed the data types of some of the columns.

One column, "level_of_education" had numerical codes representing their answers, so I created a new column "education" that replaced the numerical codes with text descriptions according to the information in the survey questionnaire.

Next, I wanted "income" to be an integer so that I can use it in calculations and charts, but the survey only gathered ranges rather than exact numbers. I decided to impute random values in each range. For "$150,000 or more" I decided to guess 150,000-300,000, and for "$10,000 or less" I used the range 1000-9,999. I then found the descriptive statistics for all the answers with numbers and used that to find the interquartile range. I used that range to impute values for those who answered "I don't know" or refused to answer.

Next I wanted to resolve any mixed type columns. For the three columns "read_printed_books", "read_audiobooks", and "read_e-books", there were a few "don't know" or refused answers so I

changed those to "No" since that is most likely. There were 301 nulls left in those columns, but they were nearly equal to the number of those who didn't read any books so those questions were not applicable. I decided to also change them to "No" since if they didn't read any books, it follows that they didn't read any of the forms of books. The other column "dem_or_rep_leaning" I found that it was missing a lot of values and the "party" column gave more information, so I decided to remove that column.

In "number_of_books_read" and "age" the interviewers entered exact numbers except for three cases: "97" was entered if the number was 97 or higher, "98" was entered if the person said they didn't know, and "99" was entered if they refused to answer the question. In both columns, I decided to leave the 97s alone because it's unlikely the real number would be much higher than that, and I imputed values for the 98 and 99 with the interquartile ranges.

Finally, I reordered the columns and did some final checks and found the summary statistics shown below.

Summary Statistics

Column Name	Mean	Min	Max
age	52	18	97
income_estimate	$87,426	$1,261	$299,870
number_of_books_read	15	0	97

Limitations & Ethics

There are several limitations to this dataset. One is that it is survey data -- we only know what people chose to tell the interviewers, so they could have lied or given an inaccurate number. People aren't necessarily good at estimating things, so unless they kept track of how many books they read somewhere and looked it up, it may be inaccurate. There's also no way of knowing if someone was lying about any of their answers. Some variables had a number of missing values where the person said they didn't know or refused to answer the question. Also since the salaries were only ranges in the survey data, they had to be imputed so they aren't the real salaries. Another limitation is that it's a small sample size. Fifteen hundred is not enough to draw reliable conclusions about the population of the U.S. Additionally, the interview was conducted over the phone, so people who don't have a phone or don't answer unknown calls will not be represented. However, it should be enough to make some interesting observations if there are any strong correlations, while keeping in mind these limitations.

As for ethical considerations, there isn't any PII in the data. There is demographic data such as age, race, sex and gender, so analysts should be aware of any possible descrimination when interpreting the data.

Key Questions

- Are there any demographic variables that correlate with how much a person reads?
    - How much do different age groups read?
    - Does income affect how much a person reads?
    - How does reading vary by state?
- Does the way a person reads (by printed book, audiobook, or e-book) correlate with any demographic variables?
    - Do younger people read e-books more?
- How many people read printed books v. audiobooks v. e-books?
- Do frequent readers (those who read 20 or more books) have a preference for how they read? Do they have any demographic variables in common?
- How has reading changed over time (will need to import more data)?
    - Are people reading more or less than previous years?
    - Are e-books and audiobooks becoming more popular?
    - Are printed books becoming more or less popular?