Data:

In the analysis, data from the General Social Survey (GSS) was utilized to focus on public sentiment surrounding the removal of certain types of books from public libraries and respondents' political affiliations. The dataset was sourced from 37 individual CSV files on GitHub, which were later consolidated into a single data frame.

The variables of interest were:

- "Librac" and "libracy": These questions assessed the respondent's stance on whether a book written by someone suggesting Black people are inferior should be removed from a public library. While "librac" was the original wording, "libracy" provided a gender-neutral version.
- "Libmslm" and "libmsly": These probed into whether a book written by an Islamic religious leader preaching hatred of the U.S. should be removed from a public library.
- "Libath" and "libathy": These variables explored if respondents would support the removal of a book against churches and religion from a public library.
- "Partyid": This variable captured respondents' political affiliations.
- "Year": This variable recorded the year that the information was surveyed

The primary challenge faced during this cleaning process was the management of missing data. The dataset exhibited a significant number of missing values, particularly in the "librac," "libmslm," and "libath" variables. To combat this issue, a useful strategy involved merging the original variables with their gender-neutral counterparts. For instance, if a response was missing in "librac," the corresponding entry from "libracy" was used to fill the gap. This approach allowed for the retention of as much data as possible.

Before this was conducted to determine if the change in gender had a significant impact on individuals' responses, a cross-tabulation analysis between the original and gender-neutral versions of the questions was conducted. The analysis determined that respondents displayed consistency in their views, regardless of the phrasing of the question. This could indicate that gender-neutral phrasing didn't significantly impact the responses, at least for these particular questions; therefore, merging the variables would likely not impact the analysis.

However, there were still large portions of missing data even after this merging technique. To maintain the integrity of the dataset, rows with missing values in the three key columns ("librac," "libmslm," and "libath") were dropped, resulting in a reduction of 26,584 rows.After identifying that there were still missing data in the columns, rows with missing values in both "libath" and "partyid" were removed, reducing the dataset to 44,890 entries. Then, rows with missing "librac" data were dropped, further bringing the dataset down to 39,732 entries. The "libmslm" variable posed a unique challenge data exploration tool on the GSS website; the question started being recorded much later than other variables. For instance, while `librac` was recorded since 1976 and `libath` since 1972, `libmslm` only began in 2008. By dropping all the missing values in the "libmslm" variable, the number of rows in the dataset would decrease by

26293. This discrepancy led to a decision to create two distinct tables that excluded missing data – one that excluded `libmslm` and one that included it, ensuring that analyses derived from each dataset would be consistent and not skewed by the lack of earlier data for the `libmslm` question.

The “partyid” variable had multiple categories that could be simplified for clarity in visualizations and analysis. Hence, a mapping was created to group the categories into broader buckets, like “independent/other party”, “democrat”, and “republican.” This decision reduced the number of unique values for “partyid” from eight to three. To gauge the implications of this transformation, cross-tabulation analysis was performed on both the original and newly mapped "partyid" categories against other relevant variables (“librac,” “libmslm” and “libath”).

Cross-tabulation was used to determine how values in the "librac", "libmslm", and "libath" columns related to different political parties in the "partyid" column. This helped inform the broader political groupings. By observing the counts it was revealed that categories like “independent, close to democrat”, “independent (neither, no response)”, “independent, close to republican”, and “other party” exhibited similar distributions across “librac,” “libmslm” and “libath”. By consolidating them under “independent/other party”, the data's original structure through joining categories with comparable behaviors. Similarly, the clear political positions of 'strong democrat' and 'strong republican' were kept as 'democrat' and 'republican'. Overall, the grouping matches common views on political affiliations, making the dataset both easy to understand and faithful to its original trends.