Hello! I’m Conie O’Malley and I’m going to walk you through my project update for Understanding Book Banning Trends in the United States regarding data collection and cleaning.

Our agenda is to go through data types, data description, data retrieval and cleaning, and initial data visualizations.

The core of our analysis relies on two primary data types: full-text book content, which captures the complete text of the banned books, and thematic metadata, which includes categorization and contextual information about each book – often themes and back cover summaries. This fits into my project because I am comparing the full text themes that are deduced from textual analysis with the written back cover themes that many people will read before opening a book – an analysis of judging a book by its cover.

Project Gutenberg is my primary source for full-text data. This platform provides access to a number of free eBooks, allowing us to retrieve the complete texts of the selected banned books.

Thematic metadata is sourced from Google Books, which genre categorizations, book summaries, and thematic classifications of books for our comparative analysis.

The goal of collecting and processing this data is to enable statistical and machine learning-based analyses. This will help us uncover patterns and trends in censorship practices across different regions - primarily focusing on the connections of cover themes and between the cover themes.

The dataset includes a total of 20 banned books, with an equal representation of 10 books from Florida and 10 from Iowa. This balanced approach allows for a comparative analysis of censorship trends in these two states but also helps prevent the project from being too unwieldy.

One of the initial challenges encountered was the limited availability of full-text content for certain banned books. Out of the initial sample, I was only able to get one full text due to copyright issues. This caused me to pivot to sample texts available through Project Gutenberg from each state’s banned book list. I was only able to download 6 texts directly from the gutenbergr package. The other 14 I had to download directly from Project Gutenberg’s website using a function I wrote in r. This allowed me to download all the texts I needed; however I had to perform extra steps and cleaning for the additional 14 texts before storing them in data frames

After gathering all of the text data, I started retrieving the thematic data from Google Books – to include the subject and descriptions of all the texts. At first I tried writing code to scrape the data, but I was running into firewalls, then I tried to use the Google Books API, but was not successful there. I then was able to query ChatGPT to pull the data from Google Books website and put it into CSV format. I spot checked a few of the descriptions to ensure accuracy and there was no hallucination.

Now that my data was assembled, I joined all of them together to form a Florida and an Iowa dataset and began preprocessing. I wrote a function that would standardize the text – converting all characters to lowercase, removing punctuation, special characters, and excessive whitespace, then removed stop words and short words like ‘a’ or ‘an’. Finally I lemmatized and tokenized the words to prepare them for future bigram analysis. I validated the final structure then began to work on my visualizations. Initially, I had to run through several iterations of NA removal – I could not eliminate them in one swoop. This was only apparent after my first TF-IDF visualization came back with NA having a very high TF-IDF rate. After 3 iterations I was able to eliminate the NA values. What I discovered in my initial analysis is that there are a high volume of verbs like will, said, and can – which may be construed as a verb or noun. This is giving me thoughts about potential parts of speech tagging – but this is an investigation for further down the road potentially. However, the TF-IIDF matrix corrects some of these words and we have some high values for words like consulate, daughter, master, madame, and mate. I will continue my analysis to try and uncover more trends and potentially preprocess the data a little further.

Lastly, I will work on my bigram analysis, subgroup comparisons between Florida and Iowa, and my classification model development.

That is the current update on my project – a description of the data, its sources, how I retrieved and cleaned them, my initial visualization and the next steps.