# Minutes 12/23/24

## Meeting notes

- Looked through news outlet data (including new news outlets) to see if it runs through web scraper properly
    - Successfully ran all 100 articles
    - Runtime: 2m 40s
- Put new data into CSV file, will be stored in Raw Data file in GitHub google sheet
- Discussed distribution of article categories
    - Originally (13): Art, Celebrity, Culture, Entertainment, Faith, Food, Health, Lifestyle, Parenting, Sports, Technology, Travel, Weather
- Article categories rearrange accordingly:
    - Lifestyle: Food, Health, Travel
    - Entertainment: Celebrity, Sports
    - Culture: Art, Faith, Parenting
    - Technology
    - Weather

## Data Cleaning Guide:

- Topics column:
    - Rename Food, Health and Travel to Lifestyle.
    - Rename Celebrity, Sports to Entertainment.
    - Rename Art, Faith, Parenting to Culture.
- Title column:
    - Ensure title was scraped correctly
- Word Count Column:
    - Fix rows that show 0 word count (rows 64, 65)
- Author column:
    - ABC News: Manually find author on article directly because it shows ‘ABC News’ for all of them

    - Fix formatting of author names to be consistent (First name, Last name)
    - AP News: Find author names and fix formatting
    - Fox news: Fix author name formatting
    - *Since we only have 100 articles, we will manually clean the data. However, we will have to use code to fix the author formatting for larger datasets in the future.*
- Release Date column:
    - YYYY-MM-DD format, manually input if needed.
- Article text column: Remove header and footnote
- Political slant column: NA those bitches
- Number of significant word column: NA these bitches too

## Assignment for Data Cleaning

Jonathan: AP, ABC
Felix: MSNBC, Forbes
Ryan: Newsmax, Breitbart
Rudd: CNN, TFP
Minh: FOX, VOX