As part of data refinement for **ics5110**, focused on the news articles csv

Step 1

- Verified each url is reachable
- Powershell script check_urls.ps1 (chatgpt)
- Produced 1.checked_urls.csv
- All url's seem reachable

Step 2

- Asked chatgpt to compute probability that row / article is related to a traffic incident, based on the title column
- Produced 2.accident_probabilities.csv

Step 3

- Asked chatgpt to refine this probability by analyzing the content of the url, which we already established is reachable in step 1
- Switched to python and beautifulsoup as powershell was too brittle

Step 4

- Looking at the generated probabilities, it misclassified rows, making it look like *title analysis* was better than content analysis. So asked for improvements

Step 5

- The data lacks publish date – created_at is not the publish date or close to the date of the potential traffic incident
- extract_publish_dates.py extracts publish date – the python script has problems with newsbook but there are only 8 rows so we can manually fetch these once newsbook is online

Step 6

- Now that we have a publish_date in the news articles, we can attempt to see if we can match up the 111 police press releases with the 321 news articles.
- Output news_merged_w_pr.csv is essentially Claudio's enhanced press releases with an added column article_id. 2 randomly checked records appear to have correctly linked the "right" records. However, there are news articles with the same date, so a press release dated 30 July 2025 ends up tagged with 5 news articles. These can be quickly cleaned MANUALLY.