Web Scraping in the Social Sciences

Conceptual Questions

Please write three to ten sentence explanations for each of the following questions. **You are only required to answer ONE of the two questions below.**

1. In the social sciences, what are two ethical or scientific risks of collecting data via web scraping (e.g., representativeness, privacy, terms of service, measurement error, scraping-induced missingness)? For each risk, briefly describe one practical mitigation strategy you would use in a reproducible workflow.
2. Web pages change. Explain two ways that changes to a website can break a scraping pipeline or alter the data you collect. What concrete steps would you take to (i) detect that the pipeline has broken and (ii) make the analysis replicable for someone running your code later?

Applied Exercises

Use the code in the week's code tutorial and the lecture slides to answer the following questions.

3. Using **ten** Penn State faculty members from your department(s) or affiliated with SoDA, create a plot of **citations over time** for each professor.
   - Try changing the plot style (e.g., line thickness, points, theme, labels).
   - Your figure should be readable with 10 people (facets are fine).
4. Visualize **or** discuss how the work of these professors overlaps.
   - One approach: use `scraped_interests` from PSU profile pages.
   - Another approach (advanced): compare their most common publication keywords.

   Provide at least one visualization **or** a short written discussion (5–10 sentences) describing the main overlap patterns you observe.
5. What is the **median citation count (per year)** for each person in the data?
   - Hint: `group_by(name)` + `summarize(median(...))`.
   - Clearly state whether your median is computed over observed years only, or whether you treat missing years as zero (and why).
6. **Challenge Question (Optional — if you finish early):** Compute each scholar's **total citations** and **h-index** using `get_profile()`, then compare those across the faculty members.
   - Present a clean comparison (a table and/or a bar chart).
   - Briefly interpret what the comparison does *and does not* tell you (2–4 sentences).