Data Sleuthing Lab

Jeremy R. Manning

🕵️ Overview

What can (and *can't*!) data tell us? In this lab, your team will explore this important issue by playing two roles: **data creator** and **data sleuth**. Your job in the "creator" role will be to (as the name implies) *create* a dataset to be shared with another team. Later, once you've passed off your team's dataset and received a new one from another team, your team will transition into a "sleuth" role, where you'll examine and analyze different aspects of the dataset.

🕵️ Learning objectives

The goal of this lab is to learn about the "results" sections of scientific articles. In general, results sections report the *results* of the analyses described in the paper's methods section. A good results section will also help its audience to understand the basic logic of the study, any critical design decisions pertaining to the analyses, and how the different findings might be interpreted. To that end, you'll:

- Practice communicating clearly and directly
- Improve your understanding of what goes into creating a dataset
- Gain intuitions about what you can versus can't conclude from a given dataset
- Practice wrangling and analyzing data, creating figures, and running statistical tests
- Practice interpreting results
- Practice thinking about study design, resources, effort allocation, and time management

🕵️ Procedure

We'll divide the class into four teams: A, B, C, and D. Each team will begin by finding or creating a dataset, following some specific guidelines (outlined below). You'll also create some "documentation" for the dataset. As part of putting your dataset's documentation together, you'll come up with a set of questions to be explored or answered about your dataset.

In the second phase of the lab, you'll pass your dataset onto another team and receive a dataset from a different team (A will give their dataset to B; B to C, and C to D; D to A). Using the "sending" team's questions as a guide, the "receiving" team will work with their designated teaching assistant to examine the dataset.

You'll conclude the lab by drafting a "results section" based on the dataset you analyzed.

🔎 Part 1: Construct a dataset!

Your team's dataset can be made-up (fake) or real (e.g., downloaded from a public repository, etc.). However, whether you create a fake or real dataset, it should be *interesting* (broadly construed), and it must follow several formatting guidelines:

1. Your dataset should contain a minimum of 500 observations and a maximum of 10,000 observations.
2. Each observation should include at least 5 features and no more than 100 features. For example, if you were constructing a dataset about fish swimming in an aquarium, each "feature" might be some aspect of a fish's position in the tank (e.g., x, y, and z coordinates, and perhaps the yaw, pitch, and roll of the fish relative to each axis).
3. You should save your dataset as a Google Sheet (set sharing permissions to "anyone with a link can view"):
    - Top row: column labels (single word or abbreviation, lowercase, no non-letter characters)
    - Each subsequent row should have a single observation
    - Each column should correspond to a single data feature
    - Data values should either be integers (1, 2, 3, etc.), real numbers (3.14, 2.71, -45.6789, etc.), dates or times (12:50:00 04/18/2022, etc.), or character strings ("happy", "sad", "angry", "excited", etc.; note: strings don't need to be enclosed by quotation marks).
    - The formatting should be consistent throughout the dataset (to facilitate analysis)
    - The formatting should be consistent within each column– e.g., all of the values in a column should have the same "type"
4. Create documentation for your dataset (save as a Google Doc; set sharing permissions to "anyone with a link can view"). Include the following information:
    - Where did the data come from? This can be made up, but it should be explained! Interesting back stories are encouraged. Help to motivate the team you're sending the dataset to.
    - How were the data collected? Briefly explain, in 1–2 paragraphs.
    - What do each of the features mean, in plain English?
    - What are the observations? (E.g., timepoints? People? Trials?)

- A list of 5 questions:
  - All questions must seem plausible
  - All questions must be answerable (or potentially answerable) with a relatively straightforward analysis (e.g., PSYC 10-style statistical test, a basic chart or figure, etc.). In other words, you should keep the scope of the questions relatively narrow and specific.
  - At least 1 question should be possible to explore/answer using the dataset
  - At least 1 question should be *im*possible to explore/answer using the dataset
  - All other questions can either be possible or impossible to explore/answer using the dataset
  - Keep track of which questions can/can't be answered with the dataset, but keep those labels hidden (i.e., don't put it in your to-be-shared documentation)

Once you've created your team's dataset, documentation, and "possible vs. impossible" labels for each question, upload your materials using this form.

🔎 Part 2: Check out the data!

Each team will analyze data from their designated source team. Answer each of the five questions included in your received dataset's documentation using any approach you deem appropriate. Figures and stats are encouraged!

For each question, document the following:

- State any assumptions you're making to answer the question
- Explain how you'll approach answering question with the given analysis. In other words, explain how what you're trying to discover relates to the analytic tools and approaches you're employing.
- Explain what conclusions can or can't be drawn from the given analysis (according to how the analysis turned out).
- If you think a particular question is impossible to answer, explain why– and what you'd need (from the dataset, analytic tools, etc.) in order to be able to answer that question. You should still carry out *some* sort of analysis for each question; the idea is to notice and document when the insights possible to attain from a particular analysis fall short of the question(s) you're trying to answer about the data.

Your team can share a common document and set of analyses, figures, stats, etc.

🕵️ Using GenAI in this lab

Generative AI can be a powerful analytical partner– but only if you understand both its capabilities and its limits. In this lab, you'll deliberately compare two ways of working with GenAI: a **hands-off** approach (where you let the AI decide everything) and a **hands-on** approach (where you specify exactly what you want).

🔎 Part A: The hands-off prompt

Take your dataset and ask your GenAI tool of choice (e.g., ChatGPT, Claude, Gemini) to **analyze it for you**, without providing further clarifications or specifications. A prompt as simple as *"Here's a dataset. Please analyze it and tell me what you find."* is fine. Then reflect carefully on what happened:

- **Which analyses did the AI choose?** Did it run a t-test? Compute correlations? Make a histogram? Do something more complex?
- **Were the chosen analyses *appropriate* for the data?** For example, did it use a t-test on continuous data when chi-squared would have been more appropriate? Did it compute correlations between variables that don't make sense to correlate?
- **Were the analyses implemented correctly?** If the AI showed you code, does the code actually do what it claims? If it just gave you results, can you reproduce them?
- **Did the AI come to its own conclusions?** Did it claim to find a "significant effect" or "important pattern"?
- **Were those conclusions accurate or hallucinated?** How can you tell? Can you verify them against the actual data?

Document what you found. Be specific: include the exact prompts you used, the AI's responses, and your verification steps.

🔎 Part B: The hands-on prompt

Now switch modes. Pick *one specific analysis* you want to perform on your dataset (it can be one of the questions from your dataset's documentation, or something new you noticed in Part A). Before opening any AI tool:

1. **Outline the analysis step by step on paper or a whiteboard.** What are the inputs? What's the output? What intermediate steps are needed? What test or visualization will you use? How will you interpret the result?
2. **Construct a detailed prompt** for your GenAI tool. Include: the structure of the dataset (column names, types), the specific question you're answering, the analysis steps you've outlined, and the format you want the output in.
3. **Run the prompt** with your dataset attached.

Then reflect:

- Did the AI follow your specification, or did it deviate?
- Was the result more accurate, more useful, or easier to verify than what Part A produced?
- What was easier? What was harder?
- Were there steps you specified that the AI got wrong? Were there steps you *didn't* specify that the AI handled well anyway?

🔎 Reflection

In your writeup, compare the two approaches. When is the hands-off approach acceptable? When is it dangerous? What does the hands-on approach require *you* to bring to the table? What does this tell you about how (and when) to use GenAI as a scientific collaborator?

The ability to use AI as an analytical tool *while maintaining rigorous verification* is one of the most valuable skills you can develop as a scientist. AI can do the computational heavy lifting, but *you* must be the one who knows whether the results make sense.

🕵️ Writing your lab report

Your writeup for this lab will mimic a "results section" of a scientific article, about the dataset your team analyzed. Your report should comprise the following elements:

1. **Overview paragraph**: Open with a paragraph (roughly 1/2 page) that summarizes the dataset you analyzed (where it came from, what the observations and features represent, how many of each), the questions you explored, and a high-level preview of what you found. This sets the stage for the rest of the report.
2. **Dataset description**: Briefly describe the dataset's structure– the number of observations, the features, the type of each feature (categorical, continuous, ordinal, etc.), and any quirks or limitations you noticed before running analyses (e.g., missing data, unusual distributions, suspicious values).
3. **Per-question analyses**: For *each* of the five questions in the dataset's documentation, include:
    - The original question (quoted verbatim from the documentation).
    - Any **assumptions** you needed to make to attempt the analysis (e.g., "we assumed missing values could be dropped," "we treated the 1–7 ratings as continuous").
    - A description of the **analytic approach** you took. Explain *how* what you're trying to discover relates to the analytic tools you used– e.g., "to test whether X differs across groups, we ran a one-way ANOVA on..."
    - At least one **figure** with a caption describing what's shown.
    - Relevant **statistical results** (test statistic, p-value, **effect size**, and a 95% confidence interval where appropriate).
    - A paragraph describing your **conclusions** or **insights**. If a question is impossible to answer with the dataset, explain *why*– and what you'd need (additional features, a different study design, more observations, etc.) in order to answer it. You should still attempt *some* analysis for impossible questions; the goal is to document the gap between what the data can support and what the question is really asking.
4. **Possible vs. impossible**: Across your five questions, identify which you ultimately judged to be possible vs. impossible to answer. How confident are you in each judgment? Were any questions ambiguous (e.g., partially answerable, or only answerable

under strong assumptions)? How does this compare to what the *creating* team had in mind? (You can ask them, or check after submitting.)

5. **Storytelling**: Organize the five question-paragraphs so that they tell a coherent "story" about the dataset. Use transition sentences between sections. The order doesn't have to match the original documentation– pick an order that makes the narrative flow.
6. **GenAI exploration writeup**: Document your two-part GenAI exercise (see "Using GenAI in this lab," above). This should include:
    - **Part A** (hands-off): The exact prompt you used, a summary of what the AI produced, and your verification steps. Identify at least one analysis the AI got *right* and at least one it got *wrong* (or did suspiciously). Explain how you could tell.
    - **Part B** (hands-on): The analytic plan you sketched out, the prompt you constructed, and a summary of what the AI produced. Compare it to Part A: which approach gave you more useful, more verifiable, or more accurate results?
    - **Reflection**: When is the hands-off approach acceptable? When is it dangerous? What does GenAI require *you* to bring to the table to be a useful scientific collaborator?
7. **Reflection on the experience**: A brief paragraph (1/2 page) reflecting on the lab as a whole. What was the hardest part of being a "data sleuth"? What did you wish your dataset's *creators* had done differently? What will you keep in mind the next time you encounter a dataset you didn't create?

🕵️ Closing discussion points

The results section of a scientific paper is, in many ways, its core. The "merit" of a paper typically rests on how accurately and effectively it communicates the main findings of the study. For example:

- Was the dataset appropriate for answering the proposed questions?
- Were the analyses appropriate?
- Were the interpretations and/or conclusions justified by the data (or analyses)?
- Are the results visualized in an intuitive way?
- Are the results communicated clearly?
- Was the "logic" of the analytic approach clear and/or easy to follow?

Achieving these ideals is not simply a matter of listing sets of numbers or the outcomes of a series statistical tests. You need to help your intended audience understand *why* the approaches you took helped you to understand the questions you explored, how the questions fit together into a cohesive narrative, and so on. In other words, you need to make your "story" easy to read, easy to believe, and easy to follow.