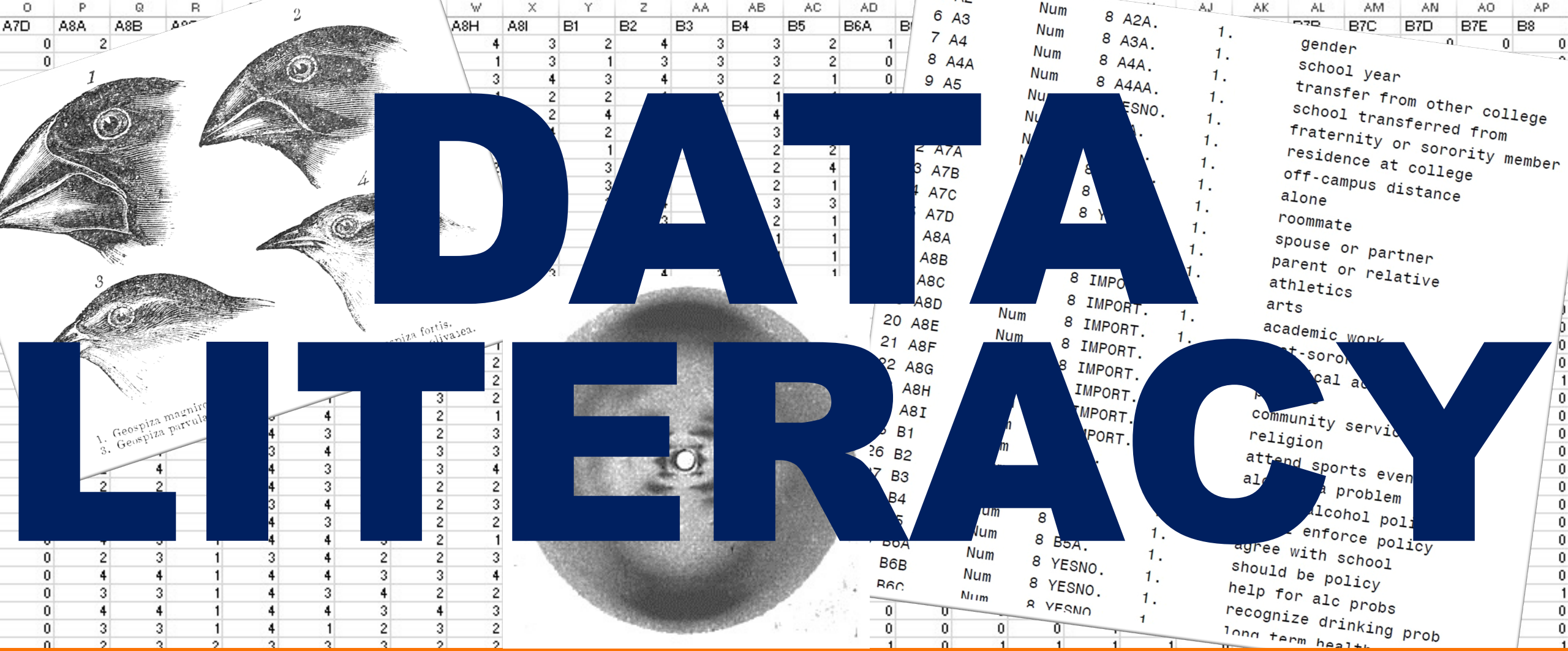
DATA LITERACY



Agnes Jasinska, Carrie Pirmann, and Claire Cahoon

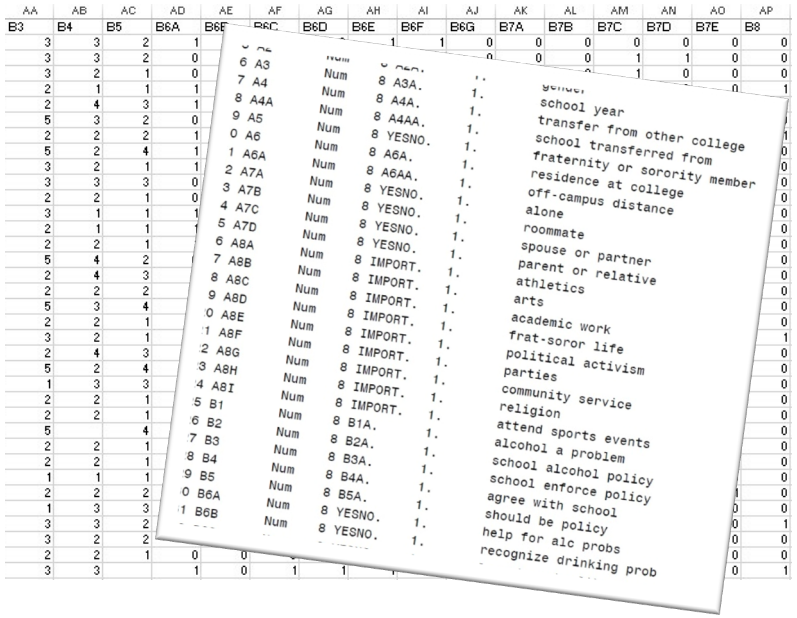
June 2023

What is data?

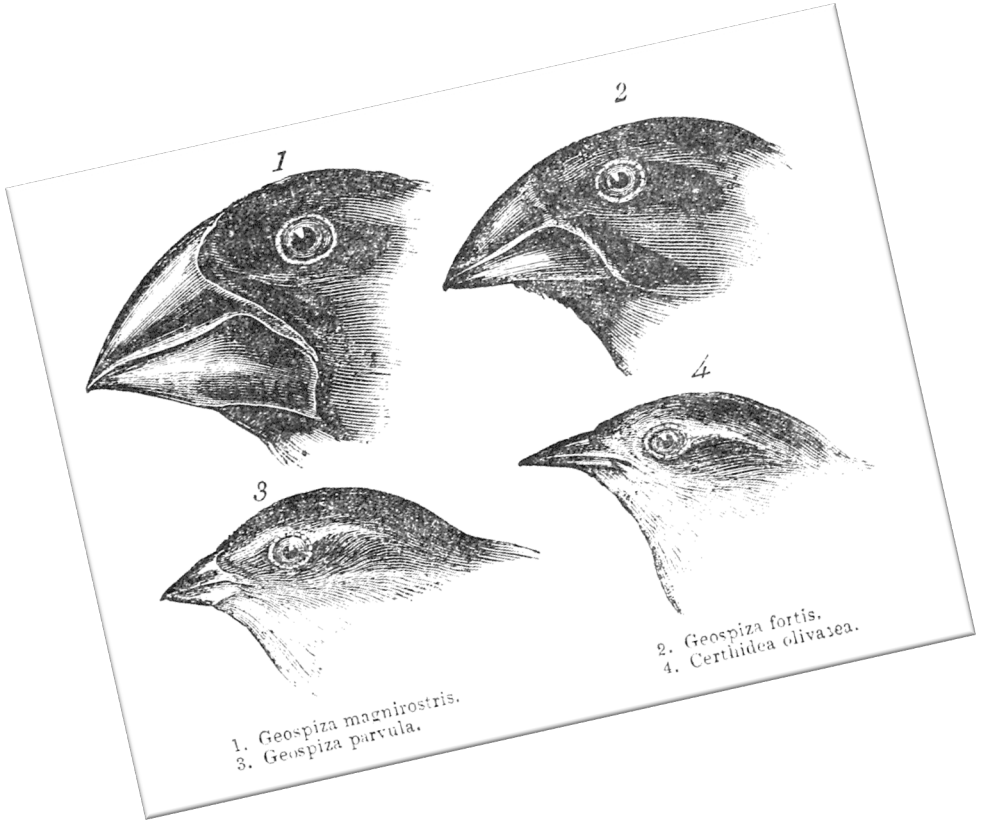
DATA* = FACTUAL INFORMATION THAT IS SYSTEMATICALLY RECORDED AND ANALYZED TO ANSWER A QUESTION

*Definition may vary by discipline

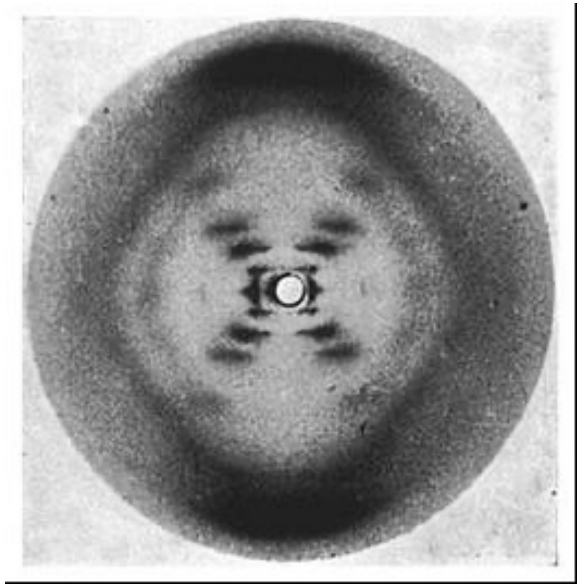
Data comes in lots of different forms!



Harvard College Alcohol Survey 2001



Darwin's finches from the Galapagos Islands (beak adaptation to specific types of foods present on different islands inspired Darwin's theory of evolution by natural selection)



Rosalind Franklin's x-ray diffraction image of crystalized DNA (evidence of a double helix structure)

Types of data

- **Quantitative data** deals with *quantities*, i.e., information that can be counted, measured, or otherwise expressed using numbers
  - Summarized and analyzed using traditional statistics or related methods

- **Qualitative data** deals with *qualities*, i.e., information that is descriptive and conceptual in nature, and cannot be easily expressed in numbers
  - Sources of qualitative data: text documents, interview transcripts, images, audio and video recordings, other
  - Requires qualitative summary and analysis (*not* statistics)

- **Mixed data** combines quantitative and qualitative data
  - Analyzed using mixed (quantitative and qualitative) methods

Data literacy & why it's important

DATA LITERACY = THE ABILITY TO READ, WORK WITH, ANALYZE, VISUALIZE, INTERPRET, ARGUE WITH, AND USE DATA TO MAKE DECISIONS AND SOLVE PROBLEMS

- **Data literacy skills are essential 21st-century skills!**
- Necessary to advance scientific knowledge, solve complex global problems, and improve human life
- *You don't need to be a data scientist – just keep developing the data literacy skills relevant to your academic & professional interests*

Activity

Tell us about your project.

1. What types of data are you planning to use?

2. Where do you plan to search for them?

(In other words, who may have already collected the data, how, and why?)

Steps (and questions to ask yourself) when working with data

1. Formulate a question or hypothesis
2. Acquire the data (collect new data or find a dataset you trust)
3. Get to know your data (incl. the research methods and ethical guidelines)
4. Prepare/“clean” the data for analyses and visualizations
5. Decide on appropriate analyses or visualizations
6. Interpret your results and tell a story about your data

1. Formulate a question or hypothesis

- What do we already know about the topic? (literature review)
- What do **you** want to know?
- What do you expect to find (e.g., pattern of results, differences between groups or conditions, cause and effect), and why?
- If you get your expected result, is there an alternative interpretation of this result? What additional questions could you ask of the data to eliminate this alternative?

2. Acquire the data (new or existing)

Two primary approaches:

- Collect new data
  - You can optimize the study design to your research question or hypothesis
  - But a new research study and data collection can be time-consuming & costly
- Find an existing dataset → OPEN DATA
  - Carefully evaluate the source: Do you trust the authors? Is the data of high quality? Does it have all required documentation (codebook, IRB approval)?
  - Make sure you have permission to use the data and present/publish the results

2 strategies to find open data

Lots of open-access, publicly available datasets online that you can use!

1. Find an **established data repository**, and search for a dataset by topic or other attributes.

2. Find **a published research article** and locate the original dataset used. (Many peer-reviewed professional journals require data sharing as a condition of publication. Some journals also curate a list of recommended data repositories.)

Open data repositories

From the Bertrand Library main page, go to Research by Subject Guides: Data Services → Data → Open Data

https://researchbysubject.bucknell.edu/c.php?g=956824&p=6906764

Some examples:

- Inter-University Consortium for Political & Social Research (ICPSR) → built-in data analysis tools!
- https://www.data.gov/ → home of U.S. government data
- Census data (people) and Economic Census data (businesses)

3. Get to know your data

Even if you didn't collect the data, understanding the research methods is critical to interpreting the results!

- How was the data collected? Who collected it? When and where? With what measures or instruments? Using what study design? Primary research question of the study? Source of funding?
- What were the relevant *ethical research guidelines*, and were they followed (e.g., IRB approval and informed consent)?
- Which *variables* will you look at to answer your question or test your hypothesis? (The Codebook is your friend.)

3. Get to know your data

Example:

Formulate and test a hypothesis about the relationship between a student's gender and their likelihood of binge drinking

(Using the Harward College Alcohol Survey 2001 dataset.)

Variables needed?

AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
B3	B4	B5	B6A	B6B	B6C	B6D	B6E	B6F	B6G	B7A	B7B	B7C	B7D	B7E	B8
3	3	2	1	0	0	0	1	1	0	0	0	0	0	0	0
3	3	2	0	0	0	0	1	1	0	0	0	1	1	0	0
3	2	1	0	0	0	0	0	0	1	0	0	1	0	0	0
2	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1
2	4	3	1	1	1	1	1	1	0	0	1	1	1	0	0
5	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	2	1	1	1	1	1	1	1	0	1	1	1	0	1
5	2	4	1	1	1	1	1	1	1	0	1	1	1	0	0
3	2	1	1	0	0	0	1	1	1	0	1	1	1	0	0
3	3	3	0	0	0	0	0	0	0	0	0	1	0	0	0
2	2	1	0	0	0	0	1	1	1	0	1	1	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
2	1	1	1	0	0	0	1	1	0	0	0	0	1	1	0
2	2	1	1	1	1	1	0	0	1	0	1	1	1	0	0
5	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	2	0	0	1	0	1	1	0	0	0	0	1	0	0
5	3	4	0	0	0	0	0	1	0	0	0	0	0	0	0
2	2	1	0	0	0	0	1	1	1	0	0	1	1	0	0
3	2	1	1	1	0	1	1	0	0	0	1	1	0	0	1
2	4	3	1	0	0	1	1	1	0	0	0	1	1	0	0
5	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0
1	3	3	1	1	1	1	1	1	1	0	1	0	1	0	0
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	0
2	2	1	1	1	1	1	1	1	1	0	0	1	1	0	0
5		4	1	1	1	1	0	0	0	0	0	1	1	0	0
2	2	1	1	1	1	1	1	1	1	0	1	1	1	1	0
2	2	1	1	1	0	0	1	0	0	0	1	1	1	0	0
1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	0
1	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0
3	3	2	1	0	0	0	0	1	0	1	0	1	1	0	1
3	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	1	0	0	0	0	1	1	0	0	0	0	1	0	0
3	3		1	0	1	1	1	1	0	1	1	0	0	0	1

A .csv data file with 483 columns (variables) and 10904 rows (subjects)

3. Get to know your data

Gender variable?

(We search the Codebook, starting with the survey instrument)

A2. Are you male or female?

◯ Male
◯ Female

5	A2	Num	8 A2A.	1.	gender
6	A3	Num	8 A3A.	1.	school year
7	A4	Num	8 A4A.	1.	transfer from other college
8	A4A	Num	8 A4AA.	1.	school transferred from
9	A5	Num	8 YESNO.	1.	fraternity or sorority member
10	A6	Num	8 A6A.	1.	residence at college
11	A6A	Num	8 A6AA.	1.	off-campus distance
12	A7A	Num	8 YESNO.	1.	alone
13	A7B	Num	8 YESNO.	1.	roommate

A2	Are you male or female?	Format:: A2A . = "Missing (.)" 0 = "female (0)" 1 = "male (1)"

3. Get to know your data

Binge drinking variable??

(We check the Codebook & survey instrument again: but no such question in the survey, and no such variable in the dataset!)

Definition: Binge drinking is a pattern of drinking that brings blood alcohol concentration (BAC) levels to 0.08 g/dL. This typically occurs after **4 drinks for women** and **5 drinks for men**—in about 2 hours
https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/college-drinking

→ Different criteria for binge drinking in men and women!

3. Get to know your data

Fortunately, there is another measure of drinking intensity that we can use.

C12	In the past 30 days, on those occasions when you drank alcohol, how many drinks did you usually have?	Format: C12A . = "Missing (.)" 0 = "none past 30 (0)" 1 = "1 (1)" 2 = "2 (2)" 3 = "3 (3)" 4 = "4 (4)" 5 = "5 (5)" 6 = "6 (6)" 7 = "7 (7)" 8 = "8 (8)" 9 = "9 or more (9)"

If we combine C12 drinking intensity + A2 gender, we get a new categorical measure of "**usual binge drinking**" which will work to test our hypothesis.

C12: 4 or higher + A2: 0 "female" → usual binge drinking for women

C12: 5 or higher + A2: 1 "male" → usual binge drinking for men

3. Get to know your data

Bottom line:

Get to know your data and the methods used to collect it!

If you understand your data, you can:

- Ask new questions and test new hypotheses
- Be creative but also rigorous in your analyses
- Understand the limitations of the data
- Better interpret your results

Data quality & integrity is key!

We use data to learn something about the world, to draw a valid and accurate conclusion, to make the best, most informed decision

Good quality data = useful and beneficial

Poor quality data = useless and potentially harmful

To help safeguard data quality & integrity:

- Manage your data carefully and document your work from beginning to end of the project (data sources, methods, processing, analyses, visualizations, tools, results, interpretation)

Break

Stretch your legs.

Steps (and questions to ask yourself) when working with data

1. Formulate a question or hypothesis
2. Acquire the data (collect new data or find a dataset you trust)
3. Get to know your data (including the research methods and ethical guidelines)
4. Prepare/“clean” the data for analyses and visualizations
5. Decide on appropriate analyses or visualizations
6. Interpret your results and tell a story about your data

Data ethics

Data ethics is part of research ethics – and it's about trust

3 principles of responsible conduct of human subject research:

- ***Respect for persons*** (a person needs to give an informed consent to participate in any research study; they have the right to know what the study is about, as well as the risks and benefits; and they can withdraw their consent at any point)
- ***Beneficence*** (minimize the risk while maximizing the benefit)
- ***Justice*** (the risks and the benefits should be fairly distributed)

Data privacy

Increasingly important (again, it's about trust!) – but the ethical guidelines and legal regulations are only evolving

- What kinds of data can be ethically and/or legally collected on people? And on what conditions?
- Who owns the data? (The person supplying the data? The researcher who collects it? The funding agency or the business who paid for the study? The government of the country?)
- Who has the right to see the data? Use the data to make decisions (and what kinds of decisions)? Sell it and profit from the data?

Be an informed and responsible data user!

Activity

Discussion of the assigned reading:

Data Feminism: The Power Chapter by Catherine D'Ignazio and Lauren Klein

1. How is data related to power and social justice?
2. What are some ways data can be harmful?
3. Ways that YOU can avoid harm and foster justice?

4. Prepare/ “clean” the data

- Save your working data file with a new name; keep the original secure
- Consider the tool you will use for data analyses or visualizations – and structure your data for that tool
- Check for missing data, and decide how to deal with it
- Be careful and consistent at each step to avoid errors

General rule: Document all the changes you make to your data files, no matter how small, so you (or someone else) can repeat/ replicate your processing steps, your analyses, and ultimately your results

5. Analyze and visualize the data

The goal is to find **the right analysis** or **the right visualization** to answer your question or test your hypothesis.

Things to consider:

- Type of data (e.g., quantitative, qualitative, etc.)
- Study design used to collect the data
- Limitations of the data
- How the results will be used and by whom
- Tool/s you intend to use (e.g., statistical software)
- *Be as simple as you can be – but no simpler*

6. Interpret the results and tell a story about the data

- Go back to your initial research question or hypothesis – and now answer it with data
- Consider your audience, their needs, interests, and level of knowledge, and how they will use the results
- The goal is to tell a clear, accurate, logical, and compelling story about your data

Activity

Data visualizations: the good, the bad, and the ugly.

1. What makes a good visualization?
2. What are some mistakes to avoid?

LinkedIn Learning is your friend!

LinkedIn Learning: https://www.bucknell.edu/linkedinlearning

- Set up your LinkedIn professional profile
- Use LinkedIn Learning online courses to learn or improve your data literacy skills (you can search by topic, skill, tool, etc.)
- Display the completed courses on your LinkedIn profile to demonstrate your learning

Thank you!

Agnes Jasinska, Data Services Specialist

ajj006@bucknell.edu

Email to make an appointment!