Lab 1: Concepts in Data Organization

Your Name: Amelia Racky

Please complete this worksheet individually as part of your lab submission. Please discuss these questions / answers with your team, but each team member should write and submit their own answers and reflections.

Part 1. Decide what you want to learn and create a data protocol

1.1. Which topic is your team working on (vaccination access, air quality, police traffic stops)?

Air Quality

1.2. Who is on your team?

Phoebe, Daniel

1.3. As a group, think of a question, related to your topic, that you can answer by collecting data. Write it below:

What agricultural practices and methods affect air quality the most? Compare regenerative, organic, and conventional farming practices.

1.4. Why is this question important from a public health, environmental, or justice perspective?

Educating farmers, sustainability, GHG emissions, environmental degradation, enhancing biodiversity, crop rotation, nutrient-rich produce.

1.5. Who or what is your unit of analysis (e.g., person, facility, neighborhood, household, incident)?

Various regenerative, organic, and conventional farms in the United States.

1.6. What are the indicators / variables that you would need to collect to answer your question (replace the sample data with your own concepts, indicators, and data types? See if you can come up with 5-10 indicators that would be important to collect. Try to

think of some data types from at least 4 of the categories we discussed (nominal, ordinal, interval, ratio, unstructured, etc.):

	Concept	**Possible Indicator(s)**	**Type (Categorical, Numeric, Ordinal)**
1	GHG emissions	CO2, Nitrous Oxide, Methane	Categorical (type of emission) Numeric (amount-ppm)
2	Pesticide & Fertilizer Use	Lingering effects/time frame of pesticide use	Categorical (type) Numerica (amount)
3	Weather/extreme weather events	Wind, dust, rain, drought, etc	Categorical (type of weather event) Ordinal (severity)
4	Location	Population density, lat/long, elevation, nearby pollution sources, size of farm	Categorical
5	Land Use	Amount of growing seasons/activity on the farm/seasonal or annual	Categorical
6	Water Quality (evaporation)	How they dispose and source	Ordinal: very poor, poor, average, good, very good
7	Public Health	farmers health, community health	Ordinal: very poor, poor, average, good, very good
8			
9			
10			

Given the list that you and your team came up with....

1.7. Which variables were hardest to decide on?

Difficulty classifying variables and ensuring we covered all of the necessary variables due to the mass amount of influences that contribute to air quality.

1.8. How would you go about gathering this data? Keep in mind that data collection can be expensive and labor intensive, so practical considerations are important. Are there any changes you would make so that your data gathering strategy could be more practical if you only have a small budget?

-We would have to determine the best farms to use for our research, and speak with the farmers about purpose and willingness to participate.
-Survey for land use and agricultural methods.
-Use sensors for measuring particulate matter.
-Measure weather (past, present, future) to determine trends.
-Conduct water quality evaluations to determine its effect.
-Survey community for ordinal health data

Changes:

- Rather than tracking all weather, we might only focus on severe weather events.
- Looking at historical weather patterns rather than current could be helpful.
- Shrink Range: start small, gather results, inferences, work our way around to different locations if possible.

1.9. What steps will you take to ensure accuracy in the data you generate?

-Make sure all sensors are aligned.
-Consider size in relation to agricultural practice.
-Consider outside air pollution.
-Consider historical records of air pollution.
-Consider preexisting health conditions.

1.10. What steps will you take to mitigate bias in your data?

-Research location beforehand
-Know population
-Include community input throughout the entire process

Part 2. Working with "Messy" Data

Download this zip file (lab01.zip), which contains 3 messy data files – one for each topic. Note that each of these CSV files (CSV stands for "comma-separated values") may have:

- Missing values
- Inconsistent date formats
- Mixed case and typos in categorical fields
- Duplicate records
- Ambiguous codes (e.g., "M" = male or "missing"?)
- Extra columns not needed for analysis

Have one person from your team upload the relevant CSV file to Google Drive. That person should then open the CSV in Google Sheets and invite the rest of the team as document editors When you're done, your team should collaboratively identify and fix:

- Missing values (decide how to handle them)
- Inconsistent formats (dates, units, capitalization)
- Typos and ambiguous entries
- Bonus (not required...just something to try): see if you can figure out how to create data entry rules so that only valid data is entered.

2.1. Summarize the decisions and changes that you made when cleaning the data:

First, we sorted cities from a-z. Then, we got rid of any missing values or nan values. Next, we made all of the text lowercase.

2.2. Which cleaning decisions felt objective? Which were subjective?

Removing all inconsistent values was subjective. Making all data lowercase to improve readability was objective. I would say sorting the cities a-z was also objective since we could have sorted by air quality designation instead.

2.3. Do you think a different team may have made different decisions when cleaning the same dataset? What might this mean for your analysis?

From what we learned today in class, there are so many ways to reorient data, so I would not be surprised if someone else were to clean this data up differently.

2.4. How can you clearly convey your decisions so that others using your dataset understand the reasoning behind your choices?

If you mean conveying the decisions of what data to collect, you can include metadata of a where, when, how, and why, as well as making sure to be as thorough as possible when gathering data. Remove personal bias and have multiple people review the work/results.

2.5. How might missing or inconsistent data affect policy recommendations?

Missing or inconsistent data could produce false conclusions or correlations. This is why no data is ever perfect, but some can be useful.

2.6. Given the cleaned dataset that you made, what kinds of conclusions can you potentially draw? Explain. Bonus:

- See if you can try aggregating or filtering the data to answer an interesting question about the data.

You could find that there are 20 “unhealthy” air quality designations across all of the cities, and 14 of those being from Franklin and Greenville. You could also conclude that Springfield has the highest amount of “good” air quality designations. You could find the average amount of PM, ozone levels, or average air quality for each city.

2.7. If you were in charge of designing the data collection protocol, what data do you wish had been collected? What data validation rules do you wish had been implemented?

Trying to link this data to some other piece of data such as average health in the community would have been interesting to see. Some metadata about time and place of collection as well as collection source could always be helpful in determining the validity of these results.

2.8. In your opinion, how much of “data analysis” is about decisions, not just numbers?

Like we’ve learned, no data is raw data. Data analysis is never just about the numbers. There is always going to be a bias based on the reasons why you chose to collect this particular data, or how you choose to interpret/display it.