Lab 1: Concepts in Data Organization

Your Name: Phoebe Schropp

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)?

Vaccination access

1.2. Who is on your team?

Angie, Gill, Ava

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

What are the strongest factors that affect vaccination rates in Western North Carolina?

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

Lower vaccination rates could indicate less access to vaccines, so seeing what impacts rates could indicate where unequal access may be occurring and where to look to make changes to ensure equal access.

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

Person and Household

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	Accessibility	Distance from clinic	Numeric (miles)
2	Accessibility/Transportation	Type of transportation	Categorical
3	State level mandates	Vaccine rate at time of mandate	Numeric
4	Politics	Political affiliation	Categorical
5	Religion	Religious affiliation	Categorical
6	Poverty	Average Household Income	Numeric ($)
7	Demographics	Age	Numeric
8	Demographics	Race/ethnicity	Categorical
9	Demographics	Gender	Categorical
10	Education	Education Level	Ordinal

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

1.7. Which variables were hardest to decide on?

Vaccine rates at time of state level mandates

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?

Census data, surveys?, CDC reports, GIS spatial analysis tools

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

Unbiased sources, use verified sources,

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

random sampling

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:

We decided to put a dash in for the missing values and set a condition that it flags the data as invalid if a non-numerical value is entered. We had 4 regions (NESW) and some were labeled nan which we renamed unknown and sorted to the bottom. We fixed the capitalization of the region column by making them all lowercase with only the first letter capitalized.

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

Moving the data with the region labeled nan to the bottom to in sorts eliminate it felt objective as that data was invalid same with making the document flag all missing data as invalid. Making all the regions only have the first letter capitalized felt subjective as well as sorting the data by region and putting dashes in for the missing data and changing all labeled nan to unknown.

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?

Yes, some may have chosen to have all the regions in all caps or all lowercase, some may have left the blanks alone, some may not have sorted the data. This might mean our analysis of the data may vary due to how we cleaned up the data set but for the most part any of the decisions should have resulted in a similar final look of the data set and thus similar analyses.

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

We could add comments about some decisions like what to do with missing values and why that makes that point invalid so others know why not to include that point in their analysis.

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

It takes away from an accurate understanding as if you don't have all the pieces of the puzzle you may not understand the entire picture it shows. This could lead to policy recommendations that do not actually work or help with the issue as it helps with the issue shown by some data but not all the data.

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.

I honestly can't draw any conclusions from this data set, it really confuses me. I can figure out how to sort it in a way to see if vaccination rate had an impact of covid cases in the region and I am confused by why there are so many of each region and

what they even really represent because how can north have so many different population numbers.

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?

I wish more details on region were collected, region of what? The US? The State? The County? And maybe showing covid cases as a percent of population and not just the number would have been more beneficial. I wish no data was accepted that had any missing values.

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

A lot of it is about decisions, maybe 50-75% because your analysis of the data depends a lot on how you organize the data and what value you accept and what you don't accept and such.

What to Submit

When you’re done answering these questions, please download this worksheet as a PDF and turn in submit it to the Moodle. Each person should submit their own worksheet (individually).