Mind the Mess: Preparing Survey Data for Public Health Analysis

Local Health Department Academy of Science Conference 2026

Lyndsey Blair and Abigail Stamm

January 28, 2026

Public Health
Prevent. Promote. Protect.

LHD ACADEMY OF SCIENCE

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Tribal-State Relations Acknowledgment Statement

The State of Minnesota is home to 11 federally recognized Indian Tribes with elected Tribal government officials. The State of Minnesota acknowledges and supports the unique political status of Tribal Nations across Minnesota and their absolute right to existence, self-governance, and self-determination. This unique relationship with federally recognized Indian Tribes is cemented by the Constitution of the United States, treaties, statutes, case law, and agreements. The State of Minnesota and Tribal governments across Minnesota significantly benefit from working together, learning from one another, and partnering where possible.

Minnesota Department of Health recognizes, values, and celebrates the vibrant and unique relationships between the 11 Tribal Nations and the State of Minnesota. Partnerships formed through government-to-government relationships with these Tribes will effectively address health disparities and lead to better health outcomes for all of Minnesota.

In the Office of Data Strategy and Interoperability, we demonstrate our commitment to Tribal-State relations by providing free assistance upon request and promoting health equity in data collection and use.

Presentation materials

Download from GitHub

- Codebook
- Survey
- Dataset

Public Health
Prevent. Promote. Protect.

LHD
ACADEMY OF SCIENCE

The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Contents

Public Health
Prevent. Promote. Protect.

LHD ACADEMY OF SCIENCE
The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Would You Trust These Data?

Age	Gender	Smoker
25	F	Yes
999	Male	Maybe
4		No
-2	Other	Only when I drink

Bad data = bad decisions

Public Health
Prevent. Promote. Protect.

LHD ACADEMY OF SCIENCE
The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Garbage In, Garbage Out

Public Health
Prevent. Promote. Protect.

LHD ACADEMY OF SCIENCE

The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Cleaning Is Most of the Job

Data scientists spend 60–80% of their time cleaning data.

Why Data Cleaning Matters

Common issues in public health survey data

- Missing or incomplete responses
- Duplicate records
- Inconsistent categories (e.g., “Male”, “M”, “man”)
- Outliers and implausible values (e.g., age=999)
- Free text instead of standardized responses
- Typos and data entry errors
- Multiple ways of recording “unknown” (e.g., NA, N/A, blank)

Effects of poor or no data cleaning

- Inaccurate results and misleading conclusions
- Biased estimates and distorted trends
- Poor decision-making and policy recommendations
- Ethical risks (misrepresenting communities)
- Inconsistent reporting across teams or agencies

Public Health
Prevent. Promote. Protect.

LHD ACADEMY OF SCIENCE
The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Definitions

Data Cleaning

- Preparing the data
- Fixing errors
- Handling missing values
- Removing duplicates
- Recoding variables
- Checking data quality

Data Analysis

- Using the data
- Finding patterns
- Running statistics
- Creating charts
- Building models
- Interpreting results

Understanding your dataset

Review the codebook

Note columns:

- **variable:** variable name in the dataset
- **label:** variable description (corresponding survey question)
- **type:** variable data type
- **options:** expected values for the variable

Review the survey

Note the following:

- Any skip logic (none in our example)
- Which questions require responses (if any)
- Which questions are open text
  - “Enter a number …”
  - “Describe your experience …”

Review basic data structure

Check the following:

- Does the number of rows equal the number of expected responses?
- Are all variables present?
- Check the first five rows of data.
  - Are columns the expected data type?
  - Do values align with expected values in the codebook?

Useful Excel functions for dataset review

Filtering: Excel's Data ribbon

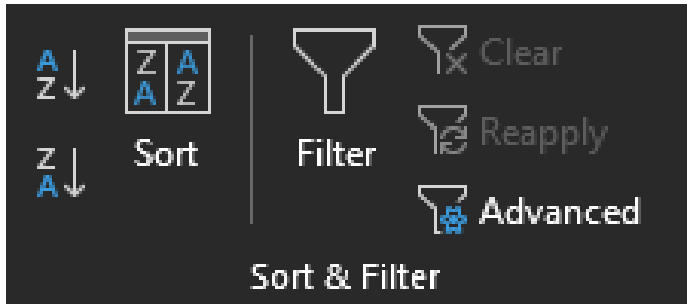
- Show only certain rows

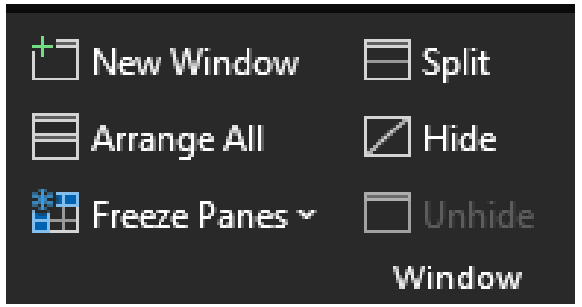
Sorting: Excel's Data ribbon

- Sort by specific columns

Freezing panes: Excel's View ribbon

- Make certain rows/columns always visible





Public Health
Prevent. Promote. Protect.

LHD
ACADEMY OF SCIENCE
The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

What if there are duplicates?

Are These the Same Person?

Name	DOB	Email
Sarah Johnson	03/12/1990	sjohnson@gmail.com
Sarah J.	03/12/1990	sarahj@outlook.com

Public Health
Prevent. Promote. Protect.

LHD ACADEMY OF SCIENCE
The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Flagging Potential Duplicates

Identify duplicates using variables expected to be unique:

- Full name + date of birth
- Email address or phone number
- Survey timestamp
- Record ID
- IP address (with caution)

Duplicates are *suspected*, not confirmed.

Handling Duplicates: What Do You Do Next?

Review all flagged records and confirm they are duplicates.

Then decide:

- Remove one record?
- Retain both records?
- Merge information across records?

No single "correct" choice

Always record and report what you decide to do!

Public Health Prevent. Promote. Protect.

LHD ACADEMY OF SCIENCE

The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Practical Tools for Managing Duplicates

Excel Tools:

- Conditional formatting
- Remove duplicates
- COUNTIF/COUNTIFS
  - IF (COUNTIF(A:A, A2), “Duplicate”, “Unique”)

What if data are missing?

Types of missingness

In analyses and reporting, these may all be handled differently.

- True (non-answer) missings
- System (skip logic) missings
- “Prefer not to say” responses

Interpreting missingness

"Prefer not to say" is different from non-answer

Possible interpretations:

- "Prefer not to say" could be "None of your business"
- non-answer could be ignoring the question

Skip logic missing is expected

- Interpret as "Not relevant"

What to do with missings

Always report number of missings and how you handle them in analyses, tables, and charts.

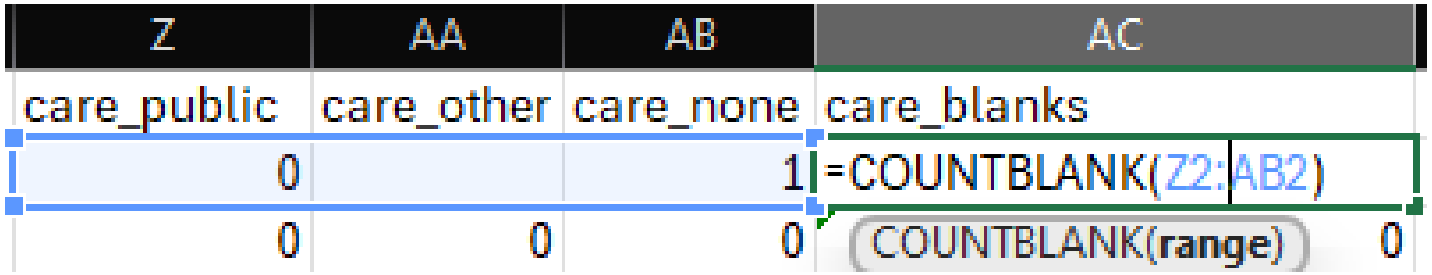
Common methods:

- Set all missings of any kind to NA
- Include only complete cases
- Exclude skip logic missings

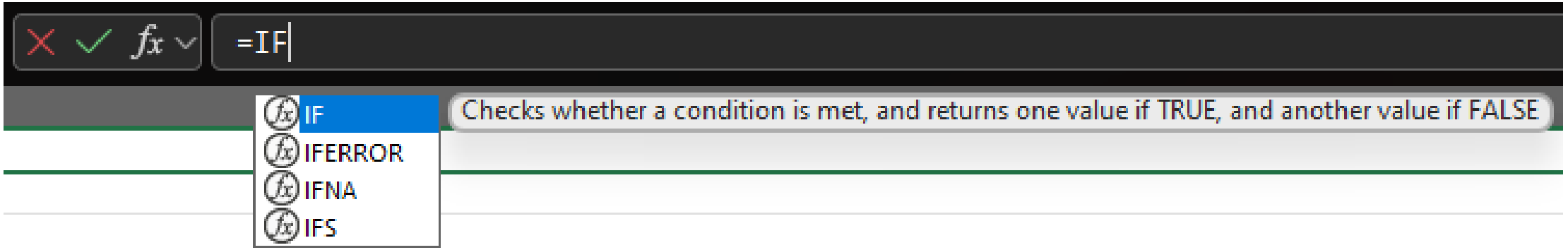
Excel tools to handle missingness

Filtering: Excel's Data ribbon

COUNTBLANK(): type into a cell, then select cells to count



IFERROR() and similar



What if there are outliers?

What Are Outliers?

Definition: Values that fall well outside the expected range for a given variable.

Example: Height (adults)

Population	Lower bound	Upper bound
Men	~5 ft 5 in	~6 ft
Women	~5ft	~5 ft 7 in

Someone **4 feet tall** would likely be flagged as an outlier.

Evaluating Outliers

In public health, removing outliers without justification can introduce bias and mask real population variation.

Ask:

- Is this value *plausible*? (review literature)
- Is this value a likely a data entry or measurement error?
- Can this value be confirmed?
- Does this value represent a real but rare case?

Outliers require **investigation, not automatic removal.**

Cleaning Outliers

Possible actions:

- Value is plausible → **retain it**
- Data entry or measurement error → **correct or modify**
  - Set to missing
  - Impute (with justification)
  - Cap (winsorize)

All decisions should be documented and justified.

Excel tools to handle outliers

- Sort & filter
- Conditional formatting
- MIN / MAX
- Descriptive statistics
- Pivot Tables

Cleaning and transforming variables

Recoding Variables

Modifying an existing variable

Changing how values are represented

Standardizing categories

Reformatting for analysis

Examples:

- Sex: “Male/Female” → 1/0
- Smoking: “Yes/No” → 1/0
- County: “jefferson / Jefferson / JEFFERSON” → “Jefferson”
- Likert scale: “Strongly agree” → 5

Excel tools for recoding variables

- Find & Replace
- IF statements
- VLOOKUP / XLOOKUP
- Text functions (UPPER, LOWER, PROPER)
- Data Validation

Always keep a record of how variables were recoded.

Public Health
Prevent. Promote. Protect.

LHD
ACADEMY OF SCIENCE
The National Association of Local Health Department Data Collectors

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY

Collapsing categories

Categorical variable types

Nominal: no order

- Example: gender

Ordinal: meaningful order

- Example: Likert scale from “Poor” to “Excellent”

Interval: named number ranges in order

- Example: BMI categories of “Underweight” to “Morbidly obese”

Categorical data definitions

- Cell count: number of responses for any one bin or value

Poor	Fair	Good	Very good	Excellent
4	8	22	36	14

- Small numbers: cell counts below a minimum threshold for privacy, confidentiality, and statistical power
- Collapse categories: combine two or more categories to handle small numbers or reduce the total number of categories

Why collapse categories

Poor	Fair	Good	Very good	Excellent
4	8	22	36	14

Address small numbers:

- “Poor” + “Fair” = “Not good”, count = 12

Interpret results more easily:

- “Good” + “Very good” + “Excellent” = “Good or better”, count = 72
- Interpretation: 86% (72/84) of respondents reported their health was good or better.

Commonly collapsed categories

Consider: does collapsing categories make sense for your goal or project?

- Variables with many categories
- Variables with at least one category with small numbers
- Race, ethnicity, or gender groups
- Income or age ranges
- Likert scales

Creating new variables

Creating new variables

Creating new variables means:

- Deriving new information
- Combining variables
- Transforming values into something analytically useful

Examples:

- Age group from age (0-17, 18-34, etc.)
- BMI from height and weight
- Risk score from multiple indicators

Excel tools for creating new variables

Binary Indicator: Trying to lose weight

=IF(ISNUMBER(SEARCH("yes",Y2)),"Yes","No")

Age from Year of Birth

=YEAR(TODAY())-S2

BMI from weight (lb) and height (ft)

=703*(V2/((W2*12)^2))

Exploring open text data

Regular expressions (RegEx)

Definition: Special character patterns used to manipulate text

Characteristics:

- Complex to learn
- Useful and powerful when evaluating free text
- Syntax varies across software and coding languages

RegEx example (Excel syntax)

Using REGEXEXTRACT() to check for private information:

- E-mail: “[a-z0-9_.-]+@[a-z0-9_.-]+.(com|us|gov|edu|org)”
- Phone number: “[0-9() -]+”

Note: this only works in some versions of Excel.

Acknowledgements

Thank you!

Dr. Lyndsey Blair: lblair3@bellarmine.edu

Abigail Stamm: abby.stamm@state.mn.us

Public Health
Prevent. Promote. Protect.

LHD
ACADEMY OF SCIENCE

DEPARTMENT OF HEALTH

BELLARMINE UNIVERSITY