Learning with Data

NC Survey on Human Health and the Environment

https://go.ncsu.edu/dsa.tsh_surveystudentlesson

Taryn Shelton
The Science House
Data Science Academy

Katlyn May
Lacey Brown
Center for Human Health and the Environment
Center for Environmental and Health Effects of PFAS

NC Survey on Human Health and the Environment

Case Study:

You are a researcher at the Center for Human Health and the Environment (CHHE) at NC State University. Your team is studying public perceptions and knowledge of certain health factors among NC residents in order to better understand how human health is impacted by certain social and environmental factors.

Data Collection (Why)

The purpose of your survey is to find two new pathways of potential research in an effort to improve the lives of NC residents. Once you have conducted your survey, you will look for trends and identify possible correlations in order to decide what your next research question and study and will be.

TASK 1: Data Collection (How)

Brainstorm with your team about how you would first go about collecting information from the public about their knowledge, perceptions, and opinions.

TASK 2: Data Collection (What)

Make a list of all the types of questions your team might want to ask residents. Think about…

- **social factors**, such as their demographics and where they get their information,
- **environmental factors**, including current public health threats.

You may use your resources to discover **current environmental health threats to NC residents**.

TASK 3a: Reviewing and Questioning the Data

The two linked datasets below contain the survey information collected from NC residents.

2020 North Carolina Survey on Human Health and the Environment **Data File**

2021 North Carolina Survey on Human Health and the Environment **Data File**

Open each file and review the data.

1) What do you see?
2) What do you wonder?

TASK 3b: Reviewing and Questioning the Data

The two linked codebooks below contain the survey questions that NC residents were asked. These questions match the Q numbers in the datasets. (Make sure you match each codebook to the correct dataset)

2020 North Carolina Survey on Human Health and the Environment **Codebook**

2021 North Carolina Survey on Human Health and the Environment **Codebook**

1) What do you see?
2) What do you wonder?

TASK 3c: Understanding Data in its Context

When given a dataset to analyze, you should ask questions about it to better understand the data in its **context**.

Consider the following questions and try to answer them as a class:

1. How was the data collected?
2. Who collected the data?
3. What was their intent (purpose)?
4. Who is represented in the data? Who is missing?
5. What is represented in the data? What is missing?
6. Are there any potential biases that you can recognize?

Good Data Practices!

If you were not a part of the team collecting the data and no codebook explaining the variables was provided, then you should go back to the team who collected it and ask questions to better understand each of the variables and the context of the data.

(Explanations about the context of these datasets are in the Teacher Guide.)

TASK 4: Data Wrangling & Cleaning

Raw datasets will have to be cleaned up so that they can be analyzed. Both of your datasets have already been cleaned.

1. Look at the data in both **tabs** (at the bottom of sheet) and, using the information on the following slides, discuss as a group specific ways this data has been tidied and cleaned.
2. Identify any additional things you could do to the data to make it easier to analyze.

Practicing data wrangling and cleaning

Key Idea: Understand how to clean up data and make important decisions during that process and how those decisions can affect analysis

Data Wrangling - formatting and structuring raw data into a more useable form

Tidy Data - organizing data so that it can be easily analyzed

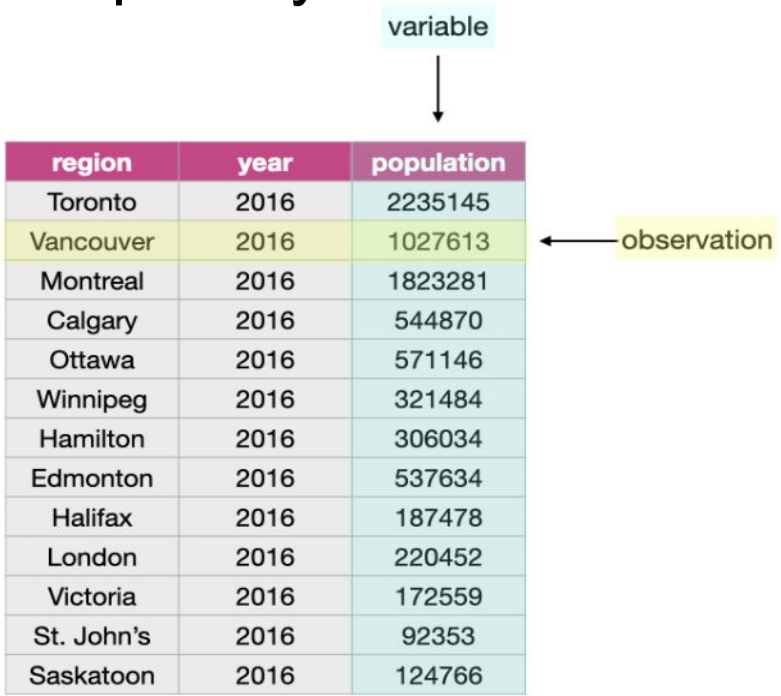
Data Cleaning - fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset

Data Wrangling - Technical Terms

variable - a characteristic, number, or quantity that can be measured.

observation - all of the measurements for a given entity.

value - a single measurement of a single variable for a given entity.



region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484
Hamilton	2016	306034
Edmonton	2016	537634
Halifax	2016	187478
London	2016	220452
Victoria	2016	172559
St. John's	2016	92353
Saskatoon	2016	124766

Source: Data Science: A First Introduction

Data Wrangling - Tidy Data

- Each row is a single observation,
- Each column is a single variable, and
- Each value is a single cell

rows = observations

region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484

columns = variables

region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484

cells = values

region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484

Source: Data Science: A First Introduction

Data Wrangling - Cleaning Data

1) Hide/Remove duplicate or irrelevant observations
    - Irrelevant observations are those that do not fit into the specific problem you are trying to analyze.
2) Fix structural errors
    - Strange naming conventions, typos, or incorrect capitalization
3) Filter unwanted outliers
    - Just because an outlier exists, doesn't mean it is incorrect!
    - Determine the validity and relevance of the outlier before deciding to delete or hide

Source: Tableau Guide To Data Cleaning

Data Wrangling - Cleaning Data

4) Handle missing data
   - Many programs will not accept missing values
   - Options:
     a) Drop observations that have missing values, but then you will lose information
     b) Input missing values based on other observations, but you will lose integrity of the data because you may be operating from assumptions and not actual observations
     c) Alter the way the data is used to effectively navigate null values

Source: Tableau Guide To Data Cleaning

Good Data Practices!

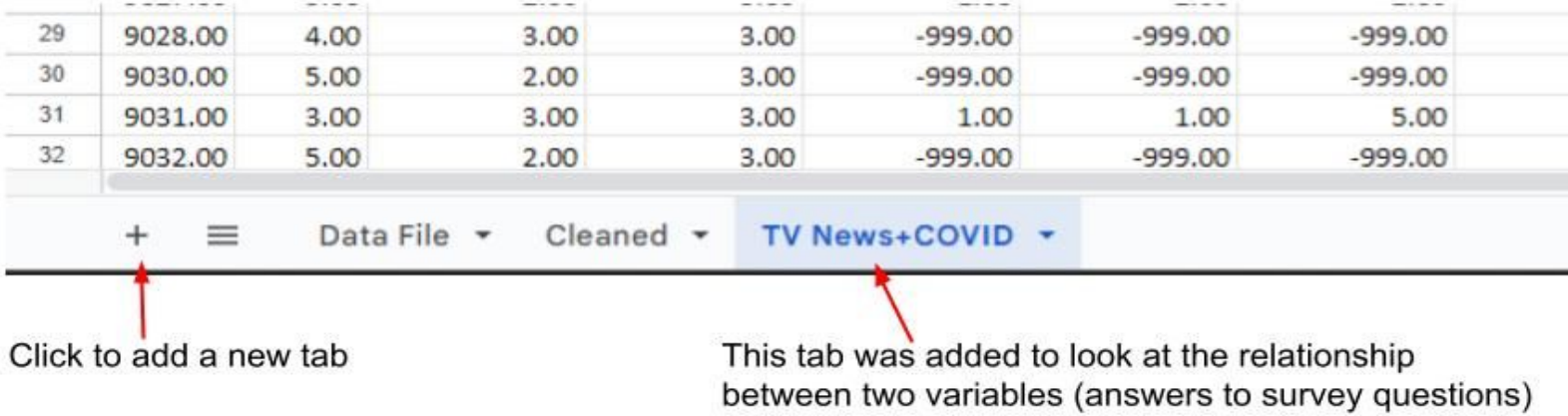
Before making any changes to the data, such as deleting observations that have missing values, first consult with those who collected the data …

a. To gain a better understanding of why values are missing
b. To better understand the importance of those observations to the overall data analysis

TASK 5: Data Analysis

When analyzing data, each time you want to make changes or pull out certain variables, you should work in a new tab so that you don't lose the original data.

1. At the bottom, click the + and add a new tab.



Good Data Practices!

A good data science practitioner ALWAYS keeps a copy of the original dataset AND tracks every step made during the data wrangling and cleaning process.

- You never want to lose your original data AND you want to record each change made because…
  a. Mistakes happen
  b. You may need to reference the original data
  c. It is important to be transparent about how you manage and analyze your data

3. Select the data from the “Cleaned” tab that you want to work with and paste it into the new tab. Label that tab.

   ** Be sure to also select the CASEID column and copy it into each new tab

4. In the new tab change the column headings to a short description of the question.

A	B	C	D	E	F	G
CASEID	Q111: Attn on TV news to govt+pol	Q110.0: Attn on TV news to sci+tech	Q111: Attn on TV news to health+med	Q118: COVID-19 symptoms are not any worse than the flu	Q120: Not really worried about COVID-19	Q121: Think COVID will harm you personally
9001.00	4.00	2.00	3.00	4.00	3.00	2.00
9002.00	3.00	3.00	3.00	-999.00	-999.00	-999.00
9003.00	-999.00	-999.00	-999.00	-999.00	-999.00	-999.00
9004.00	5.00	5.00	4.00	2.00	2.00	2.00

TASK 5: Data Analysis

Now you are ready to analyze!

Think about what types of trends or possible correlations you want to investigate. What do you need to do with the data first?

2. Select the data that you want to work with from the “Cleaned” tab and paste it into the new tab. Label that tab.

** Remember to include the CASEID column in each new tab

3. Decide how you want to manipulate the data.
   a. Ex: Find the median or mode for each response type.
   b. Calculate totals for each response option (how many number 1’s, how many number 2’s, etc.)

Good Data Practices!

You should always track your processing steps.

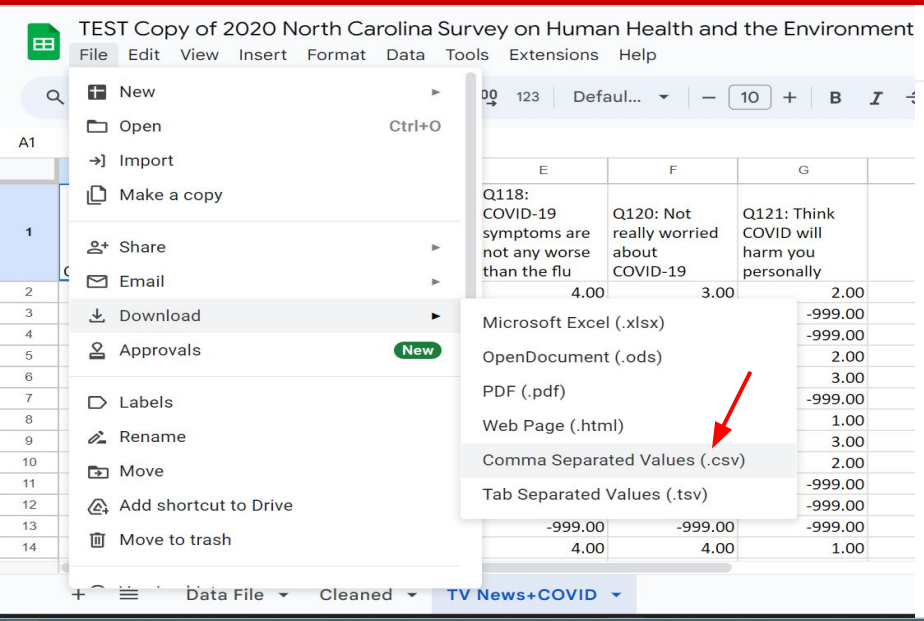
Option 1: Create a new tab and record each step you take in processing the data (such as when you hide or delete **variables/observations**.)

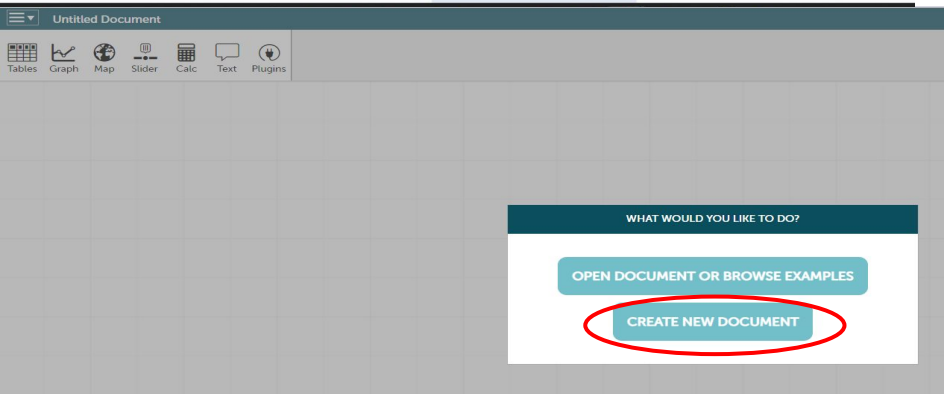
Option 2: Make a new tab each time you make changes to analyze the data in a different way.

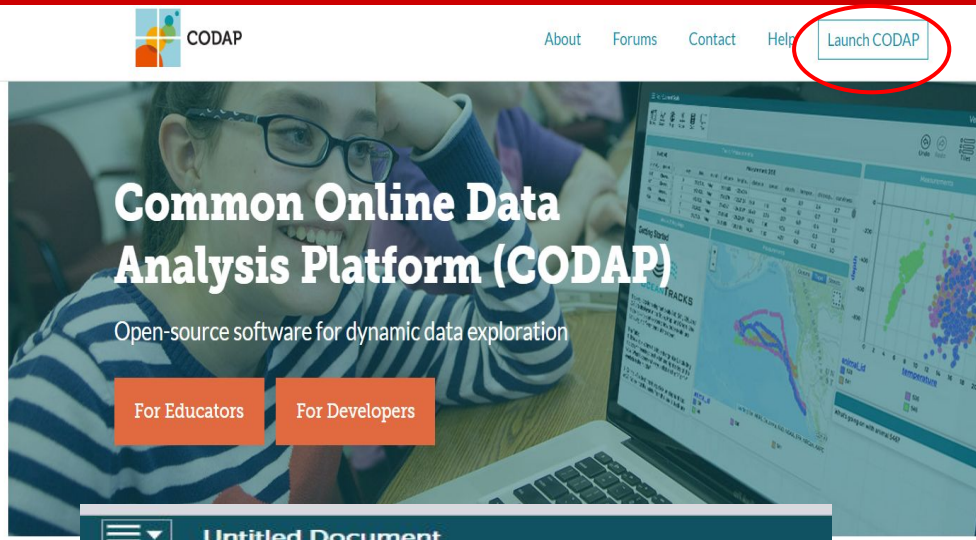
a. Decide on a naming convention so that the steps are clear when referred back to.
b. For each new tab add a note beneath the new data that describes what was done from the previous step.

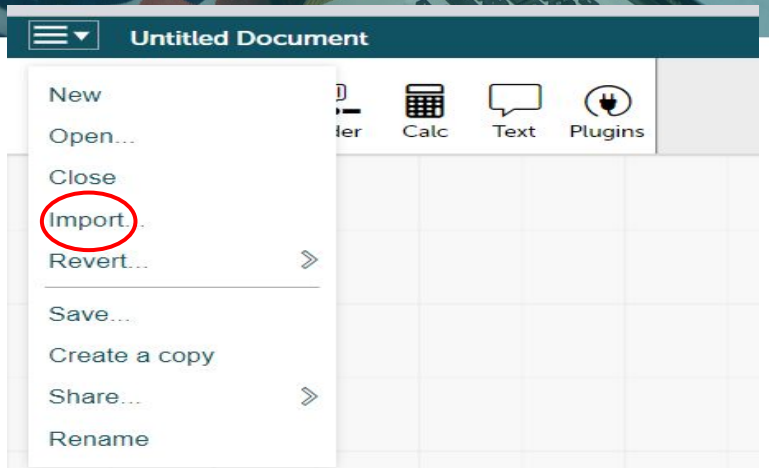
TASK 5: Data Analysis

- When manipulating your data, such as counting entries or calculating averages, you should work in a spreadsheet, such as Excel or Sheets.
  - Just remember to make any changes in a new tab
- You can then visualize the data either in Excel/Sheets or you can upload the data that you want to visualize into CODAP.
  - ➢ First download the data tab you want to work with as a CSV file. Then **IMPORT** that file into CODAP


TEST Copy of 2020 North Carolina Survey on Human Health and the Environment
File Edit View Insert Format Data Tools Extensions Help
New
Open Ctrl+O
Import
Make a copy
Share
Email
Download
Approvals New
Labels
Rename
Move
Add shortcut to Drive
Move to trash
Microsoft Excel (.xlsx)
OpenDocument (.ods)
PDF (.pdf)
Web Page (.html)
Comma Separated Values (.csv)
Tab Separated Values (.tsv)
Q118: COVID-19 symptoms are not any worse than the flu
Q120: Not really worried about COVID-19
Q121: Think COVID will harm you personally
Data File
Cleaned
TV News+COVID


Untitled Document
Tables Graph Map Slider Calc Text Plugins
WHAT WOULD YOU LIKE TO DO?
OPEN DOCUMENT OR BROWSE EXAMPLES
CREATE NEW DOCUMENT


CODAP
About
Forums
Contact
Help
Launch CODAP
Common Online Data Analysis Platform (CODAP)
Open-source software for dynamic data exploration
For Educators
For Developers


Untitled Document
Calc
Text
Plugins
New
Open...
Close
Import...
Revert...
Save...
Create a copy
Share...
Rename

TASK 5: Data Analysis - Visualizations

- What do you see? What do you wonder?
  - Explore the data, look for trends or patterns, and ask questions.
  - Exploring the data through a variety of visualizations may reveal different patterns, interpretations, or insights.

★ *Note: Data might be viewed in different ways, but the data point never changes or disappears.*

Breakdown into Smaller Datasets

The large dataset has been broken down into sections:

2020 COVID

2020 PFAS

2021 COVID

2021 Heart & Lung Health

TASK 6: Communicating Your Findings

1) What interesting trends or patterns did you find?
2) What impacts could those trends have on certain communities or populations?

- Your communication piece should include both written explanations and graphic visualizations.
- Visualizations should be interesting enough to grab the reader's attention, but clear enough to understand the trends you are trying to convey.

TASK 6: Communication - What do we do with the insights we have learned from the data?

3) What decisions can be made based on your findings?
   a) What additional information would you need first?
   b) Who is not included in the data?
   c) How can this information be attained in the future?
4) What questions are there that the given dataset can't answer or can only partially answer?

★ It is good to acknowledge that the data is not perfect but can still be used for modeling

TASK 6: Communication - What do we do with the insights we have learned from the data?

- You can discuss and argue certain ideas, while realizing that datasets cannot provide pure answers

★ Remember:

Data is information - not truth - with error, variability, and degrees of inclusion/exclusion

TASK 7: Data for Social Good

1) Can this data reveal any disparities in different geographic or social groups?
   - ➢ If not, what further information do you need?
2) How can we use this data to make decisions that can affect the health and well-being of communities? (Give specific ideas.)
3) How can we use this data to make decisions about how to improve public spaces, access to information, or access to services?
4) How can we use this data to make decisions that can affect the health and well-being of communities, including public policies?