GETTING THE MOST OUT OF YOUR DATA

Infectious Disease Epidemiology Bootcamp

Session 4

July 28, 2020

Meghan Warren, Brettania O'Connor, Ricky Camplain

NAU NORTHERN ARIZONA UNIVERSITY

INFECTIOUS DISEASE EPIDEMIOLOGY BOOTCAMPS OBJECTIVES

1. Explain the basics of infectious disease epidemiology, including transmission and conceptual model
2. Evaluate infectious disease measures (e.g., R-naught, case fatality, incidence).
3. Explain the importance of control infectious disease spread
4. Describe the process of testing, case investigation, and contact tracing for infectious diseases
5. Compare sensitivity, specificity, and positive and negative predictive value of diagnostic tests
6. Understand the concepts of database construction and data entry for quality data reporting
7. Interpret data tables and charts related to infectious disease measures

INFECTIOUS DISEASE EPIDEMIOLOGY BOOTCAMPS OBJECTIVES

1. Explain the basics of infectious disease epidemiology, including transmission and conceptual model
2. Evaluate infectious disease measures (e.g., R-naught, case fatality, incidence).
3. Explain the importance of control infectious disease spread
4. Describe the process of testing, case investigation, and contact tracing for infectious diseases
5. Compare sensitivity, specificity, and positive and negative predictive value of diagnostic tests
6. **Understand the concepts of database construction and data entry for quality data reporting**
7. **Interpret data tables and charts related to infectious disease measures**

DISCUSSION QUESTION 1

What issues do you have with your health data?

DISCUSSION QUESTION 2

What could be improved with your data collection, management, and analysis process?

IF YOU HAVE ADDITIONAL QUESTIONS

- Use the chat function
  - **We want to more hear from you ☺**
- Questions will be answered at the end during a discussion period in the order they come in

DATA ISSUES

Meghan Warren, PT, MPH, PhD

NAU NORTHERN ARIZONA UNIVERSITY

DATA

- Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

https://www.merriam-webster.com/dictionary/data

- Data are measured, collected and reported, and analyzed, whereupon it can be visualized using graphs, images or other analysis tools

- Forms of data:
  - **XML**
  - **CSV**
  - **TAB**

WORKING WITH DATA FROM OTHER PEOPLE

Pros

1. You do not have to collect the data
2. Projects can be completed more quickly
   - **Sometimes**

Cons

1. You do not know ***how*** the data were collected
2. You cannot control the database set-up
3. Data are not always collected for research/program evaluation

Issues may be OK when the data are small numbers
But increasingly become untenable

A FEW ‘TYPOS’ (N = 10)

	A	B	C
1	ID	Age	Sex
2	1	22	M
3	2	47	M
4	3	34	M
5	4	5.5	F
6	5	42	F
7	6	65	Male
8	7	87	M
9	8	0.22	F
10	9	97	F
11	10	75	F
12			

Are about to see

NAU NORTHERN ARIZONA UNIVERSITY

FUN WITH DATES (N ≃ 750)

03/13/2020	03/13/2020		
3/26/2020			
03/25/2020	03/26/2020		03/25/2020
3/22/20	n/a	n/a	3/22/20
03/21/2020	03/21/2020		
03/21/2020	03/21/2020		
03/15/2020			
3/25/20	3/25/20		
3/22	3/28		
03/16/2020	03/16/2020		03/16/2020
3/28/2020	3/28/2020		3/28/2020
	03/27/2020		03/27/2020
3/15/2020	3/19/2020	3/8/2020	3/19/2020
03/28/2020	03/29/2020	03/29/2020	
3/28/2020	3/30/2020	3/30/2020	
March 25th, 2020	March 25th, 2020		March 25th, 2020
	3/23/2020		3/23/2020
March 28th, 2020	March 31st, 2020		March 31st, 2020
03/27/2020	03/28/2020		
03/31/2020	03/31/2020	03/31/2020	
Yes, 3/30/2020	Yes, 4/2/2020		
03/17/2020	03/18/2020		

NAU NORTHERN ARIZONA UNIVERSITY

WHAT DOES THIS MEAN? (N = TOO MANY TIMES TO COUNT)

T	U	V	W	X	Y
X1	X2	X3	X4	X5	X6
Yes	Yes	Yes	F	No	1
No	No	No	M	No	1
Yes	Yes	Yes	F		1
Unsure	Unsure	Yes	M	No	1
No	No	No	M	No	2
No	No	No	V	No	2
Yes	Yes	No	M	No	2
No	No	No	M	No	2
Unsure	No	No	M	No	2
No	No	No	M	No	2
Unsure	Unsure	No	V	No	2
No	No	No	M	No	2
Yes	Yes	No	M	No	2
No	No	No	M	No	2
No	No	No	M	No	2
No	No	No	M	No	2
No	Yes	Yes	F	No	1
No	No	No	F	Yes	2
No	No	No	M	No	2
No	No	No	M	No	2
No	No	No	V	No	2
No	No	No	M	No	2
No	No	No	M	No	2
Yes	Yes	No	M	No	2

NAU NORTHERN ARIZONA UNIVERSITY

WHERE'D THE DATA GO? (N ≃ 2,200)

AG	AH	
Ext_ROM_1st_PT_visit	Flex_ROM_1st_PT_visit	D
4	90	
2	92	
1	97	
4	88	
7	105	
1	99	
6	95	
3	93	
8	84	
10	100	
12	20	
0	110	
6	85	

The SAS System

The MEANS Procedure

Analysis Variable : Flex_ROM_1st_PT_visit Flex_ROM_1st_PT_visit				
N	Mean	Std Dev	Minimum	Maximum
317	91.4731861	13.5563781	0	125.0000000

Level 4;
Nonmotorized Equipment;
No Equipment Needed;
No Equipment Needed;
Level 4;
Level 3;
Level 1;
No Equipment Needed;
Level 4;
Total Body Lift Equipment;
Level 1;
Level 1;
Level 1;
Level 1;
Level 1;
Level 4;
Level 3;
Total Body Lift Equipment;
Level 4;
Level 3;

SYSTEMS CHANGE (N ≃ 55,000)

Electronic health records changed from company A to company B

COLLECTING YOUR OWN DATA

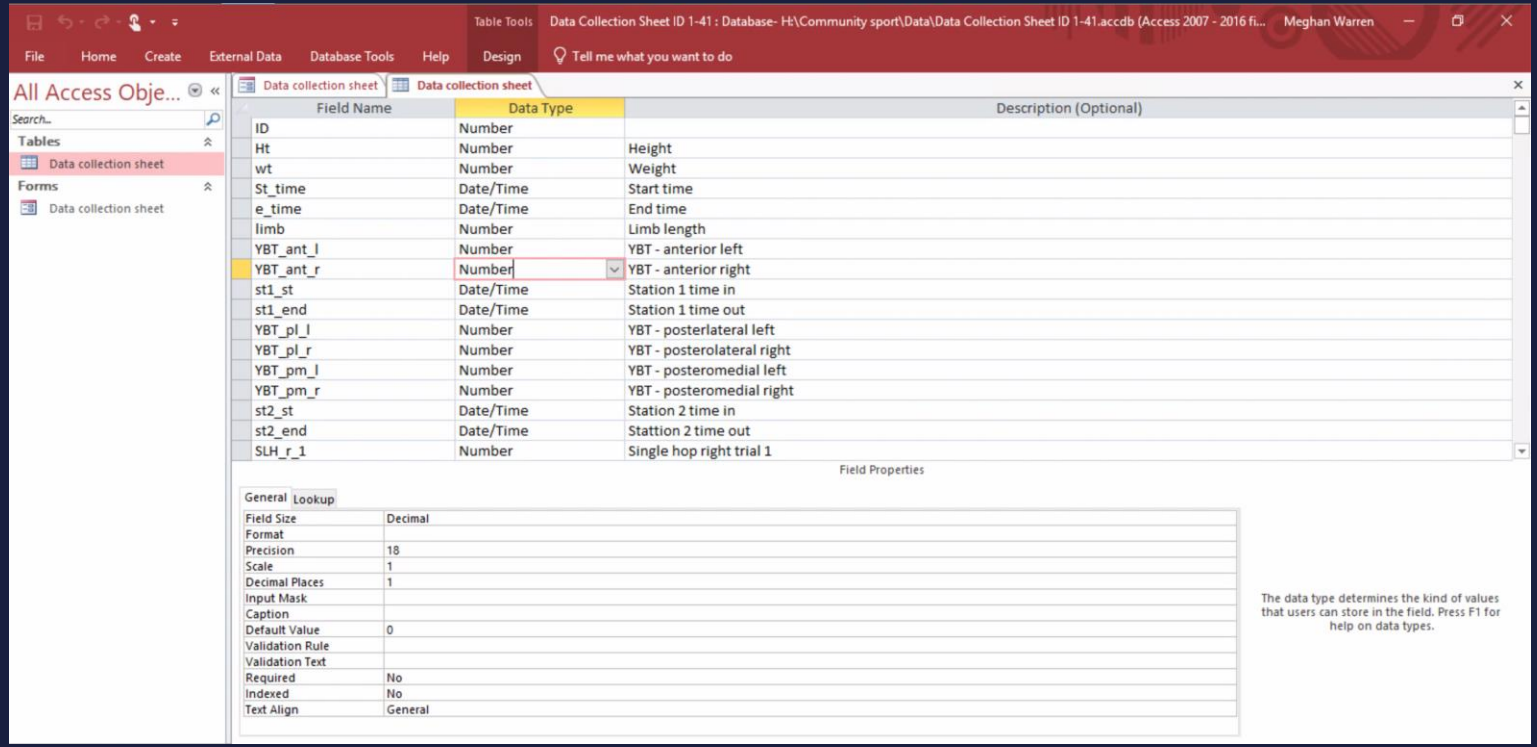
Pros

1. Control, control, control
2. Can set up database/data entry
   - **Build in quality checks**

Cons

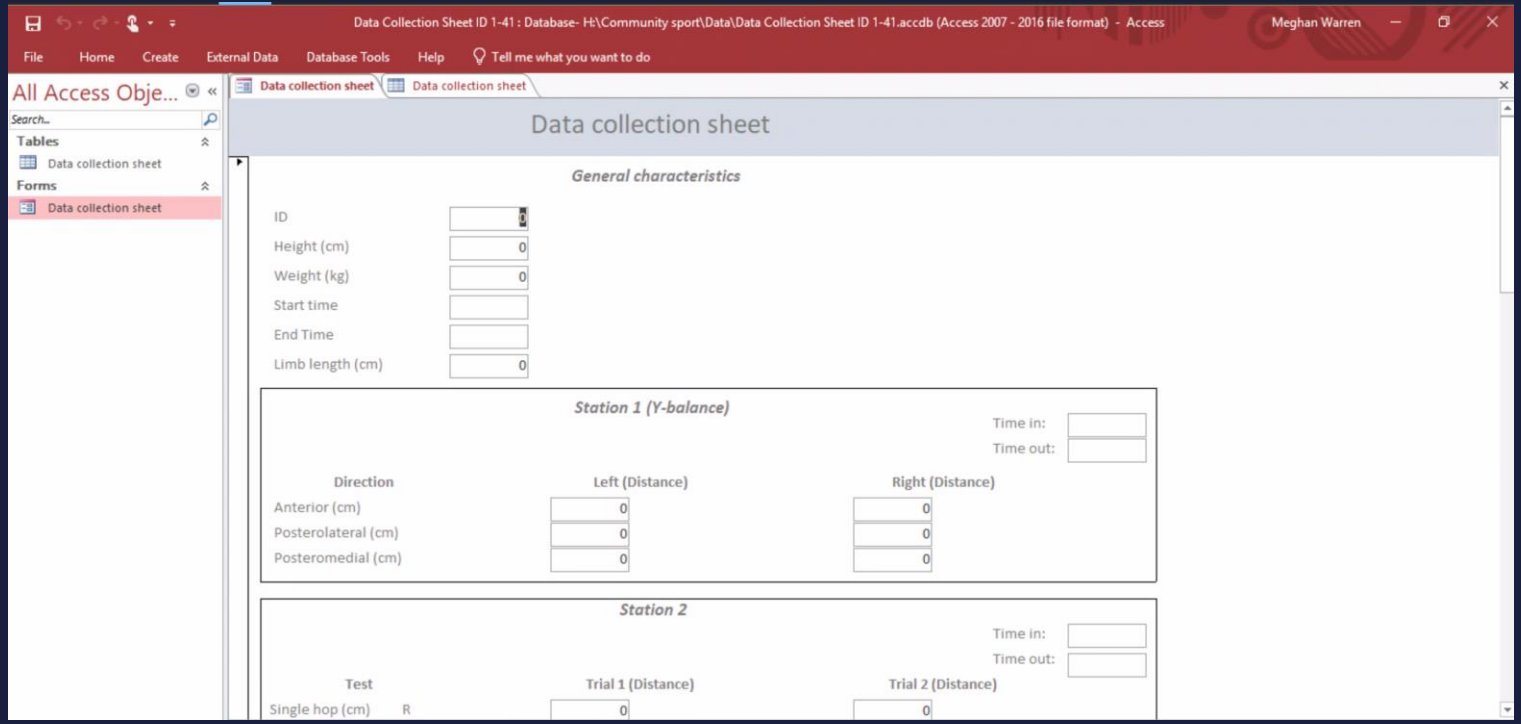
1. Time
2. Money
3. Expertise

DEFINE VARIABLES AT THE BEGINNING



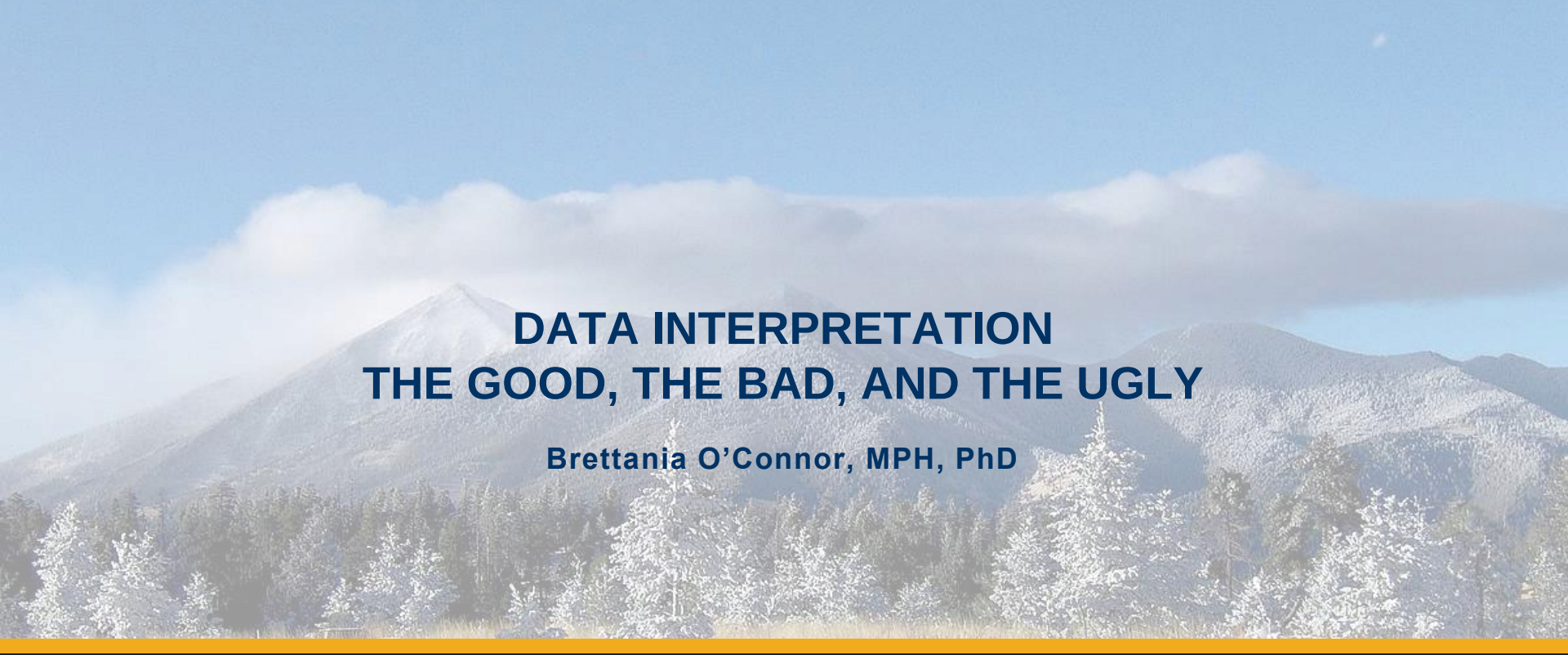
NAU NORTHERN ARIZONA UNIVERSITY

MAKE DATA ENTRY EASIER



NAU NORTHERN ARIZONA UNIVERSITY

NAU NORTHERN ARIZONA UNIVERSITY



DATA INTERPRETATION
THE GOOD, THE BAD, AND THE UGLY

Brettania O'Connor, MPH, PhD

NAU NORTHERN ARIZONA UNIVERSITY

Some possible Problems with data presentation

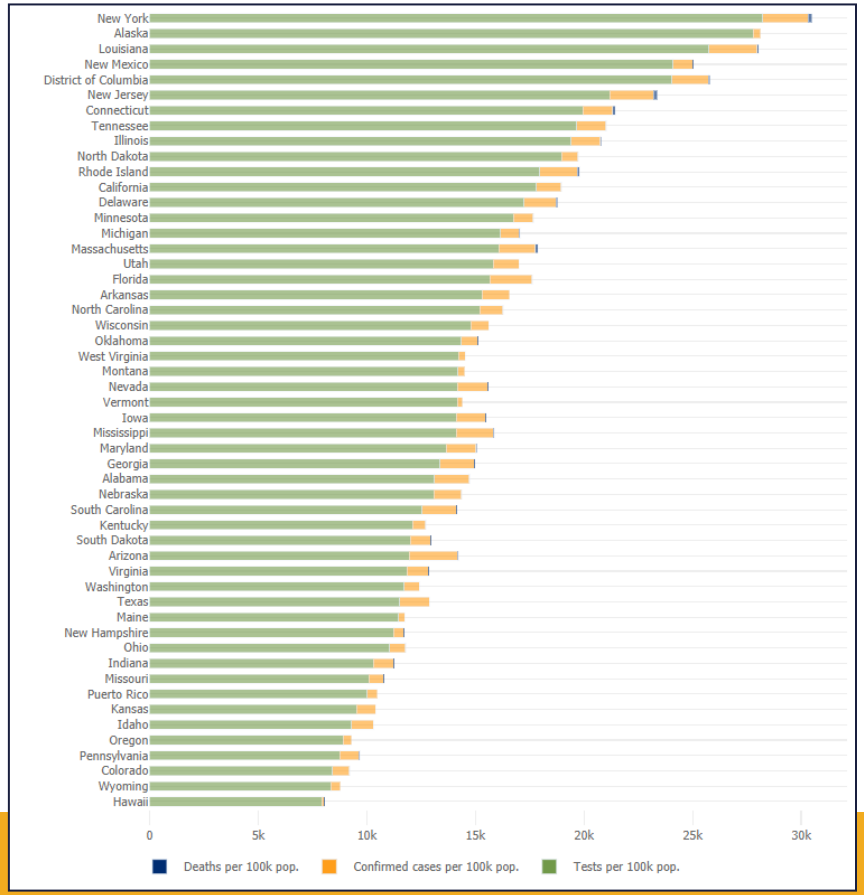
Some things that can go wrong when data are presented:

- *Lack of* titles and labels with units for graphs and tables
- *Inaccurate* titles and/or labels and/or units for graphs and tables
- Not presenting all graphs on the same numeric scale (without making this clear)
- Changing the orientation of the graphs in a way that may lead the viewer to come to the wrong conclusion
- Not presenting data per capita (where appropriate) and instead just comparing count data for different population sizes
- Showing trends over time in a way that is difficult for the viewer to process

WHY WOULD DATA BE PRESENTED IN A CONFUSING OR INACCURATE WAY?

- Misunderstandings related to the data and/or how to present it
- Lack of training in data analysis and presentation
- Honest mistakes that were not noticed (e.g. software program automatically colored maps according to the data distribution)
- Manipulation of data with the intent to mislead the viewer

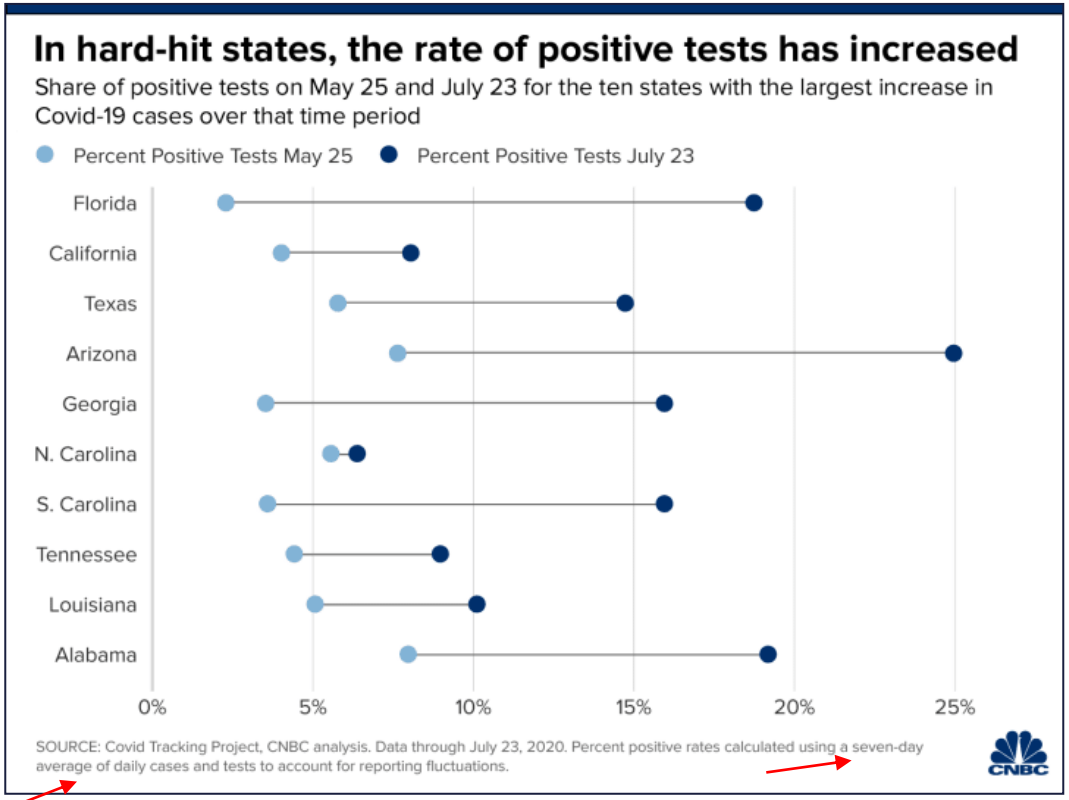
CASES, DEATHS AND TESTING BY STATE, PER 100,000 PEOPLE



These data are presented per 100,000 people living in each state.

Source: https://coronavirus.jhu.edu/testing/states-comparison
Data updated July 27, 2020

Test positivity, May 25th versus July 23rd



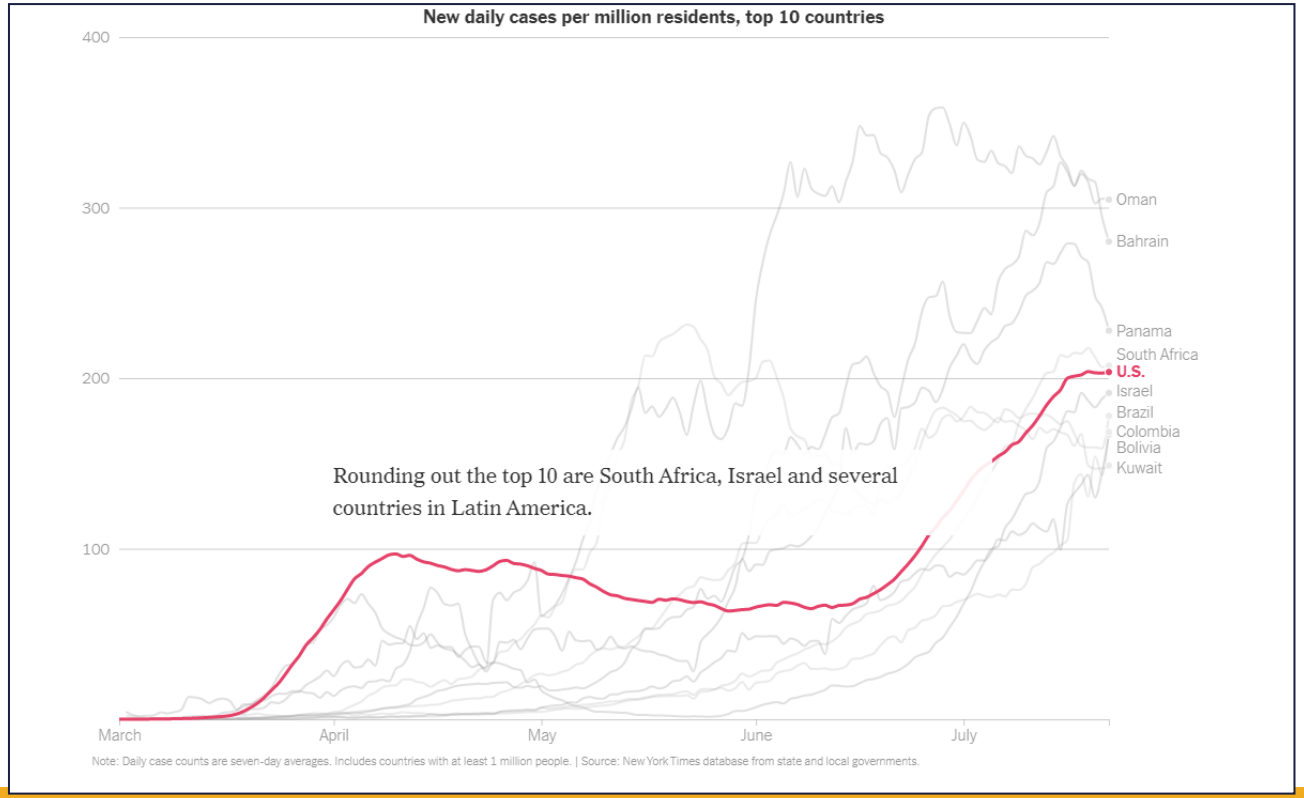
These data are not actually rates.

The testing positivity refers to the # of positive tests / the total # tests done in a state.

This graphic shows changes over time.

Source: https://www.cnbc.com/2020/07/25/coronavirus-outbreak-cant-be-attributed-to-expanded-testing-capacity.html Data through July 23, 2020

CAN YOU INTERPRET THIS CORRECTLY?



New daily cases per million residents

Source: https://www.nytimes.com/interactive/2020/07/23/us/coronavirus-hotspots-countries.html Data through July 24

WHAT IS MISSING HERE?

Here we just have count data.

How do we decide if these counts indicate a large number of cases, hospitalizations and deaths?

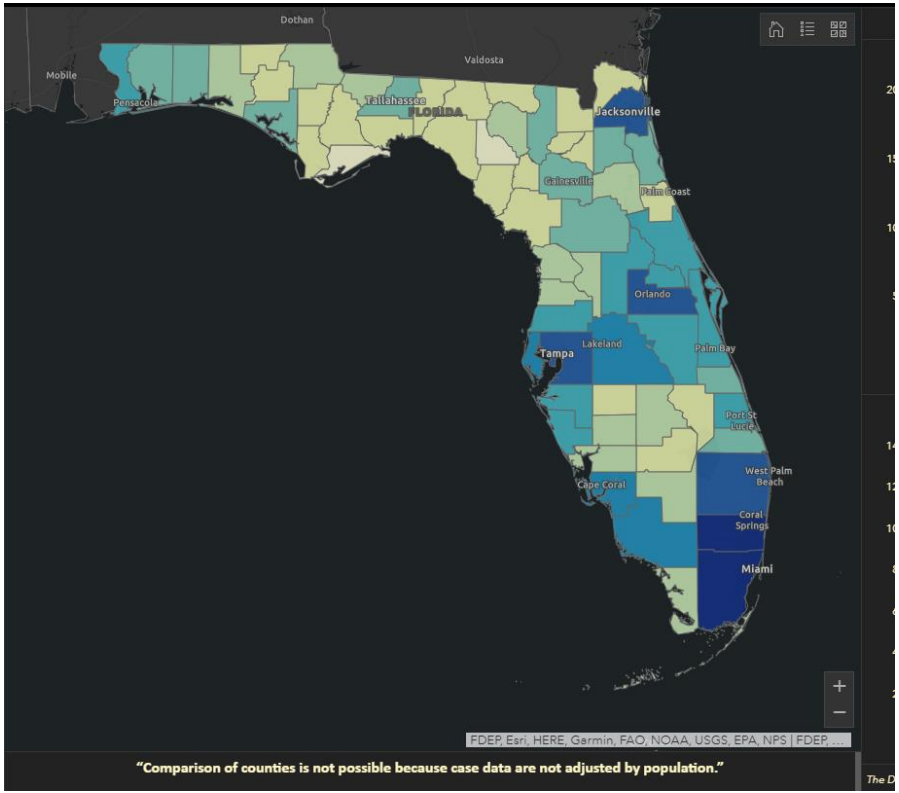
How do we compare these data with data from other states?

What do we need to know to put the data into context?

We need to know the population size.

Source: Florida Department of Health
https://experience.arcgis.com/experience/96dd742462124fa0b38ddedb9b25e429
Data through July 27, 2020

WHAT IS MISSING HERE?



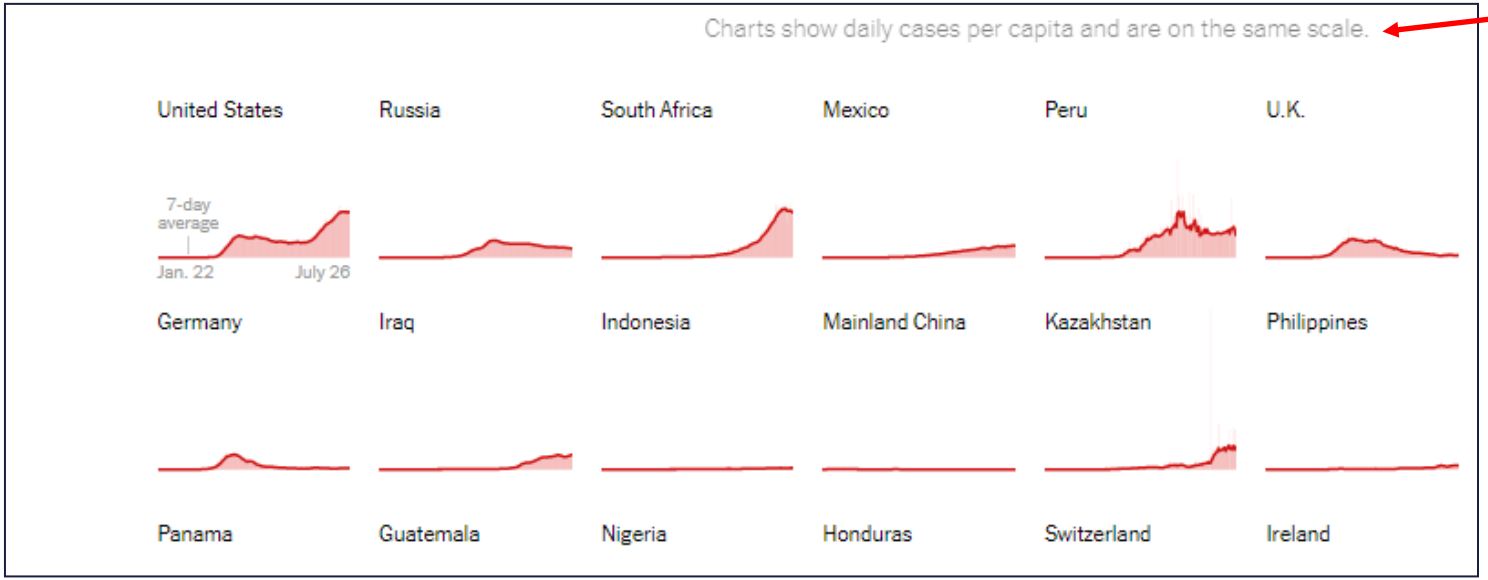
Here we have shading by county.

But we are not given a key to tell us what the shading means. We have to click on each individual county to get the total # of cases (count data)

We are told at the bottom "comparison of counties is not possible because case data are not adjusted by population"

Source: Florida Department of Health
https://experience.arcgis.com/experience/96dd742462124fa0b38ddedb9b25e429
Data through July 27, 2020

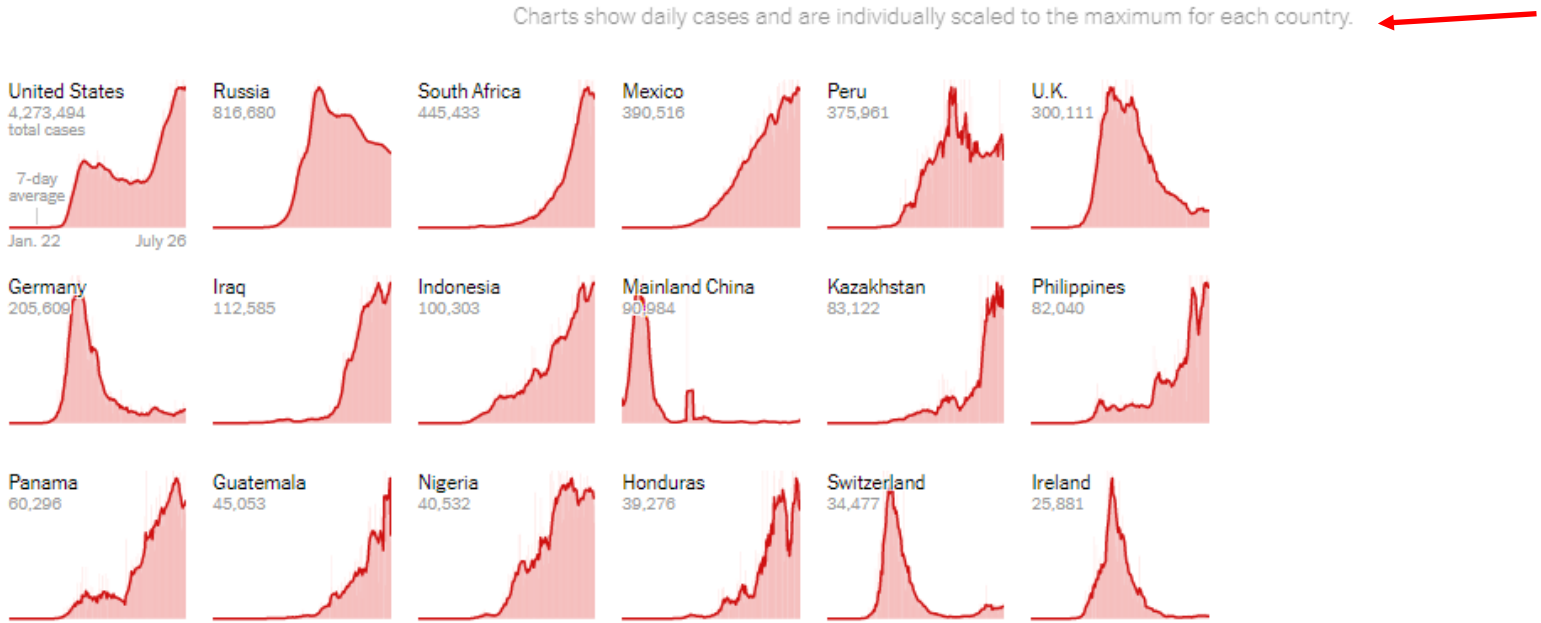
What can you tell from this graphic?



These graphs were produced on the same scale
(even though we don't see the numerical value of the scale).

Source: NYT
https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html#cases
Data through July 26, 2020

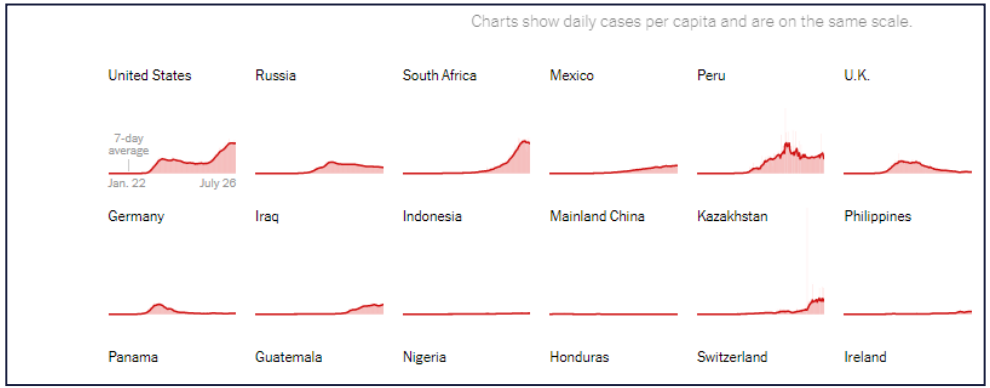
WHAT CAN YOU TELL FROM THIS GRAPHIC?



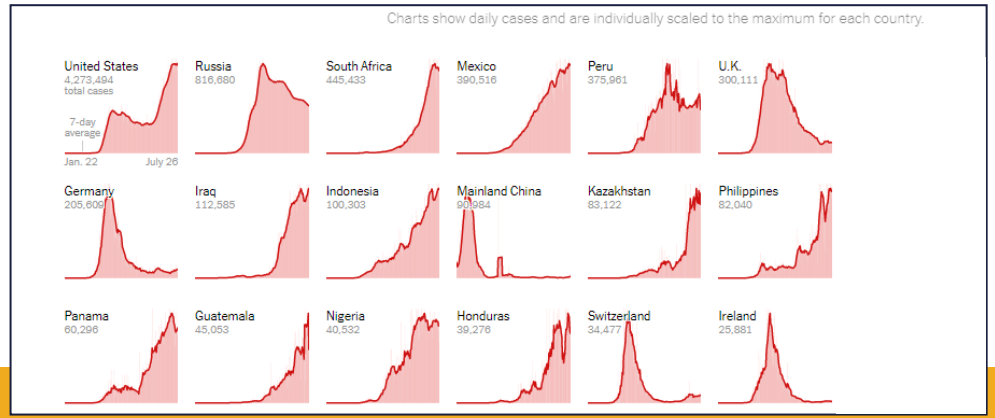
These graphs were NOT produced on the same scale
(We don't see the numerical value of the scale on the entire axis, we just see the maximum daily cases for each country).

Source: NYT
https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html#cases
Data through July 26, 2020

WHAT CAN YOU TELL FROM THIS SAMPLE OF A GRAPHIC?



Charts on the same scale and data per capita.



Charts individually scaled.

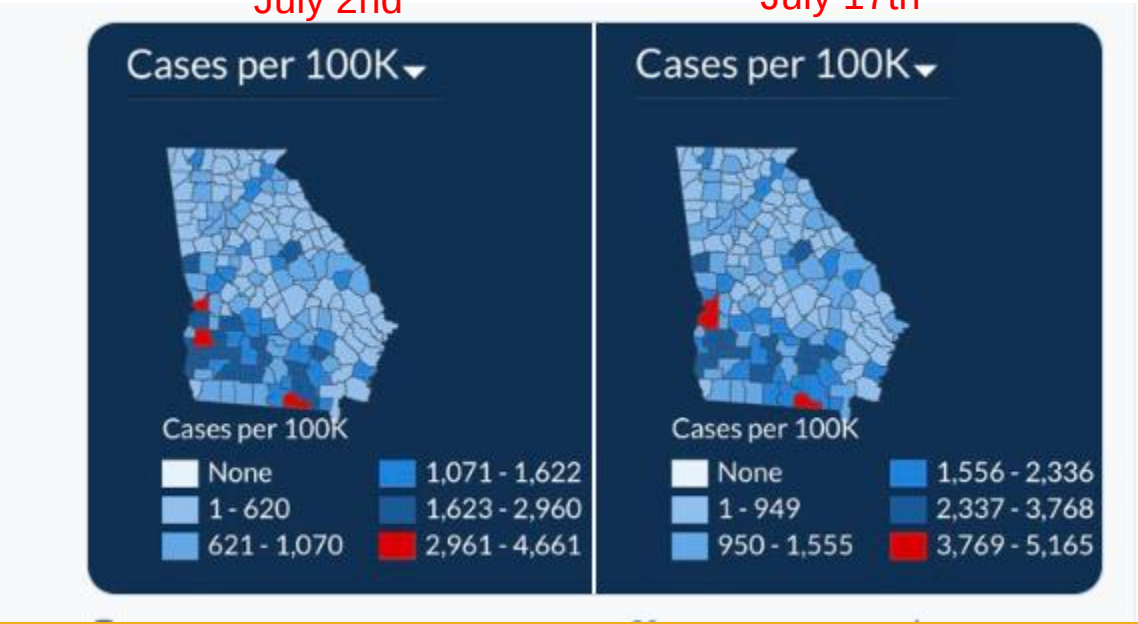
Even though the total cases are given in small print, visually the cases from many countries look as if they reached a similar maximum case count at some point in time. We also do not have data presented per capita here.

Source: NYT
https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html#cases
Data through July 26, 2020

Posted on twitter July 17th

Here we have a screenshot with state data from Georgia from 2 different dates. What do you notice about the key for the map colors?

July 2nd



July 17th

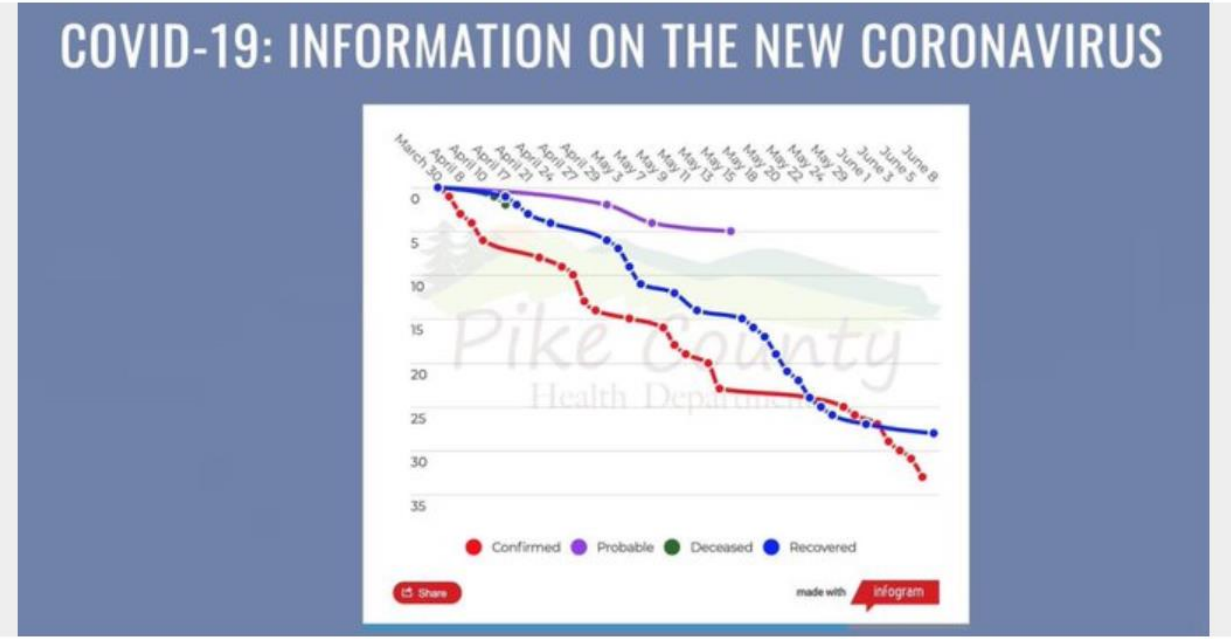
Are “issues” such as this done on purpose?

A lot of mapping software will automatically choose colored bins based on data ranges.

How could we present these data to allow people to see trends over time?

Source:https://twitter.com/andishehnouraee/status/1284237474831761408?lang=en

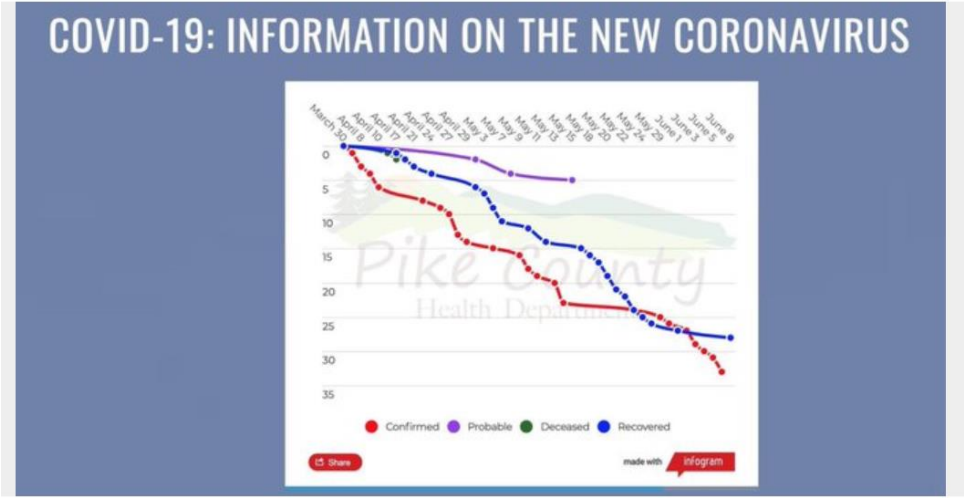
Can you interpret this correctly?



Here we see, confirmed, recovered and probable cases over time. Look carefully at the y-axis.

Source: https://policyviz.com/2020/06/16/the-inverted-vertical-axes/

CAN YOU INTERPRET THIS CORRECTLY?



Flipped version

Source: https://policyviz.com/2020/06/16/the-inverted-vertical-axes/

DATA COLLECTION, MANAGEMENT, ANALYSIS BEST PRACTICE SUMMARY

Ricky Camplain, PhD

NAU NORTHERN ARIZONA UNIVERSITY

WHY DOES FOLLOWING DATA COLLECTION BEST PRACTICES MATTER?

- Improve the usability of the data by you or others
- Your data will be “computer ready”
- Your data will be ready to share with others

DATA COLLECTION
DESIGNING THE DATA COLLECTION INSTRUMENT

- Provide clear and detailed instructions for completing the data collection instrument
  - **E.g., disease investigation, contact tracing protocols**
  - **Provide definitions and clarifying information for questions (items) and terms**

- Make questions consistent with standard definitions and conventions when appropriate and feasible
  - **In a perfect world: use questions developed nationally to ensure data is comparable**
  - **If standard definitions and/or analytic conventions are used, indicate the sources of the definitions or conventions used**

https://nces.ed.gov/pubs99/1999191.pdf

DATA ENTRY

- Use a data entry program
  - **Data are usually entered into a database program or a spreadsheet program**
- Double entry of AT LEAST some of the completed data to confirm accuracy of data entry
  - **Preferably ALL**
- Manually check 5-10% of data records

DATA ORGANIZATION

- Lines or rows of data should be complete
  - **Designed to be machine readable, not human readable**

- Include a header (variable name) on first line
  - **Should be unique**
  - **Use letters, numbers, or “_” (underscore)**
  - **Do not include blank spaces or symbols**

- Columns (responses) should include only a single kind of data with standardized formats
  - **Text or “string” data**
  - **Integer numbers**
  - **Dates**

DATA MANAGEMENT

- Refers to the entire process of record keeping
- First step: create a data management plan
- Document everything!
  - **Crete a project document file**
  - **Start at the beginning (planning) and continue through data collection and analysis**
    - Why you are collection data
    - Exact details of methods of collecting and analyzing data

DATA MANAGEMENT – CODEBOOKS

- Also knows as data dictionaries
- Describes each variable and specifies how the collected information will be entered into a computer database
- A codebook also documents how data problems like illegible handwriting and missing responses are handled

DATA MANAGEMENT – CODEBOOKS

PHQ9		Over the last 2 weeks, how often have you been bothered by the following problems? - Thoughts that you would be better off dead or of hurting yourself in some way?
N	**Value**	**Description**
	0	Not at all
	1	Several days
	2	More than half the days
	3	Nearly every day

DATA MANAGEMENT – DATA CLEANING

- Process of correcting any typographical or other errors in data files
  - **Check for out-of-range values**
  - **Check for missing or impossible values**
  - **Perform statistical summaries**

- Make notes on transformations
  - **Can be in the codebook**

- Typically after this step you create a final “analytic file” that should not be edited

DATA ANALYSIS

- Use a scripted program (R, SAS, SPSS, Matlab, etc.)
  - **Steps are recorded in a textual format**
  - **Can be easily revised and re-executed**
  - **Helps sharing and repetition**
  - **Easy to document**

- KNOW YOUR DATA BACKWARDS AND FORWARDS

- Do not “data dive” or “data mine”

- Do not “TORTURE” the data
  - **Careless or intentionally misleading data analyses are WRONG!**

EVALUATION

http://nau.co1.qualtrics.com/jfe/form/SV_4ZsJ5Gx5xZQjZtP

ACKNOWLEDGMENTS

- Lisa Dahm
- Kate Compton-Gore
- Dr. Samantha Sabo
- Dr. Julie Baldwin
- The Southwest Health Equity Research Collaborative (SHERC)
  - **nau.edu/sherc**

THANK YOU!

QUESTIONS?

NAU NORTHERN ARIZONA UNIVERSITY