Data visualizations

CIHR Course Week 5

Izzie Fulcher

Teaching Objectives

- Introduction
- Best practices
- Syndromic surveillance
- Recap

Introduction

Reasons for data visualizations

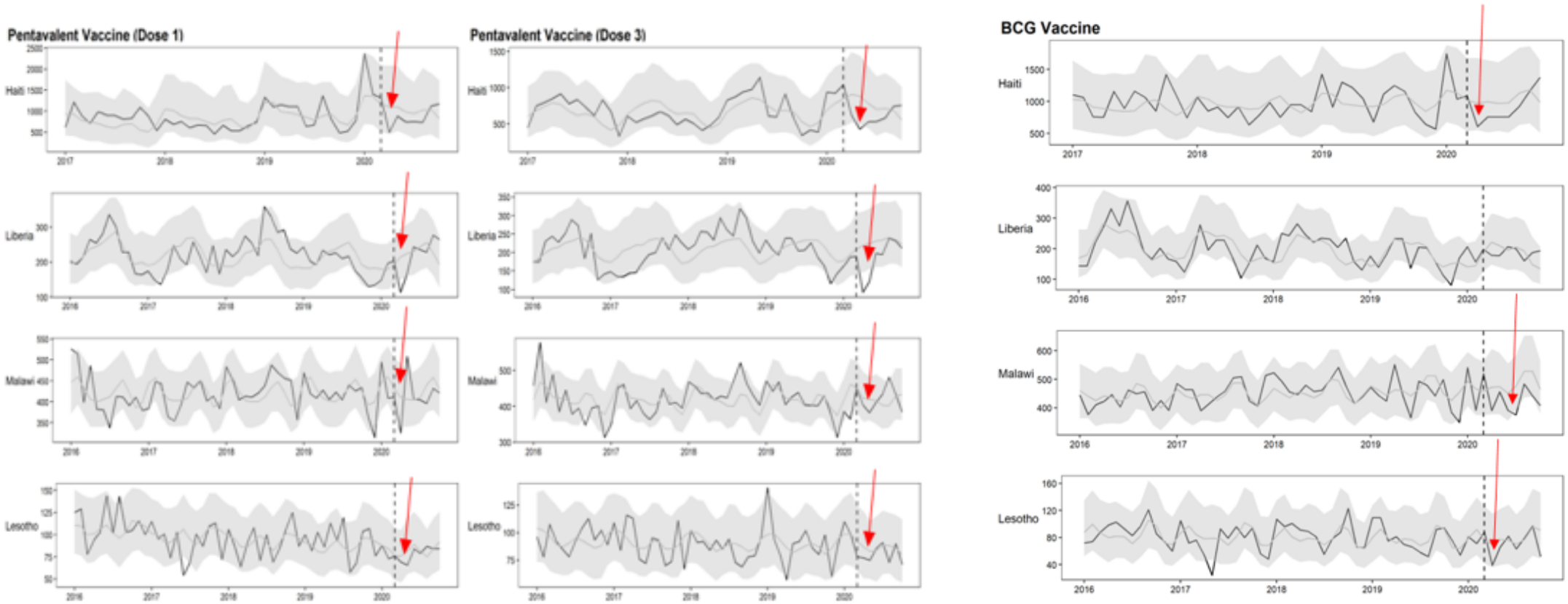
- **Exploring the data**
  - Distribution of data / transparency
  - Identify patterns, outliers, missing data
- **Analyzing the data**
  - Identifying deviations
  - Goodness of fit
  - Checking model assumption
- **Communicating results**

Communicating results

- **What is the goal?**
  - Facilitate the user's analysis
  - Communicate a specific message

- **Who is the audience?**
  - Clinicians, public health officials, or the general public
  - Keep in mind data and visualization literacy

Communicating results: specific message



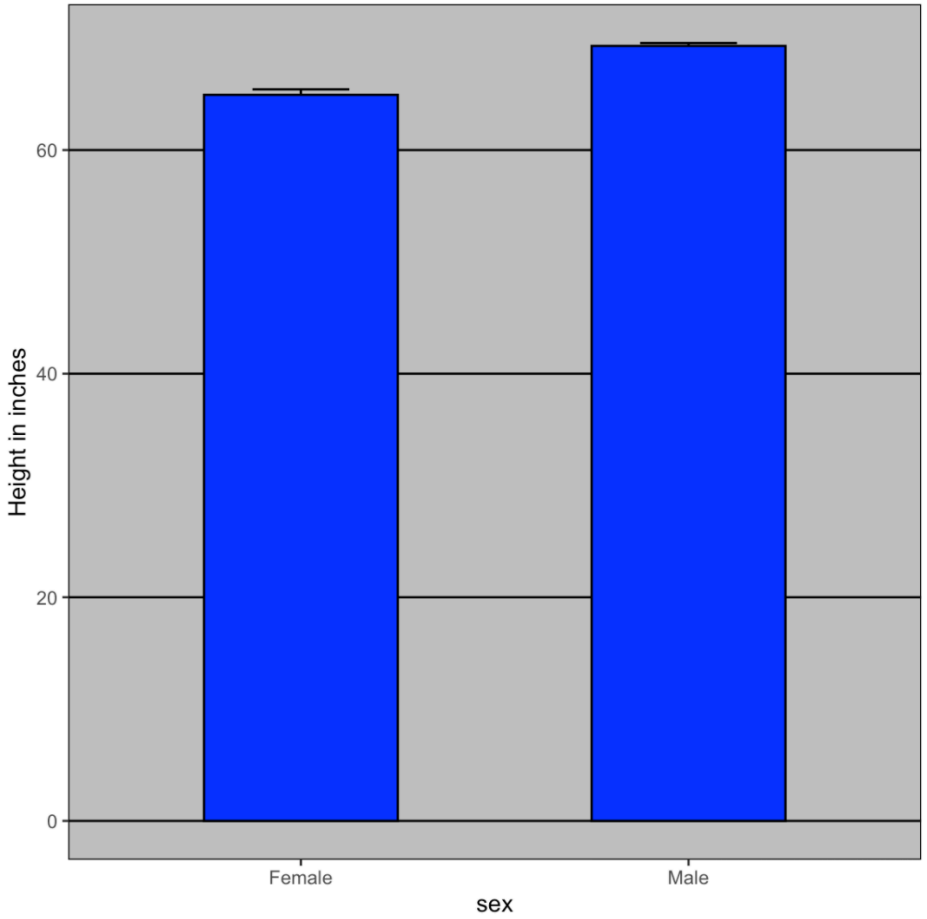
From Emma Boley and Emilia Connolly February 24, 2021 Presentation

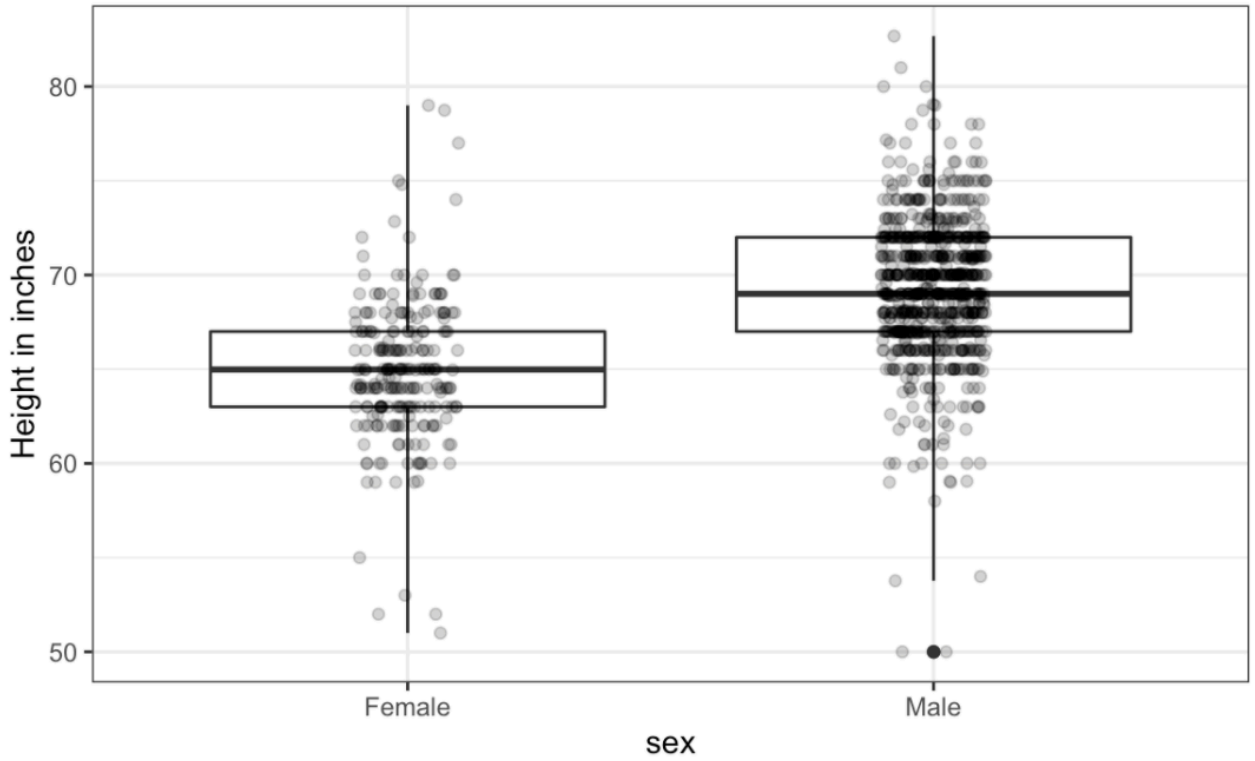
Best practices

Lots of things to consider

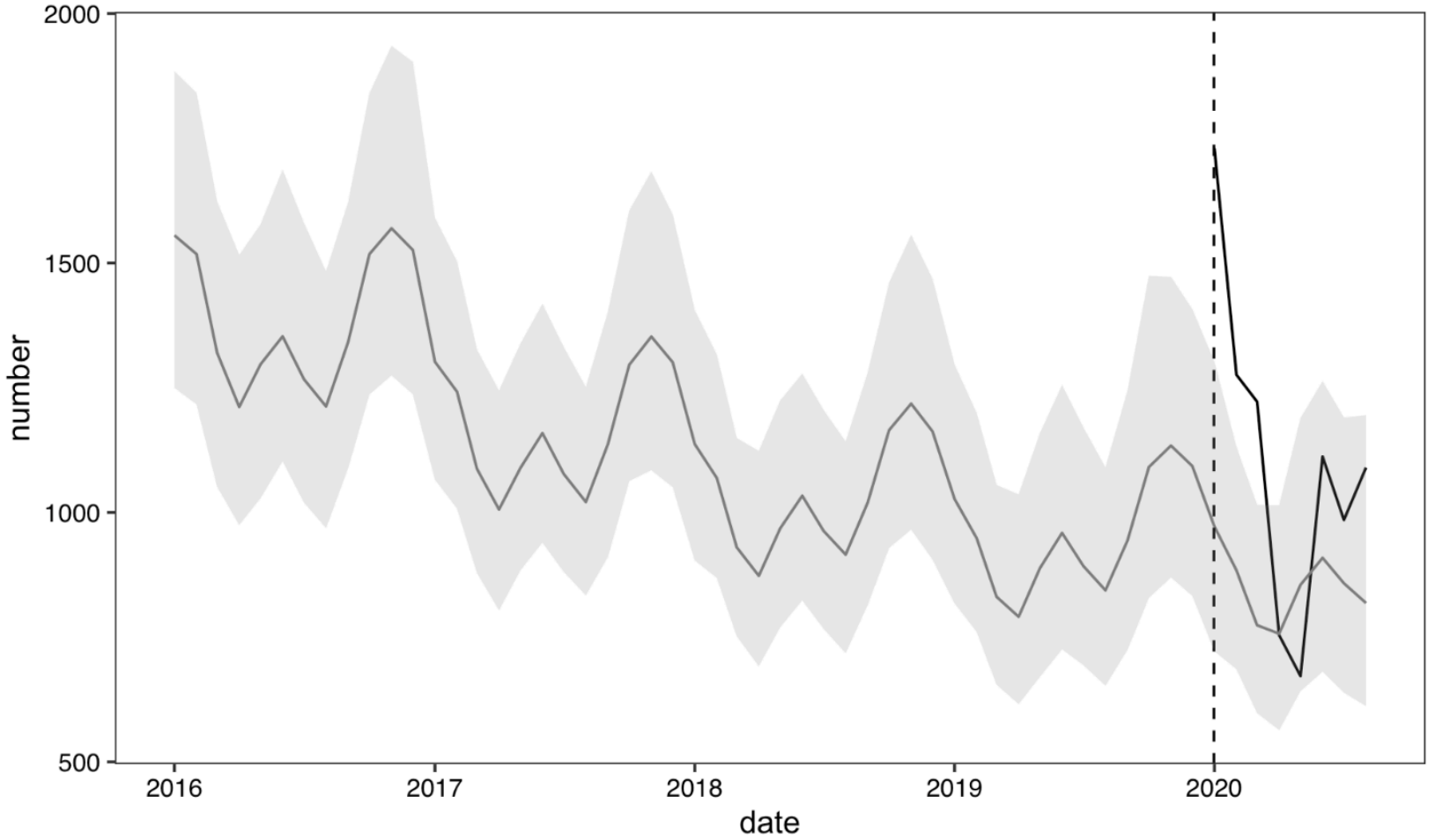
- Show the data
- Ease comparisons across groups
  - Add color
  - Put on same scale
  - Align figures vertically / horizontally
- Appropriate axis limits
  - Do not always *need* to include zero
- Descriptive labels for plot features (axis, plot title, legend)
- I really like this source (specific to ggplot in R): https://rafalab.github.io/dsbook/data-visualization-principles.html

Show the data: “let the data speak!”

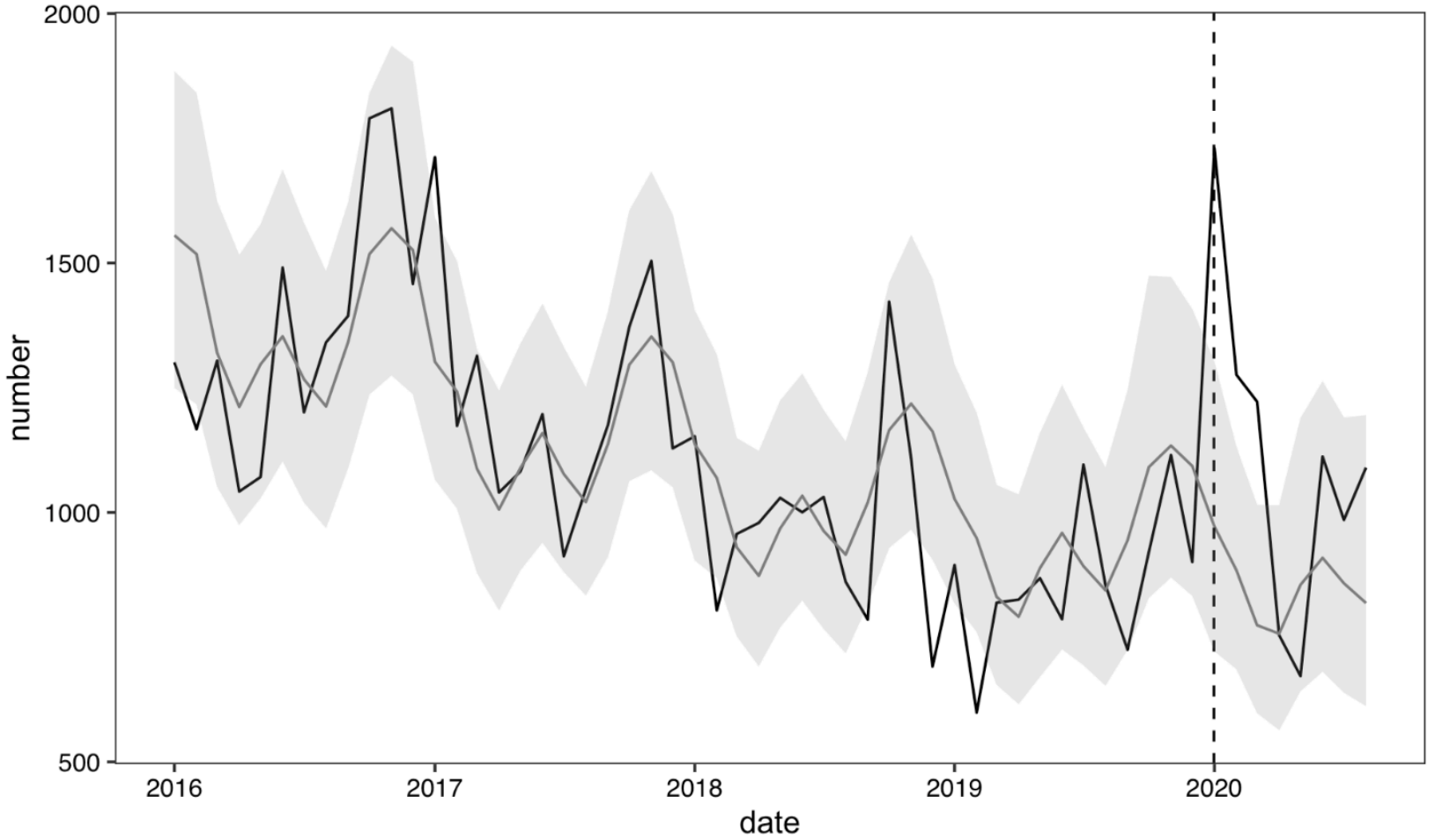




Show the data: be transparent

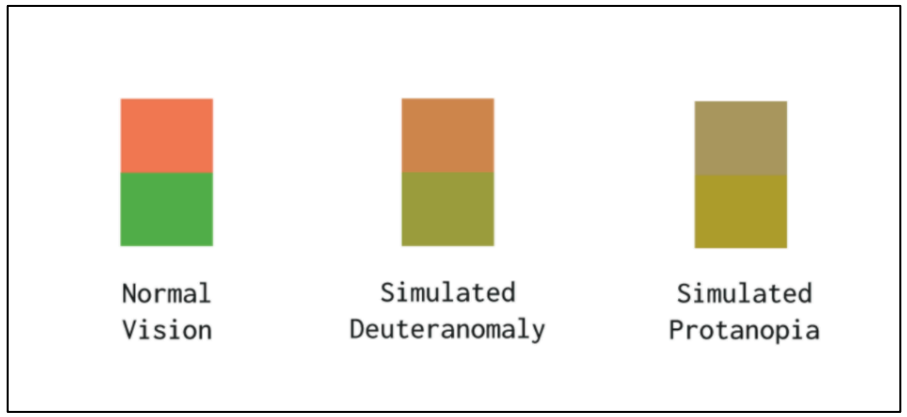


Show the data: be transparent



Choice of colors (or not!)

Consideration #1: Individuals with color vision deficiency



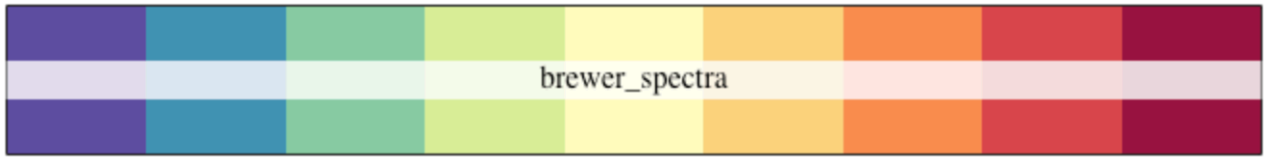
You can use a "vision simulator" to see what your plots would look like to people with color vision deficiency:

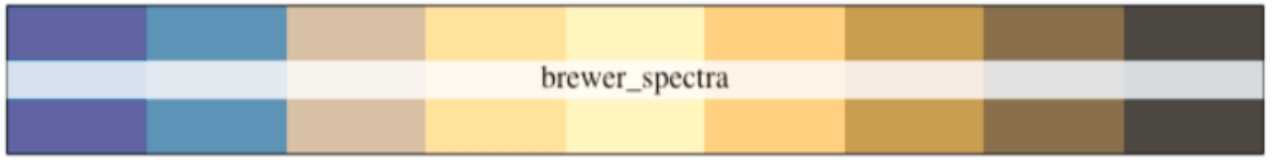
https://asada.website/webCVS/

Consideration #2: Color palettes are available online and can spruce up a figure!

Consideration #3: If you are publishing, color figures cost more.

Choice of colors (or not!)





protanopia

tritanopia

I will <u>not</u> be using this color palette anymore!

Using the https://asada.website/webCVS/

Choice of color scales

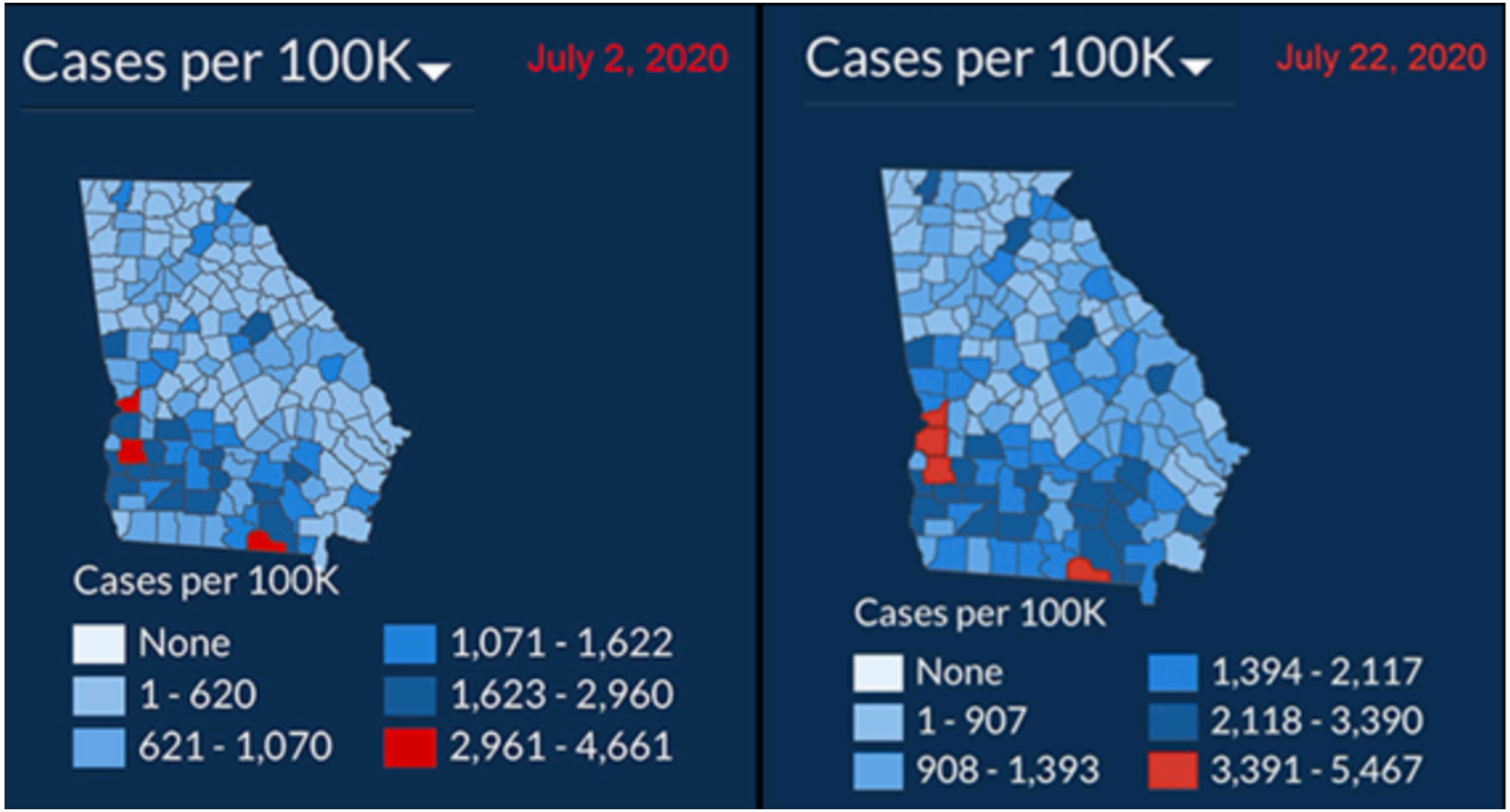
	Is there a "neutral" value with one extreme?	Is there a "neutral" value with two extremes?
Best for continuous variable	 sequential color ramp (smooth)	 diverging color ramp (smooth)
Best for categorical variable *(but could be used for continuous)*	 sequential color ramp (stepped)	 diverging color ramp (stepped)

Define min

Define max

https://www.storytellingwithdata.com/blog/2020/5/6/picking-the-right-colors

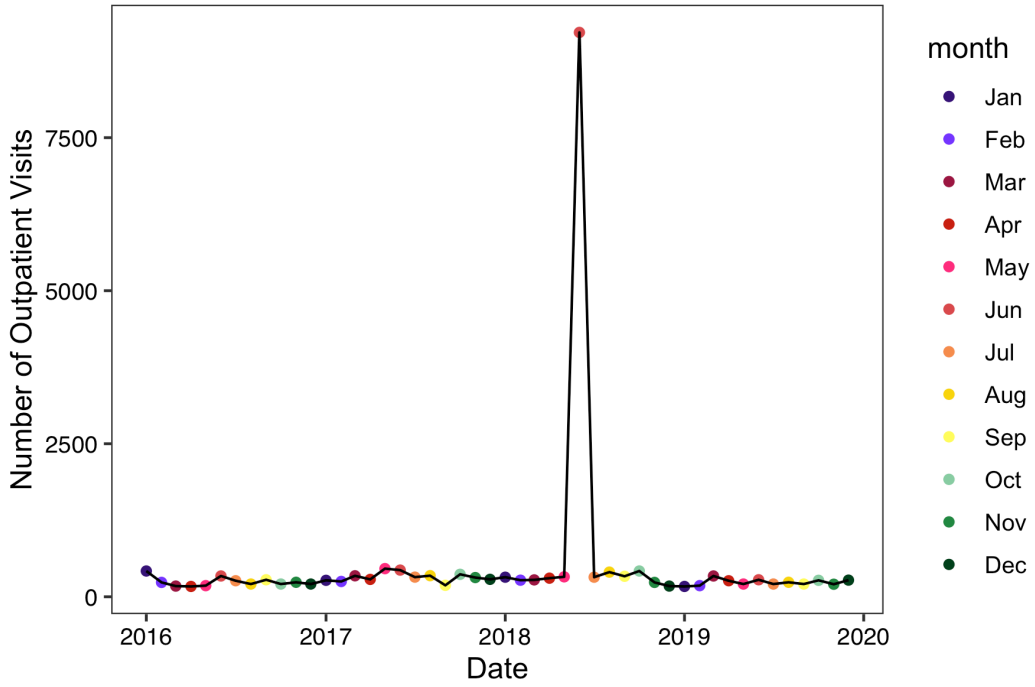
Choice of color scales



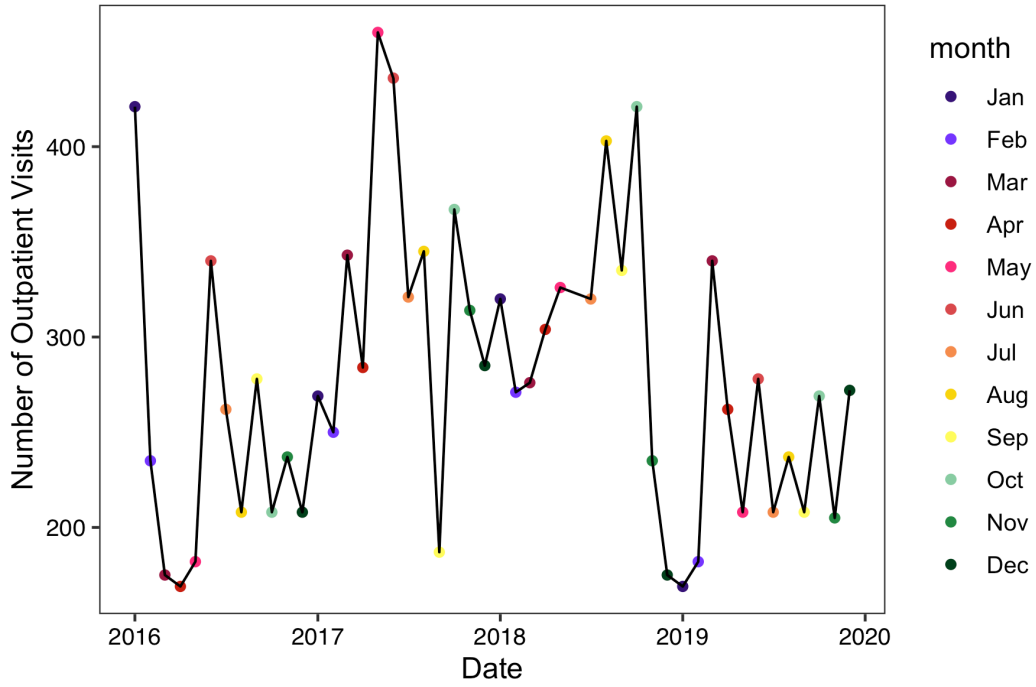
https://www.dailyposter.com/p/georgias-misleading-covid-map

Appropriate axis limits

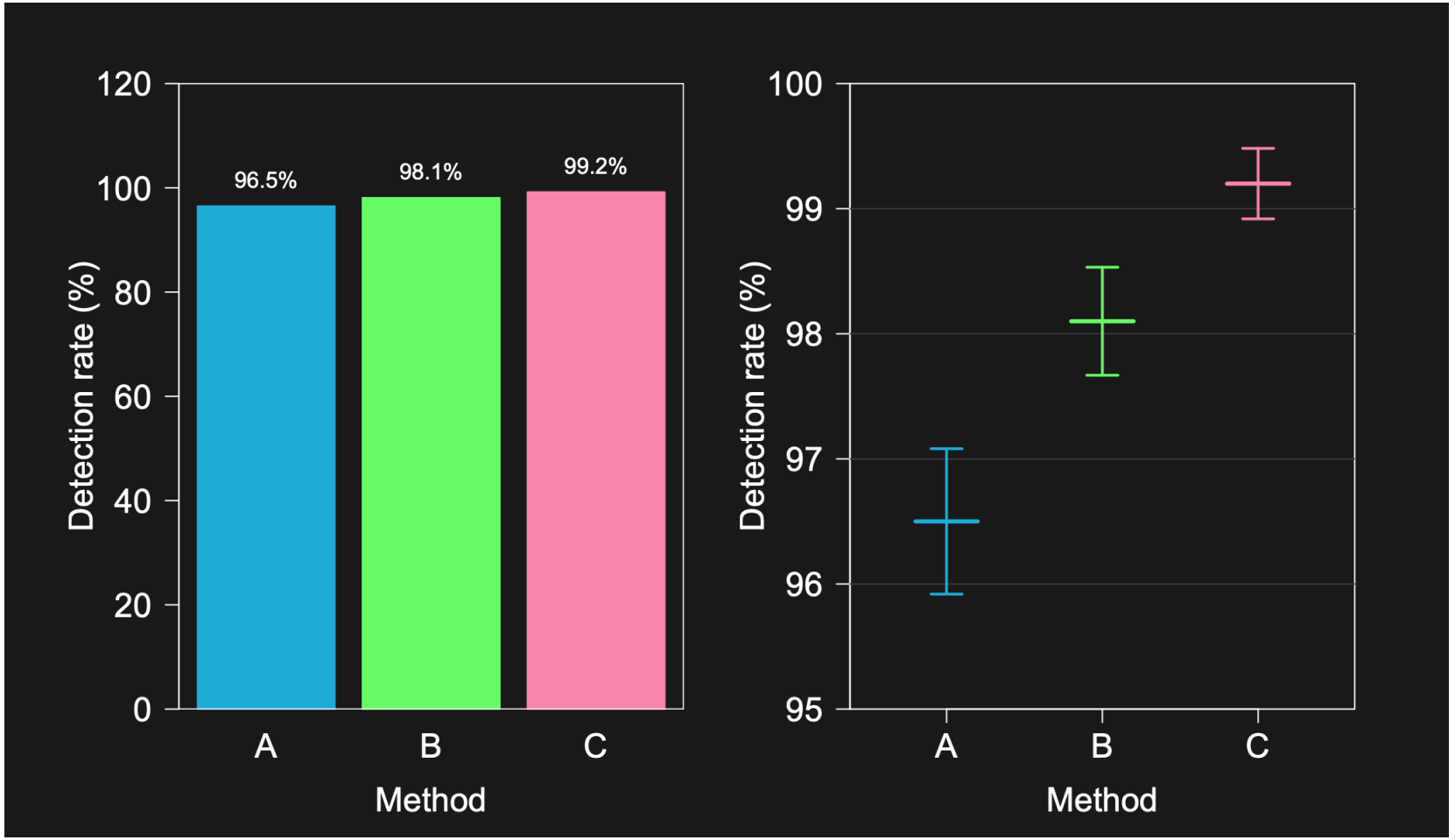
Outlier example from last week



Outlier removed



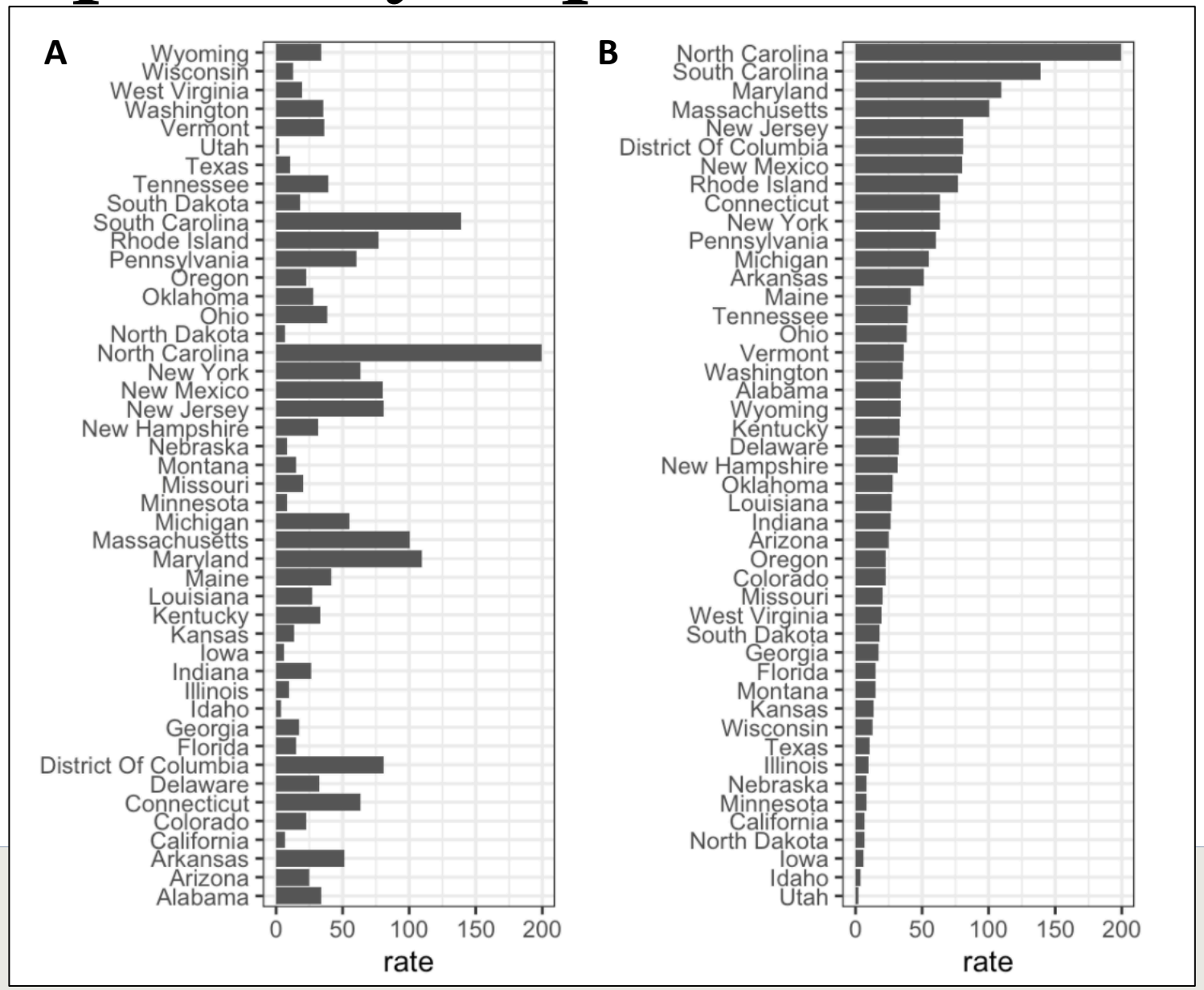
Appropriate axis limits



https://www.biostat.wisc.edu/~kbroman/presentations/graphs2017.pdf

Activity: what plot do you prefer?

<u>GOAL:</u> Determine which are the best and worst states in terms of measles rates.

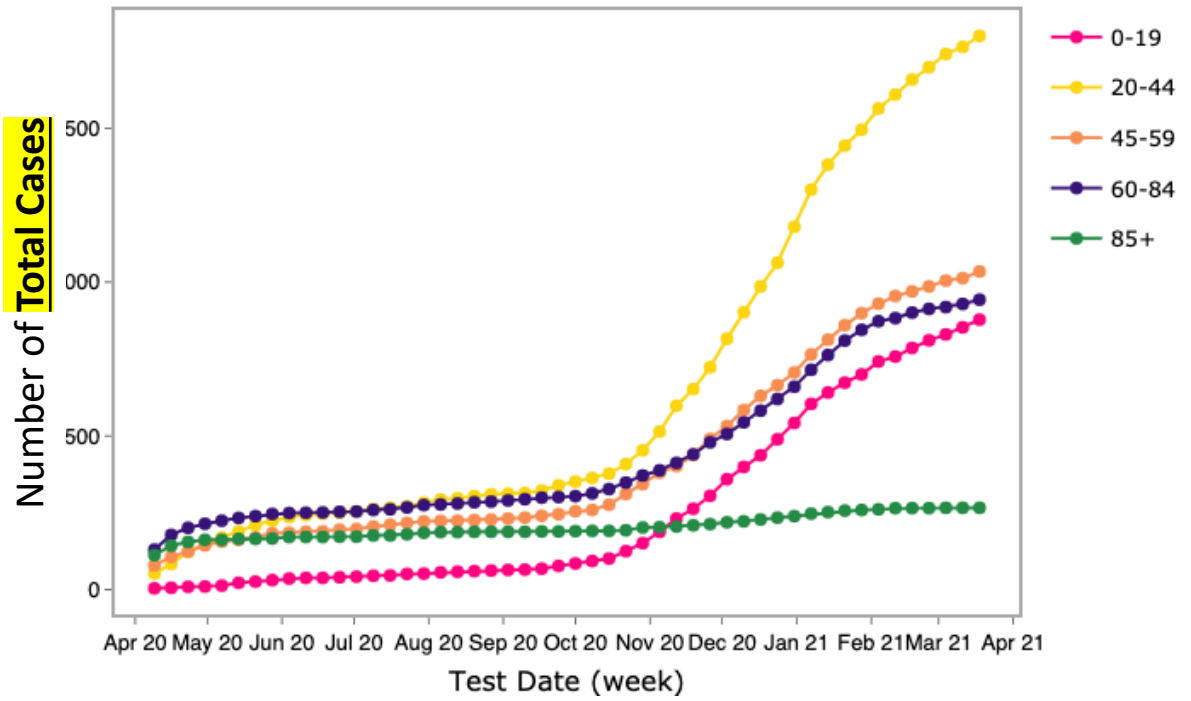


https://rafalab.github.io/dsbook/data-visualization-principles.html

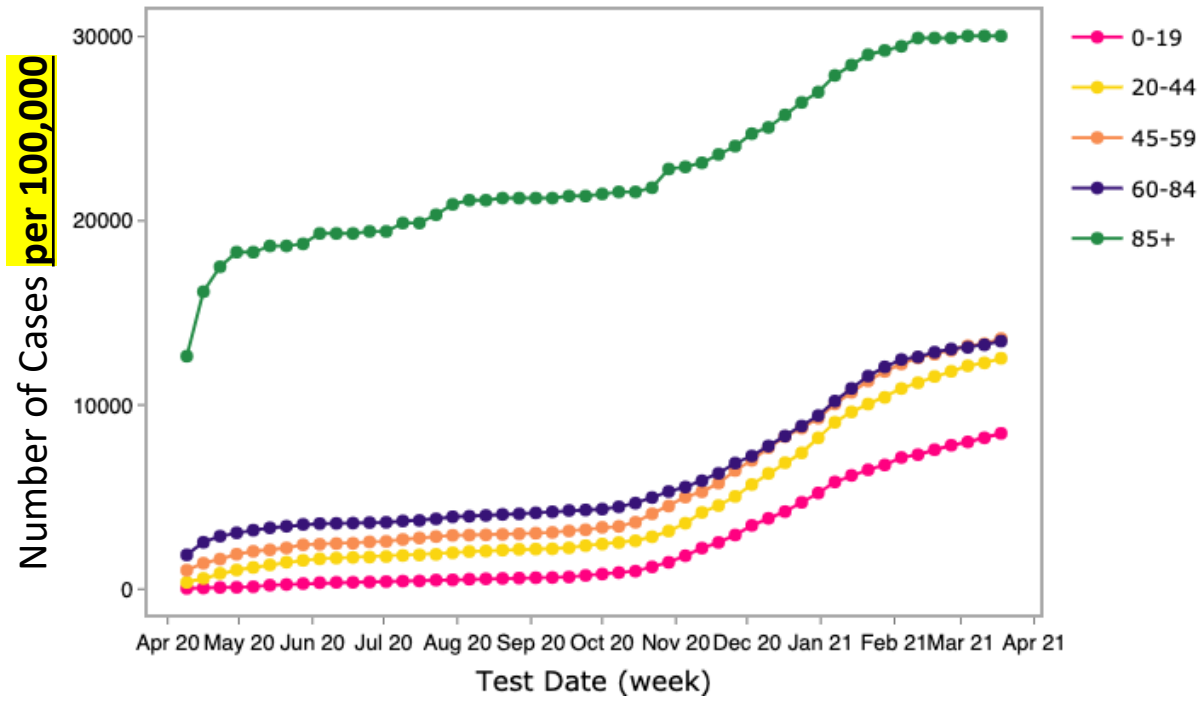
Activity: what plot do you prefer?

<u>GOAL:</u> Display which age groups have been the most impacted by COVID-19 during the entire pandemic.

A



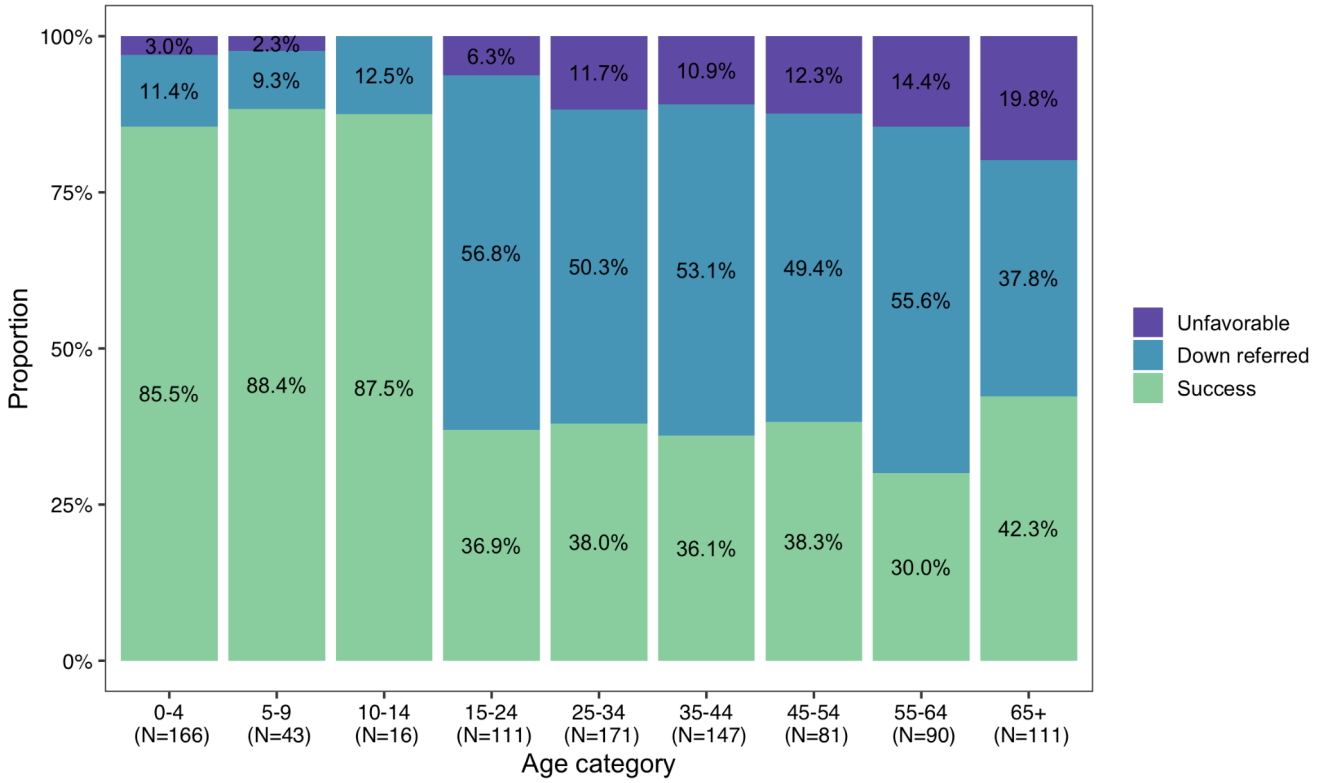
B



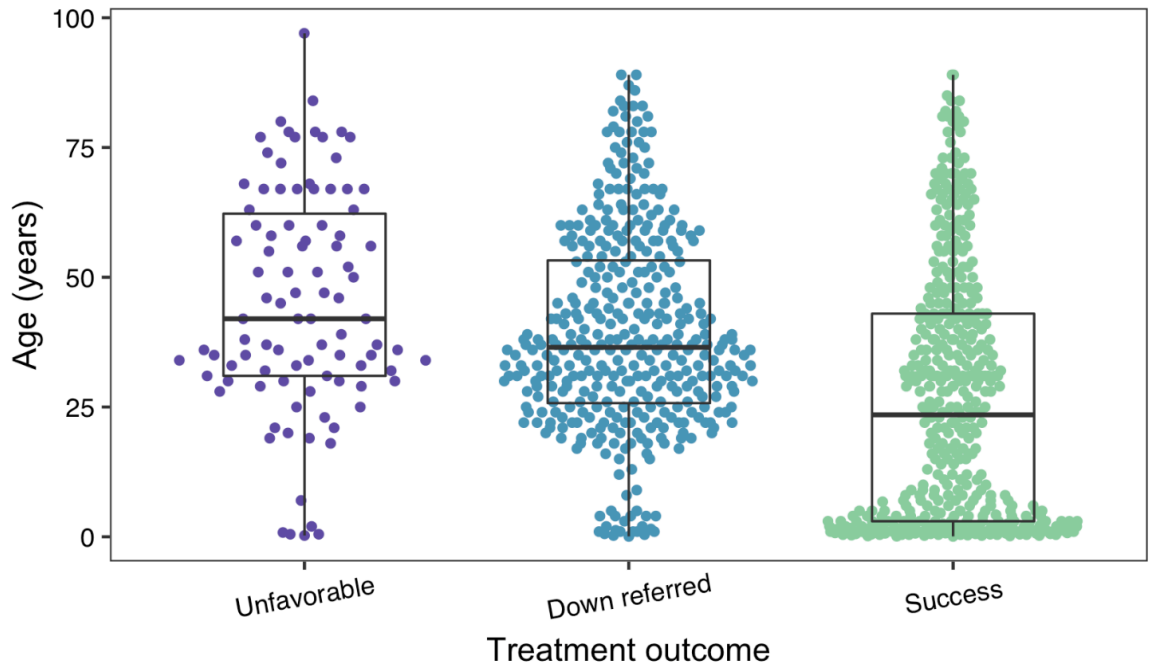
Activity: what plot do you prefer?

<u>GOAL:</u> Show how the rates of TB treatment outcomes differ by age.

A



B



Syndromic surveillance

Reasons for data visualizations

- **Exploring the data**
  - Distribution of data / transparency
  - Identify patterns, outliers, missing data
- **Analyzing the data**
  - Identifying deviations
  - Goodness of fit
  - Checking model assumption
- **Communicating results**

Sessions 1 & 4

Sessions 2 & 3

Today's focus!

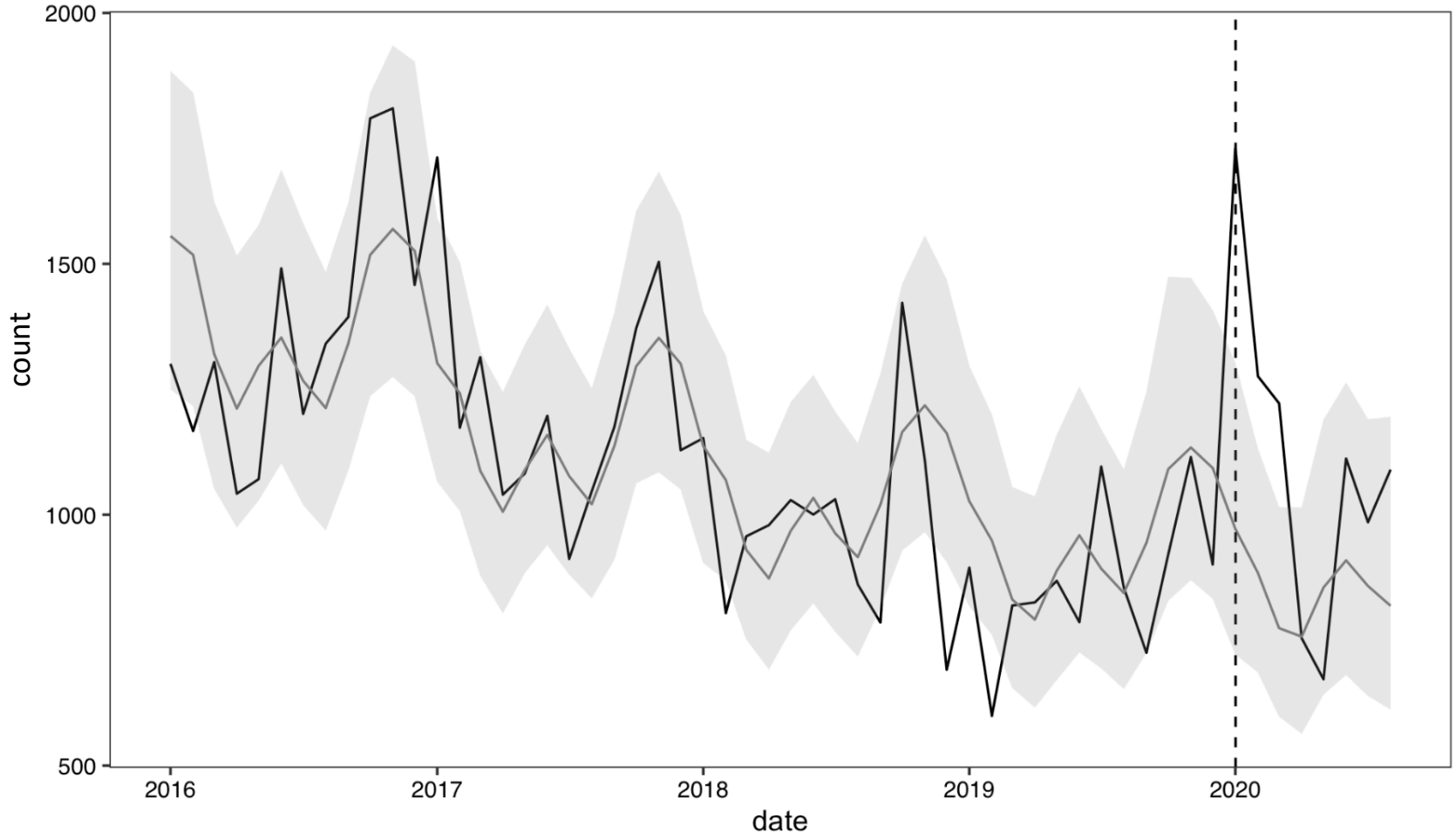
Syndromic surveillance: communicating

- The goal is to **detect** areas that have a potential uptick in cases

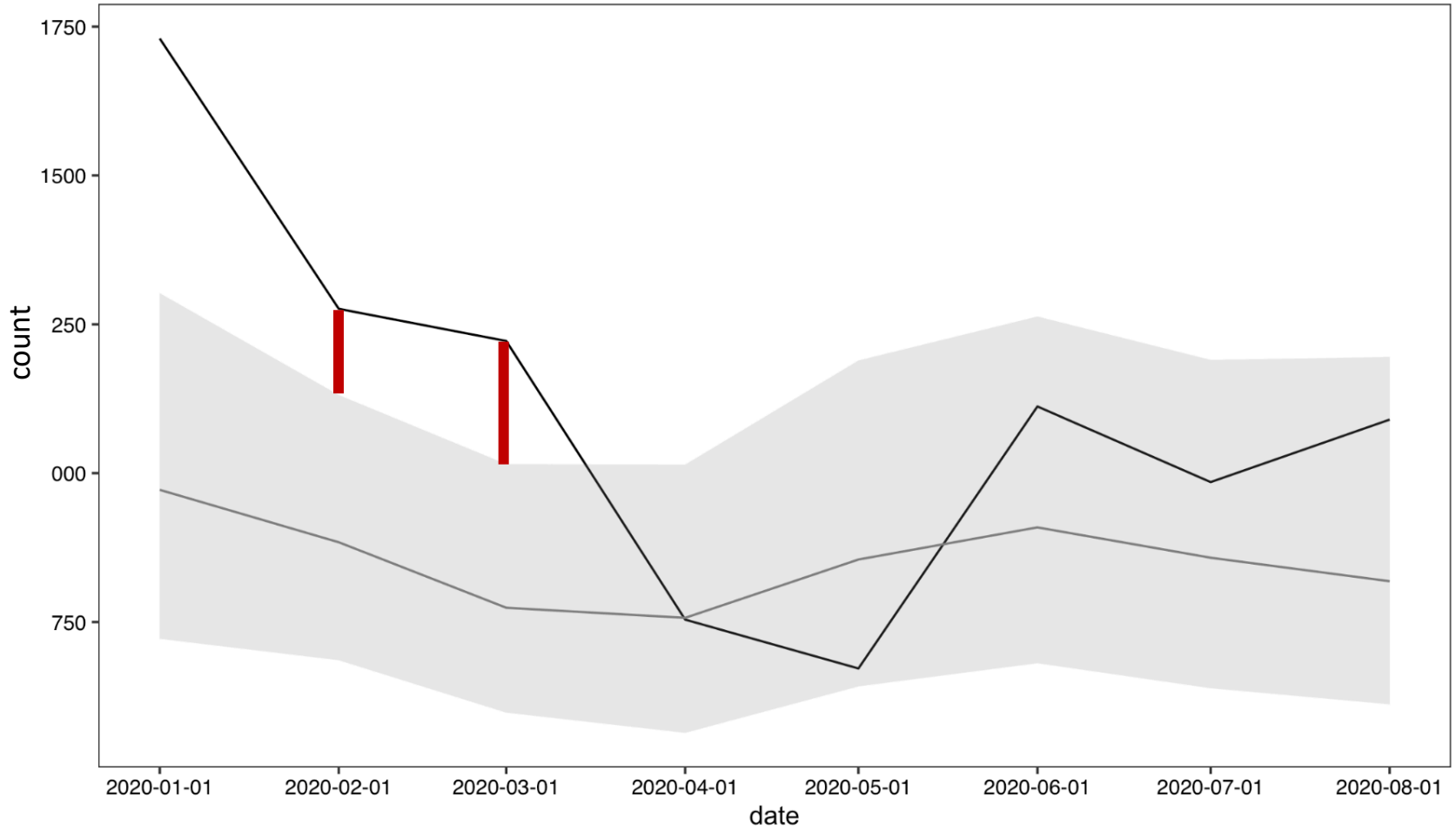
- Want to communicate:
  - Is the deviation larger than expected?
  - Magnitude of the deviation
  - *Is it important to show raw data and/or model fit?*

- Potential issues:
  - Need to contextualize the magnitude of deviation
  - Many areas or indicators to show (*how to best compare?*)

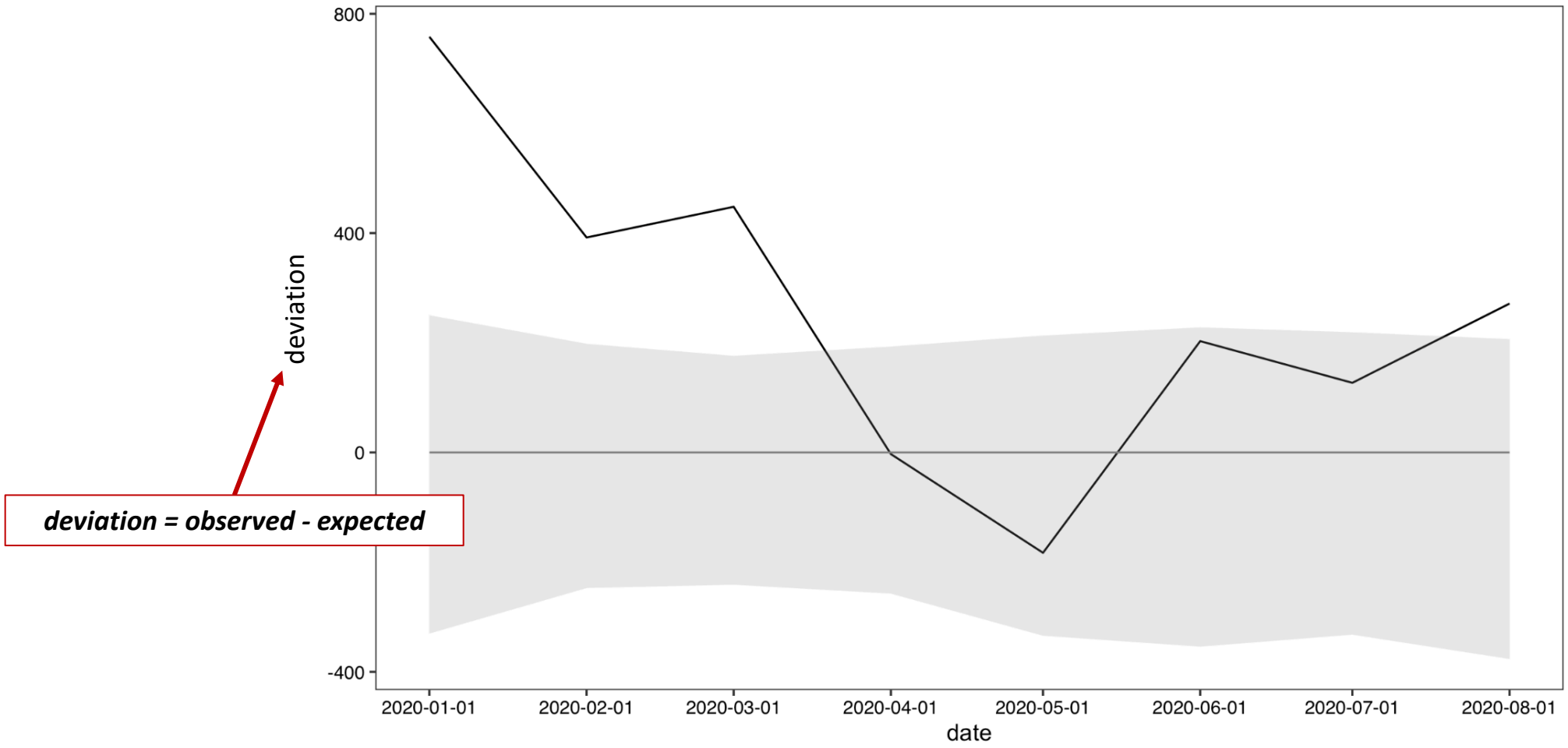
Time series plot – *all information*



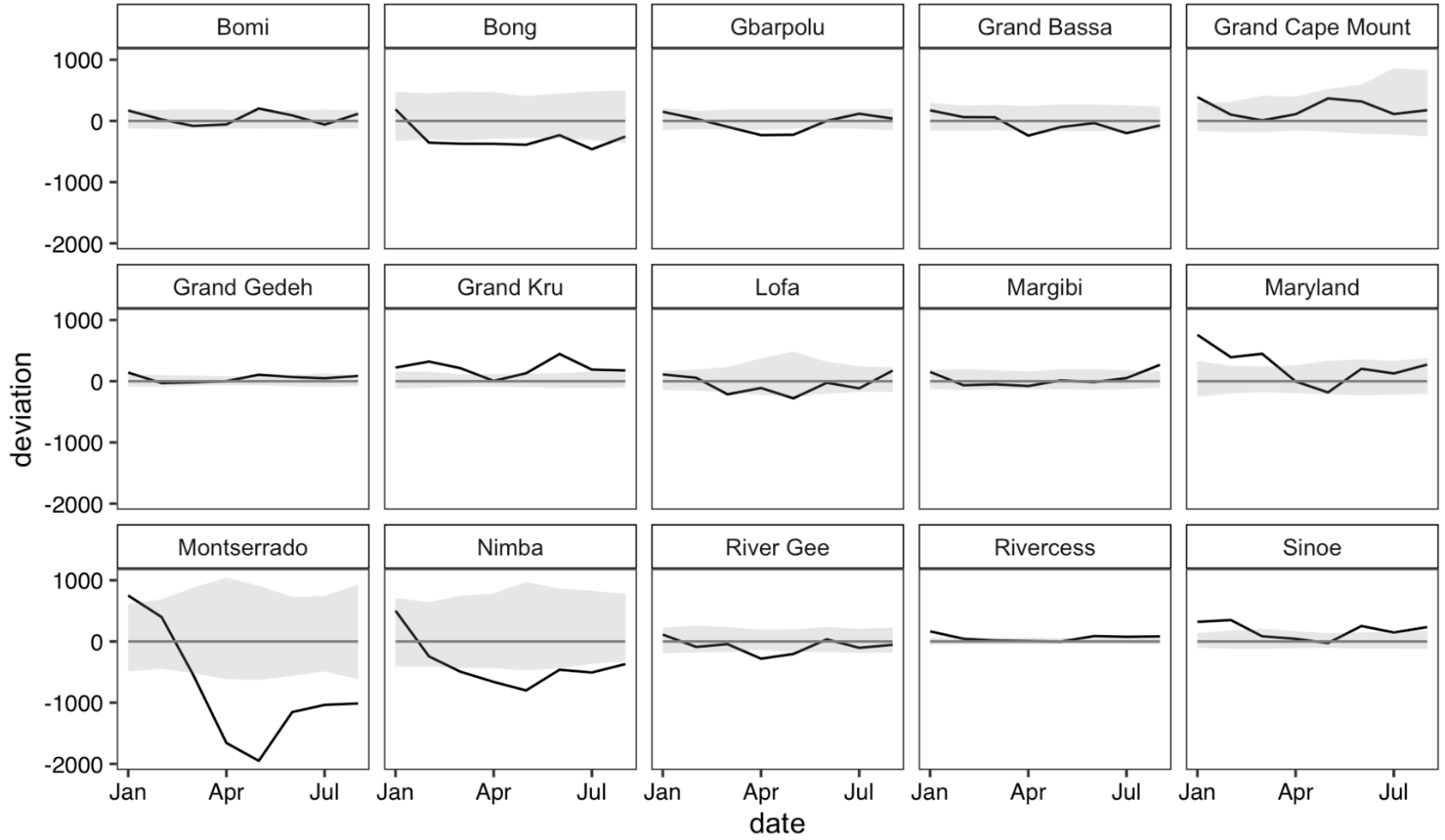
Time series plot – *evaluation only*



Time series plot – *deviations*

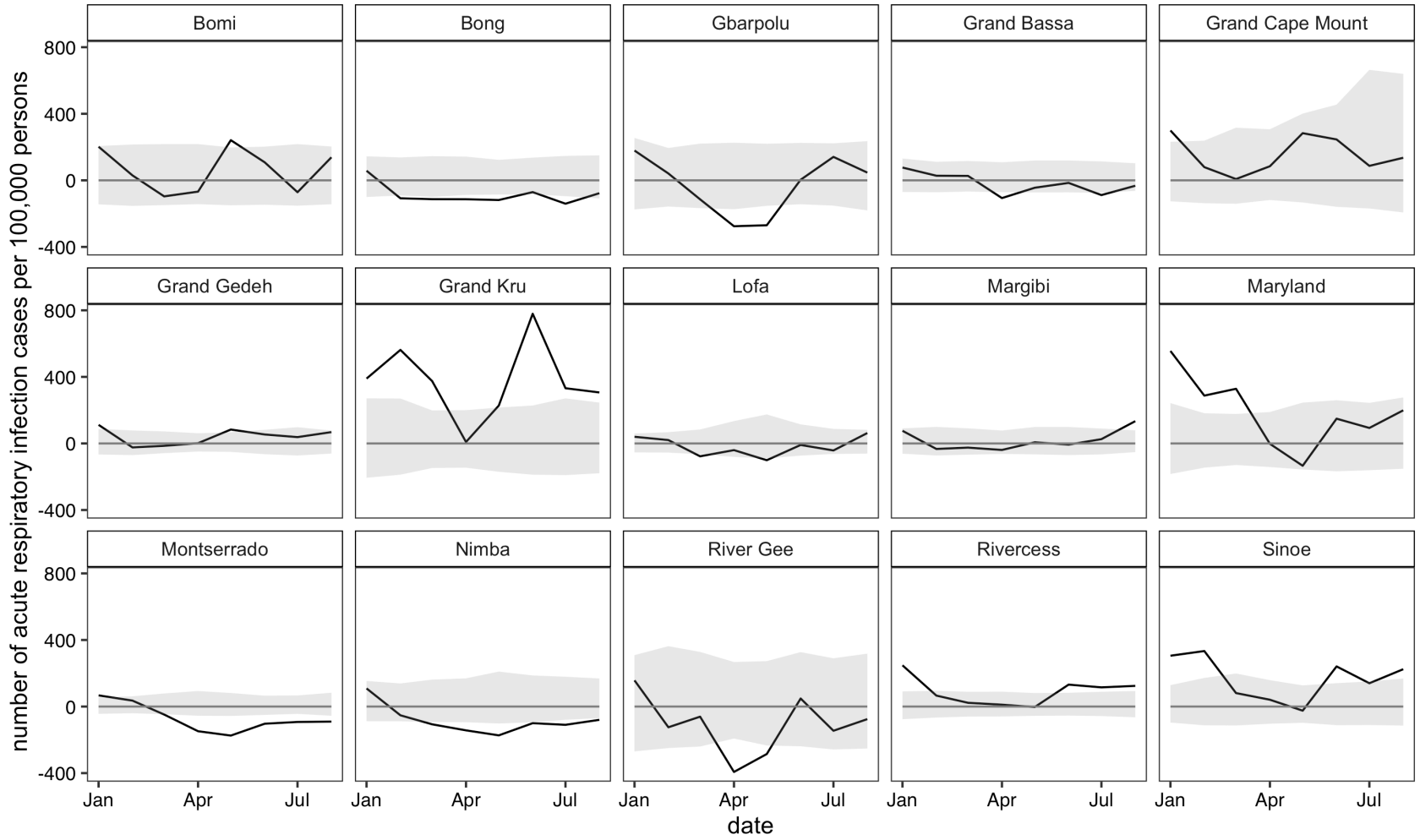


Multiple plots – *deviations*



deviation = observed − expected

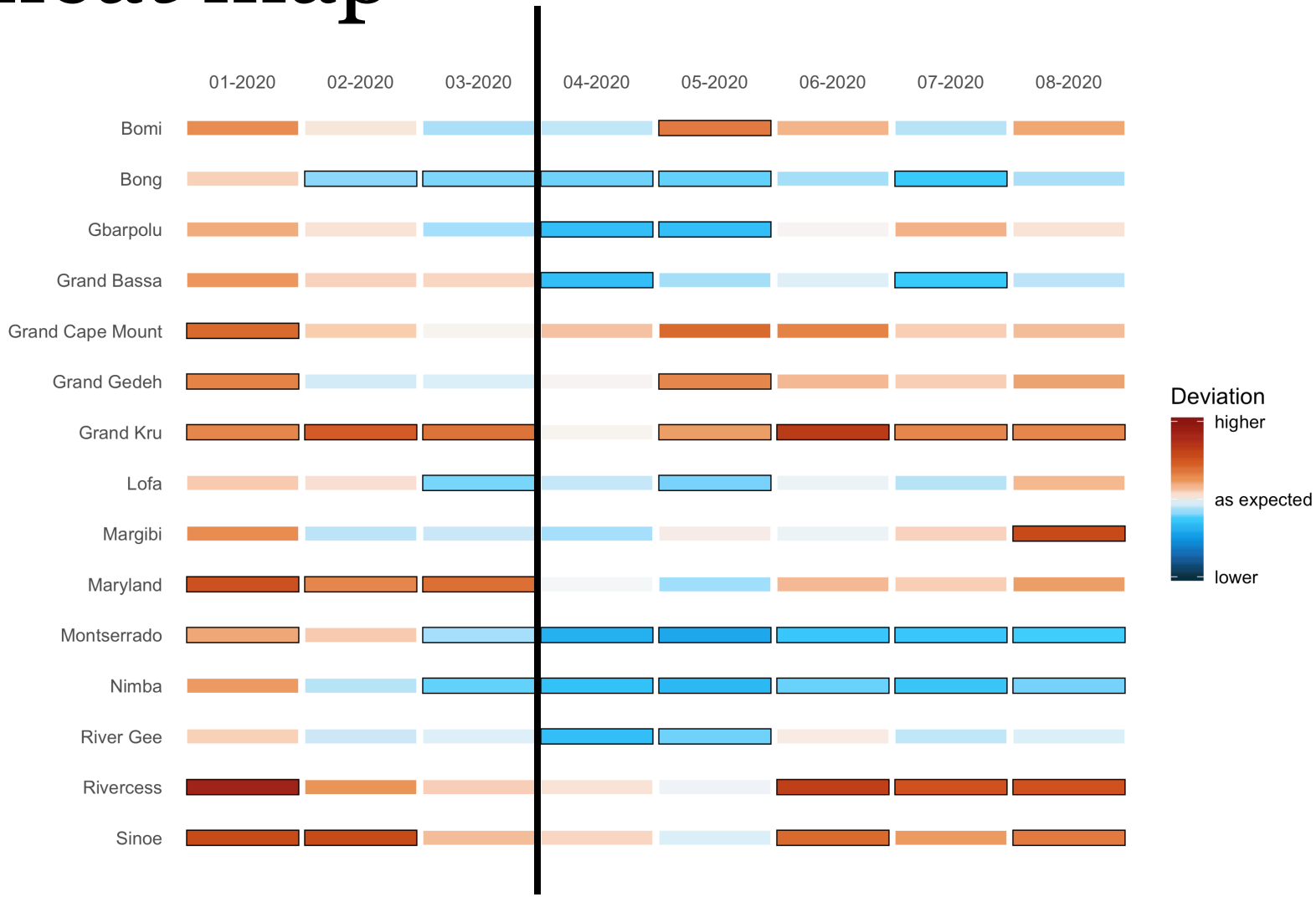
Multiple plots – *standardized*



$$\textbf{deviation} = \frac{(observed - expected)}{population} \times 100{,}000$$

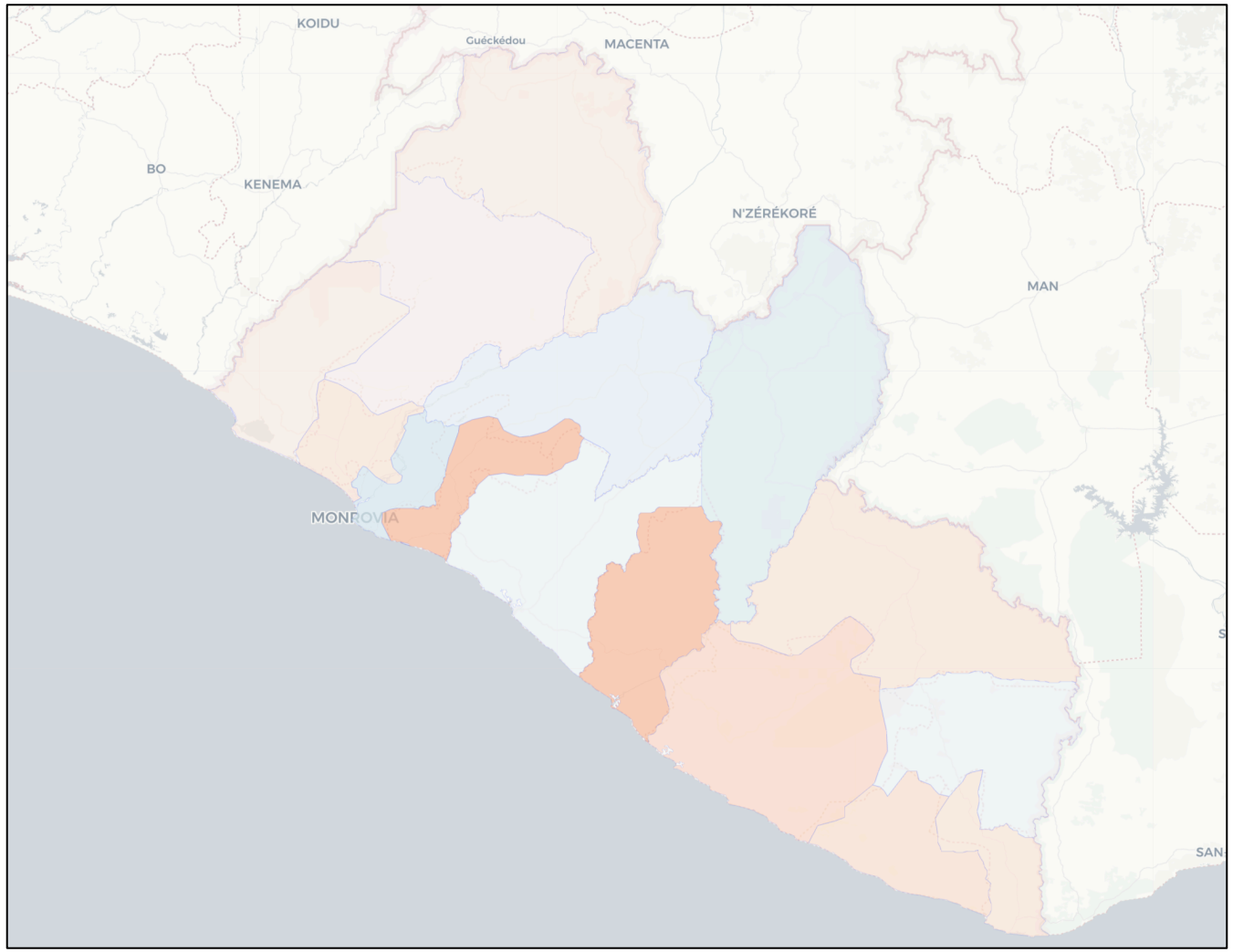
GLOBAL HEALTH RESEARCH CORE

Tiled heat map



$$deviation = \frac{(observed - expected)}{expected}$$

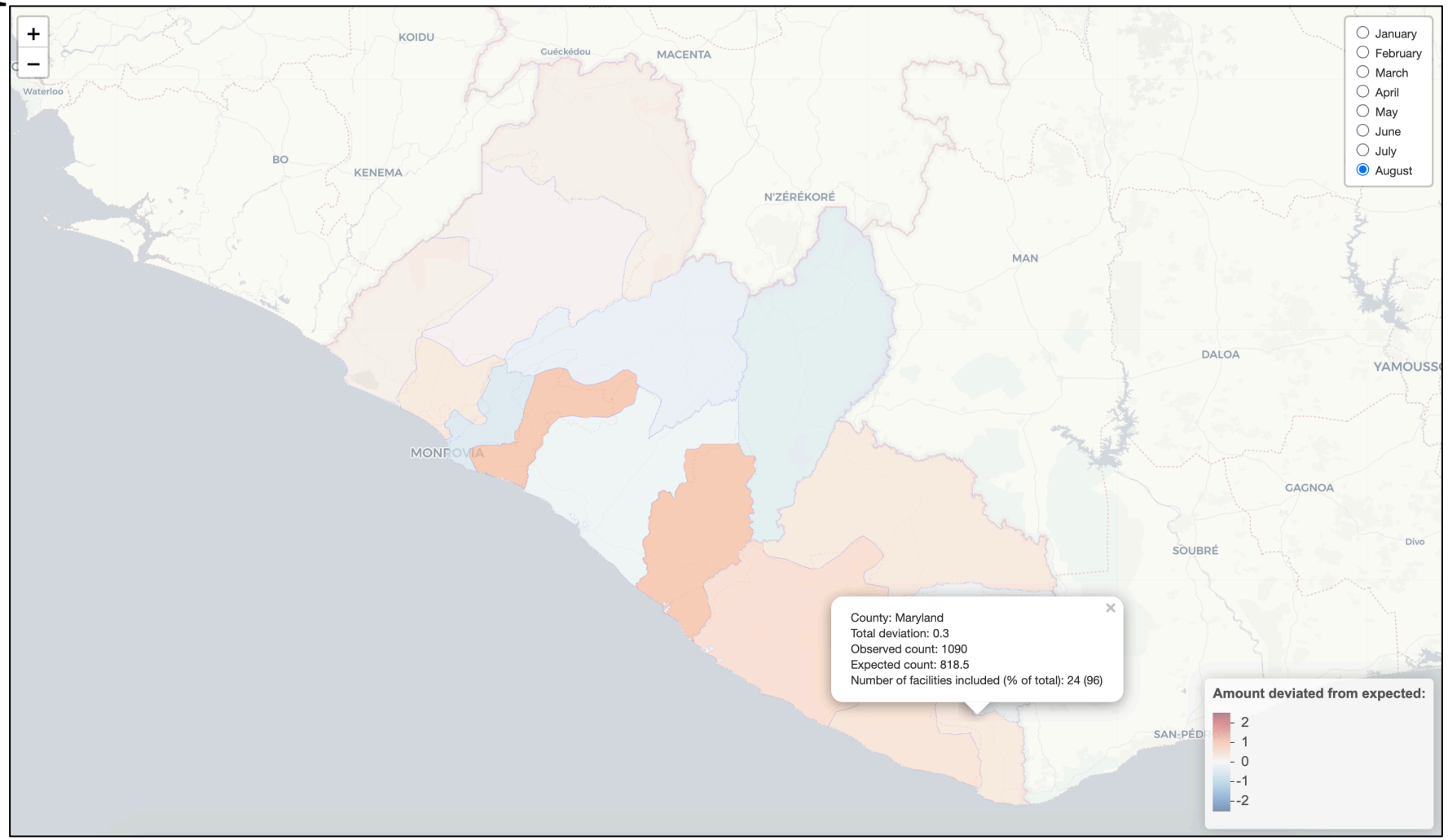
Map – static



$$deviation = \frac{(observed - expected)}{expected}$$

Map – interactive

Created by Nichole



$$deviation = \frac{(observed - expected)}{expected}$$

Software for data viz

- Figures can be generated in most software we use for data analysis (R, Python, Stata, SPSS, SAS, and Excel)

- R is free **and** has the *best* data viz options via the `ggplot2` package

- Interactive options in R include:
  - *Rmarkdown* to create HTML files (`plotly` package for interactive plots)
  - *Shiny* app
  - `leaflet` package for mapping (Nichole is an expert)

- When creating dashboards, *Shiny* requires strong knowledge of R. Another more user-friendly option is Tableau ($$).

Lab: Tying it all together!

- Lab will be a large activity to work through all skills from course
- Syndromic surveillance: Choose from three options
- Ample time to fine tune data visualizations & ask specific questions

Thank you!