Data Science Basics in R

Day 3: Exploratory data analysis

Goals for today

- Define descriptive statistics & exploratory data analysis
- Create your first data visualization in R
- Identify options for visualization in R, including ggplot2
- Get creative and have fun exploring datasets

Descriptive statistics

Goals for today

- Define descriptive statistics & exploratory data analysis
- Make a repository on github for your work
- Create your first data visualization in R
- Identify options for visualization in R, including ggplot2
- Get creative and have fun exploring datasets

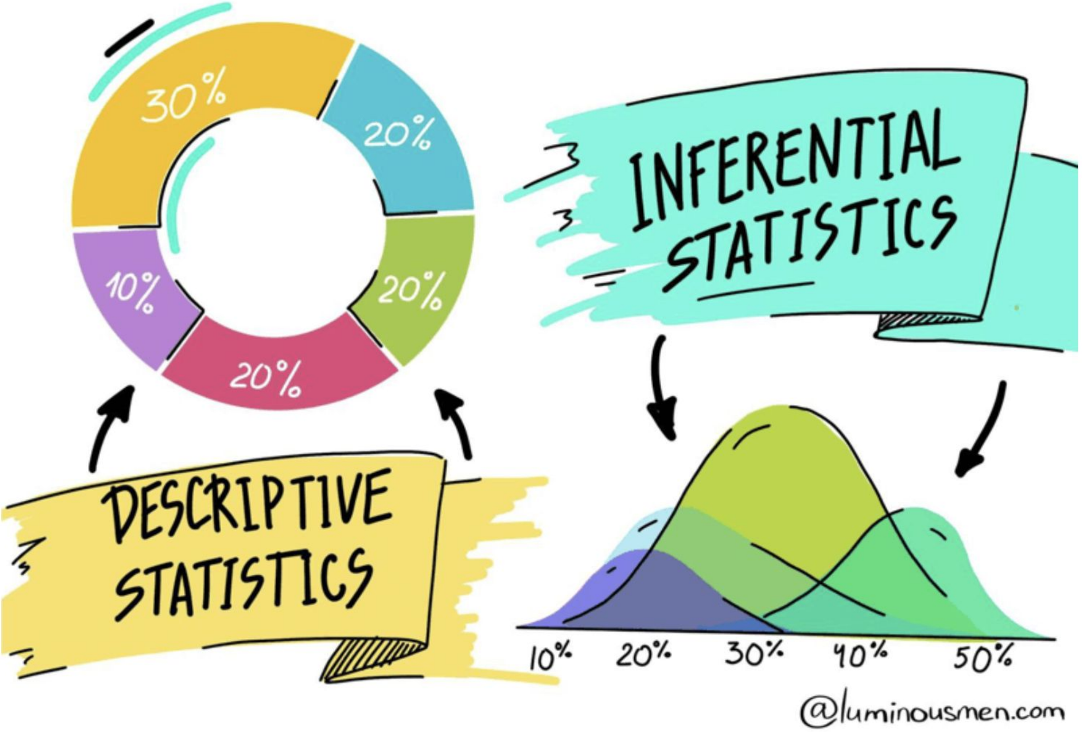
Descriptive statistics

Descriptive statistics summarize data, and typically describe three types of things:

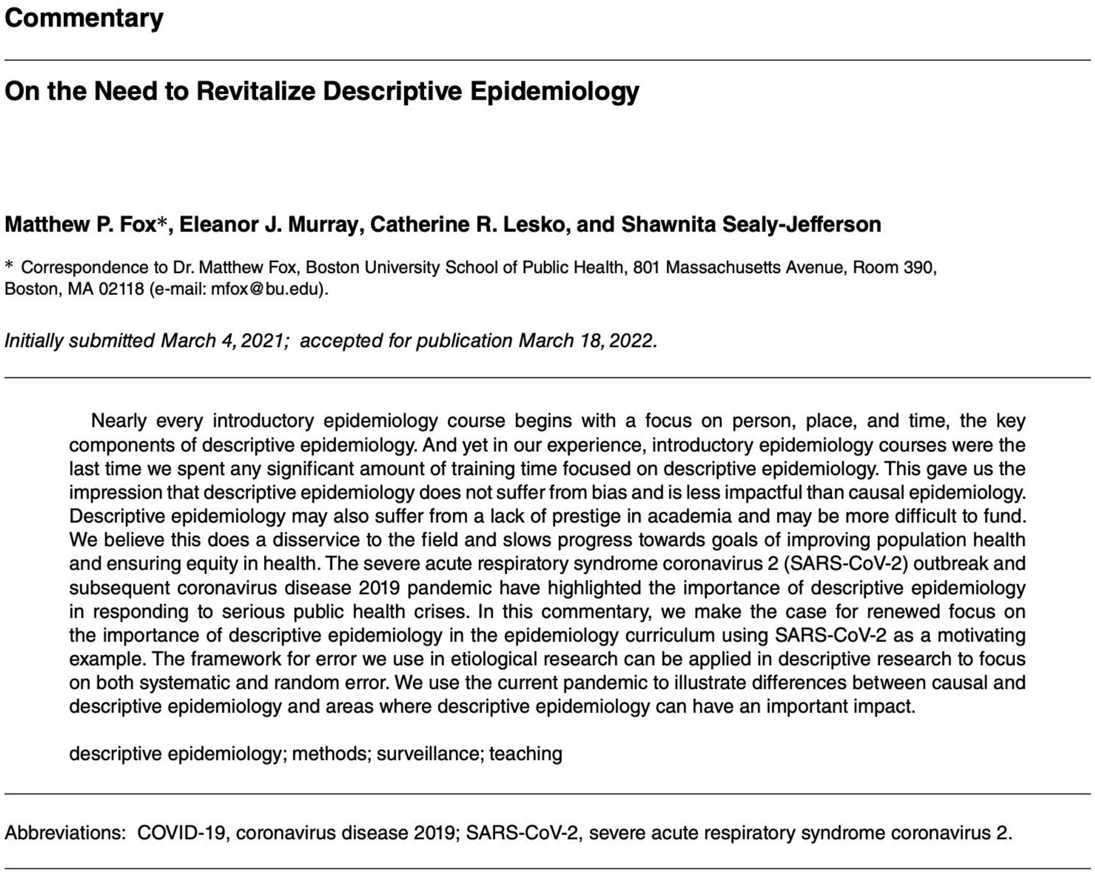
- center (e.g., mean, median)
- spread (e.g., min, max, standard deviation, interquartile range)
- counts & rates (e.g., summary tables)

In a typical data analysis workflow, we explore these first! It’s helpful to better understand your data, and to identify potential surprises.

Descriptive statistics



Graphic by luminousmen.com
https://luminousmen.com/post/descriptive-and-inferential-statistics



Commentary

On the Need to Revitalize Descriptive Epidemiology

Matthew P. Fox*, Eleanor J. Murray, Catherine R. Lesko, and Shawnita Sealy-Jefferson

* Correspondence to Dr. Matthew Fox, Boston University School of Public Health, 801 Massachusetts Avenue, Room 390,
Boston, MA 02118 (e-mail: mfox@bu.edu).

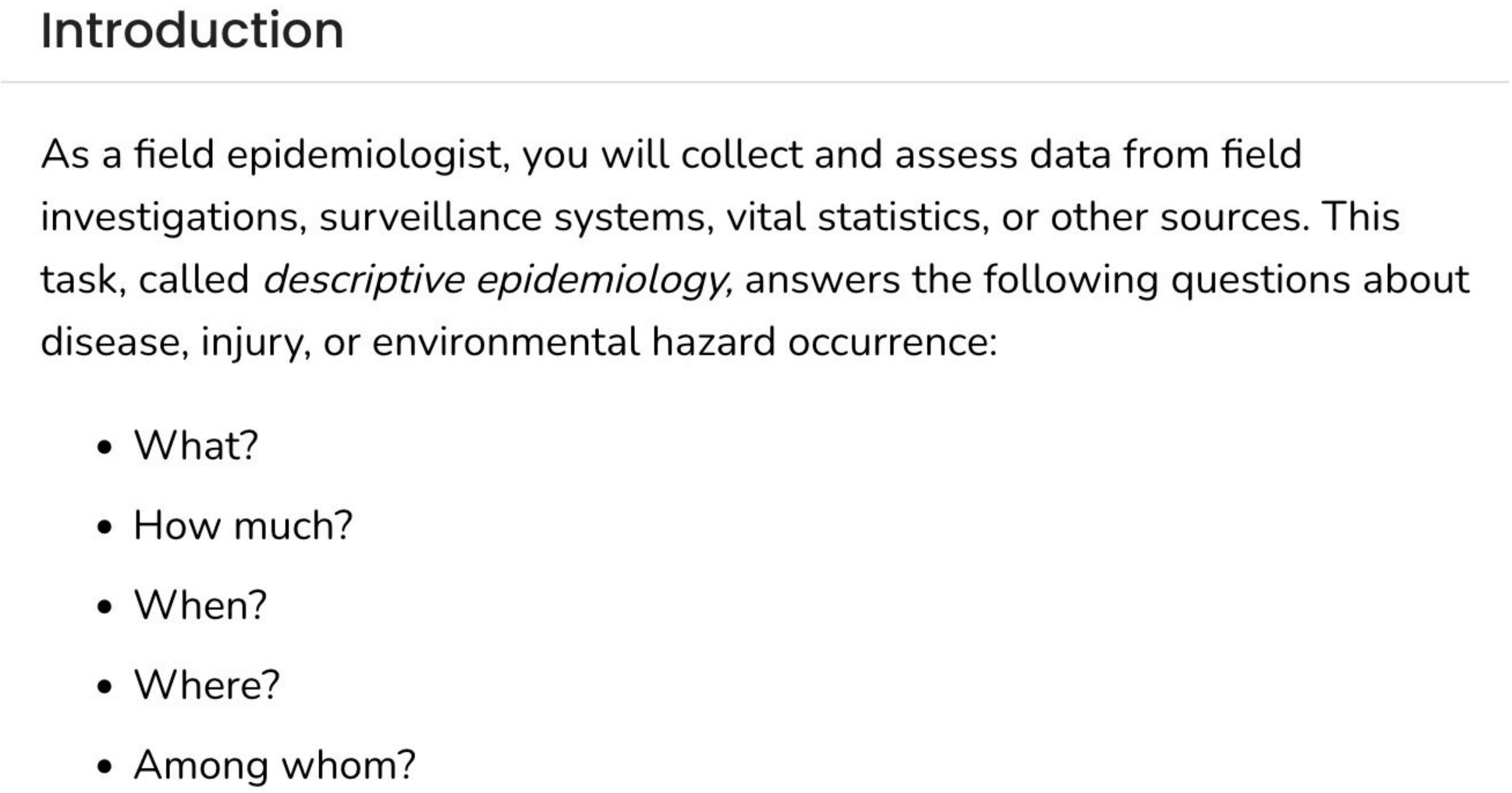
Initially submitted March 4, 2021; accepted for publication March 18, 2022.

Nearly every introductory epidemiology course begins with a focus on person, place, and time, the key components of descriptive epidemiology. And yet in our experience, introductory epidemiology courses were the last time we spent any significant amount of training time focused on descriptive epidemiology. This gave us the impression that descriptive epidemiology does not suffer from bias and is less impactful than causal epidemiology. Descriptive epidemiology may also suffer from a lack of prestige in academia and may be more difficult to fund. We believe this does a disservice to the field and slows progress towards goals of improving population health and ensuring equity in health. The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) outbreak and subsequent coronavirus disease 2019 pandemic have highlighted the importance of descriptive epidemiology in responding to serious public health crises. In this commentary, we make the case for renewed focus on the importance of descriptive epidemiology in the epidemiology curriculum using SARS-CoV-2 as a motivating example. The framework for error we use in etiological research can be applied in descriptive research to focus on both systematic and random error. We use the current pandemic to illustrate differences between causal and descriptive epidemiology and areas where descriptive epidemiology can have an important impact.

descriptive epidemiology; methods; surveillance; teaching

Abbreviations: COVID-19, coronavirus disease 2019; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Fox MP, Murray EJ, Lesko CR, Sealy-Jefferson S. On the need to revitalize descriptive epidemiology.
American Journal of Epidemiology. 2022 Jul;191(7):1174-9.



Introduction

As a field epidemiologist, you will collect and assess data from field investigations, surveillance systems, vital statistics, or other sources. This task, called *descriptive epidemiology,* answers the following questions about disease, injury, or environmental hazard occurrence:

- What?
- How much?
- When?
- Where?
- Among whom?

Robert E. Fontaine. CDC Field Epidemiology Manual. Describing Epidemiologic Data.
https://www.cdc.gov/eis/field-epi-manual/chapters/Describing-Epi-Data.html

Your turn!

Think about the measles policy datasets we started to explore yesterday.

What are five different, specific questions that you could explore based on those data?

Calculating summary statistics in R *(code in github for live demo)*

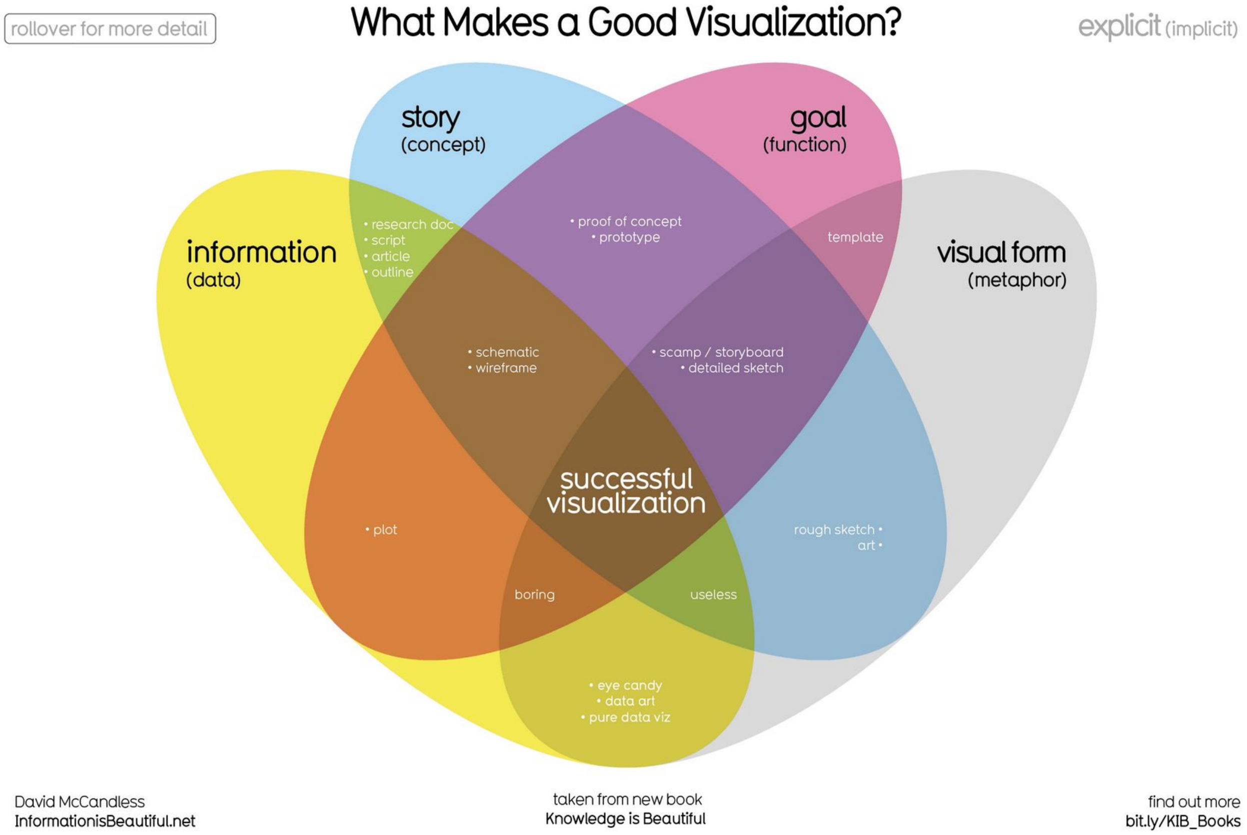
Goals of data visualization

Goals of data visualization

People use data visualizations for all types of reasons and audiences. Information is often easier to quickly understand in visualization as compared to other forms of communication (for example, listed in a table or described out loud)

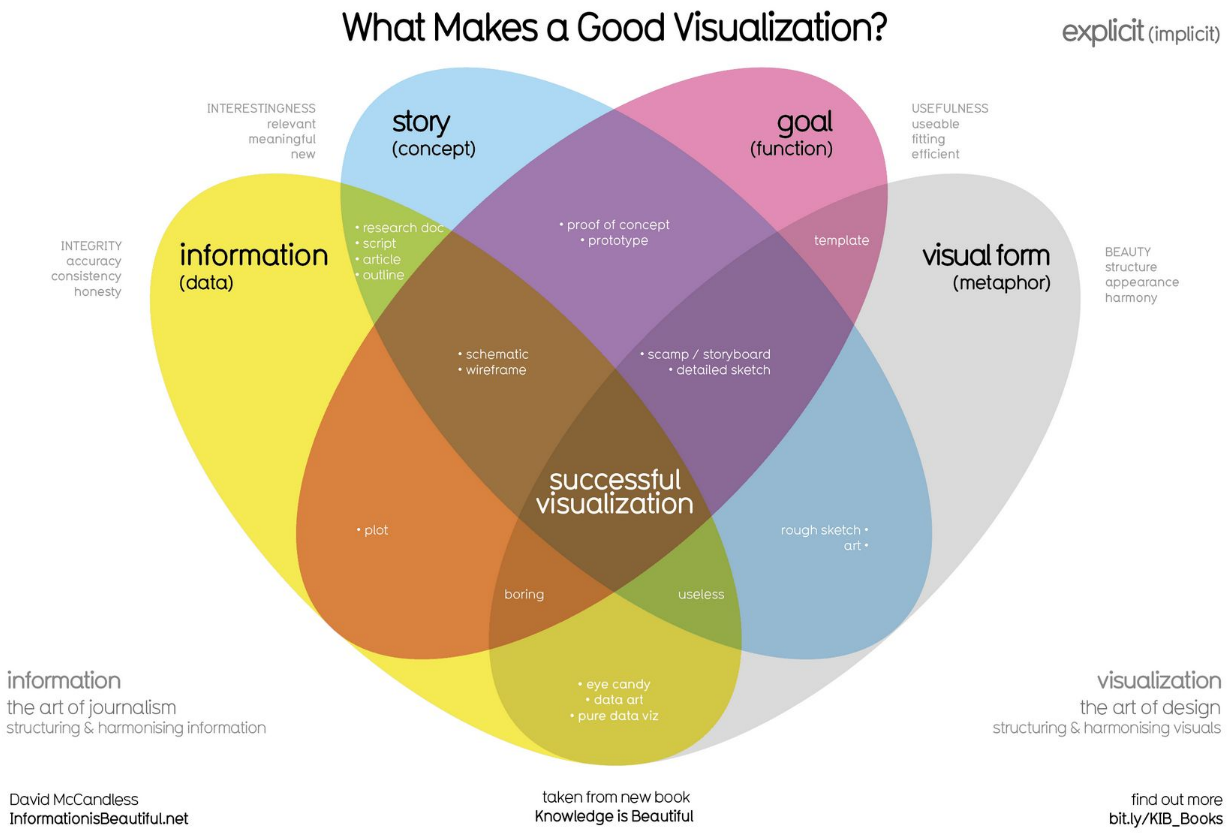
- **Understand** what is happening in a new dataset or situation
- **Communicate** information quickly and rapidly
- **Make decisions** based on an understanding of what is currently known

What makes a data visualization *good*?



https://informationisbeautiful.net/visualizations/what-makes-a-good-data-visualization/

What makes a data visualization *good*?

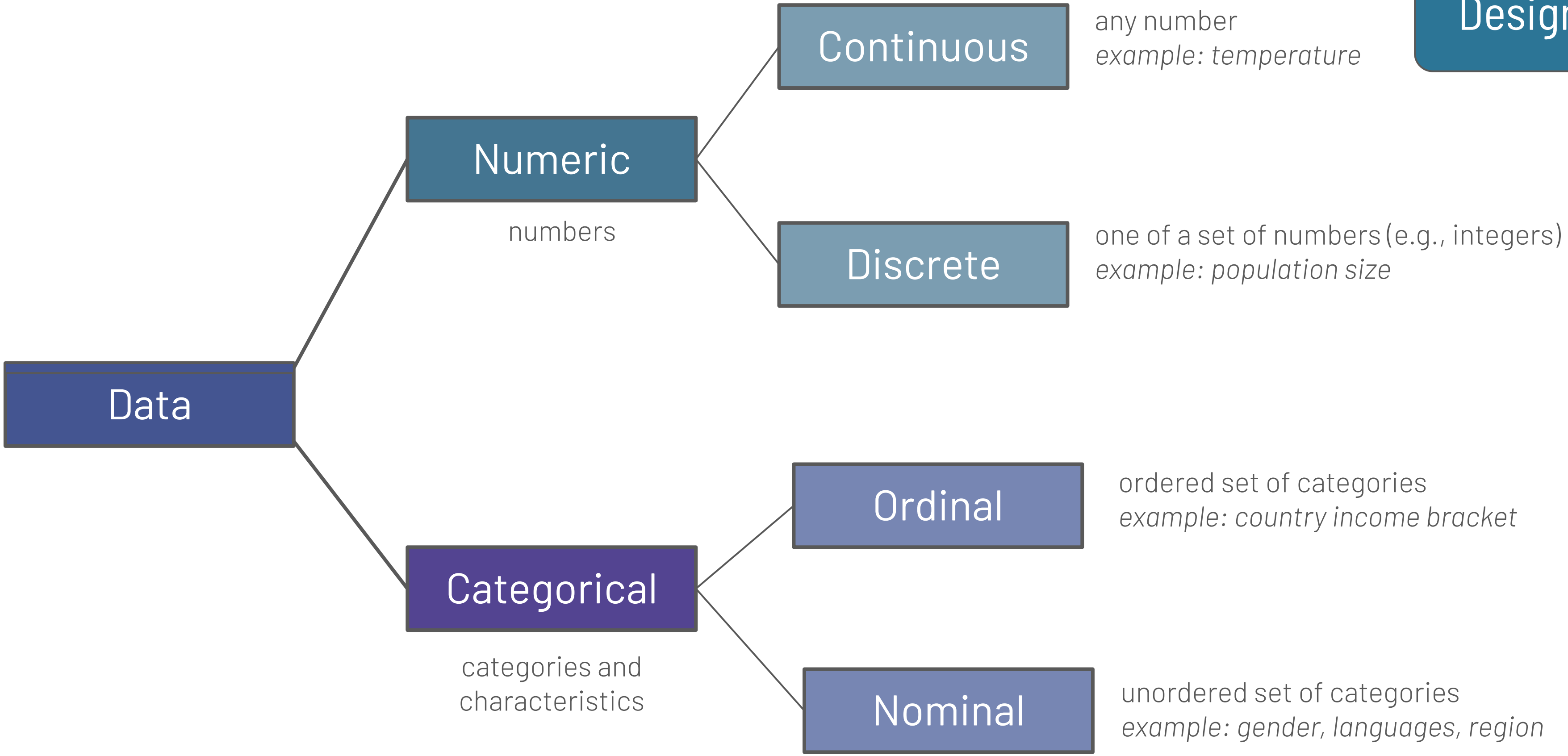


https://informationisbeautiful.net/visualizations/what-makes-a-good-data-visualization/

Choosing a data visualization

(we’ll talk more about this tomorrow)

Design



Data

Numeric

numbers

Continuous

any number
example: temperature

Discrete

one of a set of numbers (e.g., integers)
example: population size

Categorical

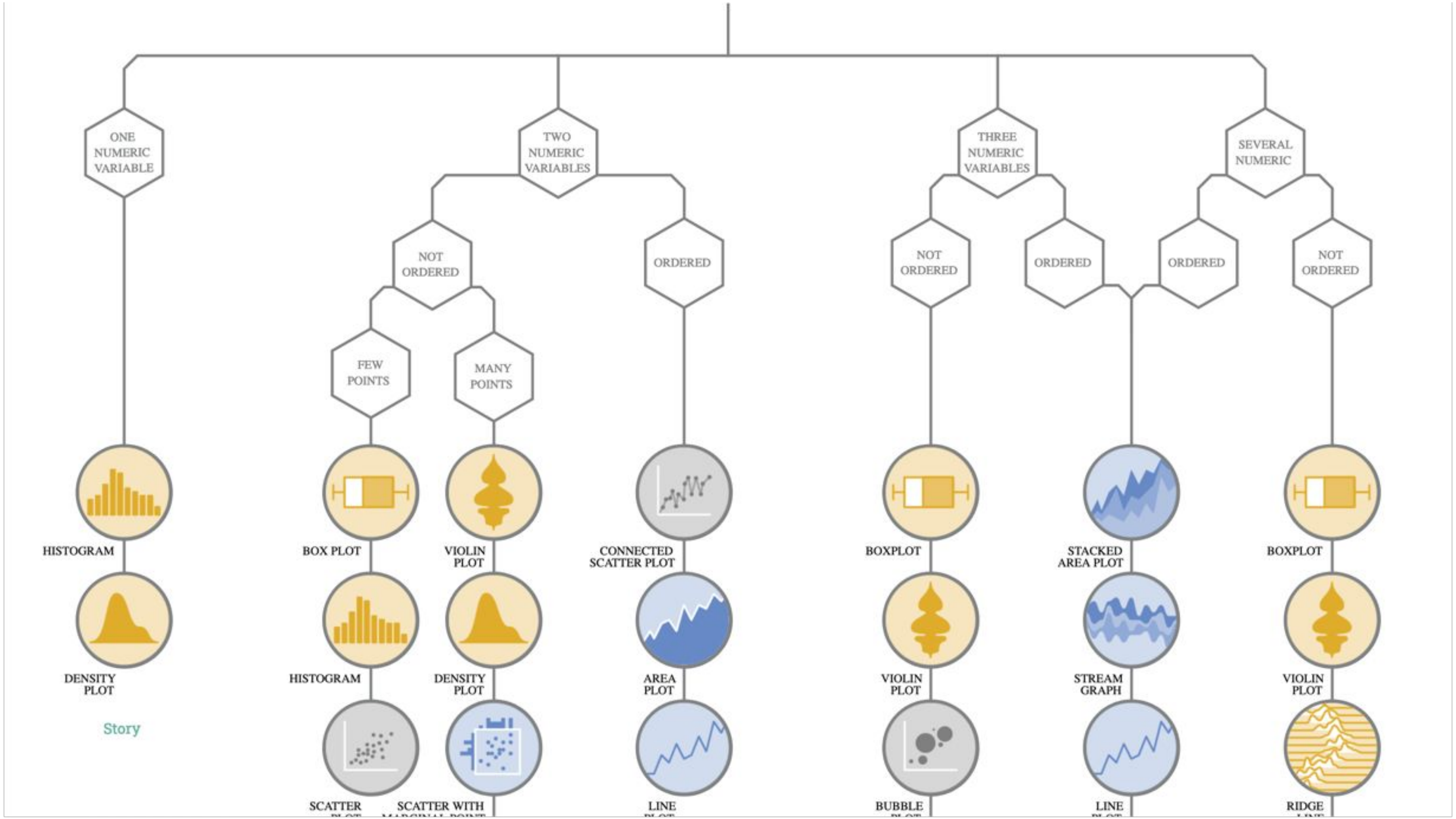
categories and
characteristics

Ordinal

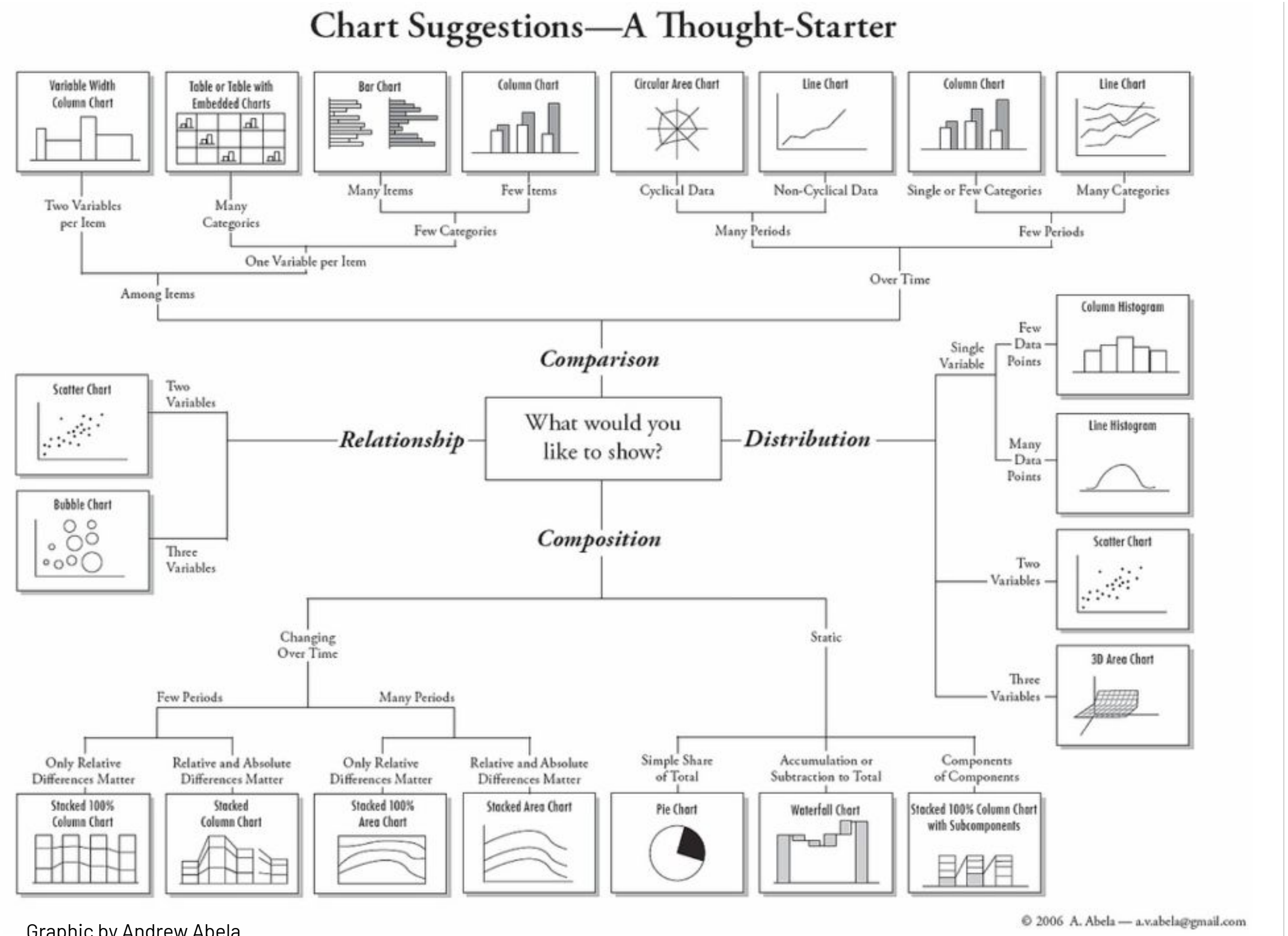
ordered set of categories
example: country income bracket

Nominal

unordered set of categories
example: gender, languages, region



Graphic by Data to Viz
https://www.data-to-viz.com/



Graphic by Andrew Abela
https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Plots in base R
(code in github for live demo)

Reminder: Course datasets

Policy data

Categorical

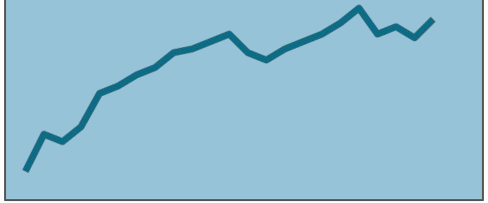
Current national policies related to measles vaccination, per country:

- Required
- Not required
- No data

Coverage data

Continuous

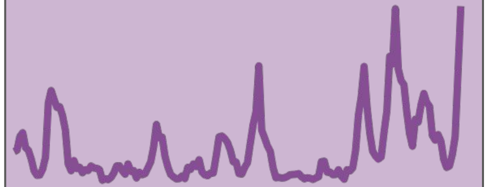
Yearly measles containing vaccine first-dose (MCV1) immunization coverage among 1-year-olds, per country, over time



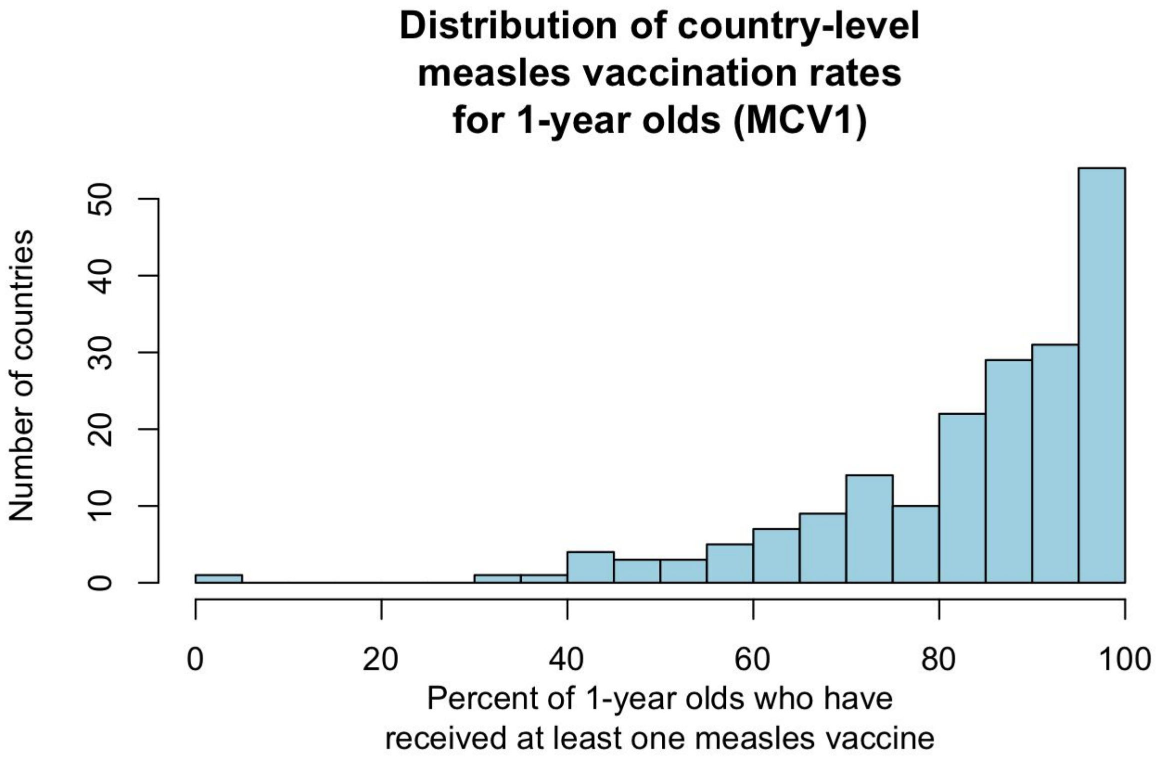
Caseload data

Continuous

Monthly data on laboratory confirmed, epidemiologically linked, and/or clinical measles cases reported to WHO per country, over time

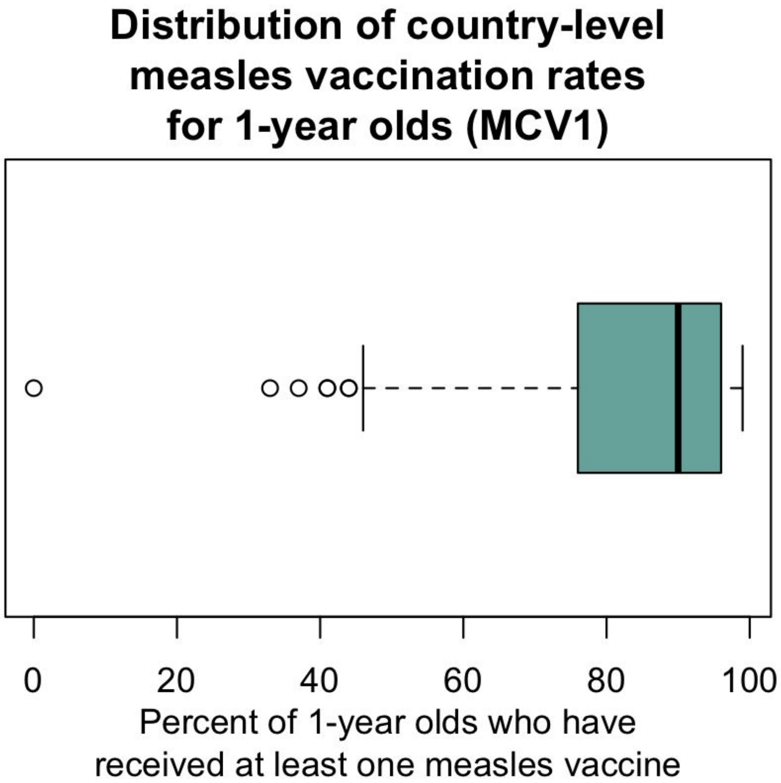


Histogram

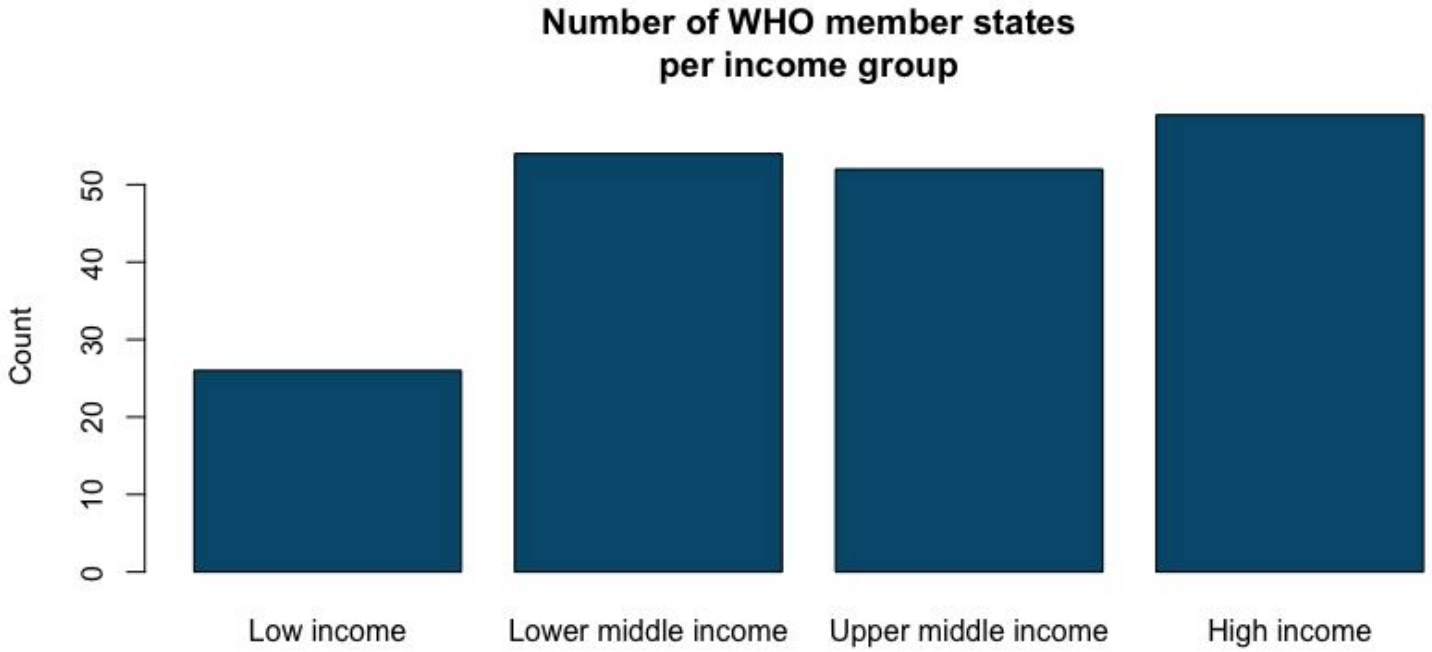


Distribution of country-level measles vaccination rates for 1-year olds (MCV1)
Number of countries
0
10
20
30
40
50
0
20
40
60
80
100
Percent of 1-year olds who have received at least one measles vaccine

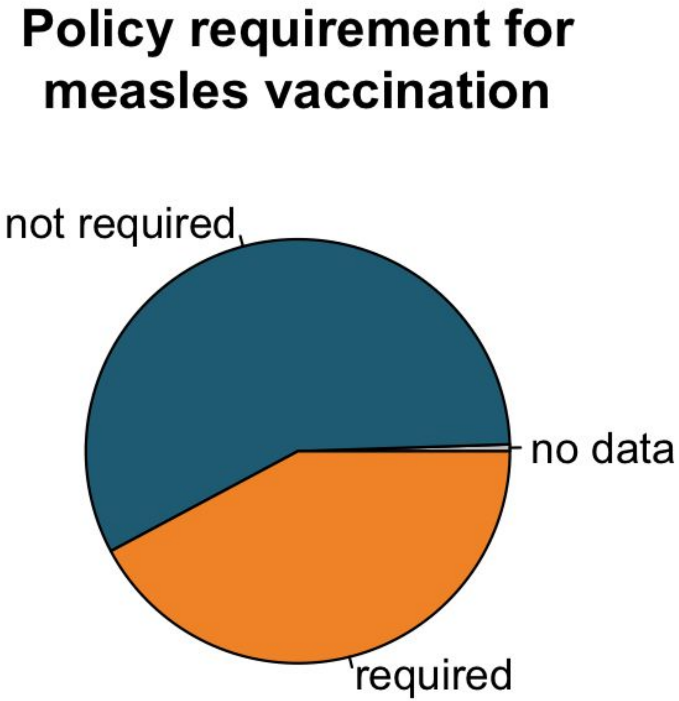
Boxplot


Distribution of country-level measles vaccination rates for 1-year olds (MCV1)
0
20
40
60
80
100
Percent of 1-year olds who have received at least one measles vaccine

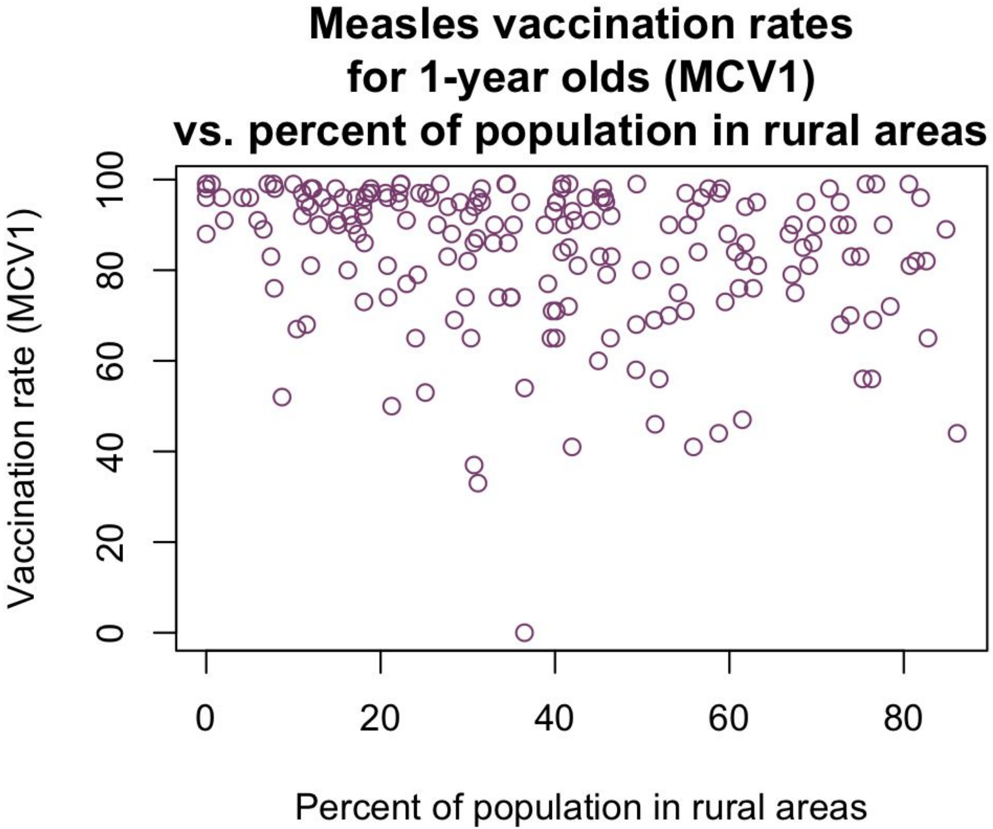
Barchart


Number of WHO member states
per income group
Count
0
10
20
30
40
50
Low income
Lower middle income
Upper middle income
High income

Pie chart


Policy requirement for measles vaccination
not required
no data
required

Scatterplot

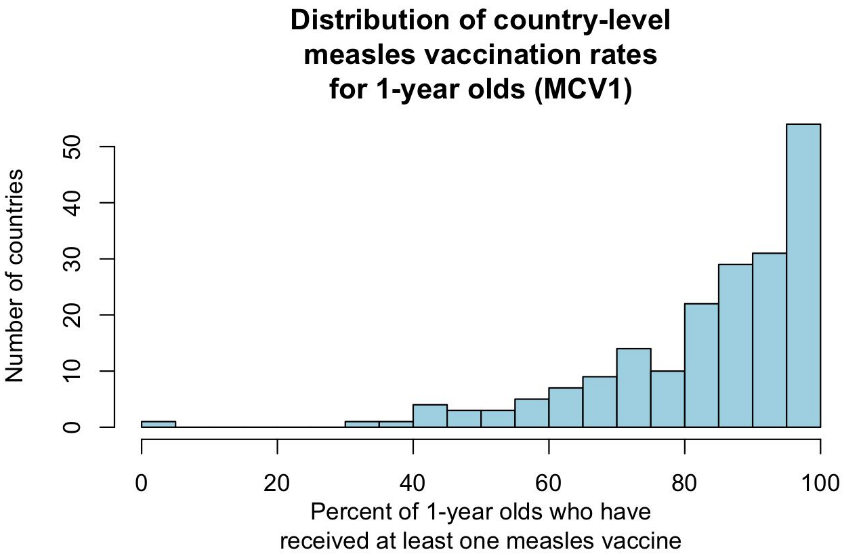


10 minute break

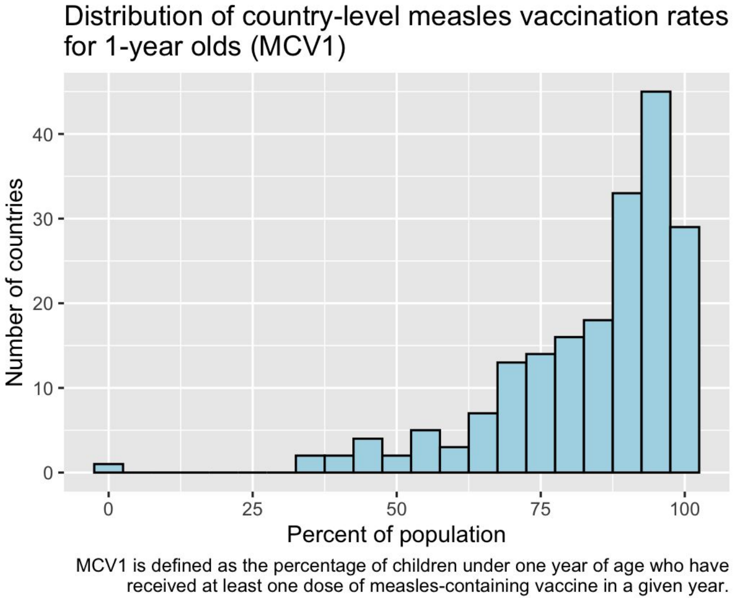
Plots in ggplot2
(code in github for live demo)

Histogram

● One variable: *continuous*



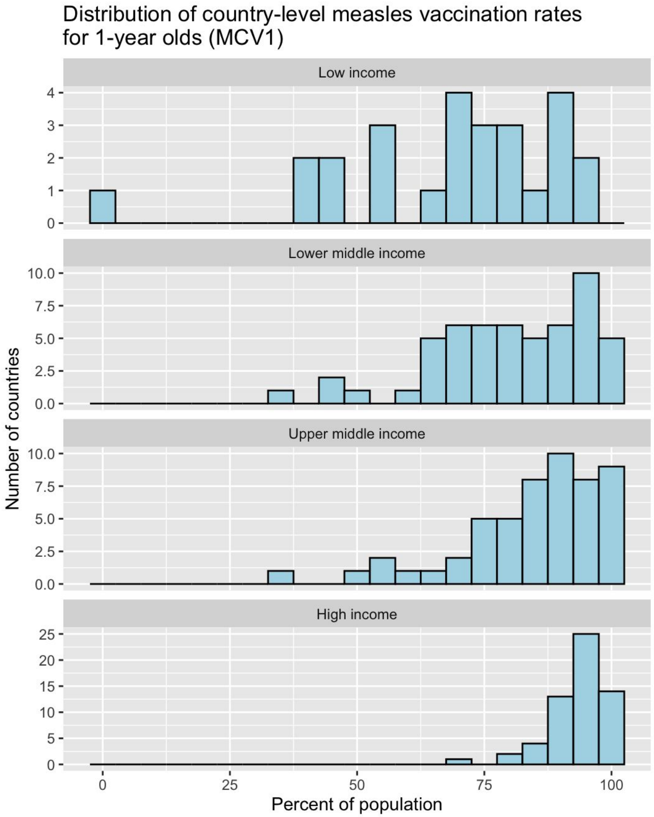
Base R



ggplot

Histogram by group

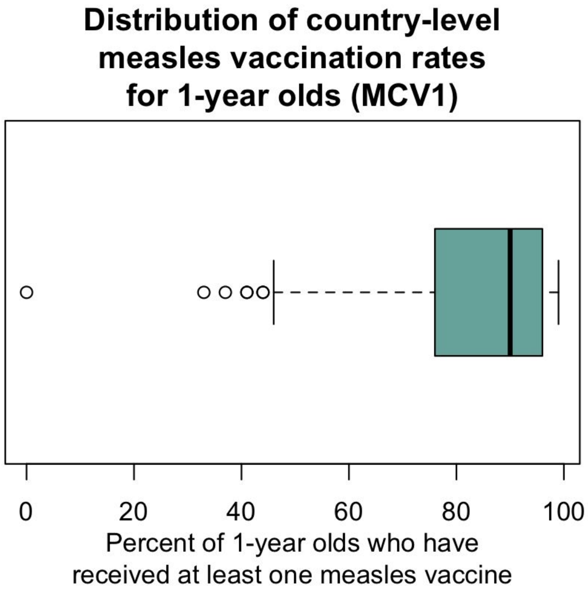
Two variables:
continuous & categorical



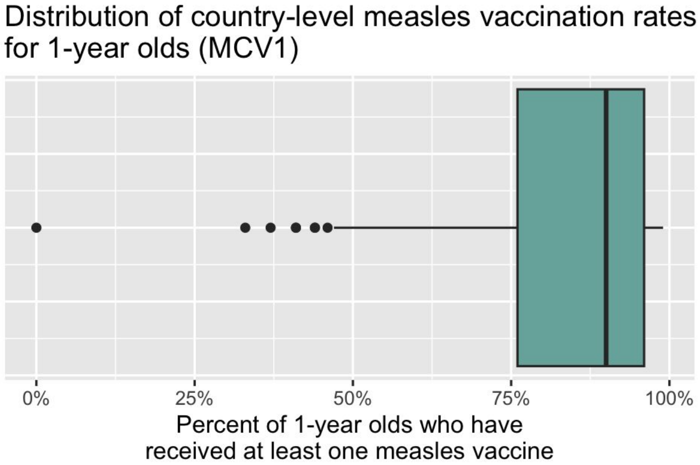
MCV1 is defined as the percentage of children under one year of age who have received at least one dose of measles-containing vaccine in a given year.

Boxplot

● One variable: *continuous*



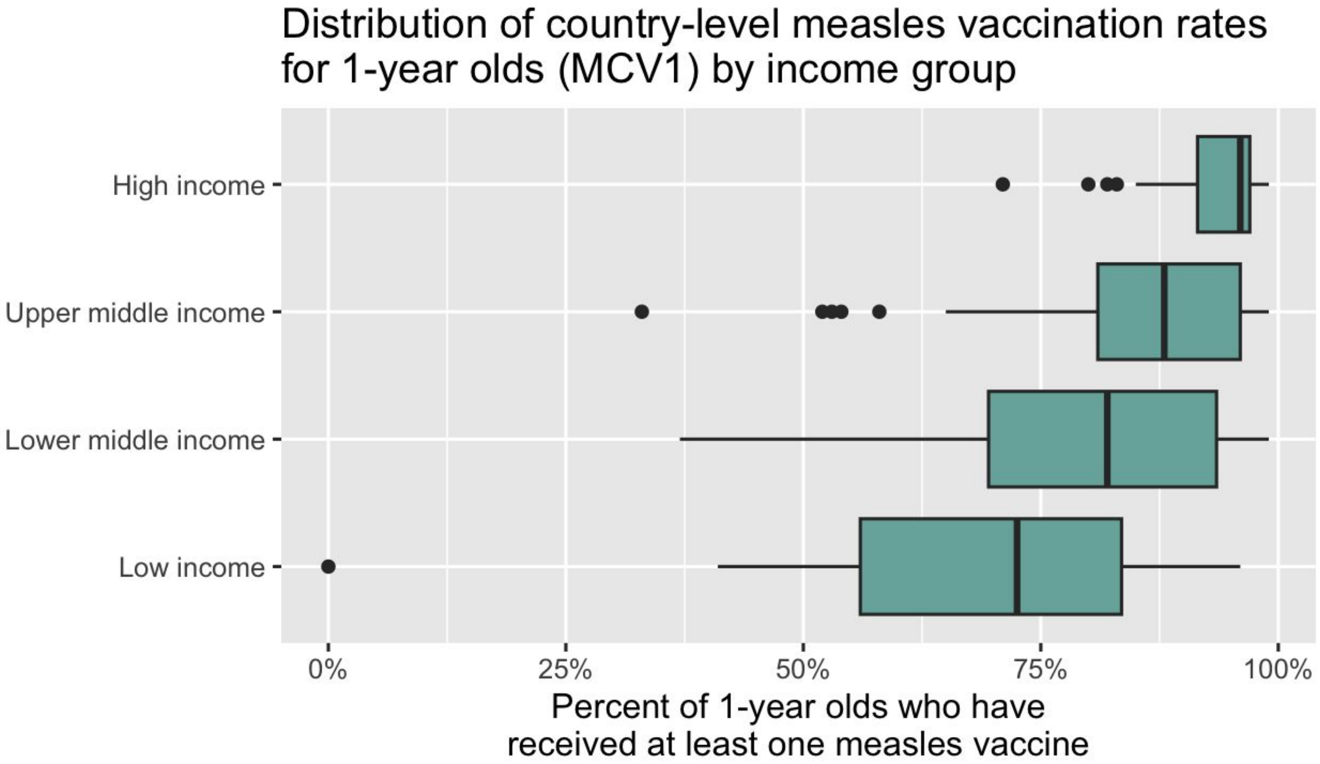
Base R



ggplot

Boxplots by group

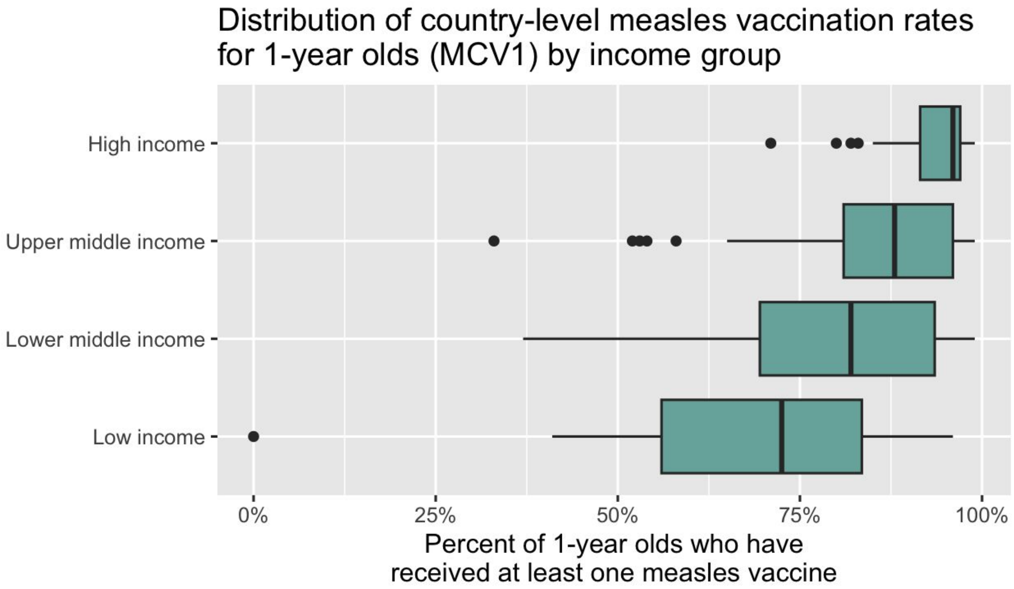
● ● Two variables: *continuous & categorical*

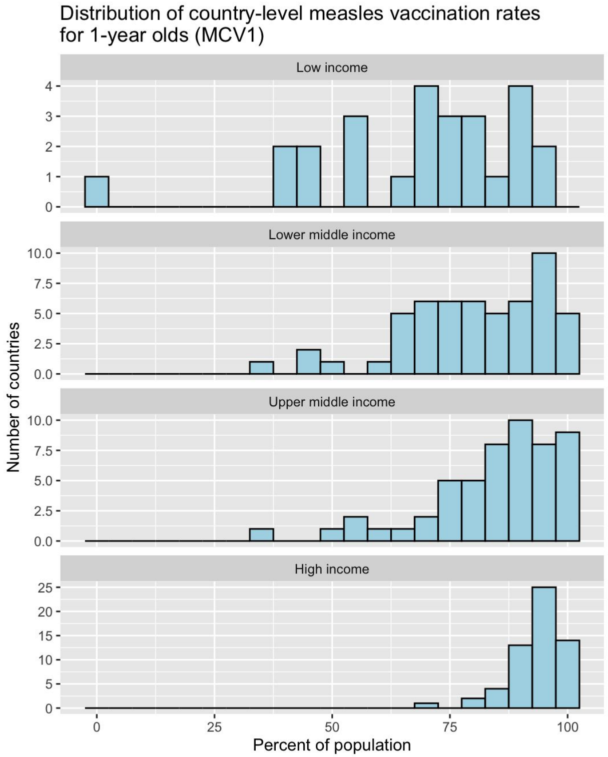


Compare and contrast

Two variables: *continuous & categorical*

These plots show similar information with the same data. Which does each emphasize? Can you think of a situation where you would want to choose one over the other?





MCV1 is defined as the percentage of children under one year of age who have received at least one dose of measles-containing vaccine in a given year.

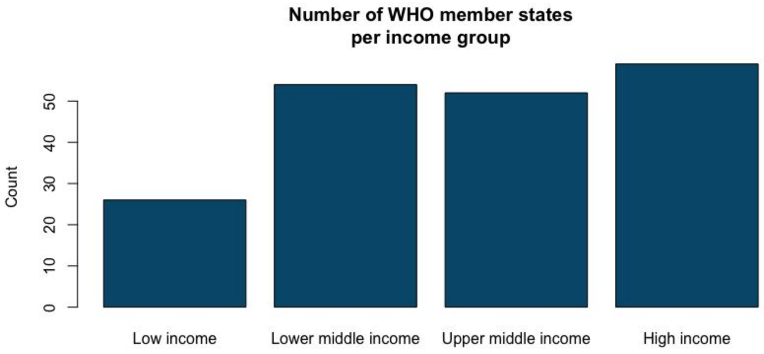
Your turn

What other ways could boxplots help us better understand vaccination coverage?

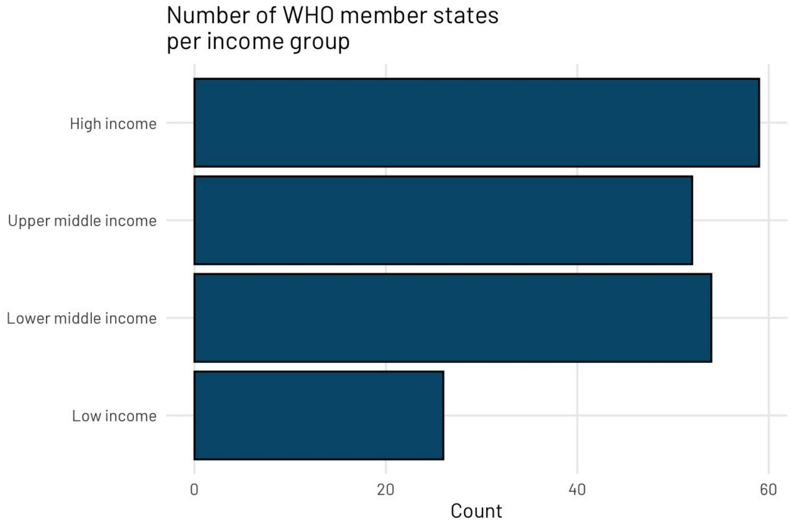
- What else might we want to group by?
- What questions would we answer by looking at those data?
- What would you expect to see?

Bar chart

● One variable: *categorical*



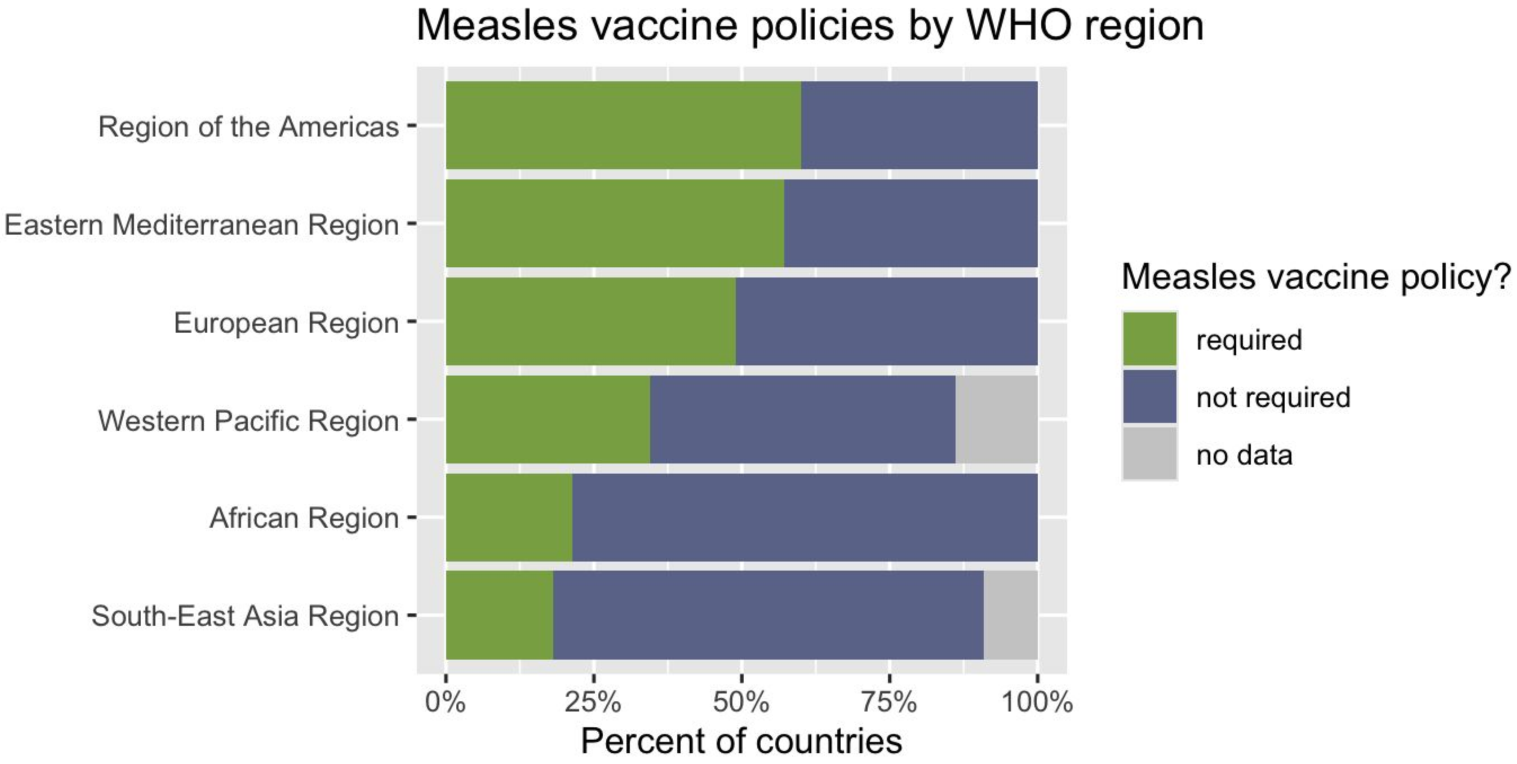
Base R



ggplot

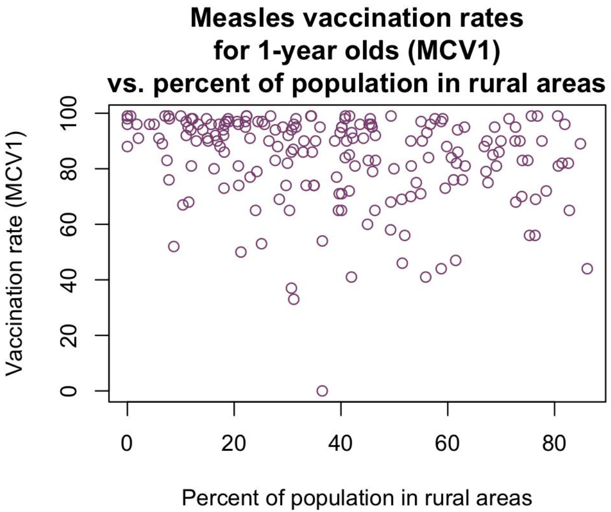
Stacked barchart

Two variables:
categorical &
categorical

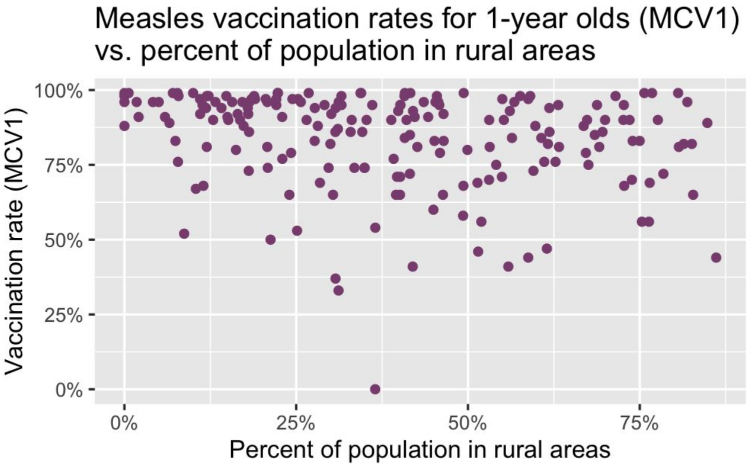


Scatterplot

●● Two variables: *continuous & continuous*



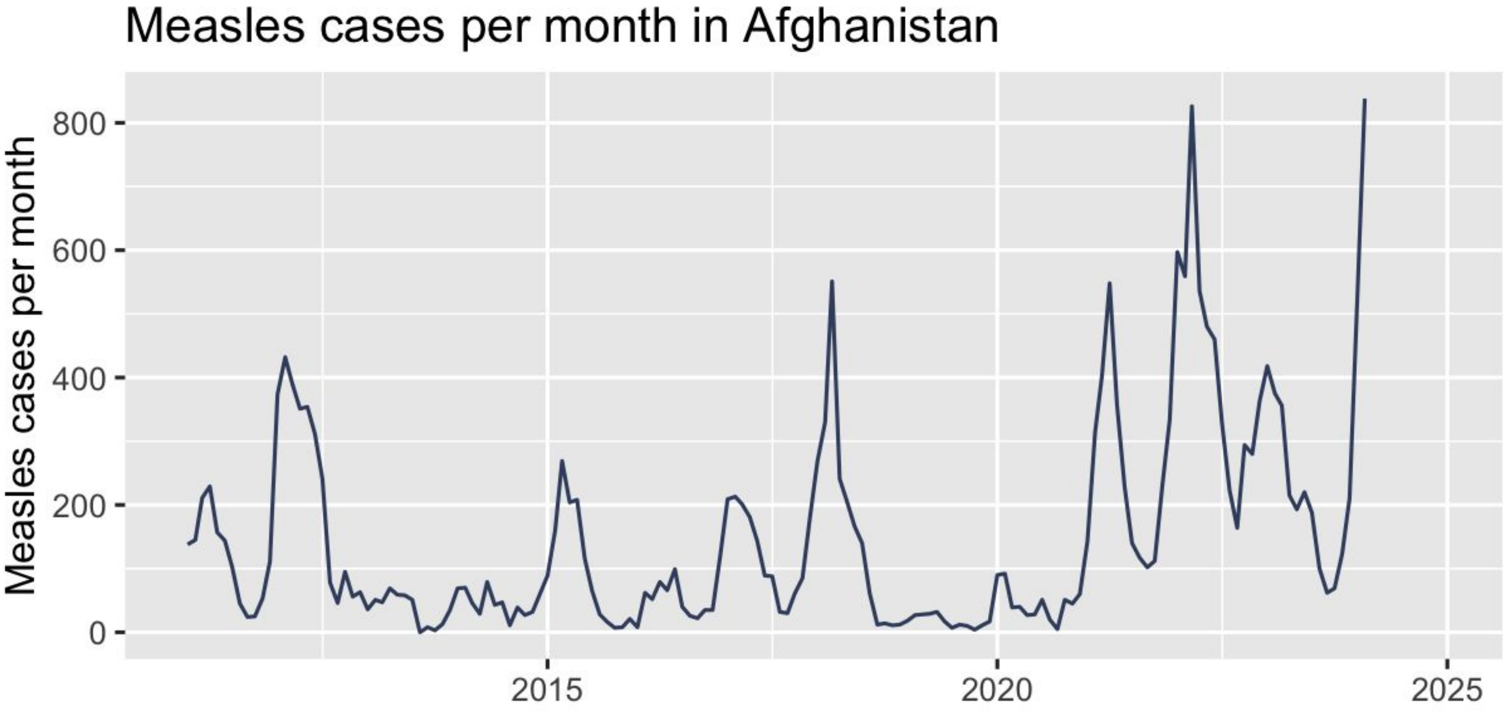
Base R



ggplot

Line chart

● One variable: *temporal & continuous*



Your turn

It’s both an art and a science figuring out which types of data work best with which types of plot. It depends on what question you are trying to answer, and how your data are structured. Please brainstorm data you could use, in the existing dataset, to generate a **histogram,** a **scatterplot**, a **barplot,** and a **line chart**

- What data points?
- What questions are answered by these plots?

Feel free to explore either using pen and paper or using the code we already have.

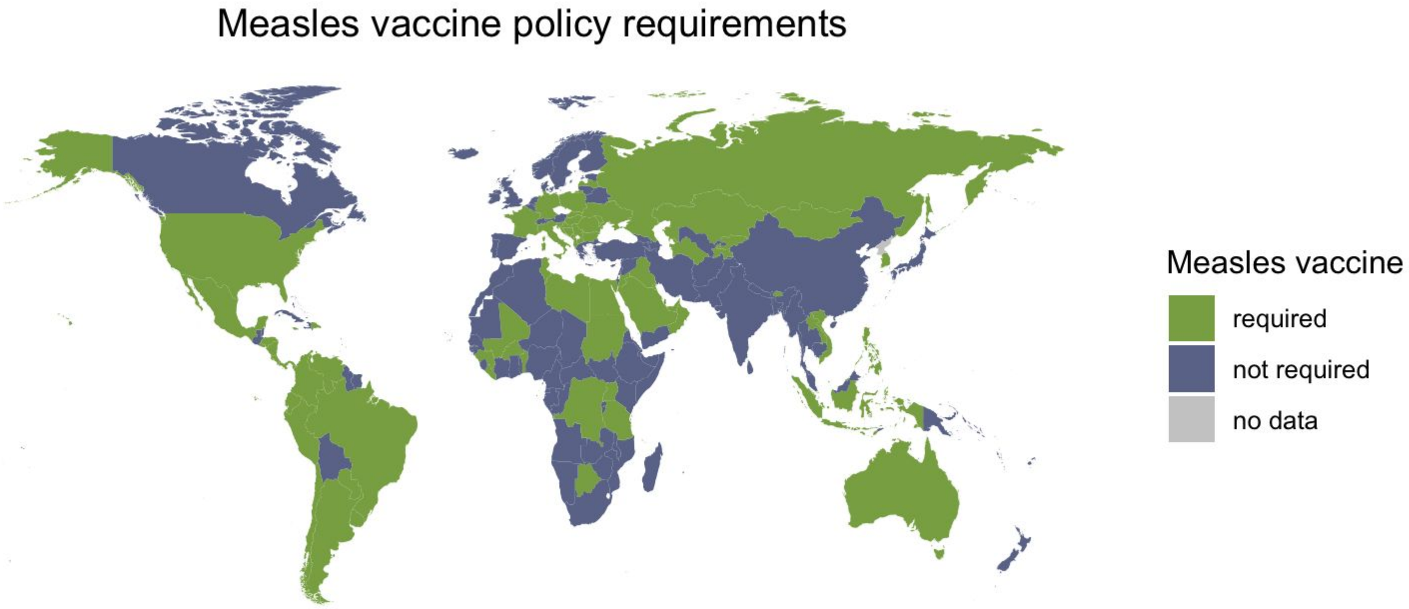
Live problem solving

Let’s whiteboard the questions we came up with earlier.

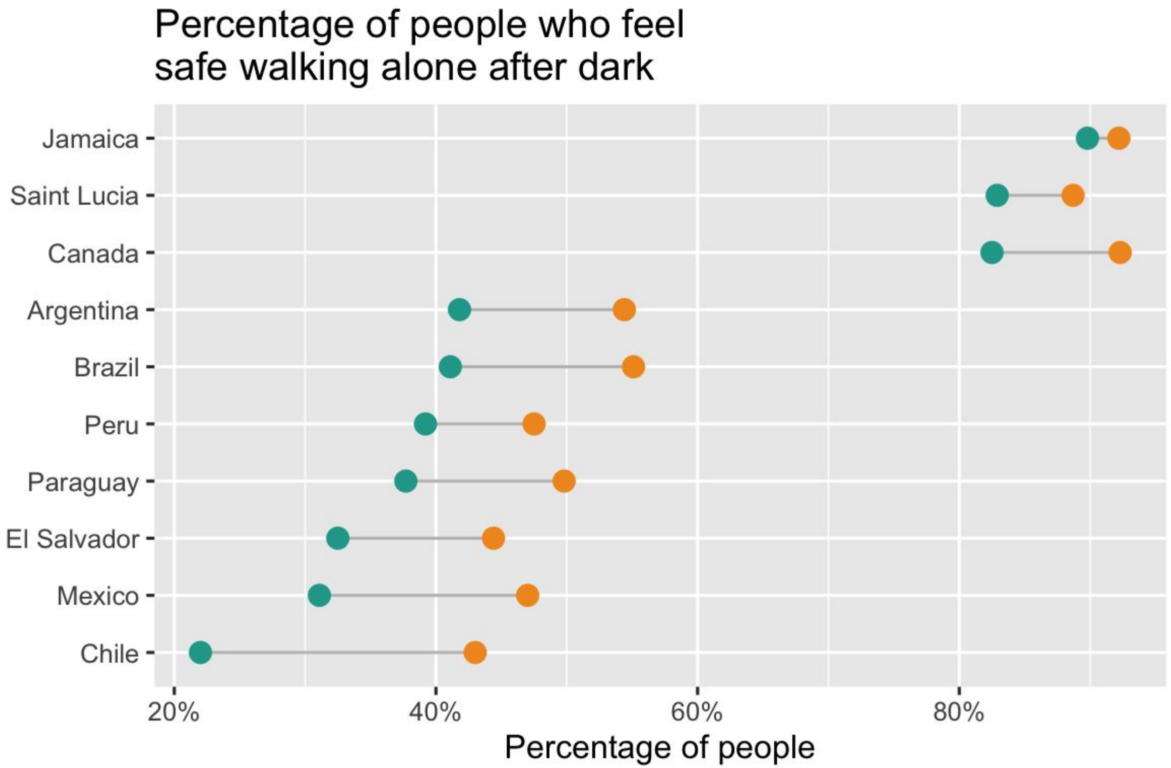
Can we make some visualizations together to help explore these questions?

Bonus charts
(code in github for live demo)

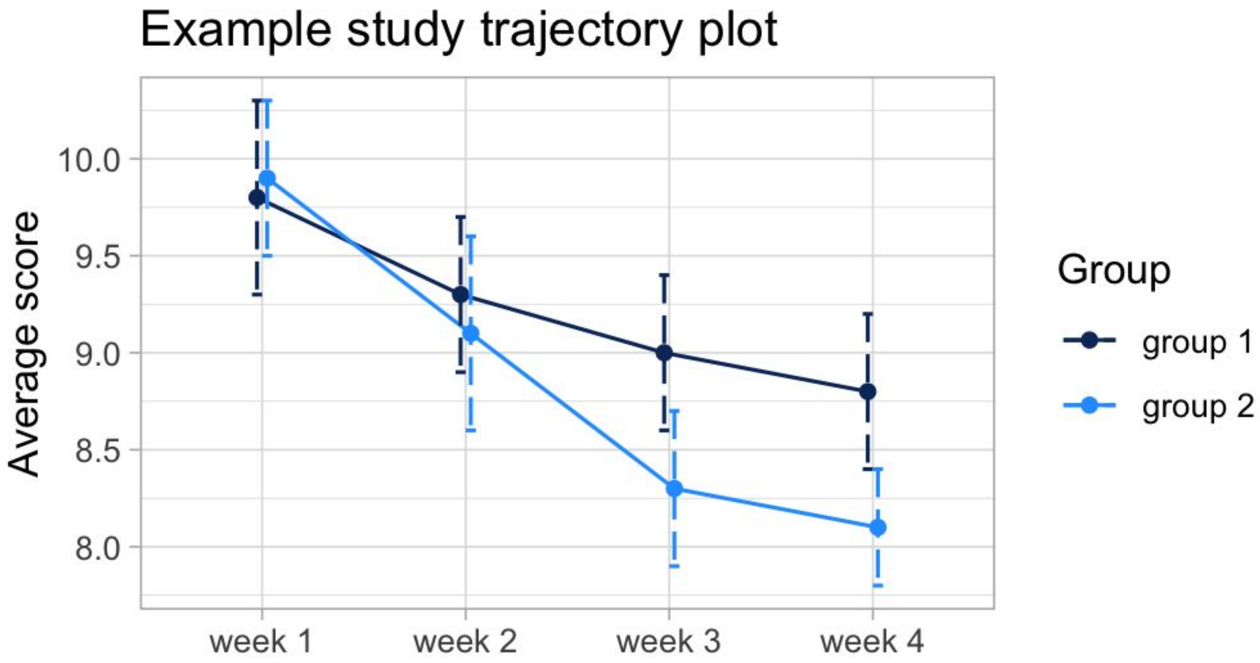
Maps


Measles vaccine policy requirements
Measles vaccine
🟩 required
▨ not required
□ no data

Cleveland dot plot



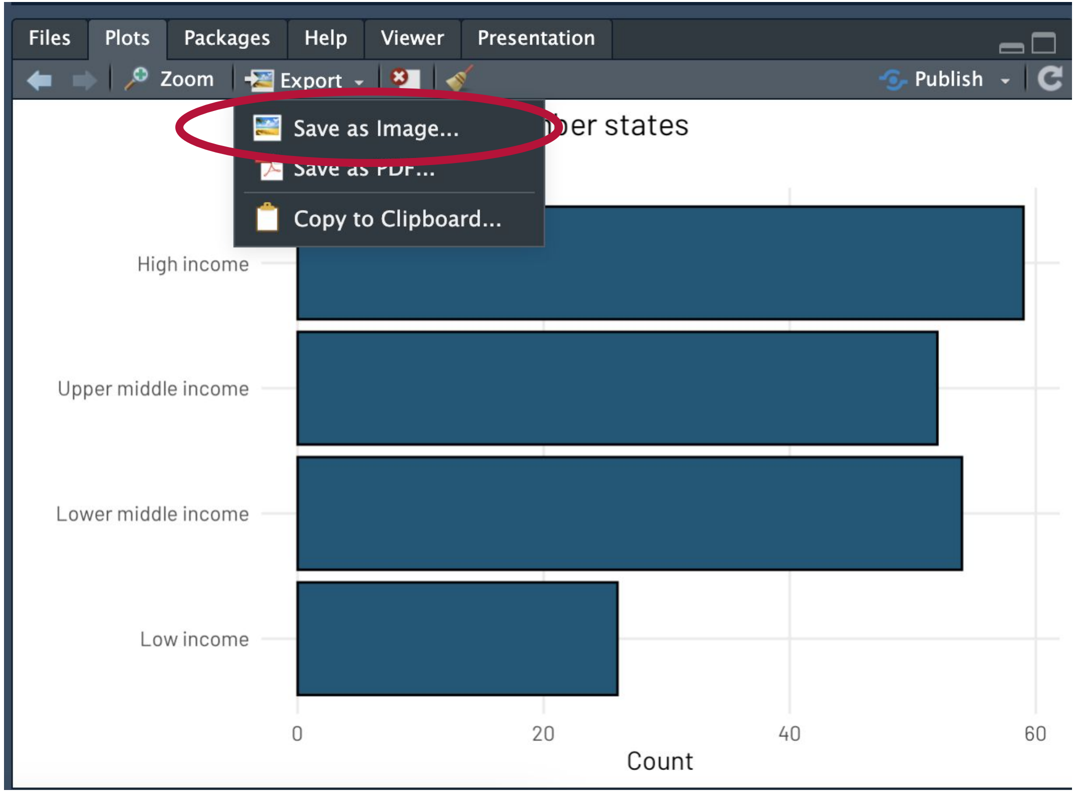
Error bars



Data are notional and do not reflect actual study data

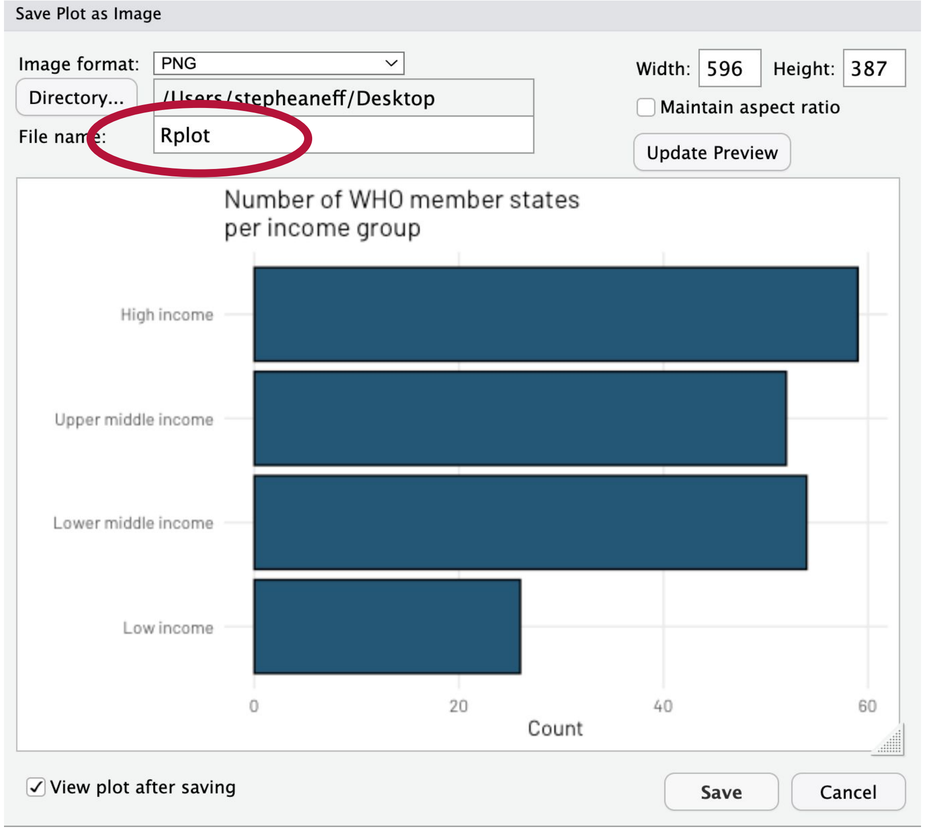
Exploring plots using RStudio

Exporting plots using R Studio



Exporting plots using R Studio

Name your file



Save Plot as Image

Image format: PNG

Directory... /Users/stepheaneff/Desktop

File name: Rplot

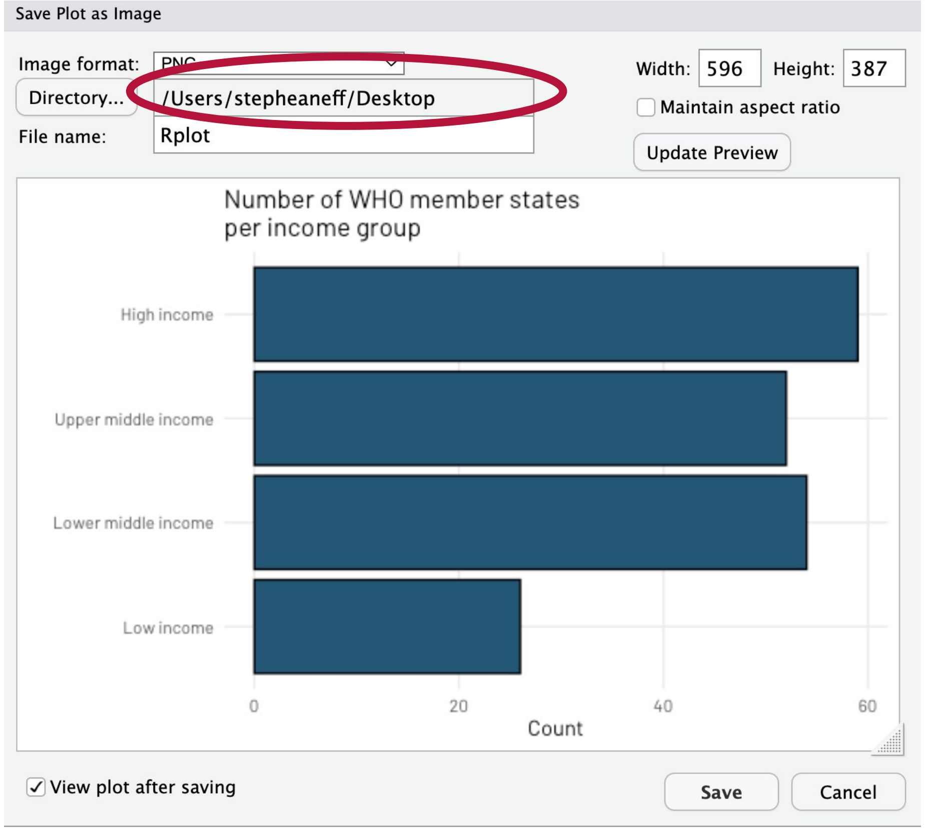
Width: 596 Height: 387

□ Maintain aspect ratio

☑ View plot after saving

Exporting plots using R Studio

Specify where you want to save it



Save Plot as Image

Image format: PNG

Directory... /Users/stepheaneff/Desktop

File name: Rplot

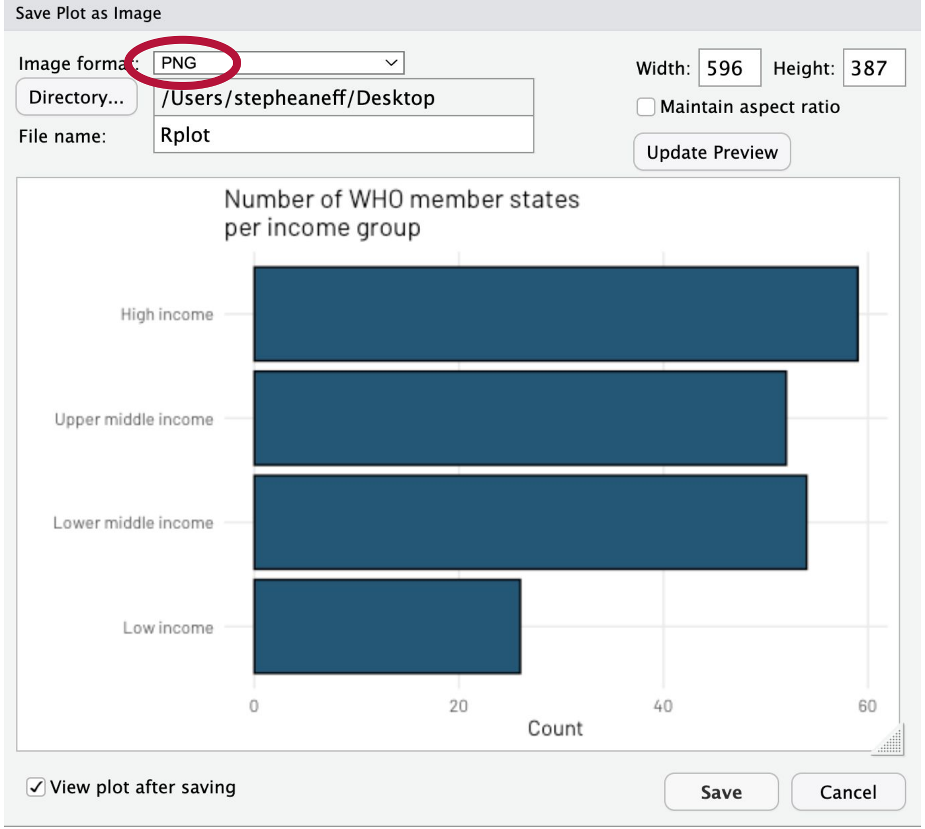
Width: 596 Height: 387

□ Maintain aspect ratio

☑ View plot after saving

Exporting plots using R Studio

Choose an image format (png, jpg, etc)



Save Plot as Image

Image format: PNG

Directory... /Users/stepheaneff/Desktop

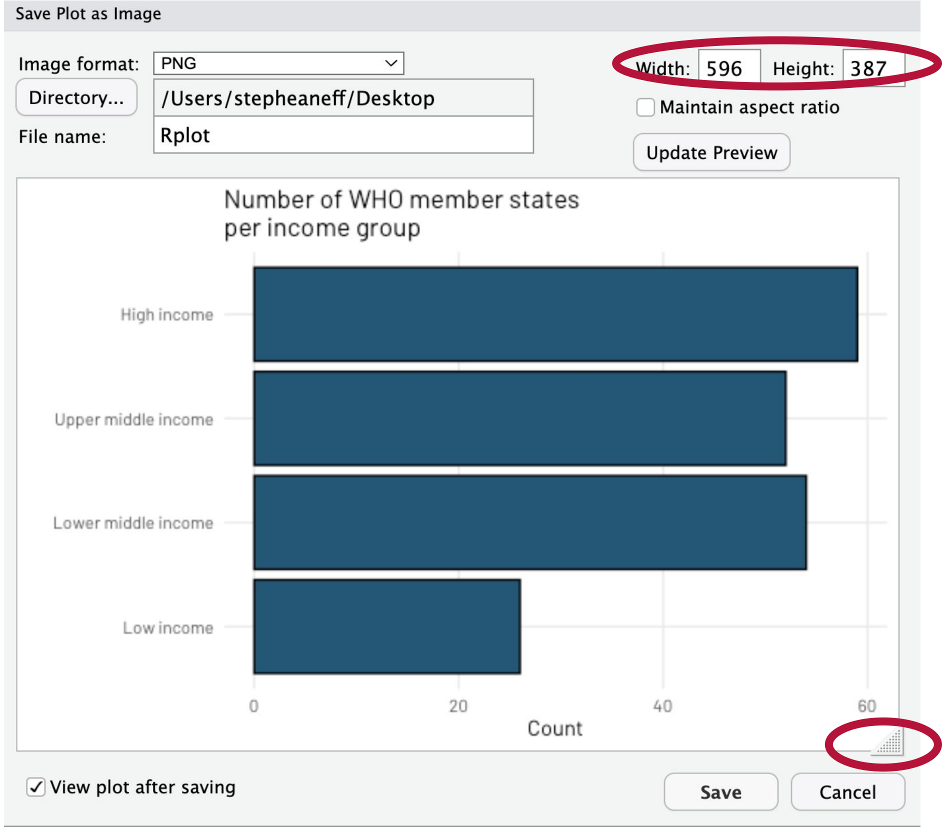
File name: Rplot

Width: 596 Height: 387

☐ Maintain aspect ratio

☑ View plot after saving

Exporting plots using R Studio



Save Plot as Image

Image format: PNG

Directory... /Users/stepheaneff/Desktop

File name: Rplot

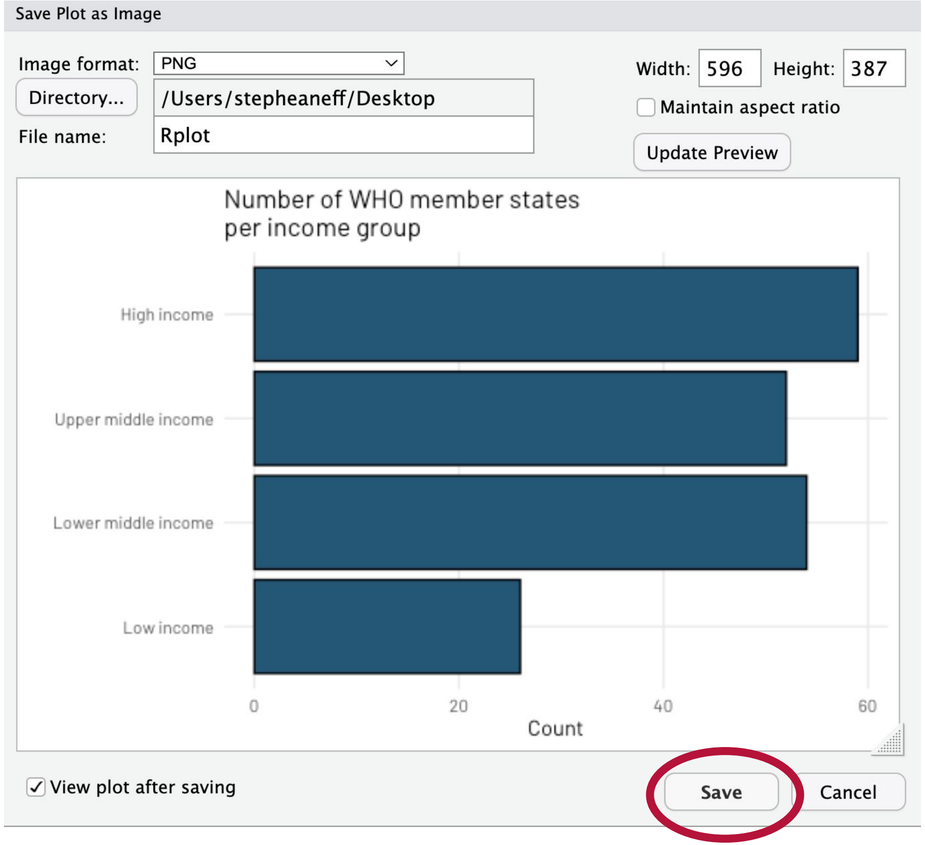
Width: 596 Height: 387

☐ Maintain aspect ratio

☑ View plot after saving

Updating sizing/ratios

Exporting plots using R Studio



Save Plot as Image

Image format: PNG

Directory... /Users/stepheaneff/Desktop

File name: Rplot

Width: 596 Height: 387

□ Maintain aspect ratio

☑ View plot after saving

Save and export

Hands-on exploration

Save your work on github

OPTIONAL homework

create your own new data visualization

- Based on the code we worked through in class today, can you create a new data visualization “from scratch”?
- I recommend starting with the code we’ve already written in github, but swapping out the data that we’re showing. That way, you have some guideposts to show you where to start, then you can branch off and explore some more on your own.

Plan for tomorrow

Designing data visualizations

- My *absolute favorite* lesson on data visualization
- Learn a step-by-step process for creating great data visualizations
- Understand your audience and your goals when visualizing data
- Design some fun and beautiful data visualizations
- Get creative and explore some new skills in R

Thank you!

See you tomorrow.

Please come with a fully charged laptop.