

# 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
- Create your first data visualization in R
- Identify options for visualization in R, including ggplot2
- Get creative and have fun exploring datasets

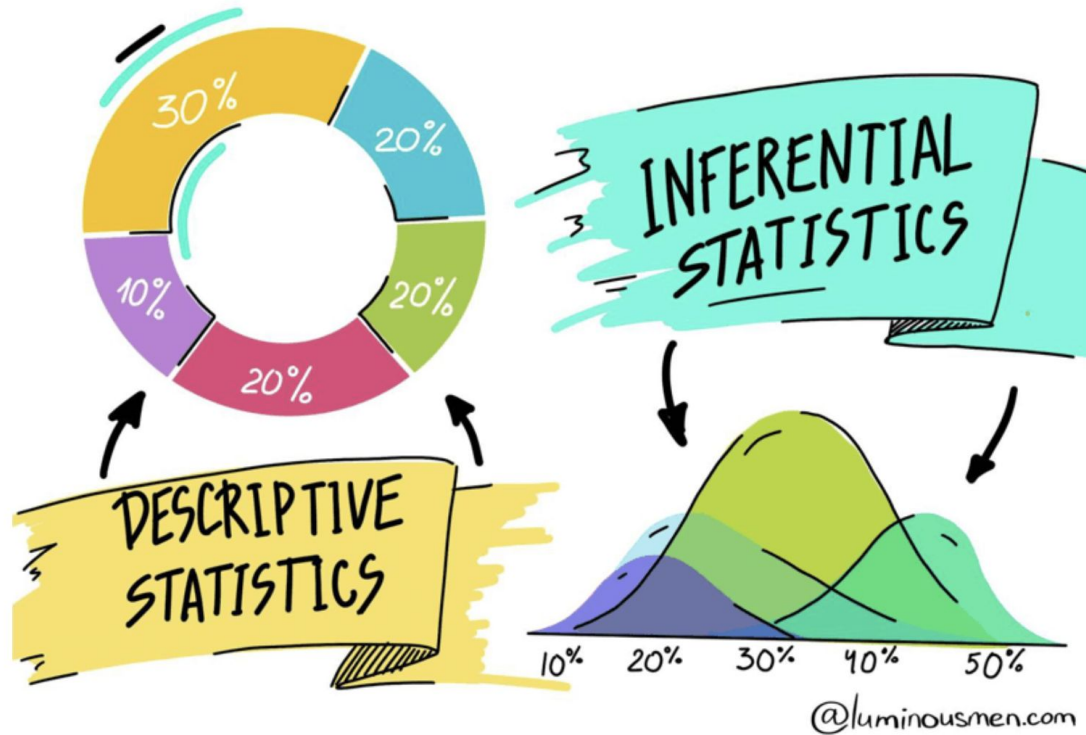
# Descriptive statistics

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



## 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.

---

**Your turn!**

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

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

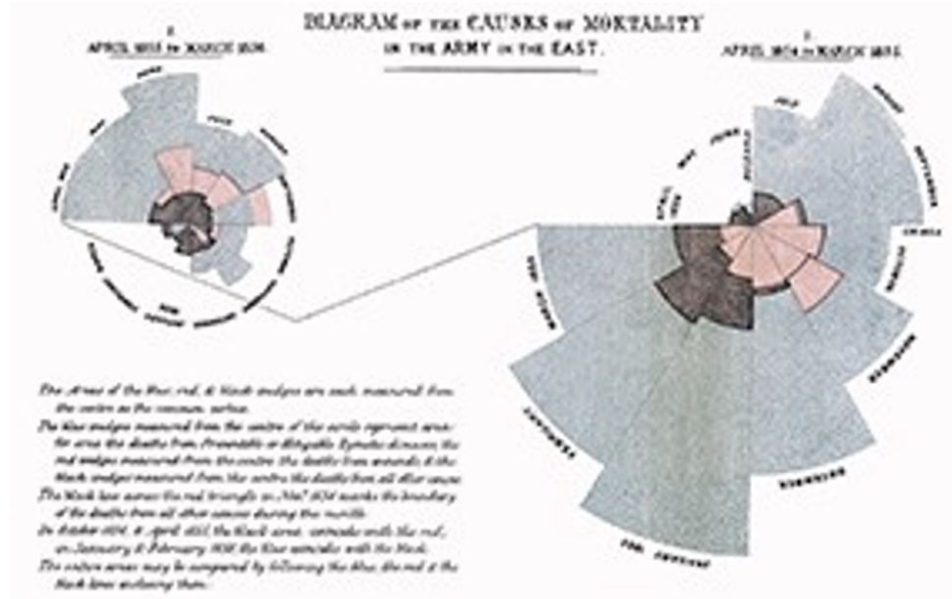


# Calculating summary statistics in R

*(code in github for live demo)*

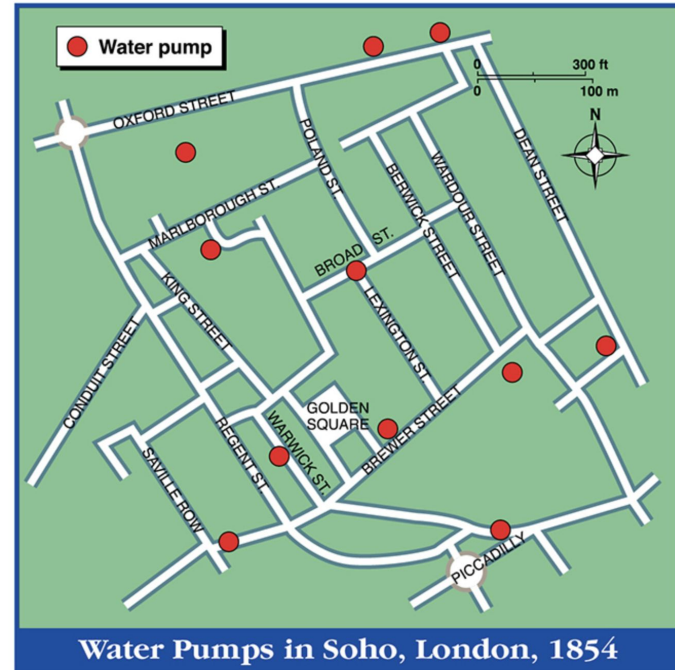
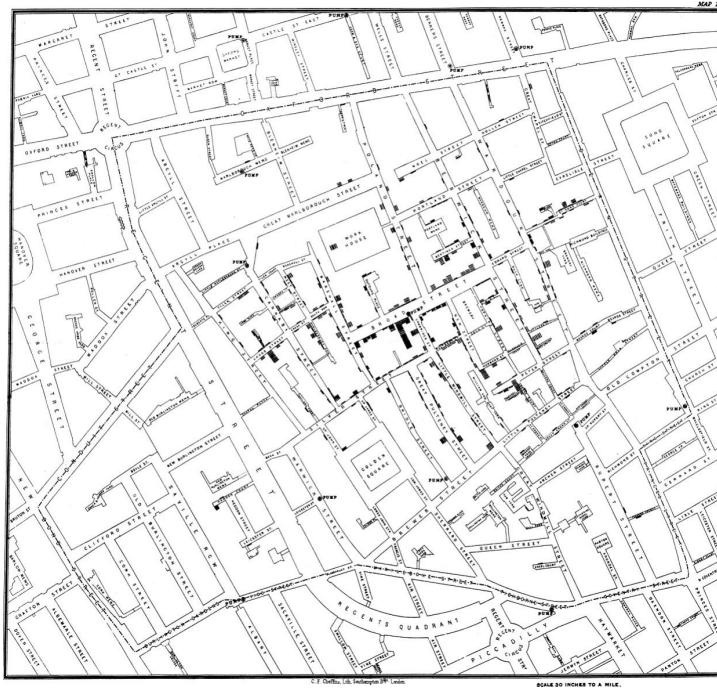
# Goals of data visualization

# Florence Nightingale's "Diagram of the causes of mortality in the army in the East".



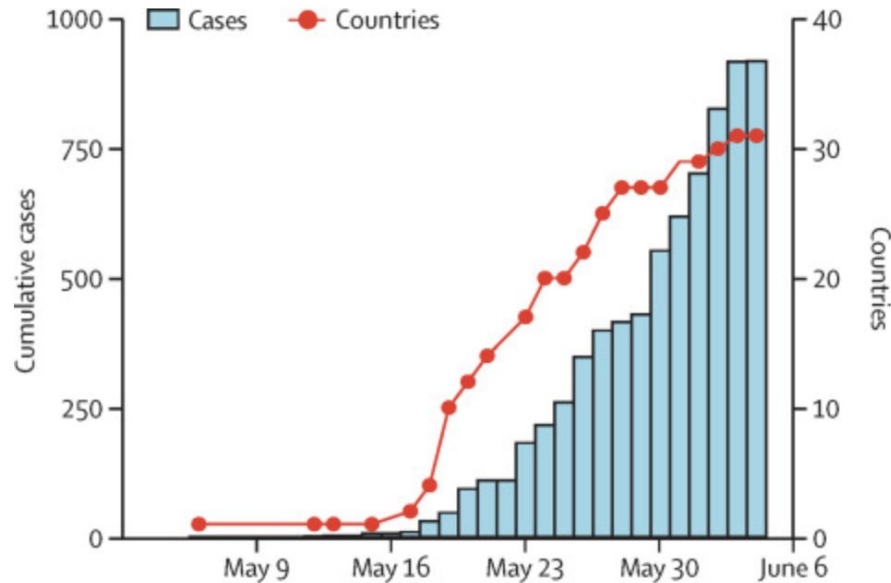
Significance, Volume 17, Issue 2, April 2020, Pages 26–30,  
<https://doi.org/10.1111/1740-9713.01376>  
 Crimean War deaths and cause of deaths, polar area chart

# John Snow's cholera map



National Geographic. Mapping a London Epidemic. Cholera deaths.  
<https://education.nationalgeographic.org/resource/mapping-a-london-epidemic/>

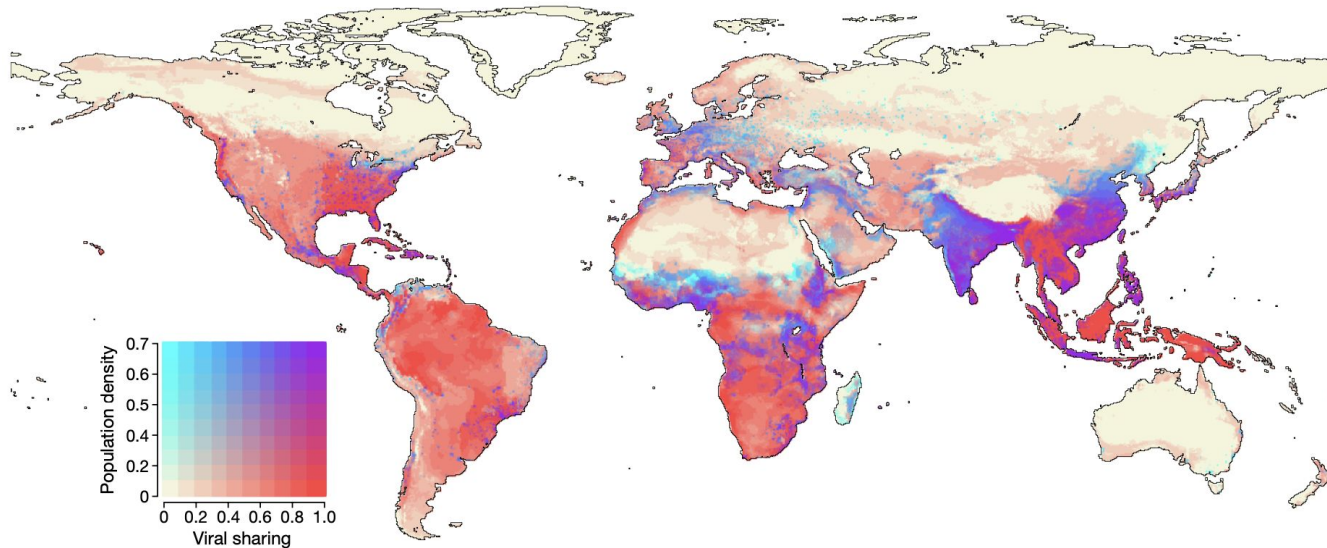
# Monkeypox global caseload



**Figure** Rapid expansion of the 2022 monkeypox outbreak

Kraemer MU, Tegally H, Pigott DM, Dasgupta A, Sheldon J, Wilkinson E, Schultheiss M, Han A, Oglia M, Marks S, Kanner J. Tracking the 2022 monkeypox outbreak with epidemiological data in real-time. *The Lancet Infectious Diseases*. 2022 Jul 1;22(7):941-2.

# Population density and viral shedding projections



**Fig. 4 | Novel viral sharing events coincide with human population centres.** In 2070 (SSP 1–RCP 2.6; climate only), human population centres in equatorial Africa, south China, India and southeast Asia will overlap with projected

hotspots of cross-species viral transmission in wildlife. Both variables were linearly rescaled to 0 to 1. The results were averaged across nine GCMs.

Carlson CJ, Albery GF, Merow C, Trisos CH, Zipfel CM, Eskew EA, Olival KJ, Ross N, Bansal S. Climate change increases cross-species viral transmission risk. *Nature*. 2022 Jul;607(7919):555-62.

# 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)

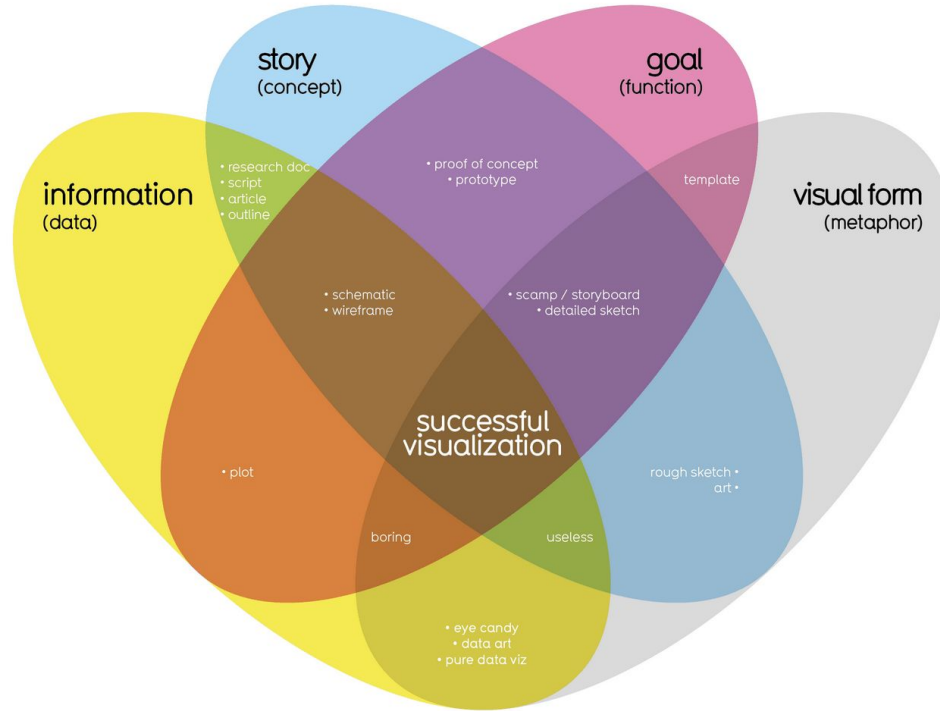
- **Understand** what is happening in a new dataset or situation
- **Communicate** information quickly and rapidly
- **Make decisions** based on an understanding of what is currently known

# What makes a data visualization good?

rollover for more detail

## What Makes a Good Visualization?

explicit (implicit)



David McCandless  
InformationIsBeautiful.net

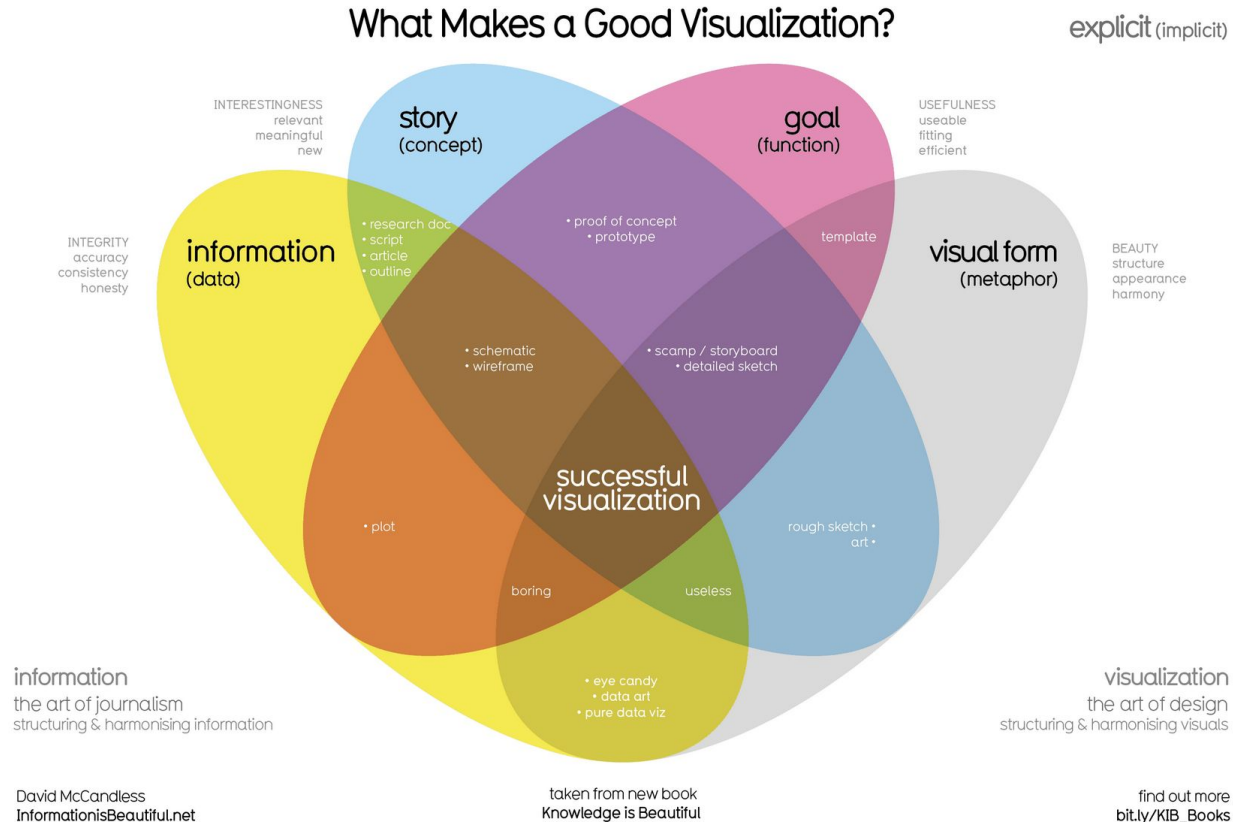
taken from new book  
Knowledge is Beautiful

find out more  
[bit.ly/KIB\\_Books](http://bit.ly/KIB_Books)

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

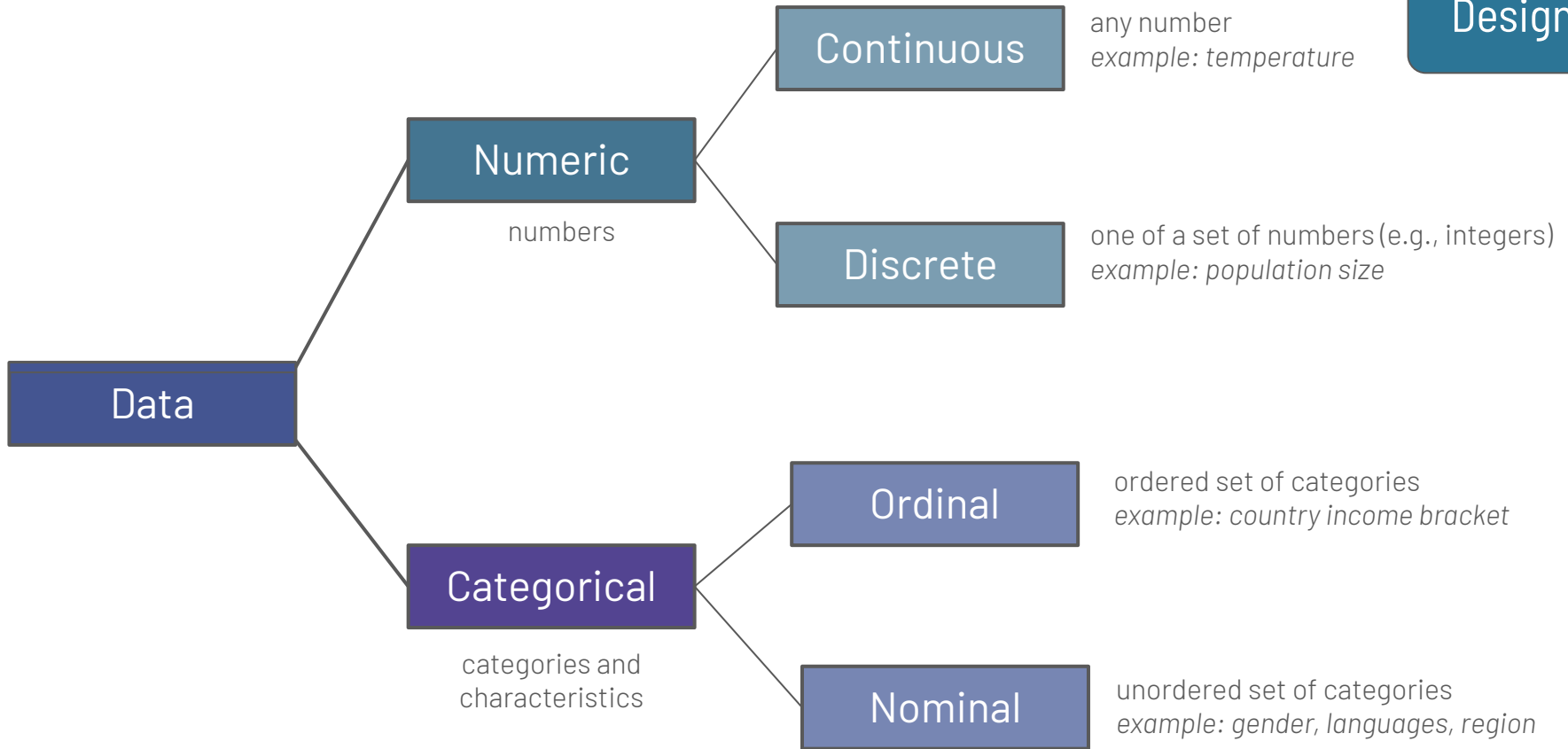


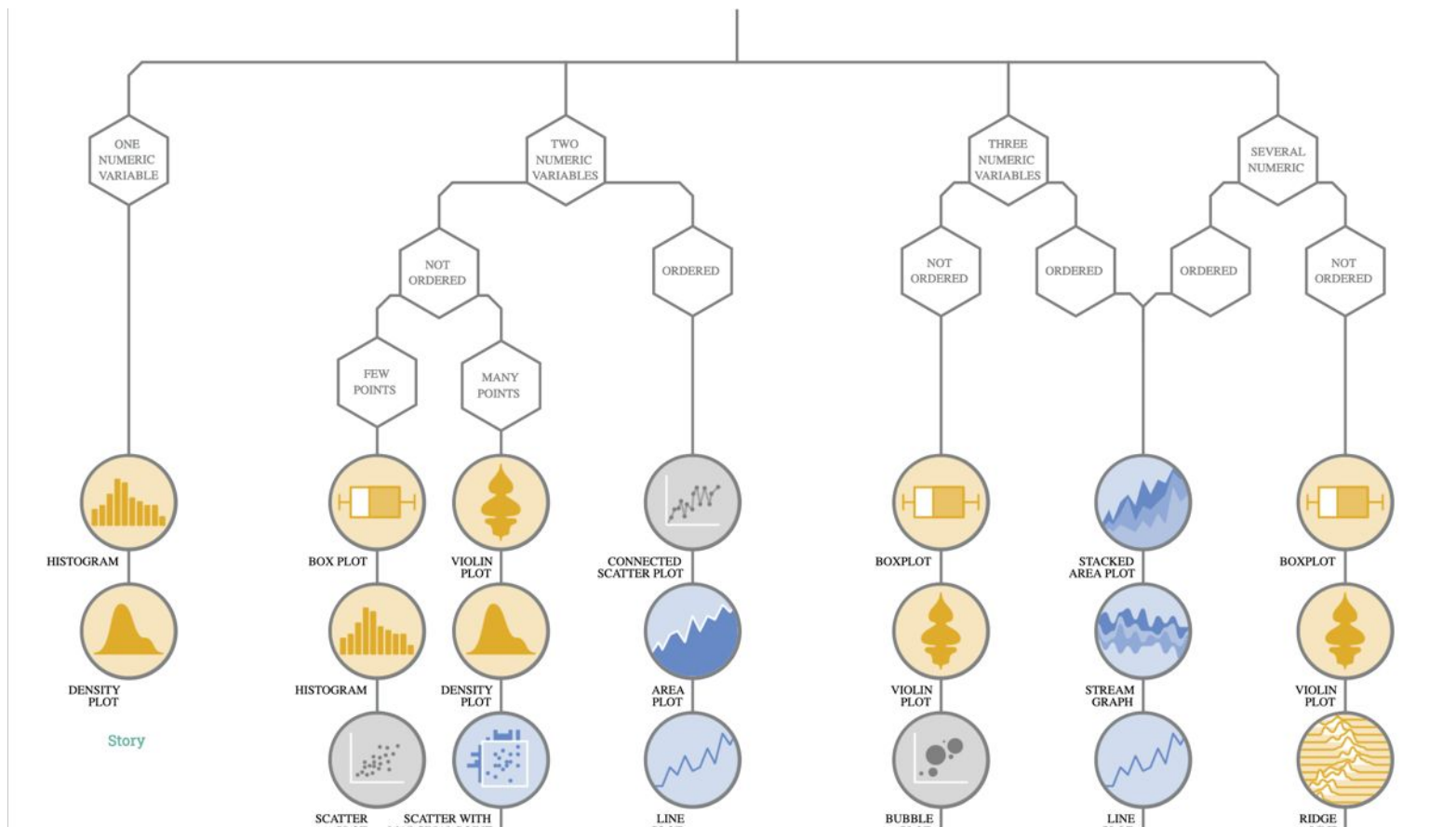
# What makes a data visualization good?



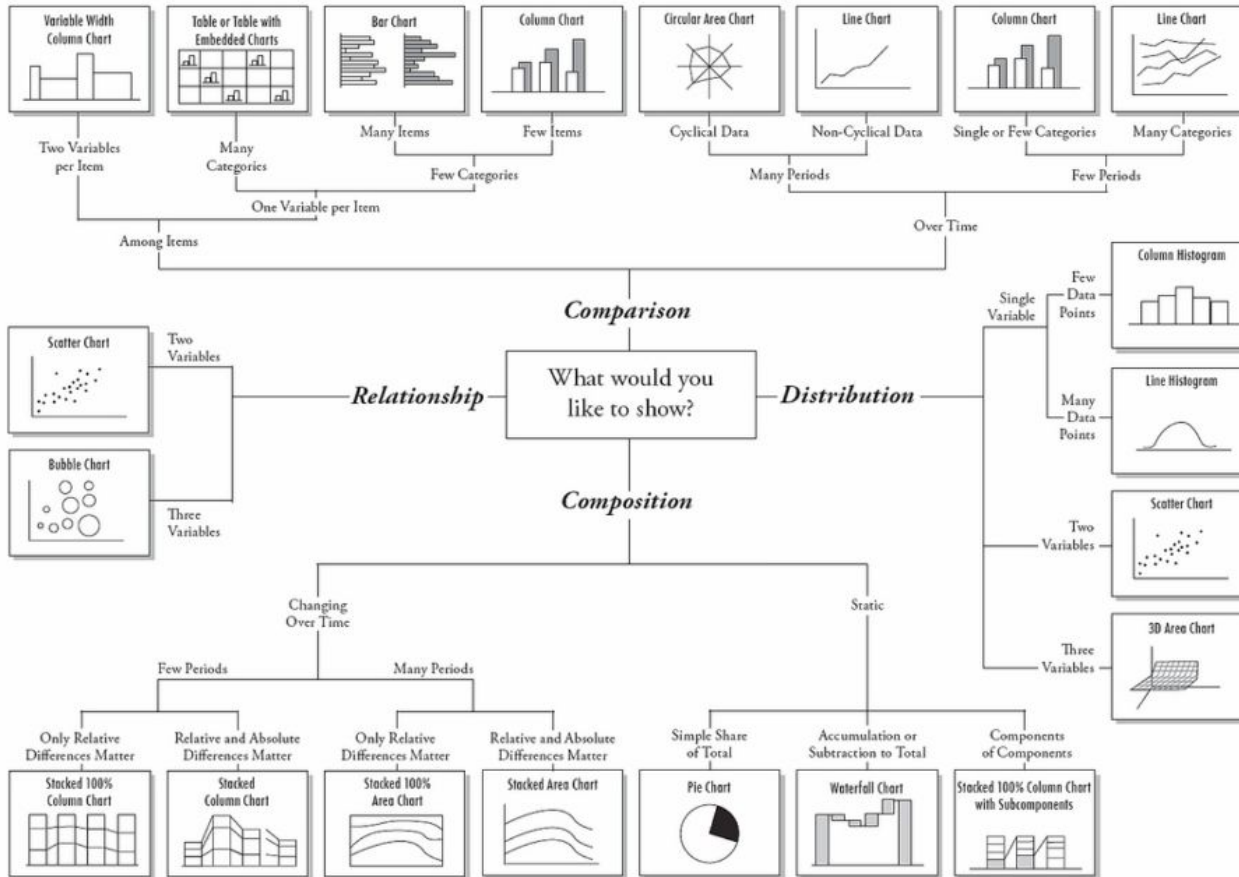
# Choosing a data visualization

(we'll talk more about this tomorrow)





# Chart Suggestions—A Thought-Starter



Graphic by Andrew Abela

[https://extremepresentation.typepad.com/blog/2006/09/choosing\\_a\\_good.html](https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html)

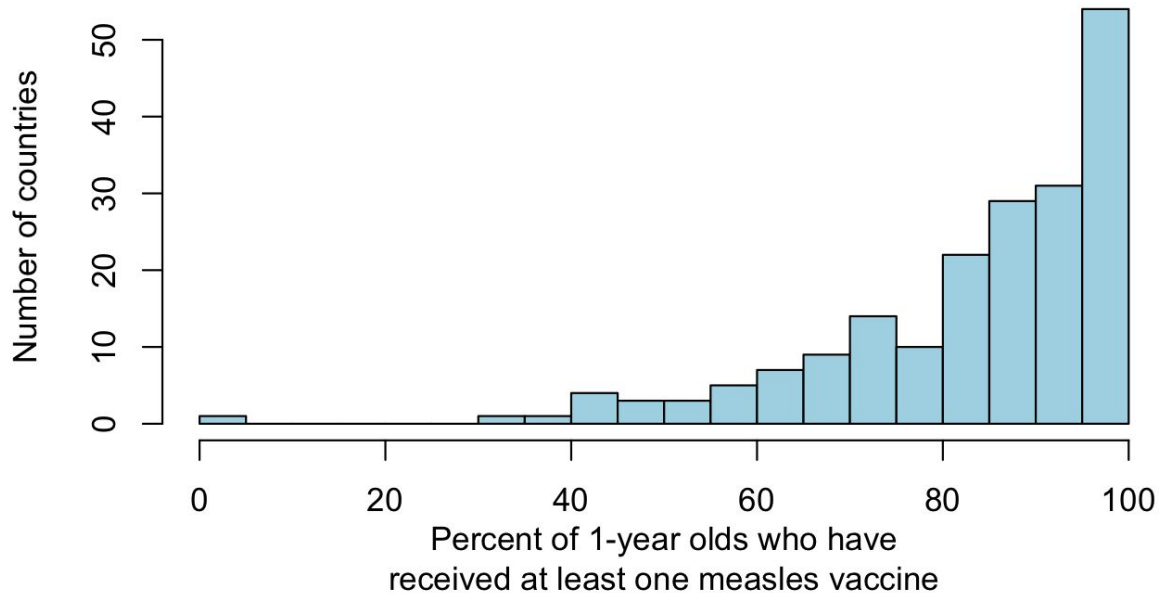
© 2006 A. Abela — a.vabela@gmail.com

# Plots in base R

*(code in github for live demo)*

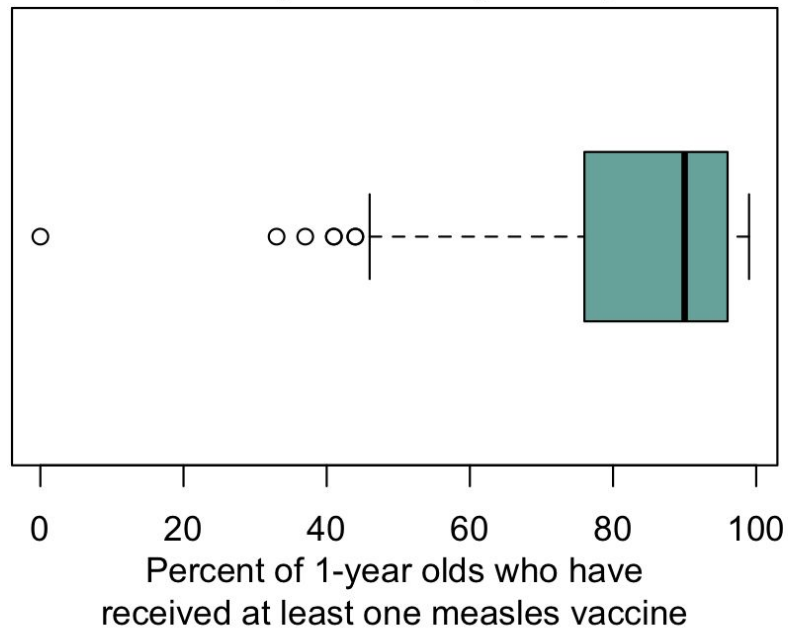
# Histogram

**Distribution of country-level  
measles vaccination rates  
for 1-year olds (MCV1)**



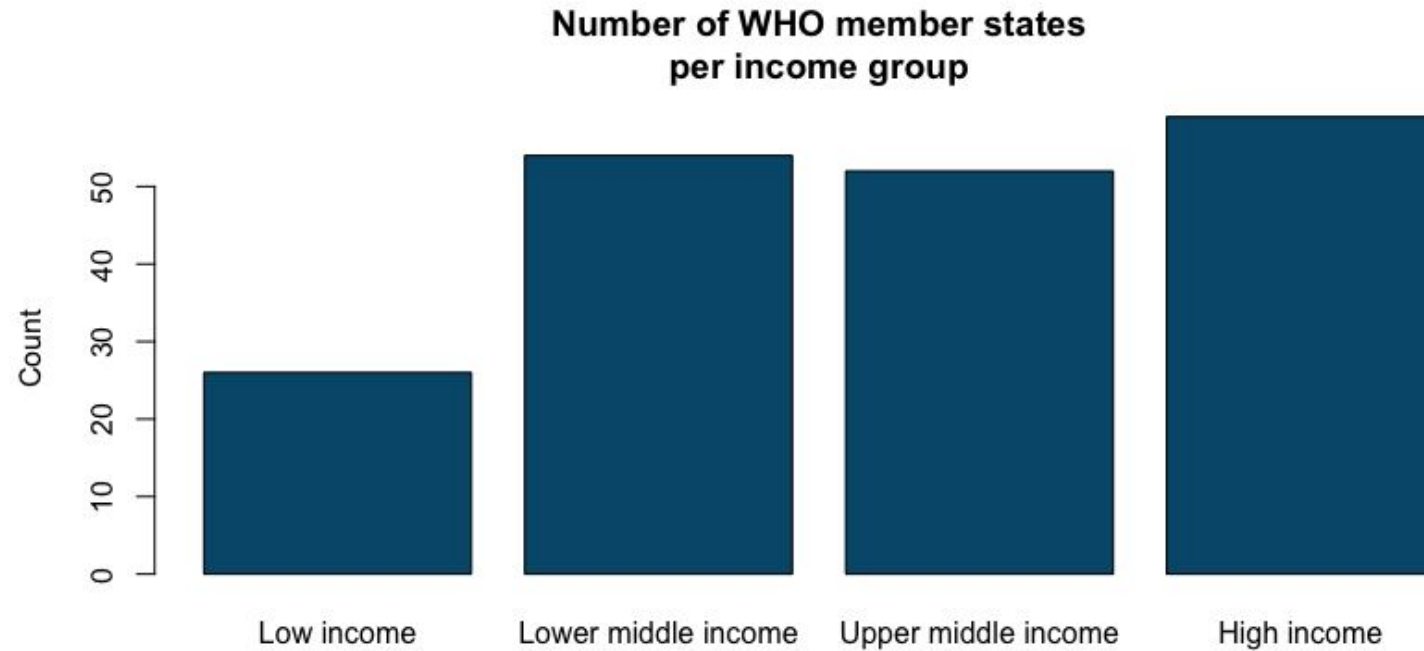
# Boxplot

**Distribution of country-level  
measles vaccination rates  
for 1-year olds (MCV1)**



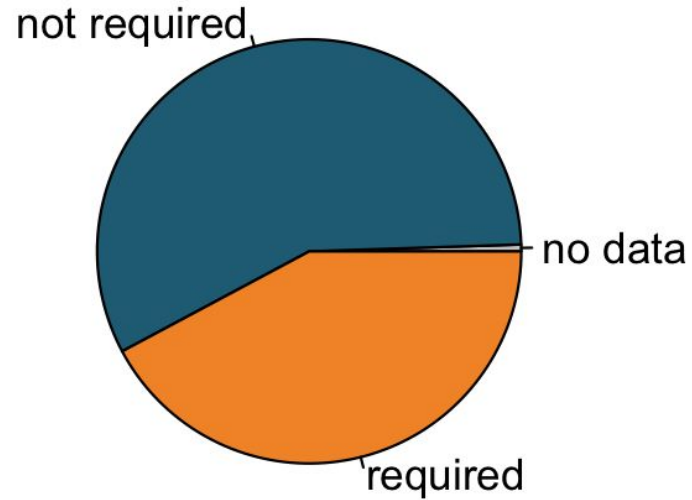


# Barchart

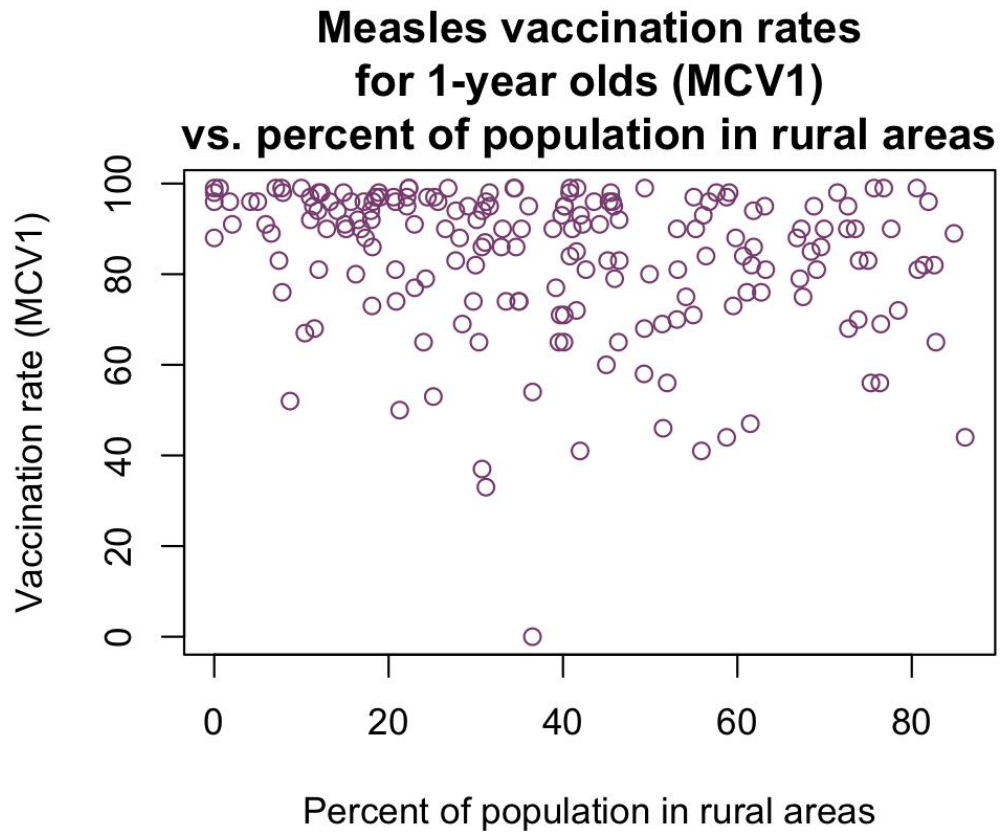


# Pie chart

## Policy requirement for measles vaccination



# Scatterplot



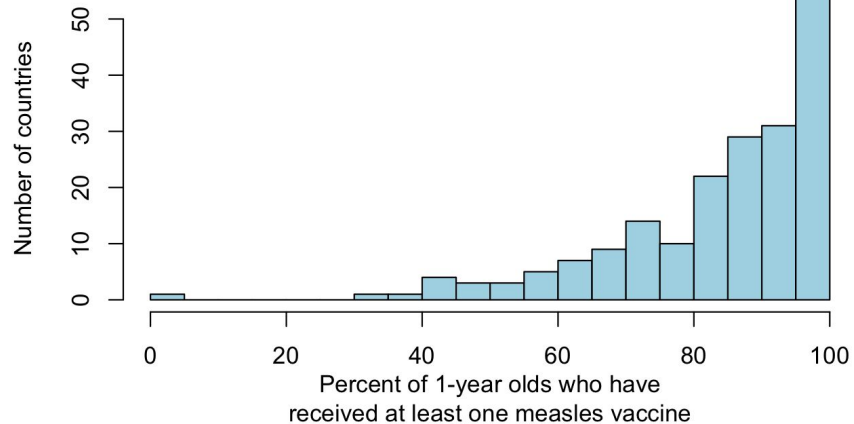
**10 minute break**

# Plots in ggplot2

*(code in github for live demo)*

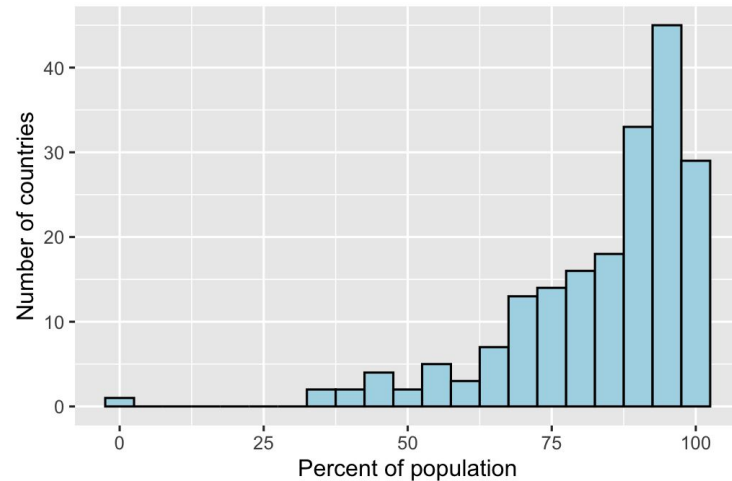
# Histogram

**Distribution of country-level  
measles vaccination rates  
for 1-year olds (MCV1)**



Base R

**Distribution of country-level measles vaccination rates  
for 1-year olds (MCV1)**

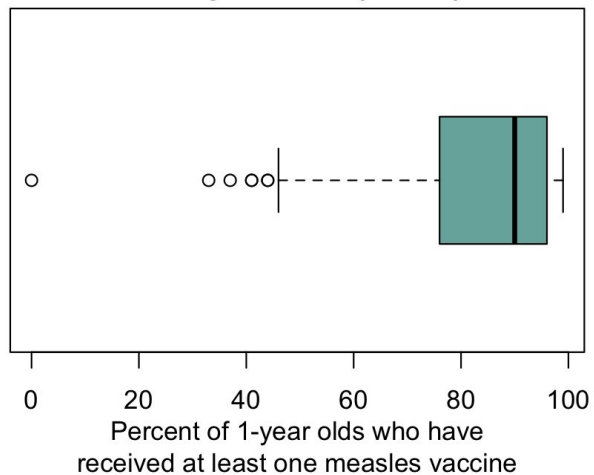


MCV1 is defined as the percentage of children under one year of age who have received at least one dose of measles-containing vaccine in a given year.

ggplot

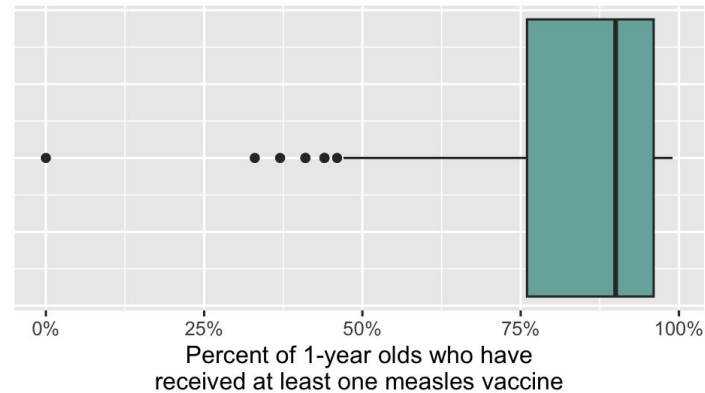
# Boxplot

**Distribution of country-level  
measles vaccination rates  
for 1-year olds (MCV1)**



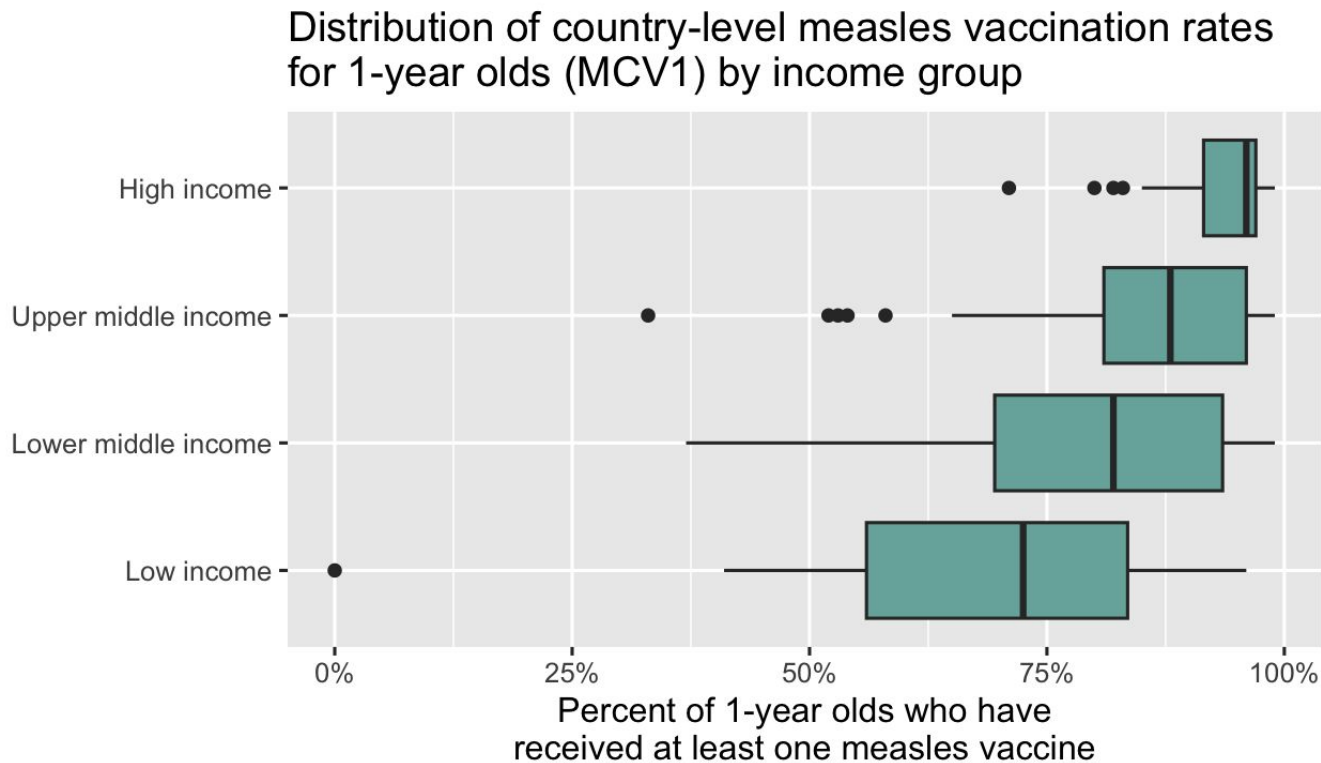
Base R

Distribution of country-level measles vaccination rates  
for 1-year olds (MCV1)



ggplot

# Boxplots by group



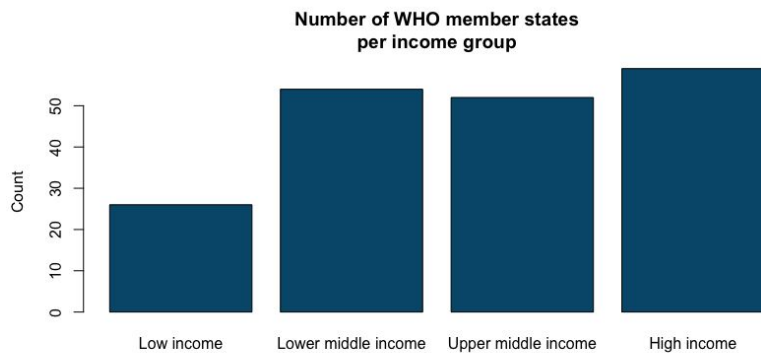


## Your turn

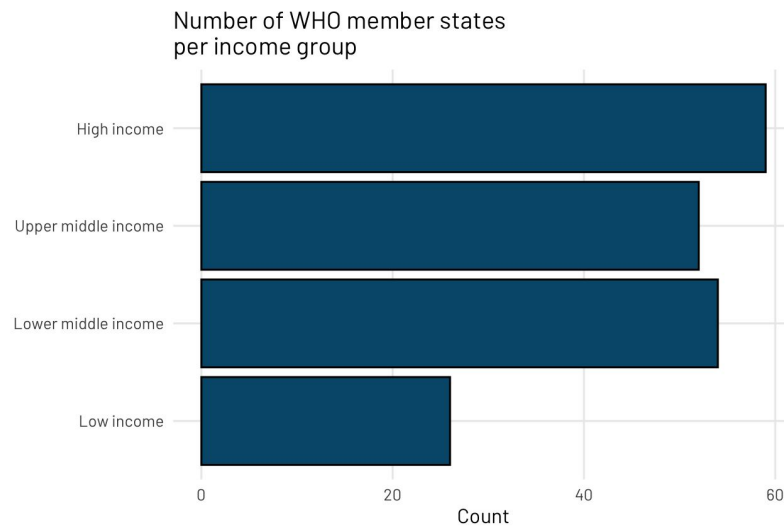
What other ways could boxplots help us better understand vaccination coverage?

- 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

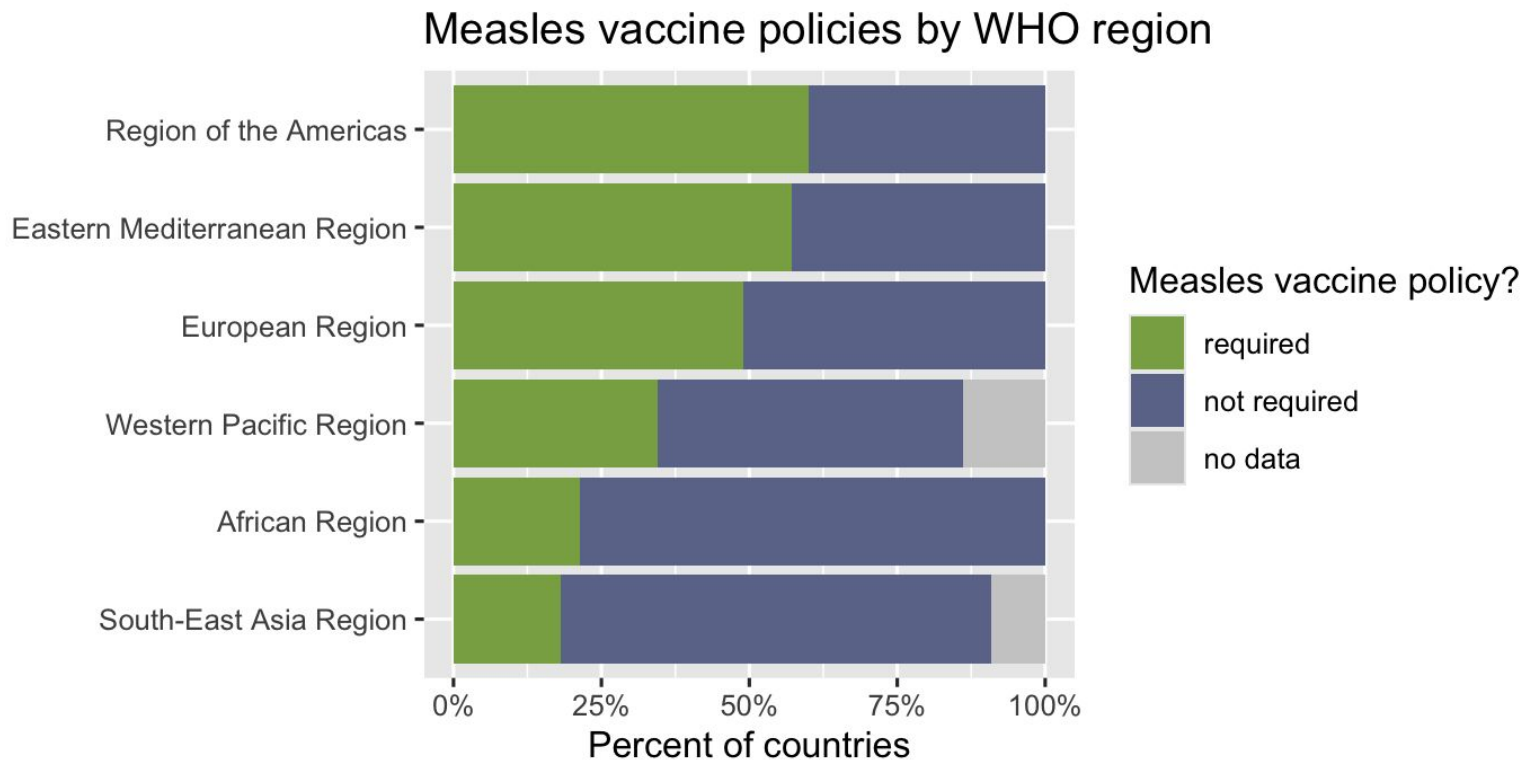


Base R

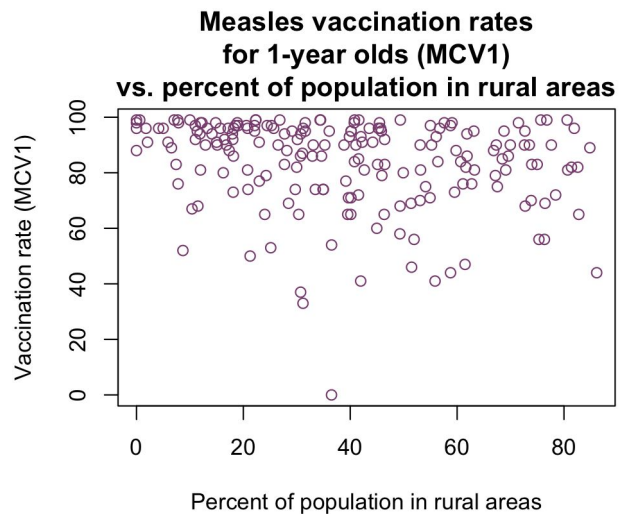


ggplot

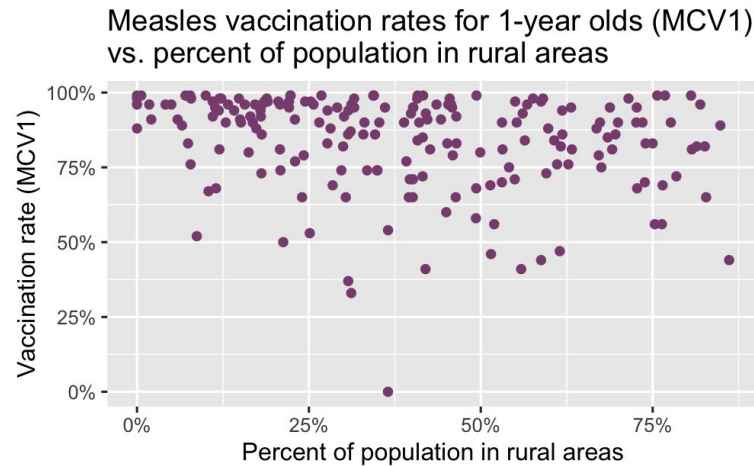
# Stacked barchart



# Scatterplot

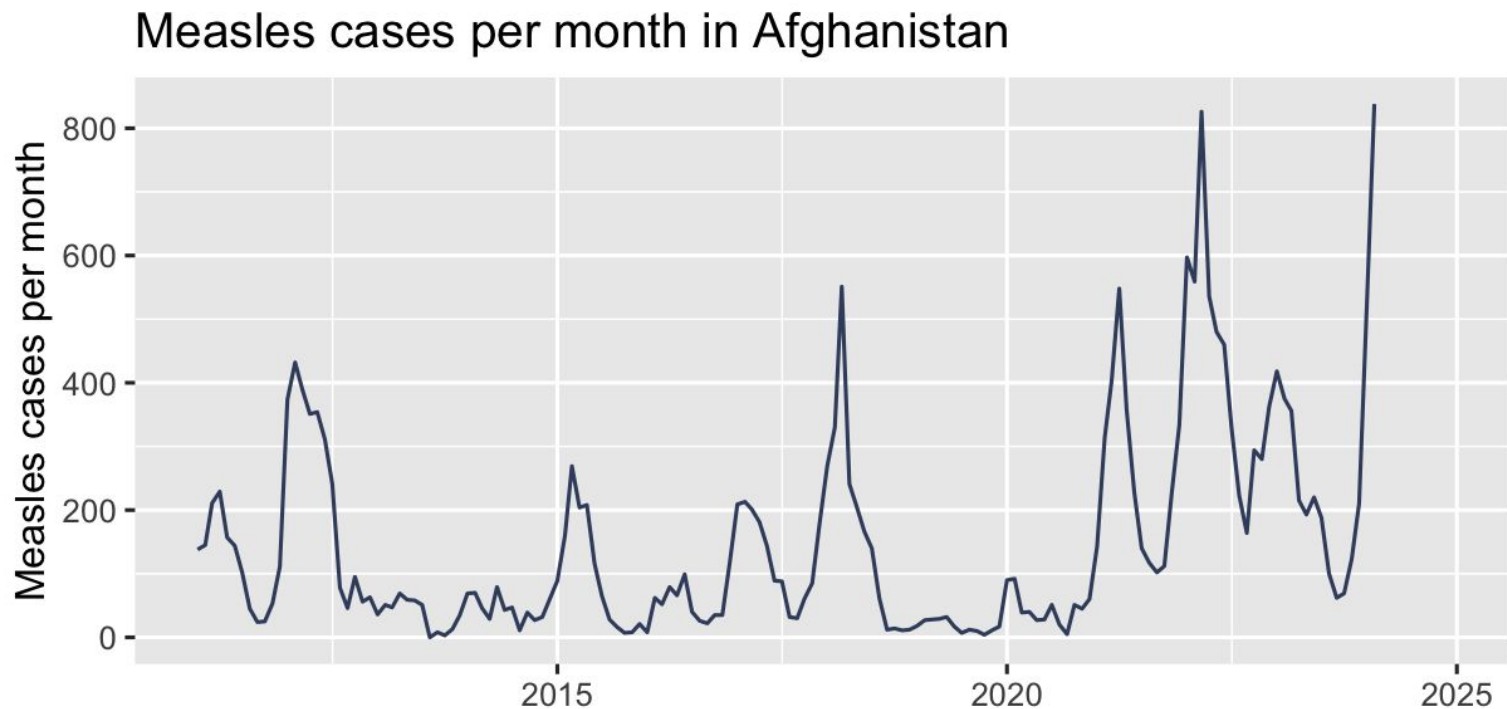


Base R



ggplot

# Line chart



## 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**

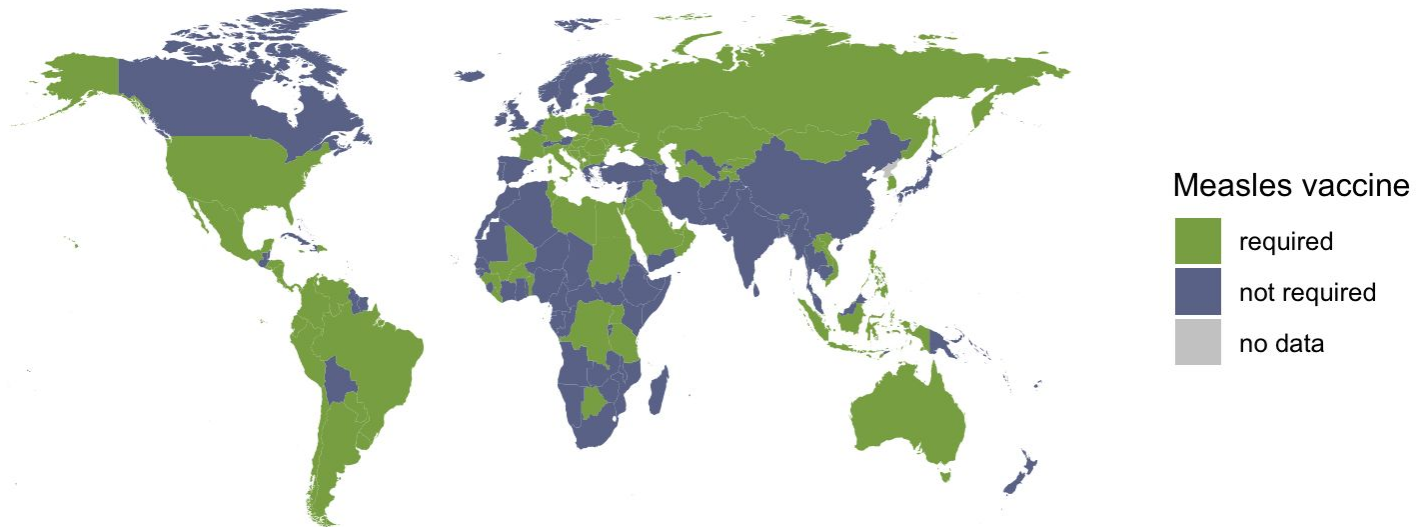
- What data points?
- What questions are answered by these plots?

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

**Bonus charts**  
*(code in github for live demo)*

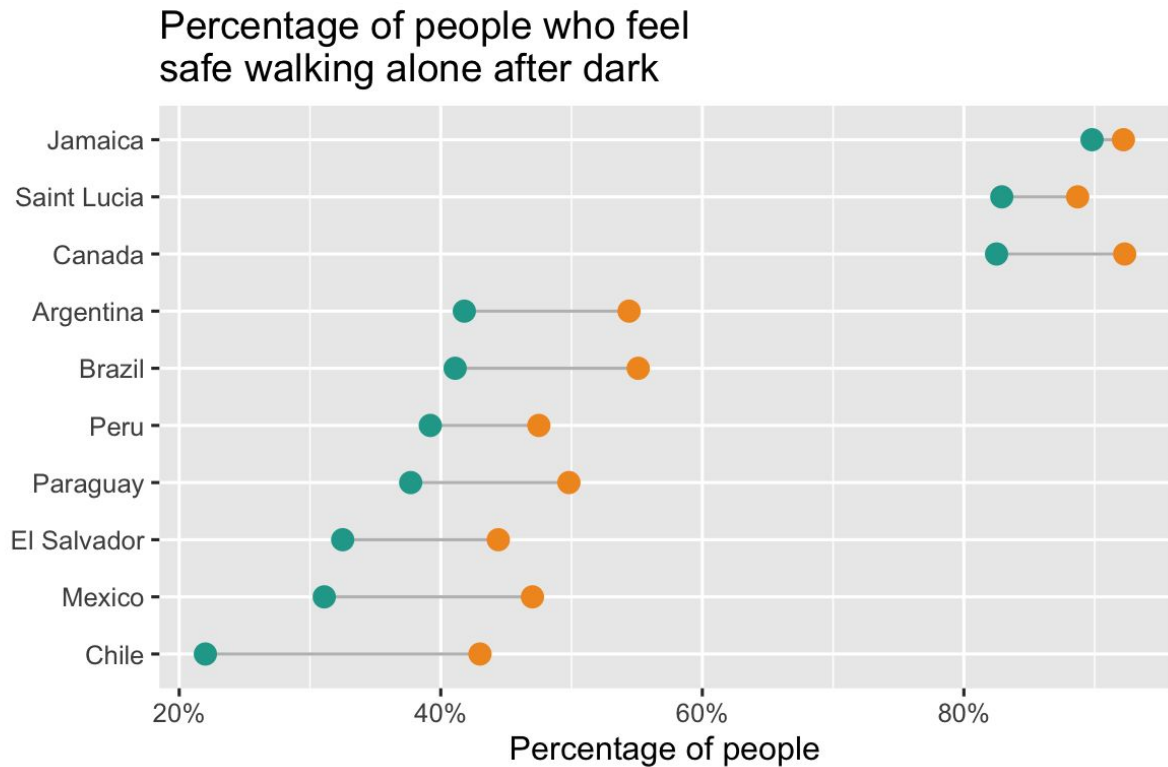
# Maps

Measles vaccine policy requirements



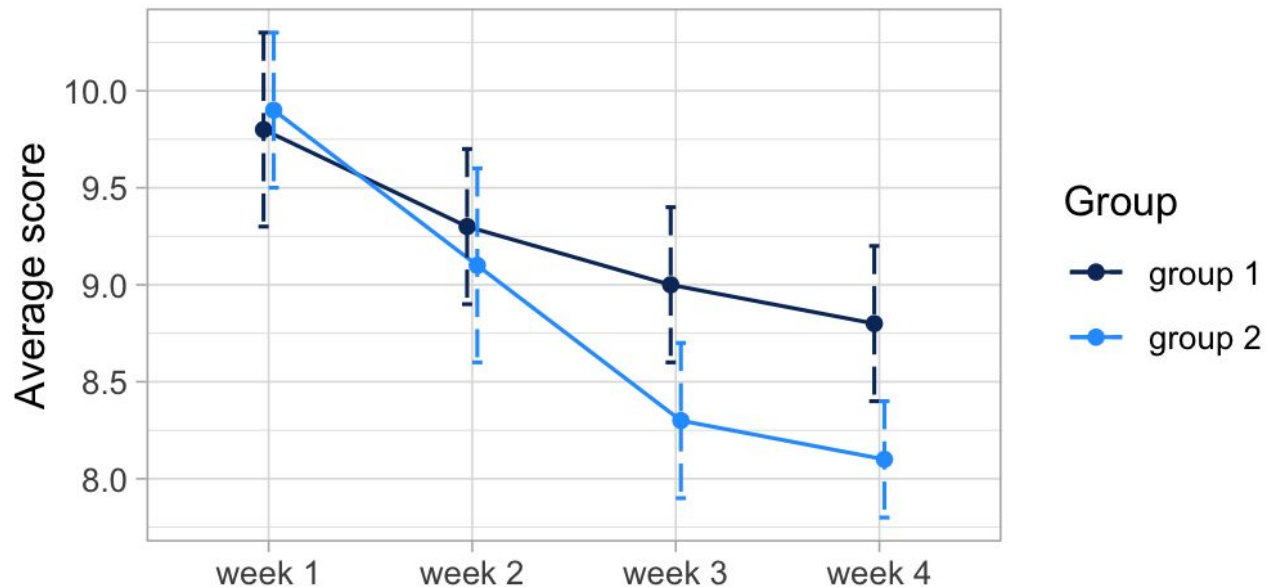


# Cleveland dot plot



# Error bars

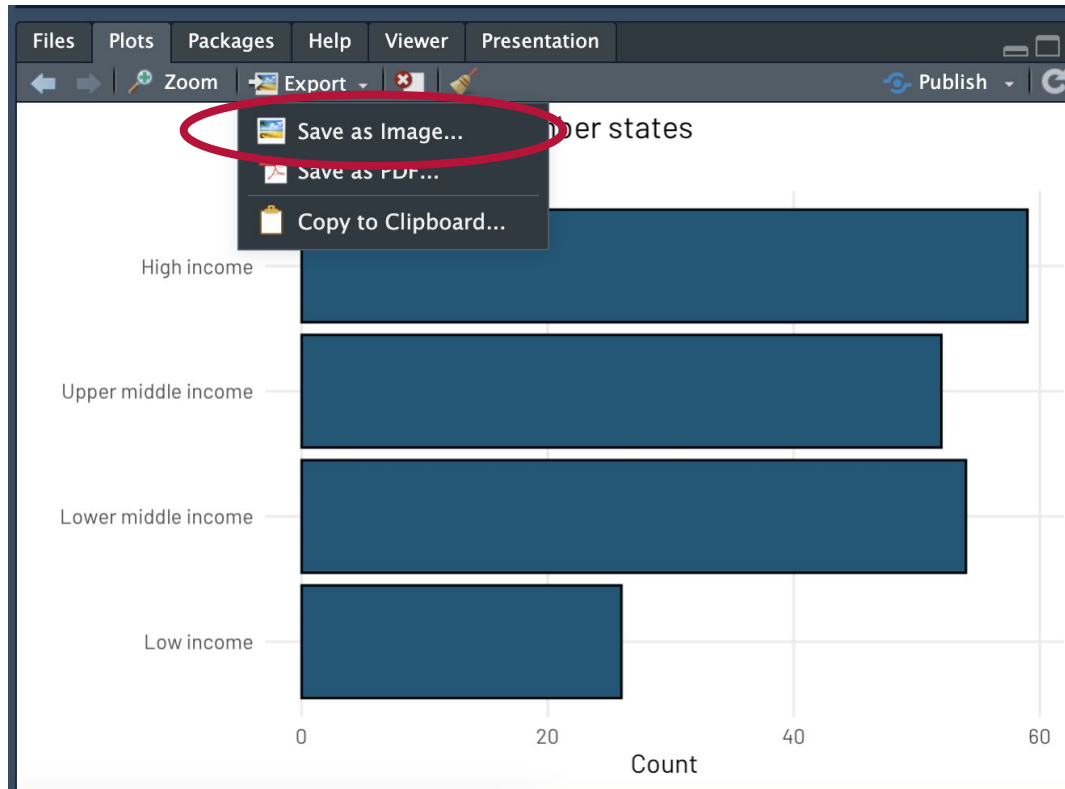
Example study trajectory plot



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:

Width:  Height:

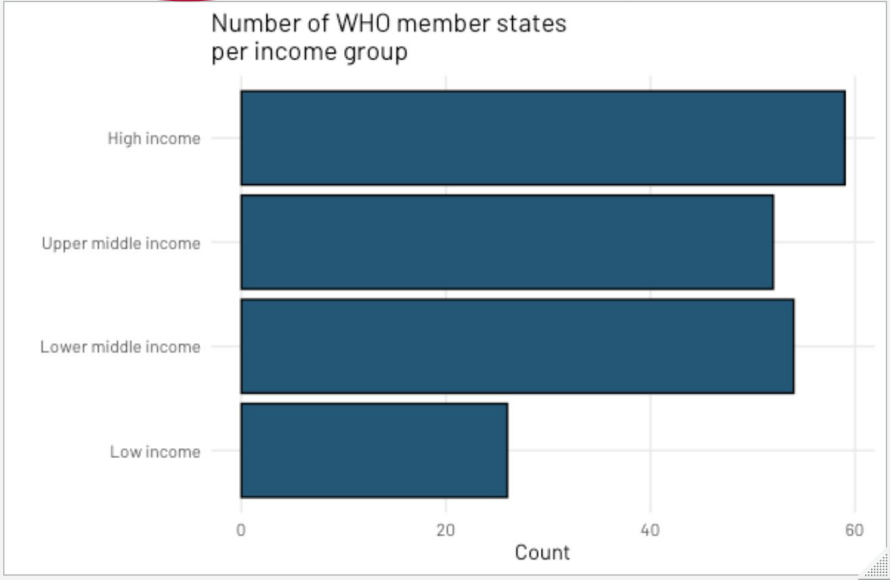
☐ Maintain aspect ratio

Update Preview

Directory...

File name:

Number of WHO member states per income group



Income Group	Count
High income	58
Upper middle income	52
Lower middle income	54
Low income	26

☒ View plot after saving

Save Cancel

# Exporting plots using R Studio

Specify where you  
want to save it

Save Plot as Image

Image format:

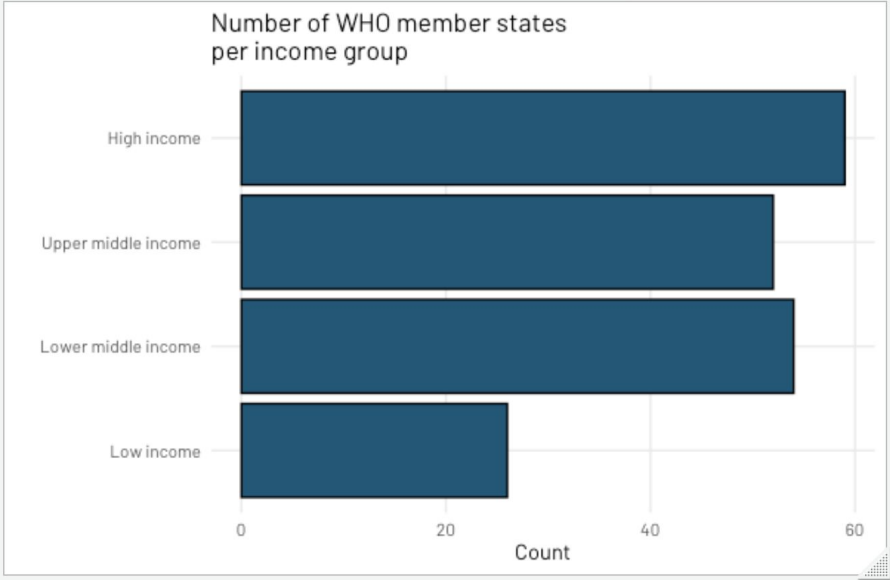
Directory...

File name:

Width:  Height:

☐ Maintain aspect ratio

Number of WHO member states  
per income group



Income Group	Count
High income	58
Upper middle income	52
Lower middle income	54
Low income	26

☒ View plot after saving

# Exporting plots using R Studio

Choose an image format  
(png, jpg, etc)

Save Plot as Image

Image format: **PNG** Width: 596 Height: 387

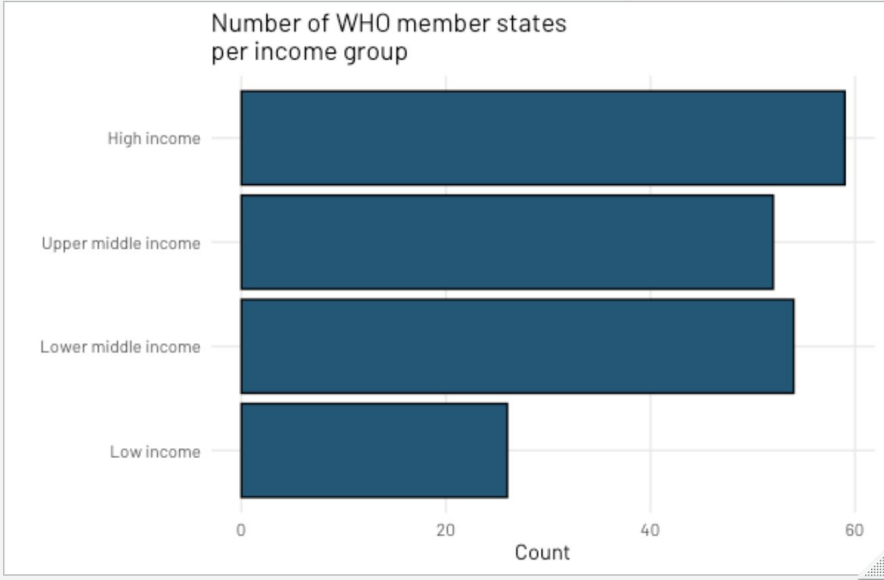
Directory... /Users/stepheaneff/Desktop

File name: Rplot

☐ Maintain aspect ratio

Update Preview

Number of WHO member states per income group



Income Group	Count
High income	58
Upper middle income	52
Lower middle income	54
Low income	26

☒ View plot after saving

Save Cancel

# Exporting plots using R Studio

Save Plot as Image

Image format:

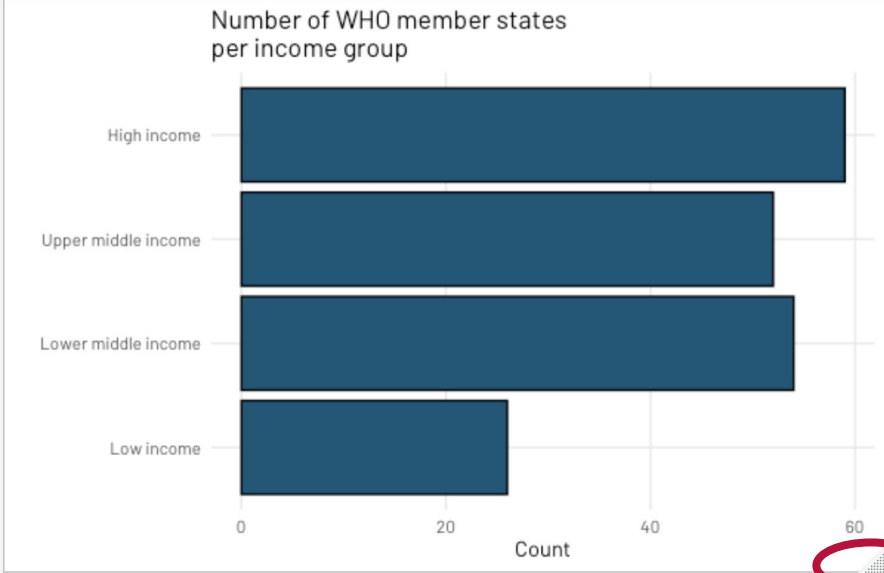
Directory...

File name:

Width: 596 Height: 387

☐ Maintain aspect ratio

Number of WHO member states  
per income group



Income Group	Count
High income	58
Upper middle income	52
Lower middle income	54
Low income	28

☒ View plot after saving

Updating sizing/ratios



# Exporting plots using R Studio

Save Plot as Image

Image format:

Width:  Height:

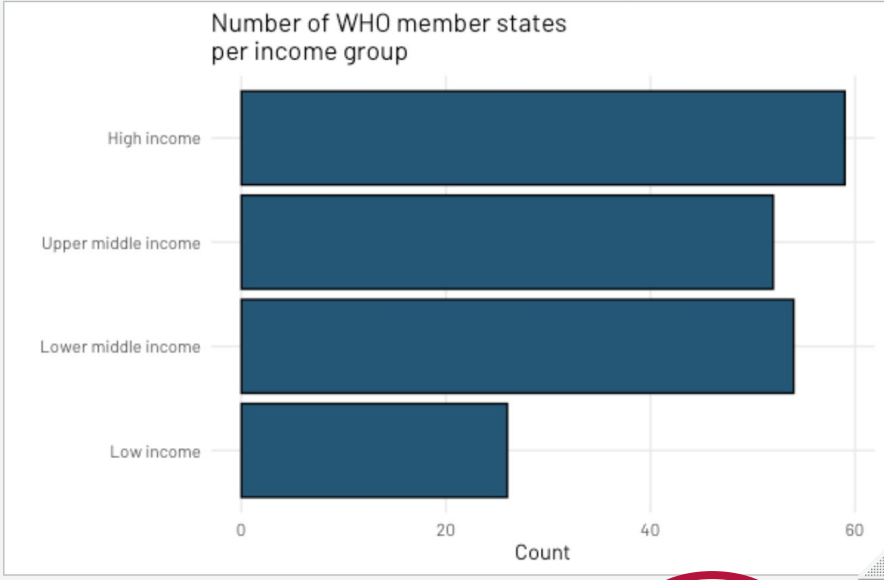
☐ Maintain aspect ratio

Update Preview

Directory...

File name:

Number of WHO member states per income group



Income Group	Count
High income	58
Upper middle income	52
Lower middle income	54
Low income	26

☒ View plot after saving

Save Cancel

Save and export

# Hands-on exploration

## 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.***