Audio Connection and Chat (Zoom)

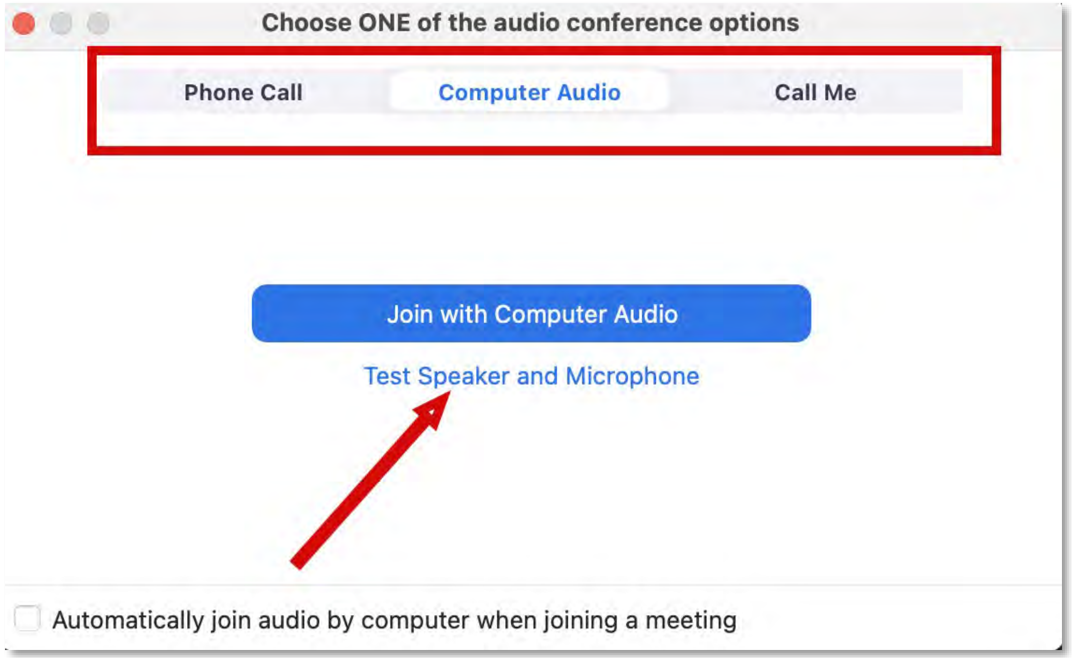
You will not hear any sound until the webinar starts.

Connect Audio

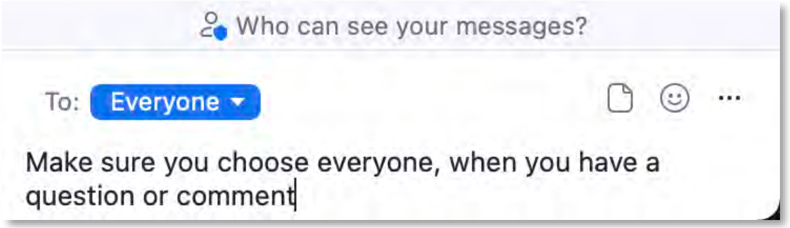
1. When you join Zoom, the ***Join Audio*** preferences box pops-up (Phone Call, Computer Audio, or Call Me)
2. Choose an option that works best for you
3. Join using that option
4. Use Test Speakers and Microphone option to optimize your webinar experience

Chat

Please send your chat to ***Everyone*** to make sure the monitor sees your question


Choose ONE of the audio conference options

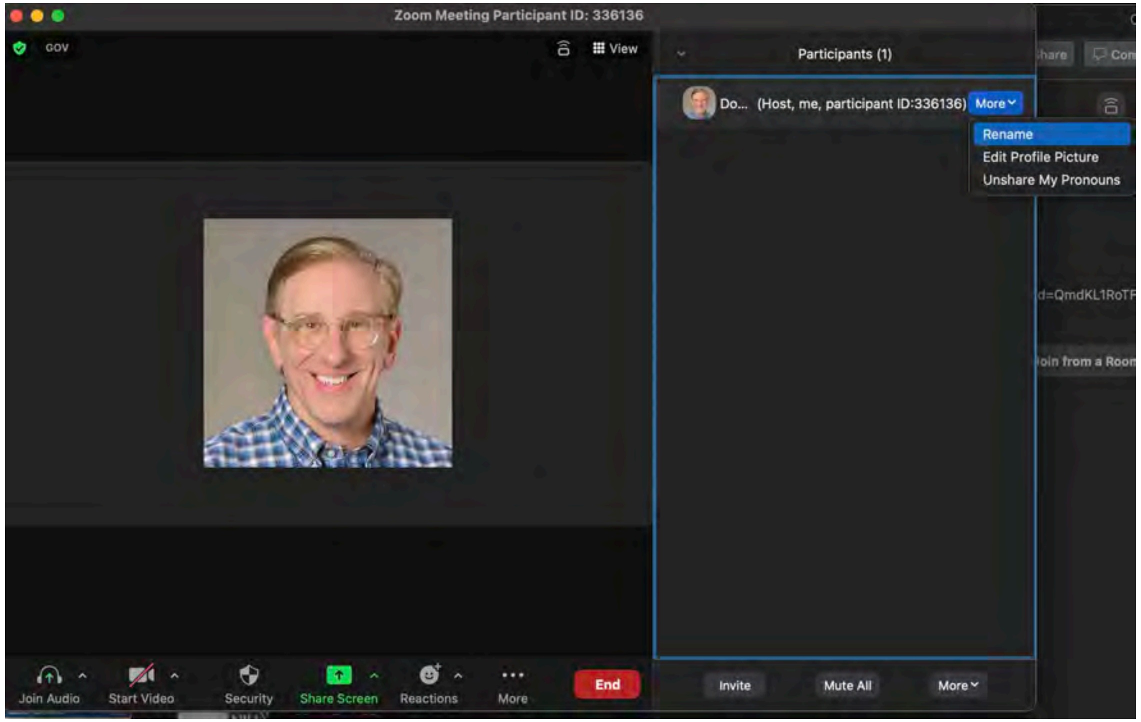
□ Automatically join audio by computer when joining a meeting

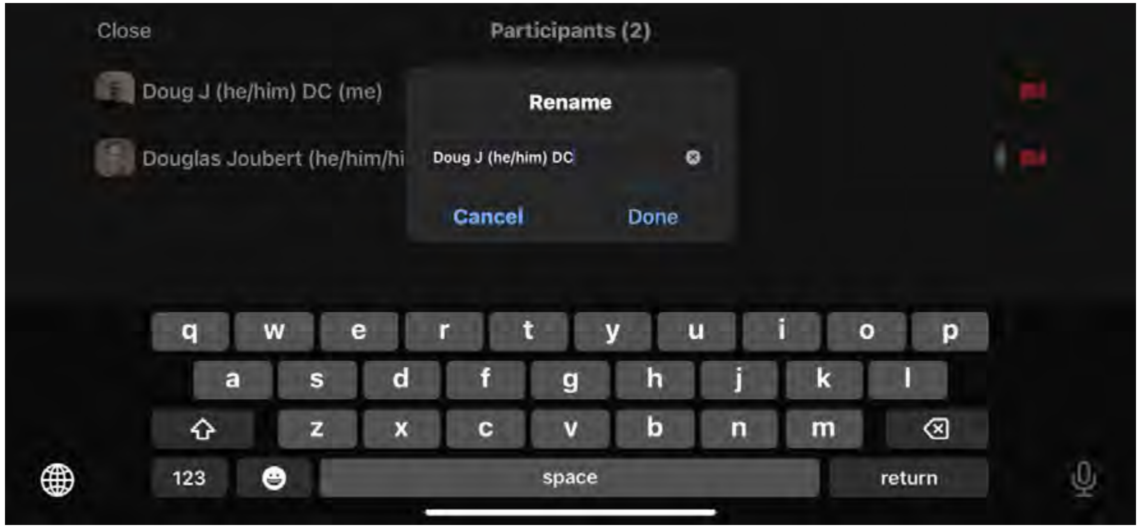


To: Everyone

Make sure you choose everyone, when you have a question or comment

Rename Yourself via Participants List





Please rename yourself, so we can:

- Send you the student version of the PowerPoint
- Add you to our list-serve

Please Fill Out Our Training Survey

- Helps me become a better teacher
- Helps me identify training gaps
- Gives you an opportunity to suggest training

Survey Link

National Institutes of Health
Office of Management

Introduction to Data Visualization in R: ggplot (Basics)

Doug Joubert

2023-03-14

Class Objectives

- Discuss the connection between data, aesthetics, & the grammar of graphics
- Describe how ggplot works
- Define geoms and distinguish between individual geoms and collective geoms
- List options:
  - Graphing one discrete and one continuous variable
  - Graphing two continuous variables

Resources from PowerPoint

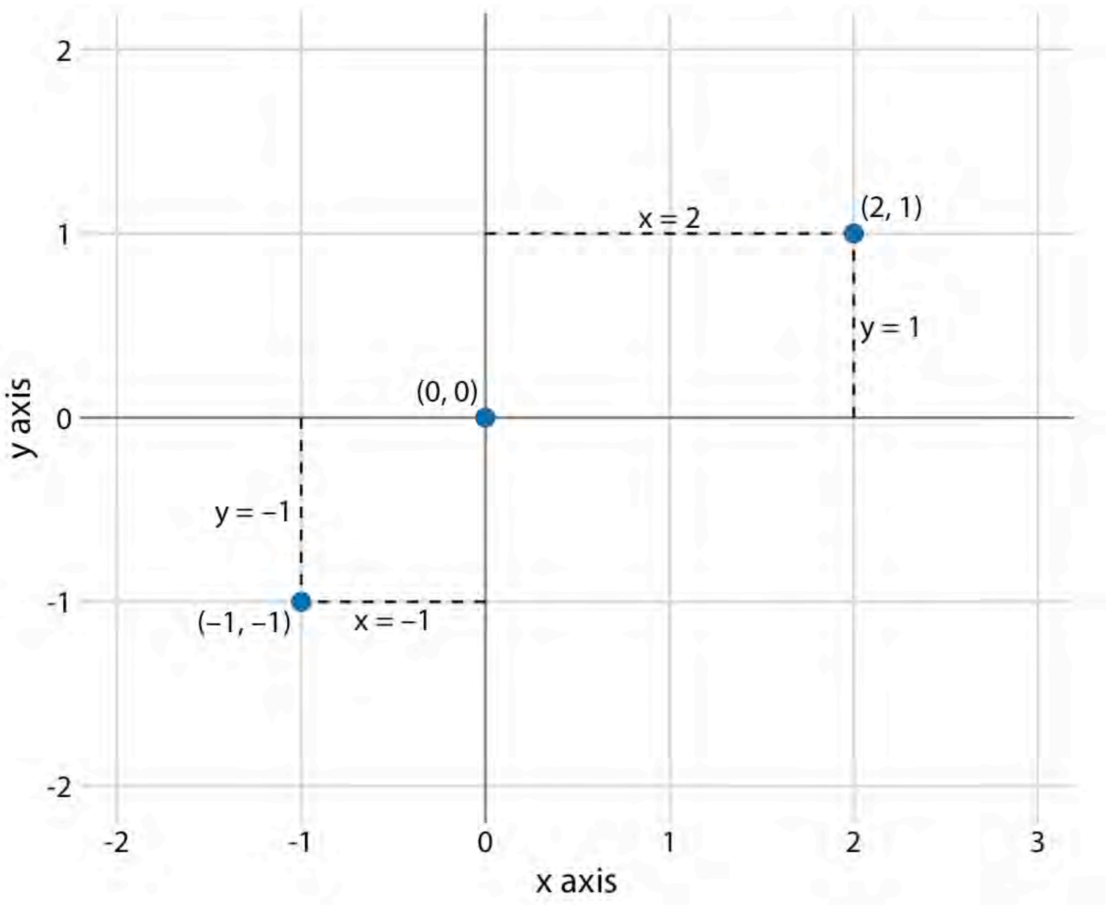
- Additional resources
- Content that I could not cover in class
- Expanded code sections

Configuration for Exercises

Coordinate Systems and Axes

Cartesian Coordinates

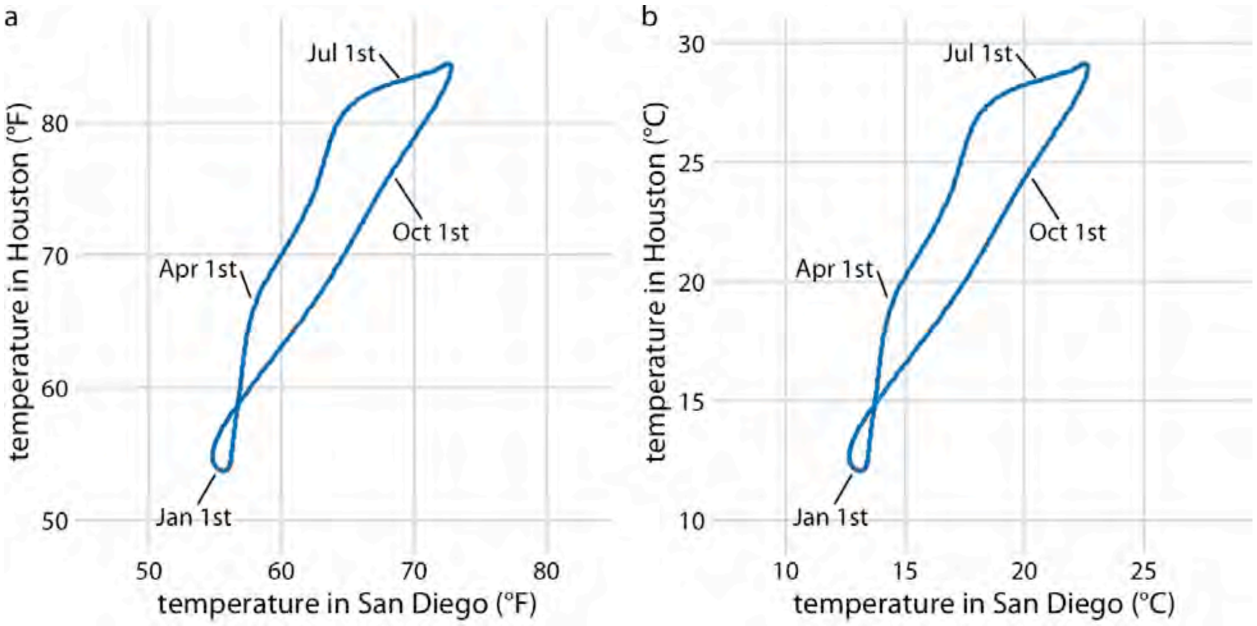
- Need to define position scales, which determine where different data values are located
- Two position scales are usually the x and y axes of the plot
- 2D Cartesian coordinate system is the most common



Wilke, 2019

Axes and Transformations

- Data values come with units:
  - Celsius or Fahrenheit
  - Kilometers or miles
- Change in units is a linear transformation:
  - Add or subtract a number from all data values
  - Multiply all data values with another number
- Linear transformation do not change the “shape” of viz



Daily (normal) temperature for Houston, TX, plotted versus the (normal) temperature of San Diego, CA (Wilke, 2019). Data source: NOAA.

Wilke, 2019

Introduction to tidyverse

Tidyverse Package

- Collection of functions that extend the use of R beyond what is capable in base-R
- Tidyverse tools are an example of packages that extend the capabilities of base-R:
  - Collection of R packages designed for data science
  - Share an underlying design philosophy, grammar, and data structures
  - Designed to work together to transform, analyze, and visualize tidy data

Tidyverse Components

- `tibble`: replaces data frames with tibbles
- `readr` and `readxl`: facilitate data import and export
- `dplyr` and `tidyr`: perform data manipulation
- `stringr`: manipulate text strings
- `ggplot2`: data visualization library

Better at Handling Data

- `readr` for example, is much better at importing data, when compared with base-R
- `readr` also supports SAS, SPSS and Stata files
- It is faster, and it is better at importing dates and formatting numbers
- Never imports columns of strings as factors

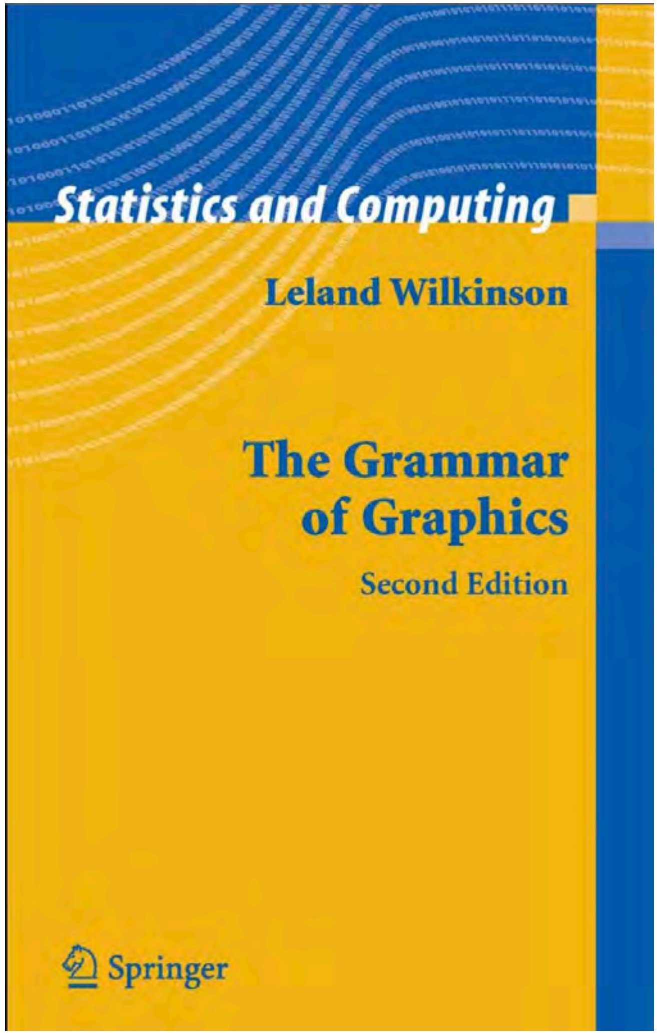
Wrangling and Data Visualization

- It facilitates data wrangling with the pipe operator `%>%`
- `Tidyr` library provides tools for reshaping, and transforming data
- `dplyr` is the main package for sub-setting, filtering, and summarizing data
- ggplot2 provides a consistent method for dealing with graphics in R, based on The Grammar of Graphics

Introduction to ggplot

ggplot Framework

- Grammar of graphics:
  - Any plot can be expressed from the same set of components
- Designed to work iteratively:
  - Data layer
  - Annotation layers
  - Statistical summaries layer



Wilke, 2019; Leland, 2005

ggplot Mapping Components

- Layer: collection of geometric elements and statistical transformations
- Scales: map values in the data space to values in the aesthetic space
- Coord: describes how data coordinates are mapped to the plane of the graphic

Wilke, 2019; Wickham, 2022

ggplot Mapping Components

- Facet: how to break up and display subsets of data as small multiples
- Theme: controls the finer points of display, like the font size and background color

Wilke, 2019; Leland, 2005

Plots are Built in Layers

- `ggplot` graphics are built step by step by adding new elements
- Key to understanding ggplot2 is thinking about a figure in layers
- May be familiar to you if you have used programs like Photoshop, Illustrator, or Inkscape
- Basic template that can be used for different types of plots

```
ggplot(data = DATA, mapping = aes(AESTHETIC MAPPINGS)) +
geom_function()
```

Adding Layers in ggplot

- Plus symbol + must be added to each new layer that you add to plot syntax
- Plus sign + must be placed at the end of each line of code
- Using second method will not add the new layer and will return an error message

```
surveys_plot +
geom_point()
```

✓

```
surveys_plot

+ geom_point()
```

✗

ggplot: Key Components

Key Components: Data

- Every layer must have associated data
- Data must be in a tidy data frame
- Tidy data is described in R for Data Science:
  - Variables in the columns
  - Observations in the rows
- You begin every plot by telling the ggplot() function defining your data

```
ggplot(data = gapminder)
```

```
gapminder %>%
  ggplot()
```

Wickham et al., 2022a

Key Components: Variable Mapping

Variables from gapminder dataset

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811

Key Components: Aesthetics Mapping

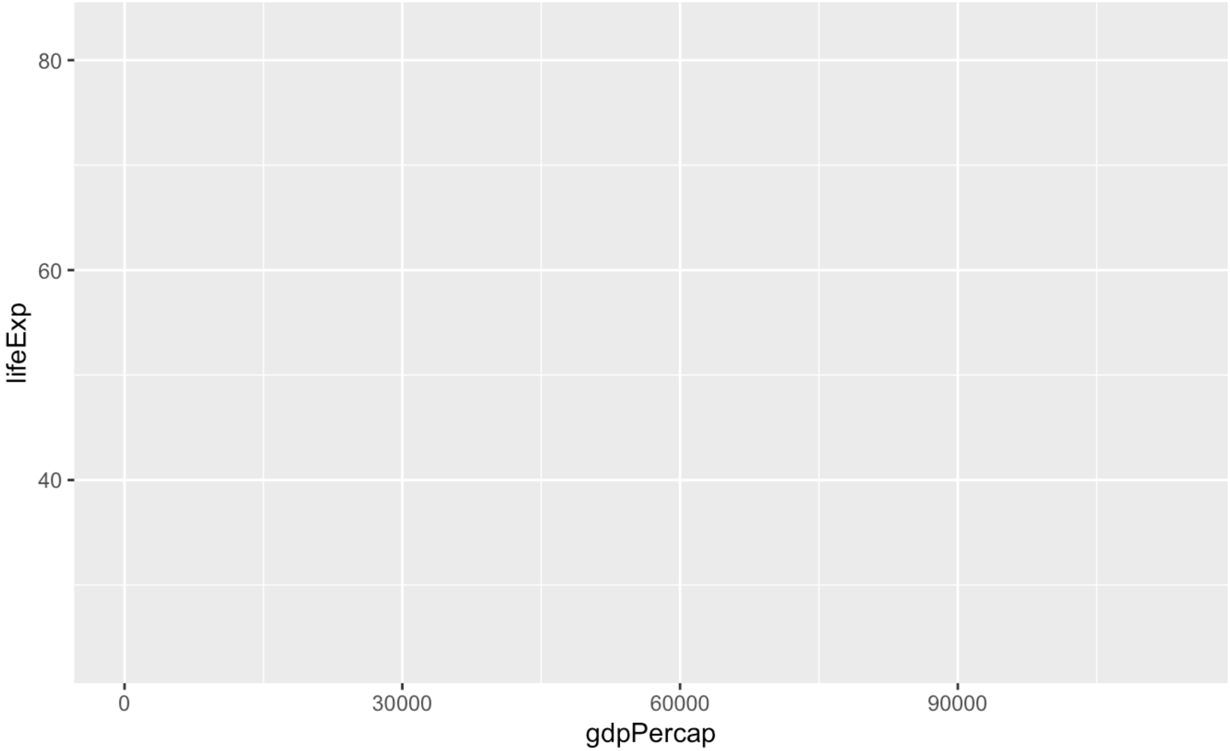
- Defines how variables in your dataset are mapped to visual properties
- Select the variables to be plotted and how to present them in the graph:
  - x/y positions
  - Size
  - Shape
  - Color

```
gapminder %>%
  ggplot(mapping = aes(x = gdpPercap,
y = lifeExp))
```

Wickham et al., 2022a

Mapping: Checking In

- Data argument tells ggplot where to find the variables
- Mapping function
  - x-axis variable is `gdpPercap`
  - y-axis is `lifeExp`
- `aes()` function does not say where variables with those names are to be found
- `ggplot()` assumes data comes from object in the data argument

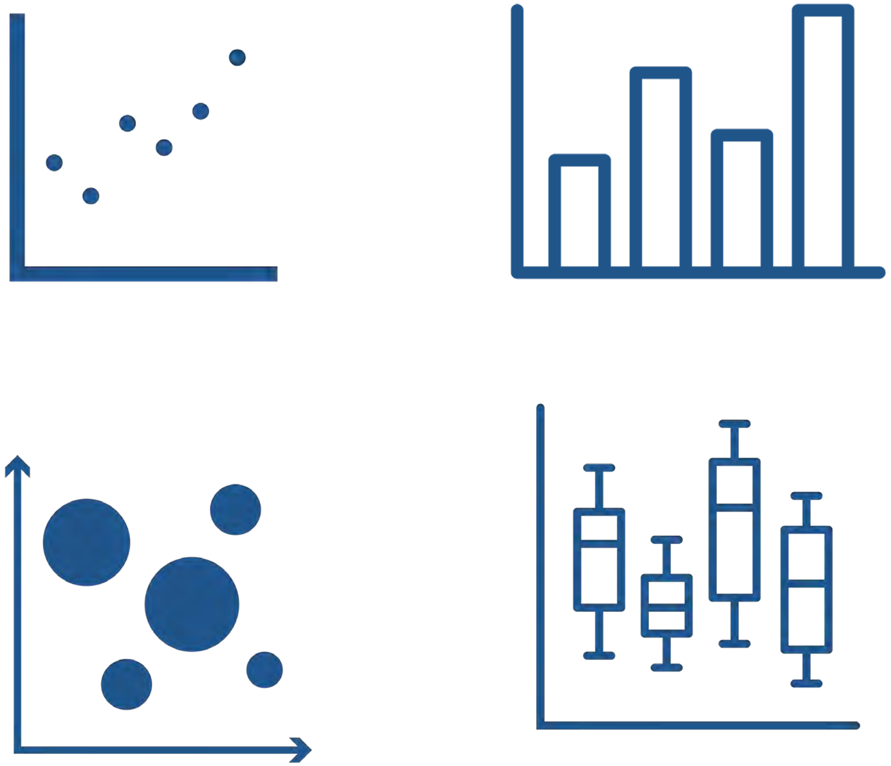


Geoms in ggplot

Geoms: Introduction

- Geometric objects, or geoms, render the layer, controlling the type of plot that you
- Point geom will create a scatterplot
- Line geom will create a line plot
- Two primary classes:
  - – Individual geoms
  - – Collective geoms

Wilke, 2019; Wickham, 2022



Geoms: Individual geoms

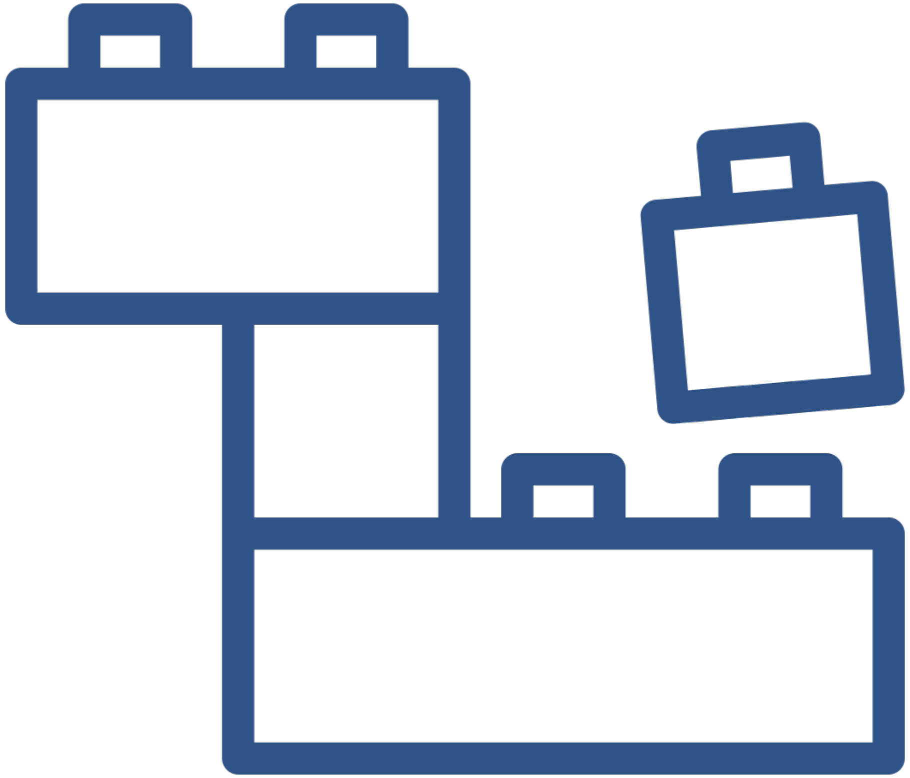
- Draws a distinct graphical object for each observation (row)
  - For example, the point geom draws one point per row
- Two dimensional and require both x and y aesthetics
- Understand color, size
- Bar, tile and polygon also understand fill

- geom_point()
- geom_bar(stat ="identity")
- geom_col()
- geom_line()
- geom_area()
- geom_polygon()

Wilke, 2019; Wickham, 2022

Geoms: Collective geoms

- Displays multiple observations with one geometric object
- May be a result of a statistical summary, like a boxplot
- May be fundamental to the display of the geom, like a polygon
- `Group` aesthetic controls the assignment of observations to graphical elements



Wilke, 2019; Wickham, 2022

Collective geoms (When to Use)

- By default, the group aesthetic is mapped to all discrete variables in the plot
- When the mapping is not automatic, or when no discrete variable is used in a plot, you'll need to explicitly define the grouping structure to a variable
- Common in longitudinal studies with many subjects, where the plots are often descriptively called spaghetti plots

Wilke, 2019; Wickham, 2022

Geoms for One variable

- Discrete
  - geom_bar(): display distribution of discrete variable.
- Continuous
  - geom_histogram(): bin and count continuous variable, display with bars.
  - geom_density(): smoothed density estimate.
  - geom_dotplot(): stack individual points into a dot plot.
  - geom_freqpoly(): bin and count continuous variable, display with lines.

Geoms for Two Variables

- Two Continuous Variables
  - geom_point(): scatterplot
  - geom_quantile(): smoothed quantile regression
  - geom_smooth(): smoothed line of best fit
  - geom_text(): text labels
- At Least One Discrete Variable
  - geom_count(): count number of point at distinct locations
  - geom_jitter(): randomly jitter overlapping points
- One continuous Variable and One Discrete Variable
  - geom_boxplot(): boxplots.
  - geom_violin(): show density of values in each group.

Saving Plots

Viewing or Saving Plots

- Export tab in the RStudio Plot pane will save plots at low resolution
- Might need higher resolution for many journals and will not scale well for posters
- `ggsave()` allows you to change the dimension and resolution of your plot:
  - Width
  - Height
  - dpi

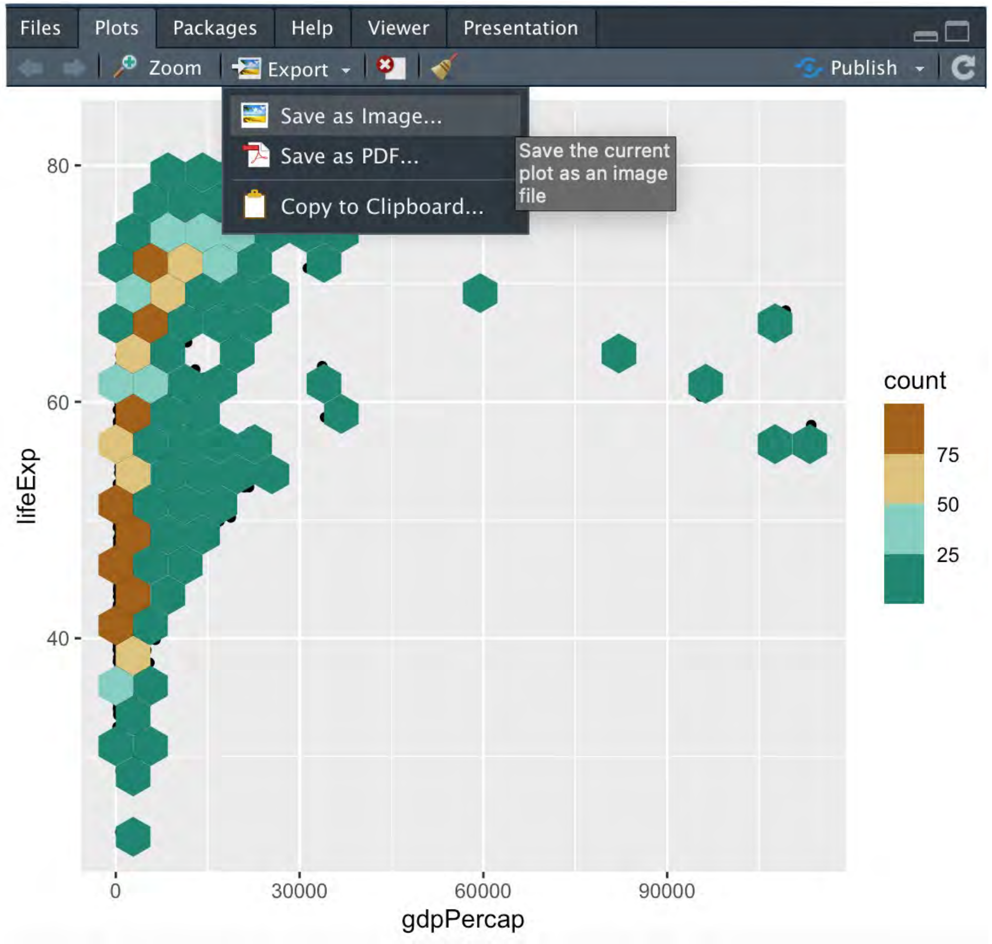


Image Formats

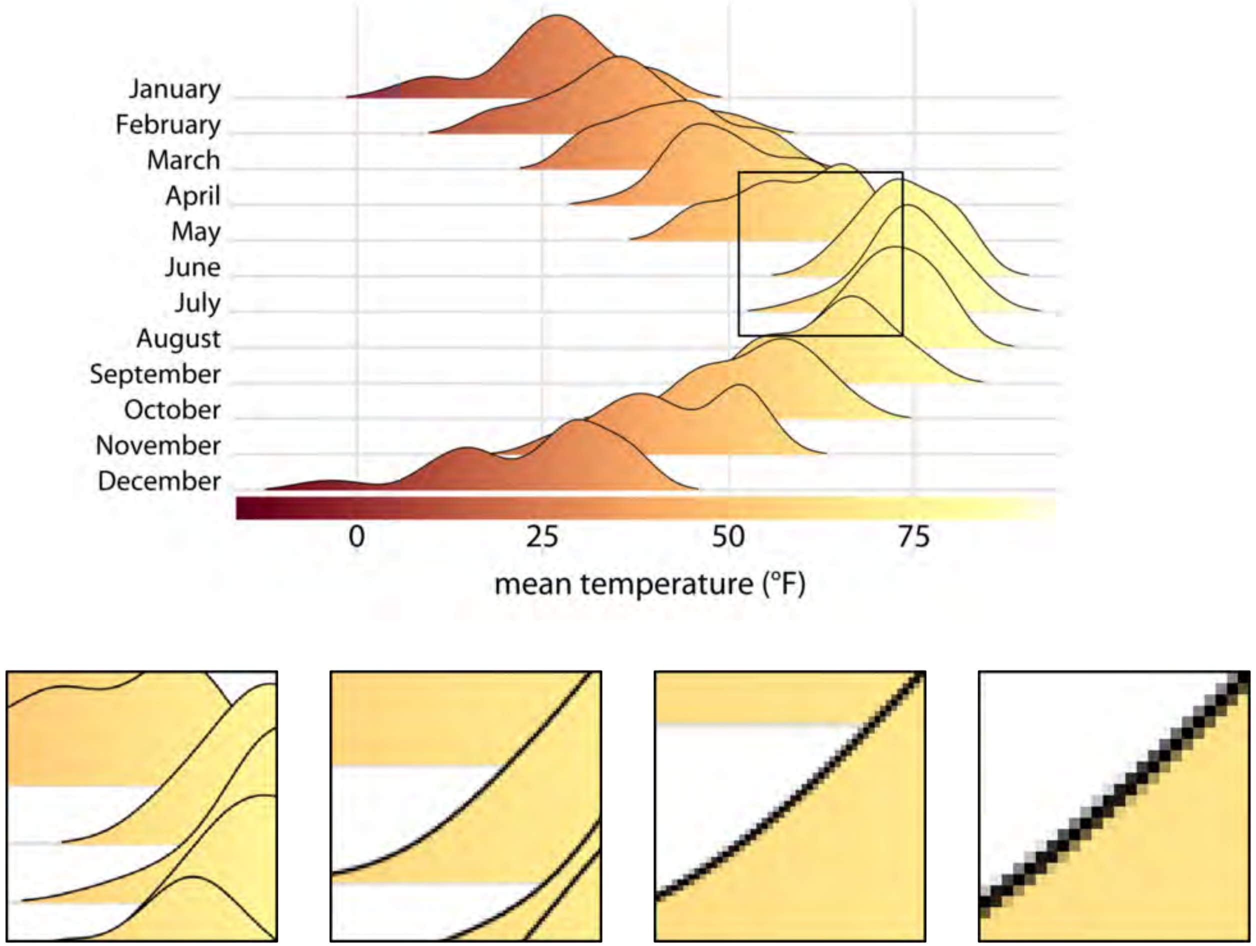
- Most important difference between graphics formats is whether they are bitmap or vector:
  - Bitmaps
  - Vector graphics

Acronym	Type	Application
pdf	vector	general purpose
svg	vector	online use
png	bitmap	optimized for line drawings
jpeg	bitmap	optimized for photographic images
tiff	bitmap	print production, accurate color reproduction

Wilke, 2019; Wickham, 2022

Image Formats

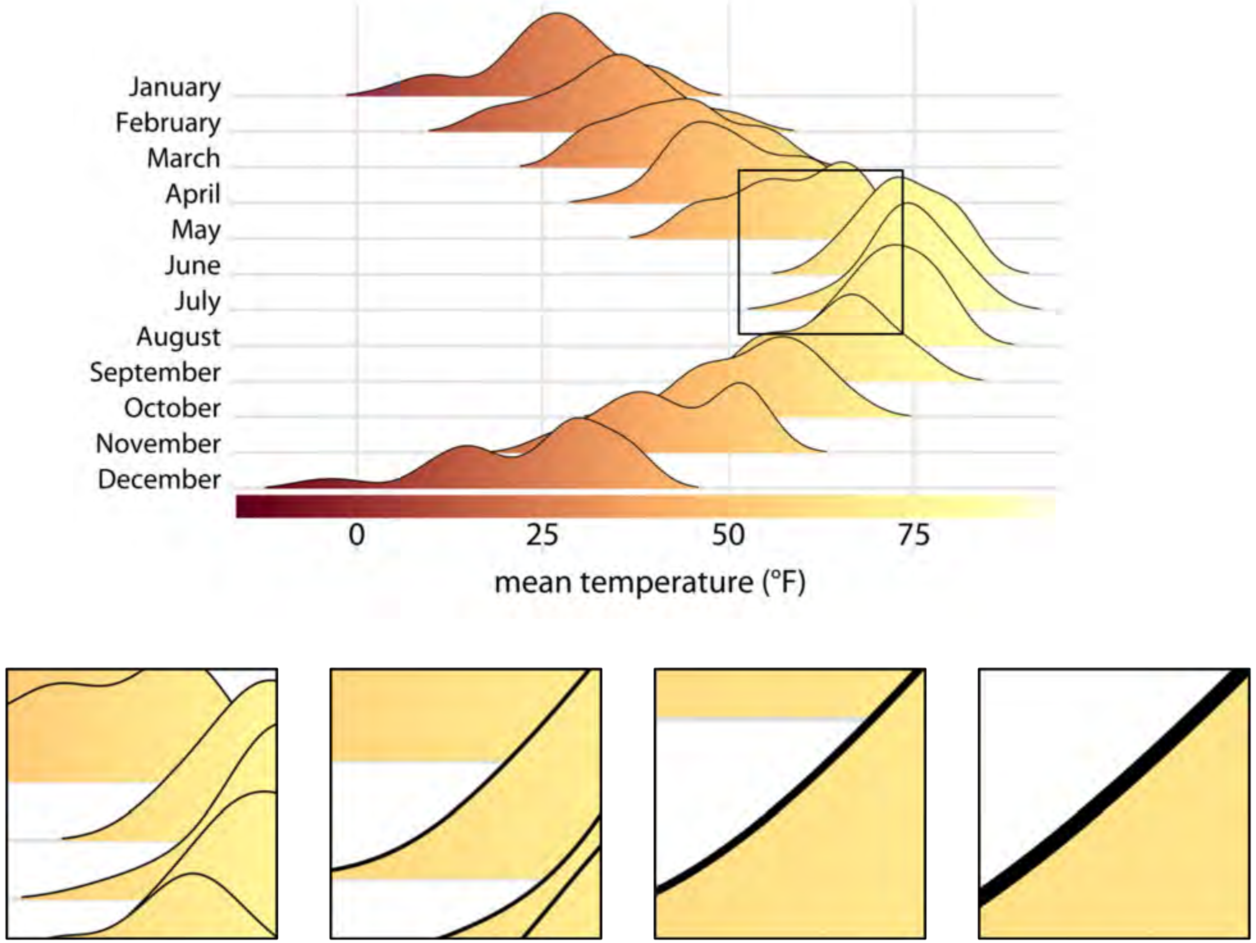
- Bitmaps or raster graphics stores the image as a grid of individual pixels, each with a specified color



Wilke, 2019; Wickham, 2022

Image Formats

- Vector graphics stores the geometric arrangement of individual graphical elements in the image
- “Resolution-independent,” because they can be magnified to different sizes without losing detail or sharpness



Wilke, 2019; Wickham, 2022

Saving Files using Tidyverse

- `ggsave()` function allows you easily change the dimension and resolution of your plot
- Save the plots in fig_output folder
- We need to do is save our plot to an object called `gapminder_scatter`

```
gapminder_scatter <- map1 +
  geom_point() +
  labs(title = "Scatterplot of GDP by
Life Expectancy",
       x = "GDP",
       y = "Life Expectancy") +
  theme_light()
```

Wilke, 2019; Wickham, 2022

Saving Files using Tidyverse

We can then export the graph that we just saved, using `ggsave`

```
ggsave("fig_output/gapminder-scatter.png", gapminder_scatter,
width = 15, height = 10, dpi = 600)
```

Please Fill Out Our Training Survey

- Helps me become a better teacher
- Helps me identify training gaps
- Gives you an opportunity to suggest training

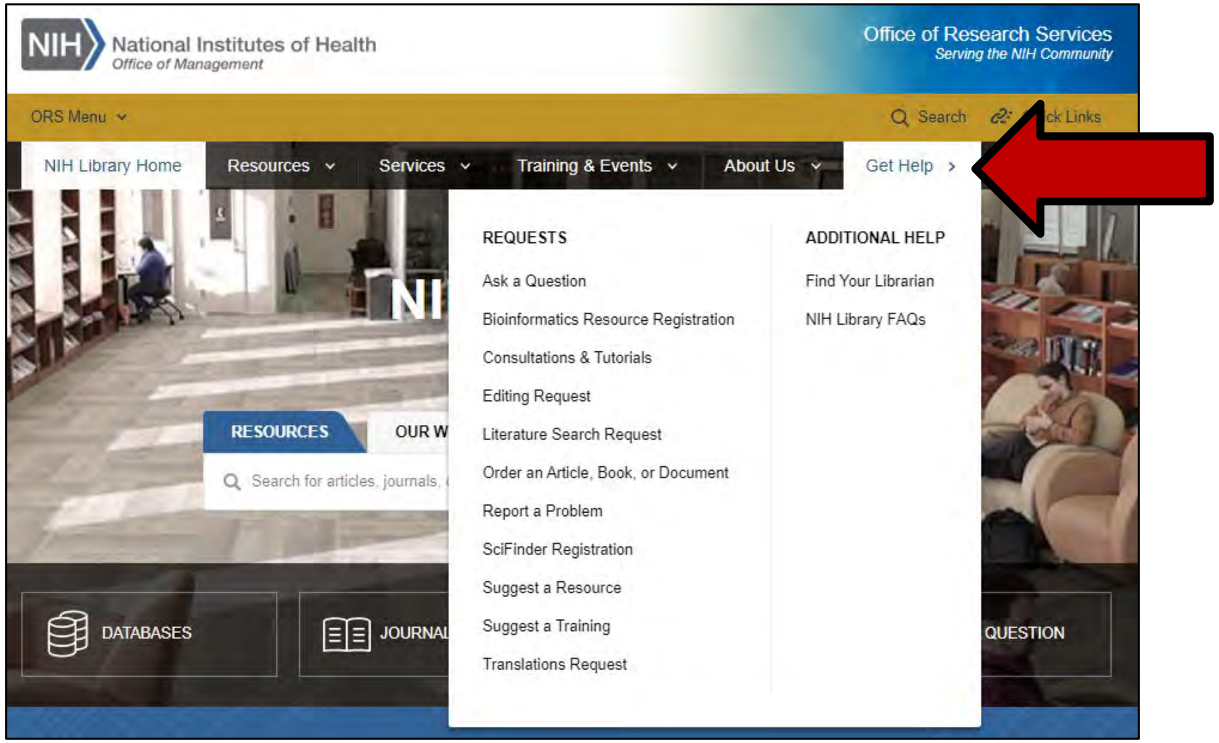
Survey Link

We Can Help

- Classes on a variety of data-related topics, including:
  - Data management
  - Data visualization
  - Data analysis
  - R and RStudio

- Computers which offers a suite of tools for data analysis, processing, and visualization

Contact Us for Ongoing Support



Doug Joubert

Bioinformatics Support Program

301-827-3829

douglas.joubert@nih.gov

NIH Library Help Desk

(301) 496-1080

- **Ask a Question:** https://www.nihlibrary.nih.gov/get-help/ask-question
- **Request a Tutorial:** https://www.nihlibrary.nih.gov/get-help/consultations-tutorials
- **Sign up for Additional Classes:** https://www.nihlibrary.nih.gov/training/calendar

Questions & Comments

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institutes of Health
Office of Management