R programming for Natural Resource Professionals

Lecture 9:

Data Visualization I:

Intro to ggplot2

Learning Objectives

- Understand guidelines for data visualization
- Understand how to use ggplot functions and calls
- Understand the use of various data visualizations

Groups

- Group 1: Dakota, Jordan, Andrew, Anasara
- Group 2: Macayla, Everett, Dan, Jeremey
- Group 3: Becca, Alicia, Ben, Eric, Keenan

Paper Discussions

- Go to the discussion doc:

R for Nat Res class discussion - Google Docs

- Each group: Discuss the papers each group member selected and then choose one of the graphs that you feel could be improved and one you feel is well made. Go online to the pdf of the paper and take a screen shot of each and paste them into the google doc. Below the graphs describe the research article briefly and why you selected this graph (e.g., a list of pros and cons about the graph)

- Run one of the figures through the colorblindness simulator by uploading a jpeg of a screen shot to the website provided. Does the figure become illegible for someone with color blindness. Grab a screen shot of the figure if it is not legible in one of the colorblind filters. How could you improve the graphic so it is now legible?

Notes on graphics

- Figures (and captions) should stand alone
  - Tables do not need to do this
- Figures convey the message of the paper
- People will grab figures for their own purposes (example here)

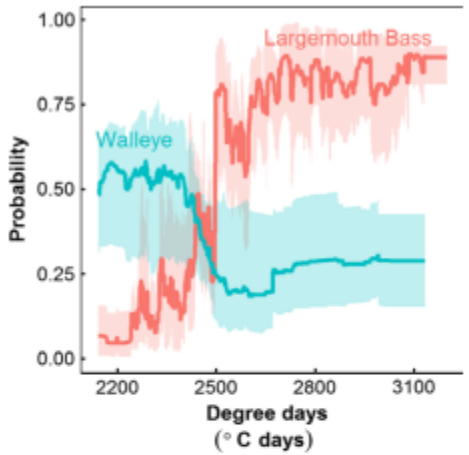
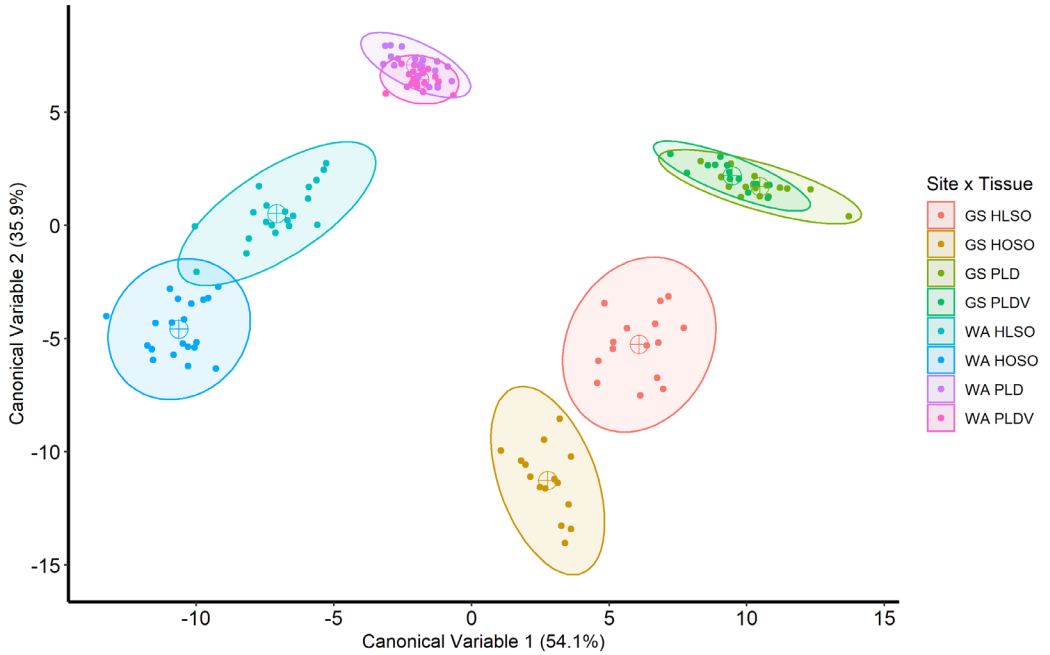


Fig. 1 Predicted probability of successful walleye recruitment (electrofishing catch rates ≥10 age-0 fish per mile; blue line) and high largemouth bass relative abundance (electrofishing catch rates ≥ season-specific median catch rates; orange line) as a function of mean water temperature degree days (base temperature 5 °C) in contemporary period (1989–2014). Predicted probabilities are based on species-specific random forest models with additional predictors: lake area, conductivity, and shoreline complexity for walleye and lake order and Secchi depth for largemouth bass. Solid lines show median probability for a given value of degree days across all possible combinations of other predictors; shaded ribbon shows 25th–75th percentile.

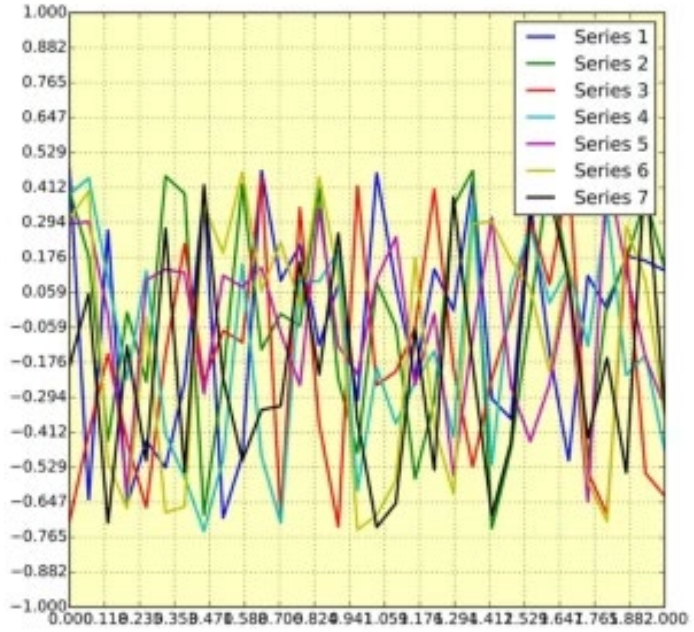
Color Blindness

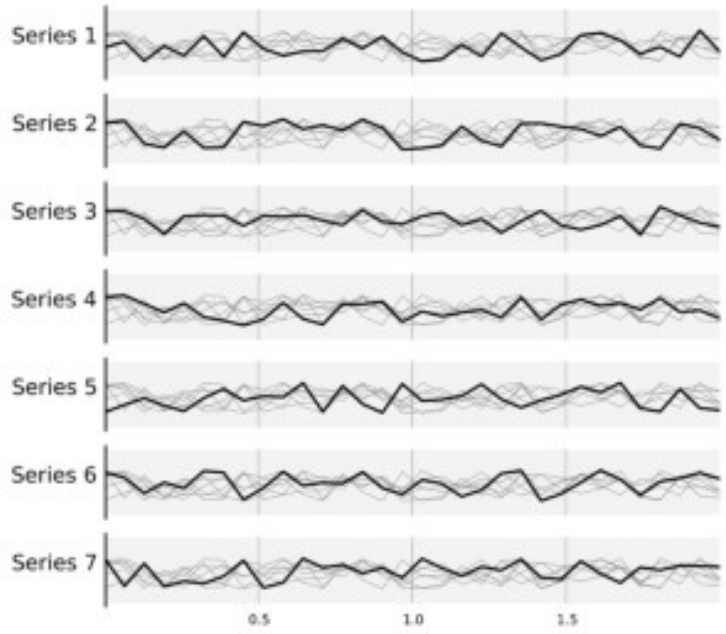
- https://www.color-blindness.com/coblis-color-blindness-simulator/



Other Pitfalls

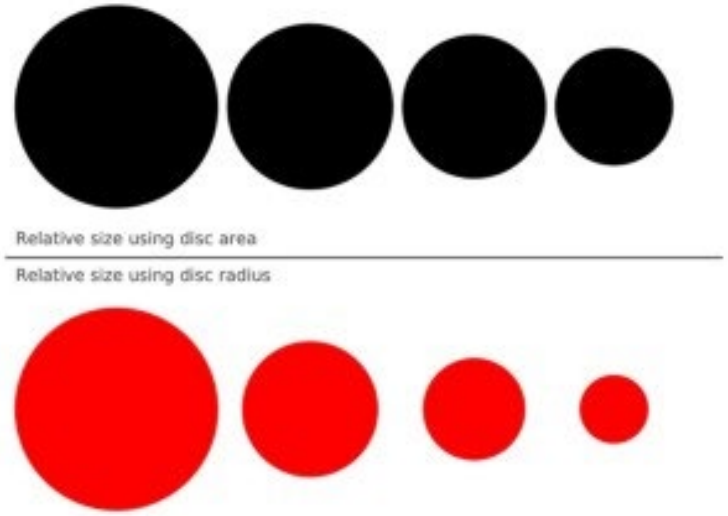
- Complex figures that nobody can understand (KISS)

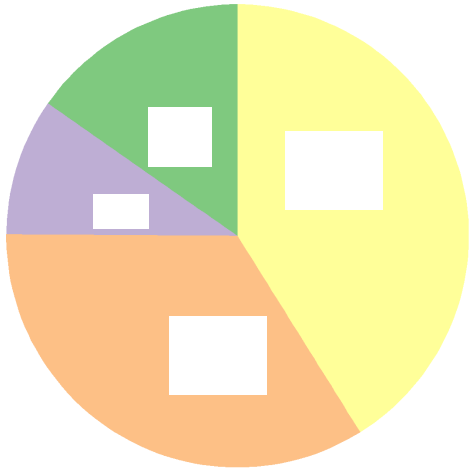




Other Pitfalls

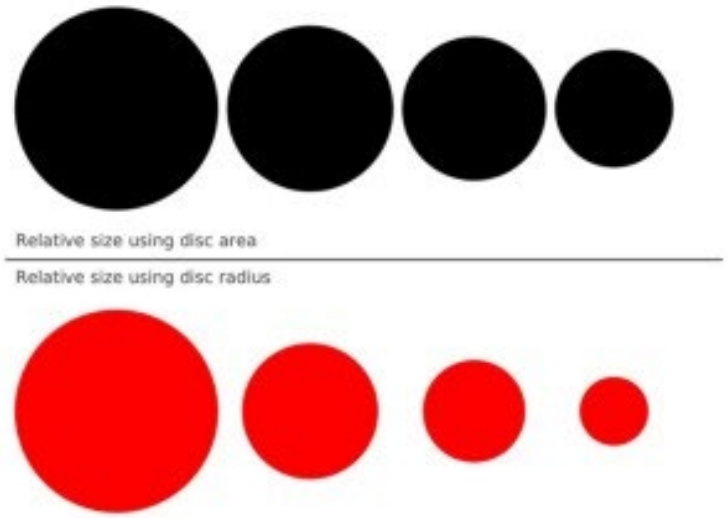
- Pie charts and circles

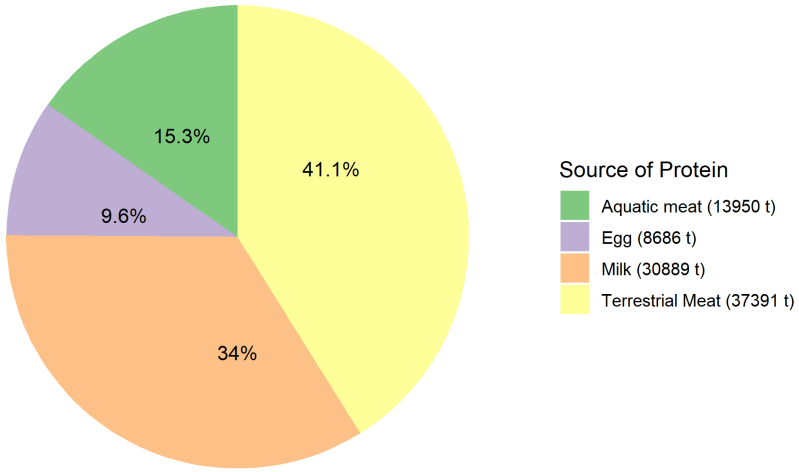




Other Pitfalls

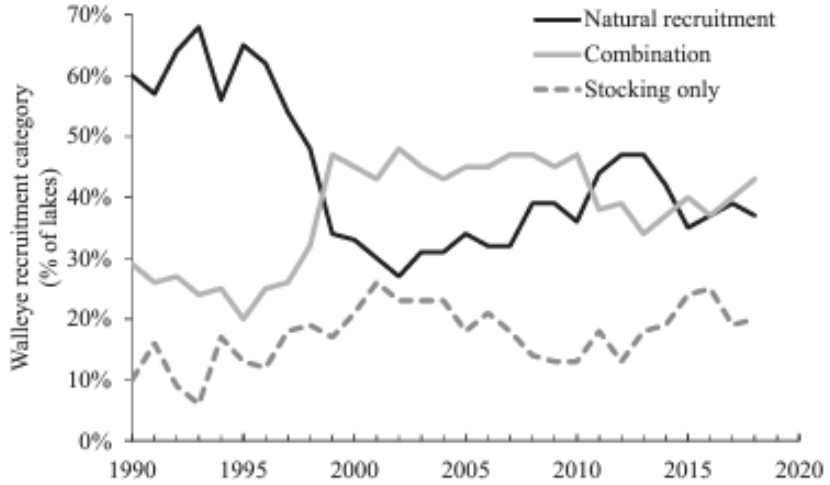
- Pie charts





Other Pitfalls

- Double axis graphs



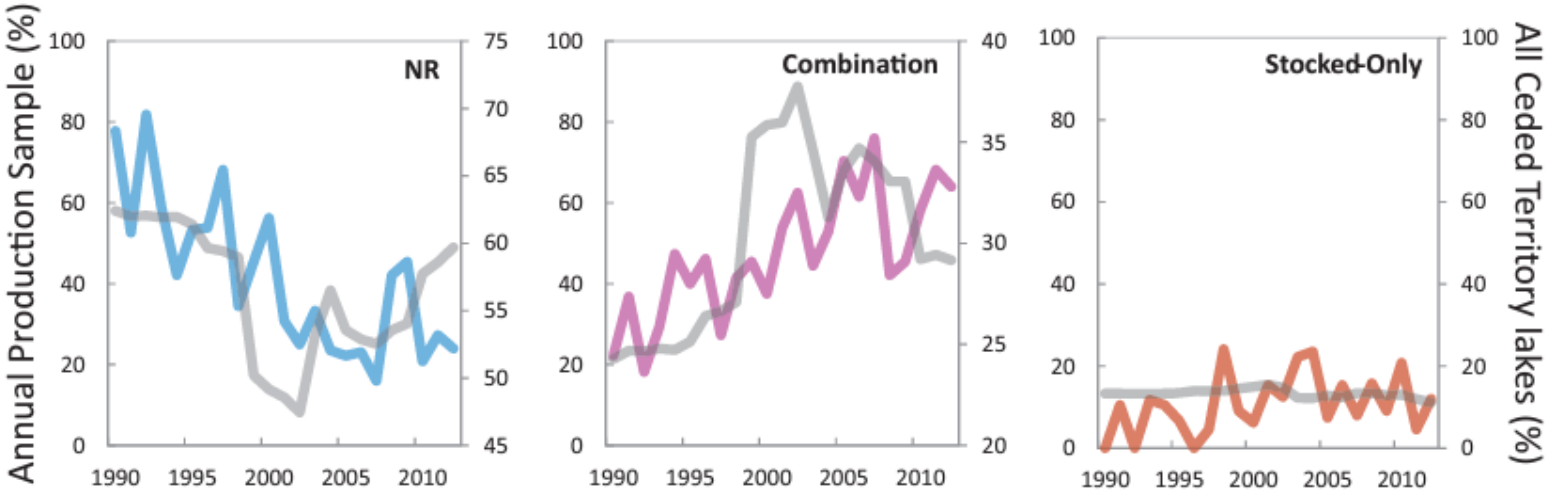


Fig. 4. Proportion of walleye (*Sander vitreus*) lakes classified into natural reproduction (NR), combination, and stocked recruitment categories. Walleye lakes across the entirety of the Wisconsin Ceded Territory are plotted on the secondary *y* axis as a dashed gray line. Production lakes (i.e., lakes where data were available for production calculations) are plotted on the primary *y* axis in color. [Colour online.]

Other Pitfalls

- 3D graph, incorrect graph type

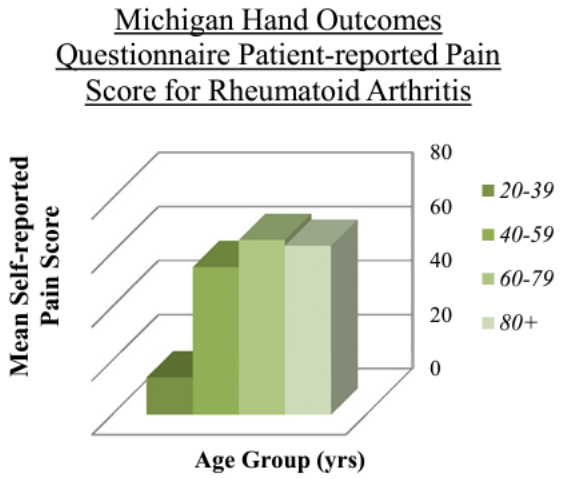


FIGURE 1: Three-dimensional bar graph.

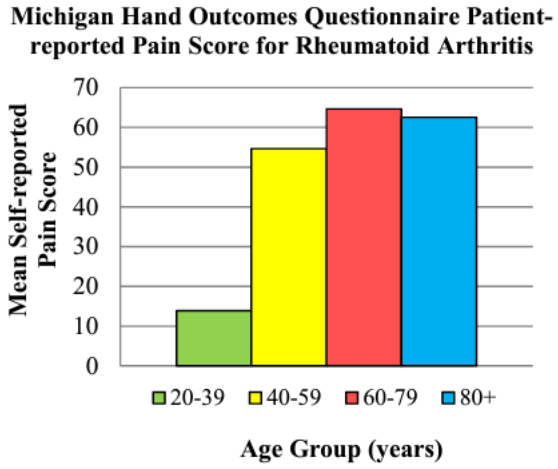


FIGURE 2: Two-dimensional bar graph.

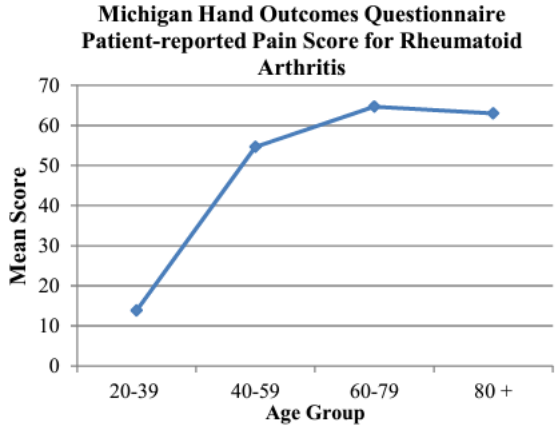
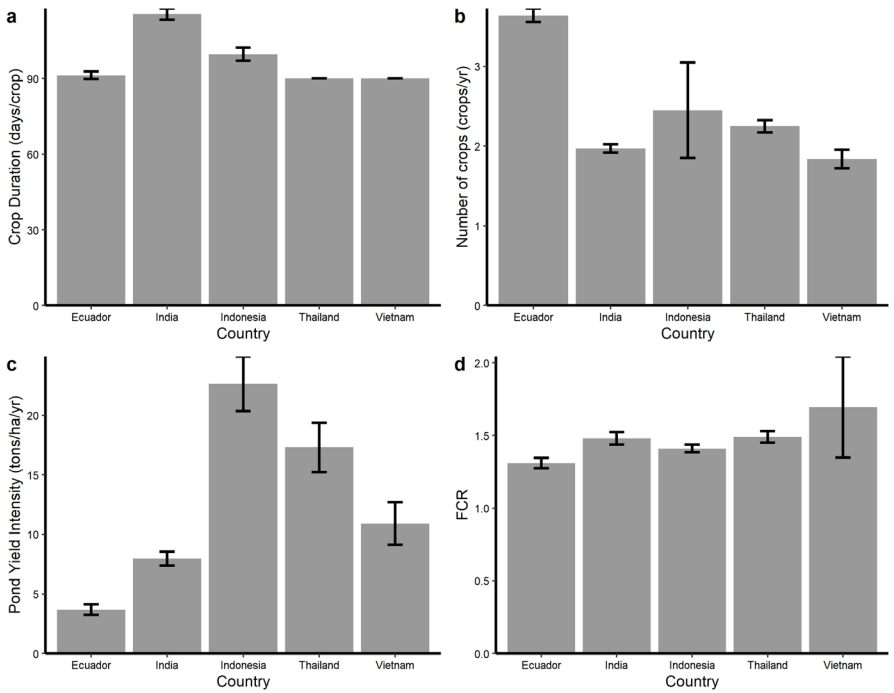
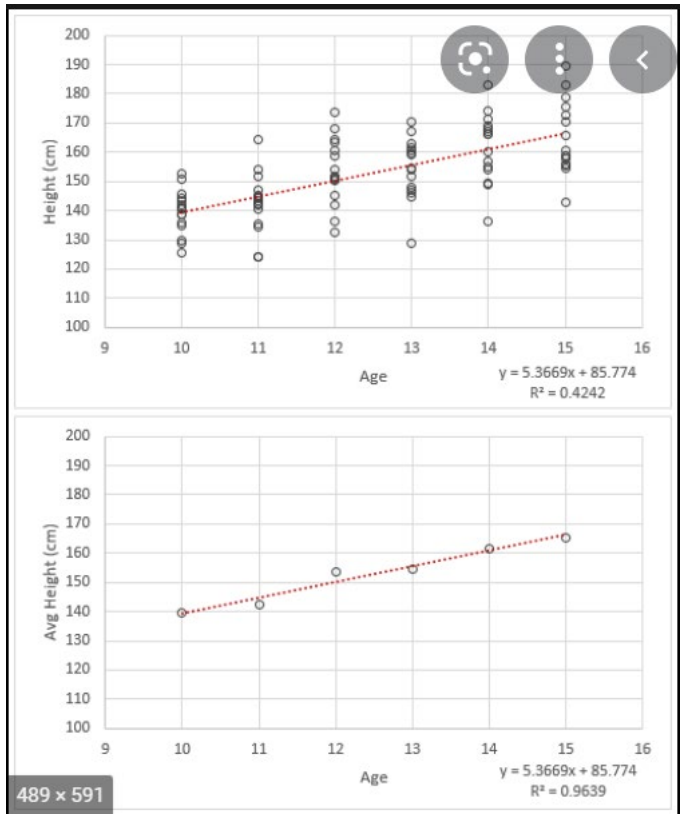


FIGURE 3: Connecting discrete data points.

Other Pitfalls

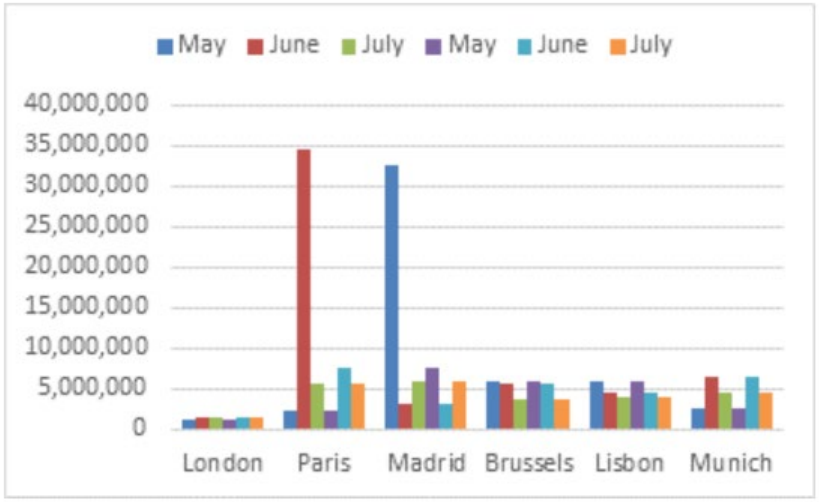
- Reducing dispersion

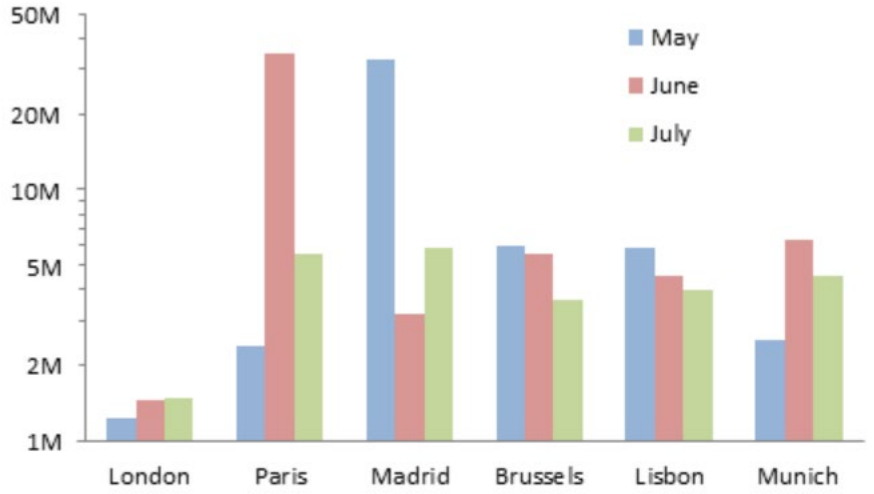


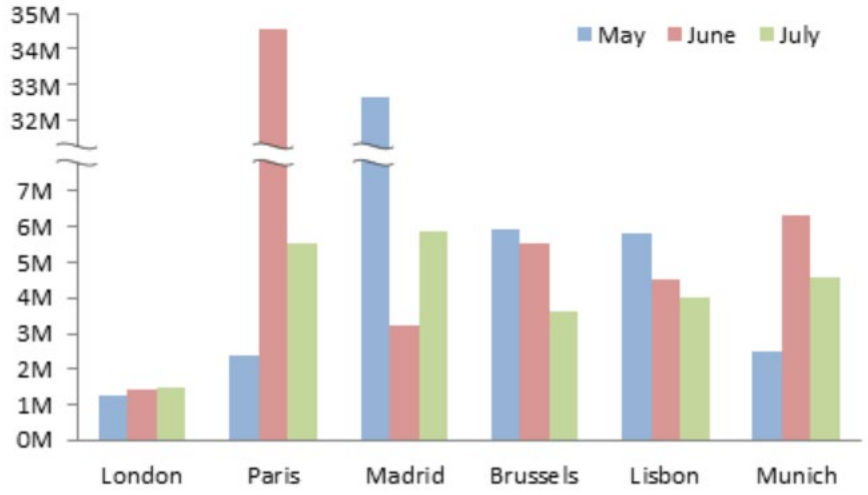


Other Pitfalls

- Misleading axis

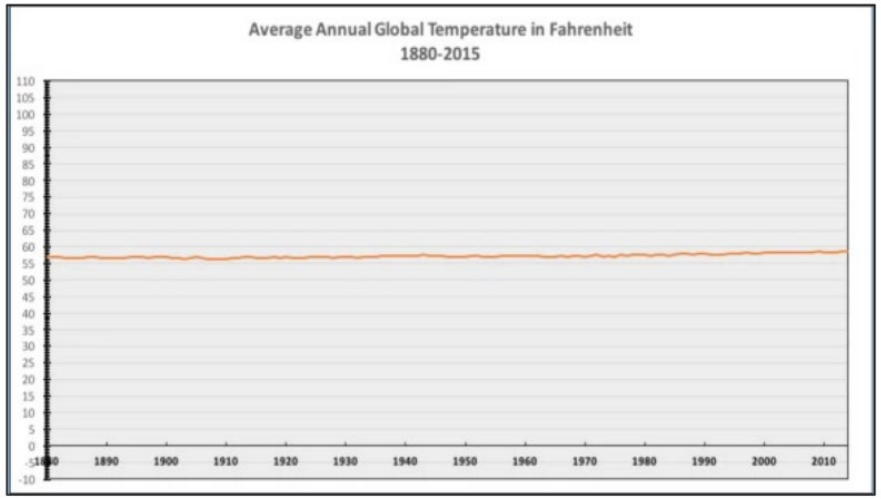


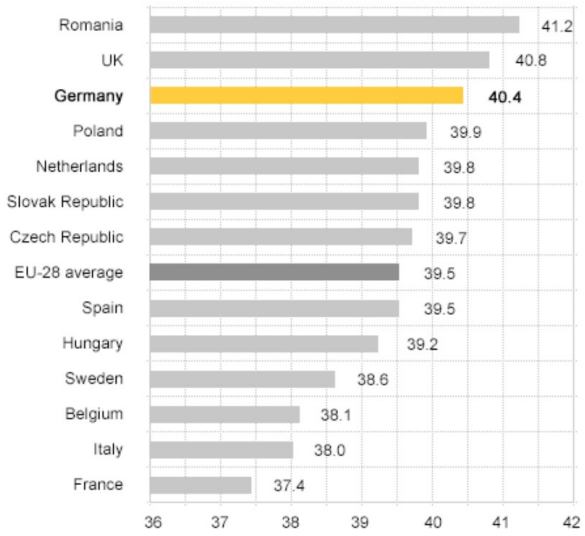


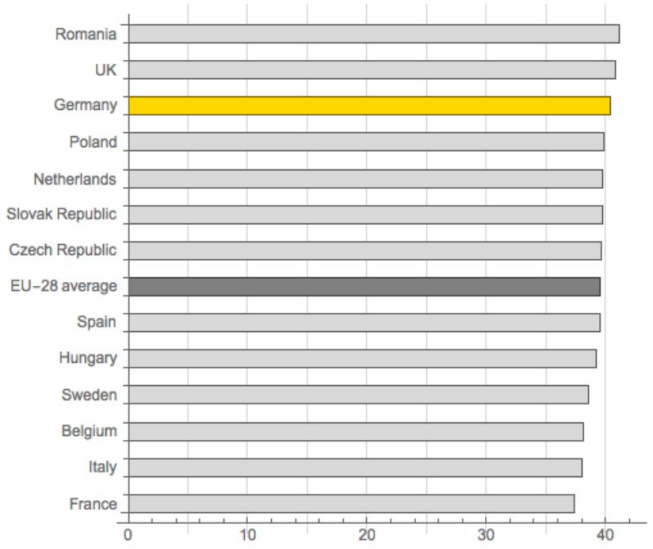


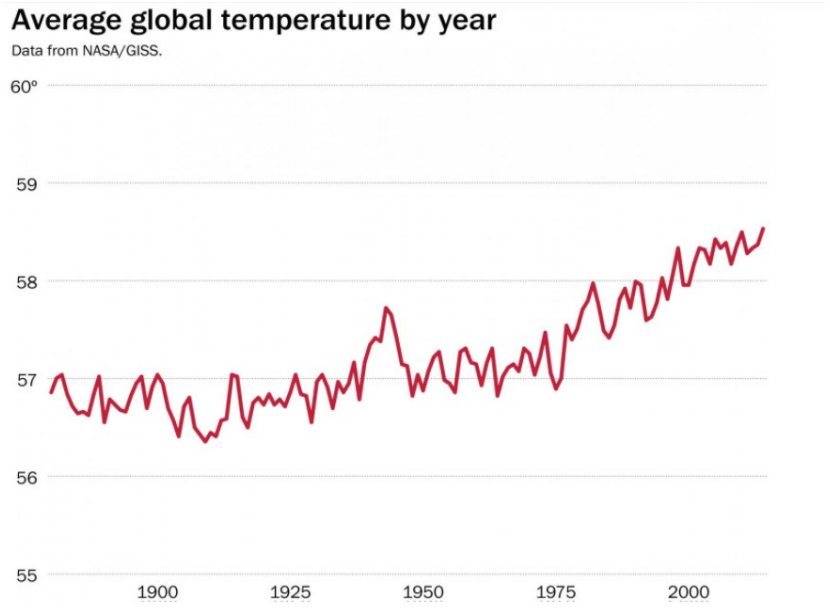
Other Pitfalls

- Misleading axis





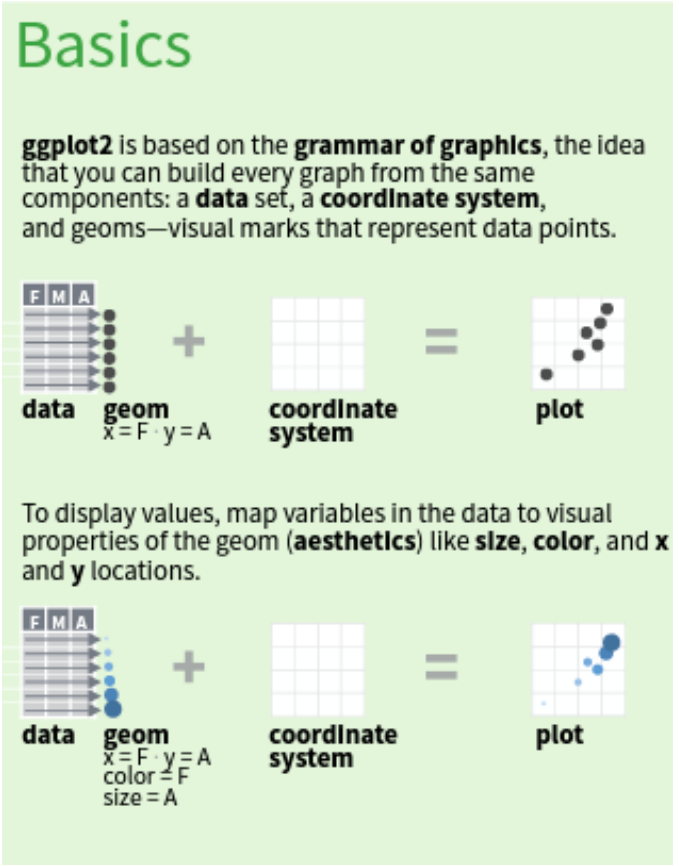




ggplot2 – The Tidyverse's plotting package

- Based on the "Grammar of Graphics" (hence the gg)
- Follows a logical system
  - data
  - coordinate System
  - geoms (shapes and objects to map data to)
  - aesthestics – colors, sizes, etc
    (How data are mapped).

  - Also "stats" and facets



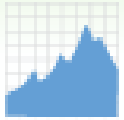
One Variable Styles

- One dimensional plots

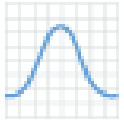
- geom_histogram()
- geom_density()
- geom_area()
- geom_bar()

ONE VARIABLE continuous

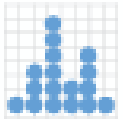
c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)



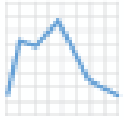
c + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size



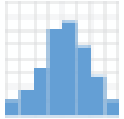
c + geom_density(kernel = "gaussian"**)**
x, y, alpha, color, fill, group, linetype, size, weight



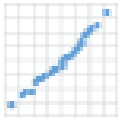
c + geom_dotplot()
x, y, alpha, color, fill



c + geom_freqpoly() x, y, alpha, color, group, linetype, size



c + geom_histogram(binwidth = 5**)** x, y, alpha, color, fill, linetype, size, weight



c2 + geom_qq(aes(sample = hwy)**)** x, y, alpha, color, fill, linetype, size, weight

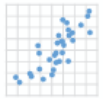
Two Variable Styles

- geom_point()
- geom_line()

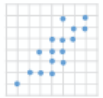
continuous x , continuous y

e <- ggplot(mpg, aes(cty, hwy))

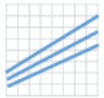
e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE**)** x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust



e + geom_jitter(height = 2, width = 2**)**
x, y, alpha, color, fill, shape, size

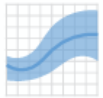


e + geom_point(), x, y, alpha, color, fill, shape, size, stroke



e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl"**)**, x, y, alpha, color, linetype, size

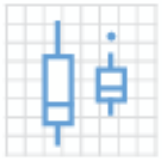


e + geom_smooth(method = lm**)**, x, y, alpha, color, fill, group, linetype, size, weight

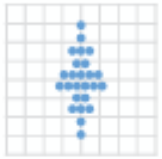
e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE**)**, x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

Data Exploration

- geom_boxplot()
- geom_violin()



f + geom_boxplot(), x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight



f + geom_dotplot(binaxis = "y", stackdir = "center"**)**, x, y, alpha, color, fill, group

f + geom_violin(scale = "area"**)**, x, y, alpha, color, fill, group, linetype, size, weight

https://blog.bioturing.com/2018/05/16/5-reasons-you-should-use-a-violin-graph/

Aesthetics (cont.)

- Can group colors, shapes, fills, etc. with aesthetic mapping
  - ggplot2 chooses colors by spacing them out evenly as possible on the color wheel
  - Can use other color scales
  - https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/
- Can change axis labels with xlab() and ylab()
- ggplot2 has “floating axes”…you can get rid of these with expand command in the scale_y_continuous or scale_x_continuous

Aesthetics

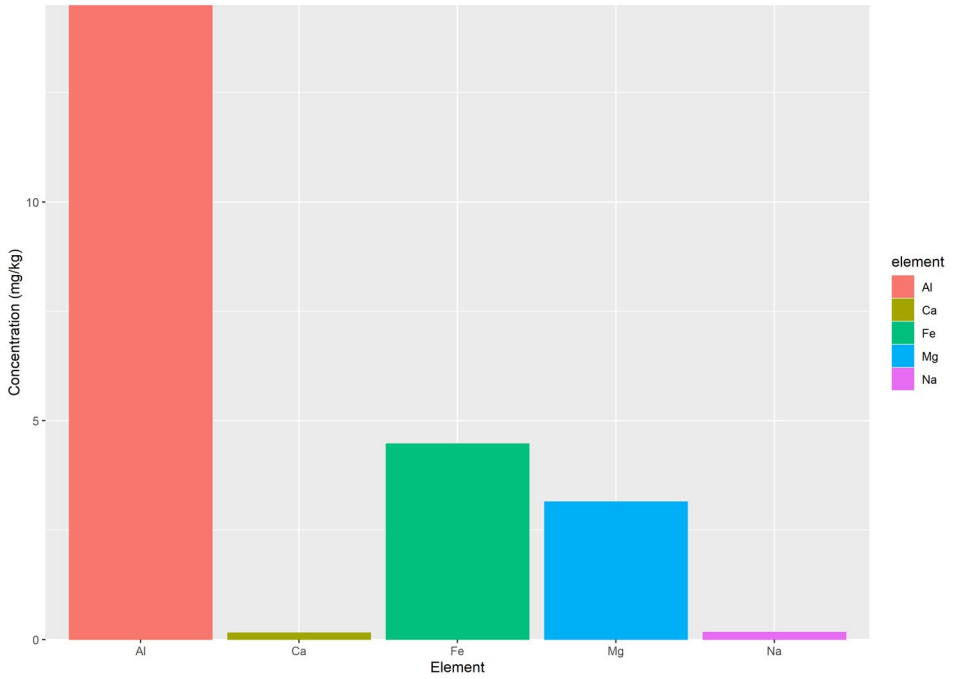
- Themes
  - Fixed graphing parameters that set a range of aesthetics
  - ggplot2 has canned themes
  - ggthemes is a separate package that contains additional themes
    - I like theme_few() with some tweaks
  - Can set individually on each plot like a geom object OR can set the entire script to one theme with theme_set()

https://yutannihilation.github.io/allYourFigureAreBelongToUs/ggthemes/

Saving

Two ways:

- With the GUI (show example)
- With code using ggsave()



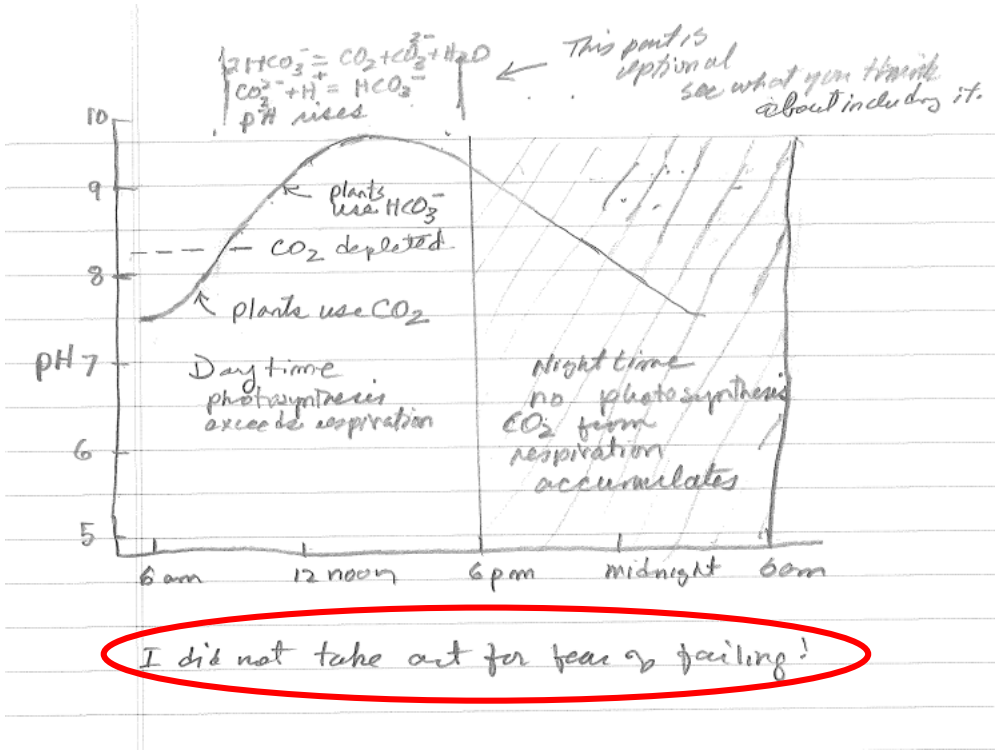
Corrplot

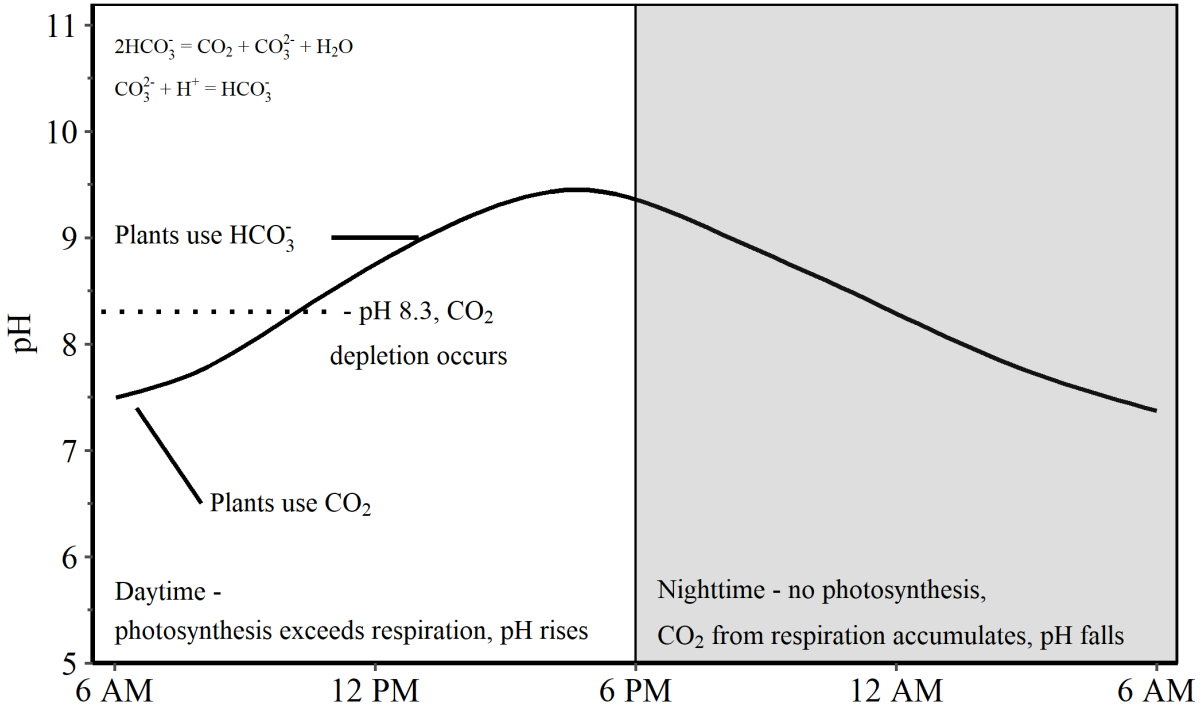
- Useful during data exploration to see which variables are related.

https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

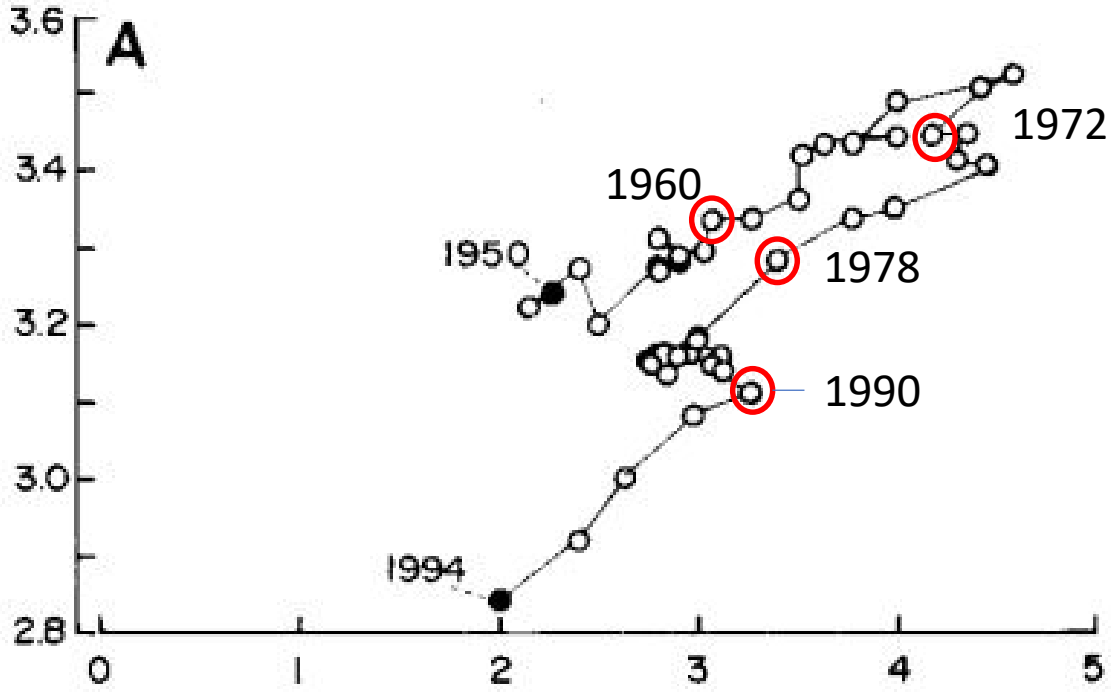
Metadigitize

- Use to extract data from figures





Metadigitize (homework)



- There are 46 years (1949 – 1994) on the graph (there is only 37 identifiable points)
- Some overlap of years makes it difficult to identify years, use the highlighted years as guidelines

- 1960 (12th point), 1971 (23rd), 1978 (30th), 1990 (42nd)