Data Visualization for Environmental Epidemiology with ggplot2

Mastering Presentation Grade Figures

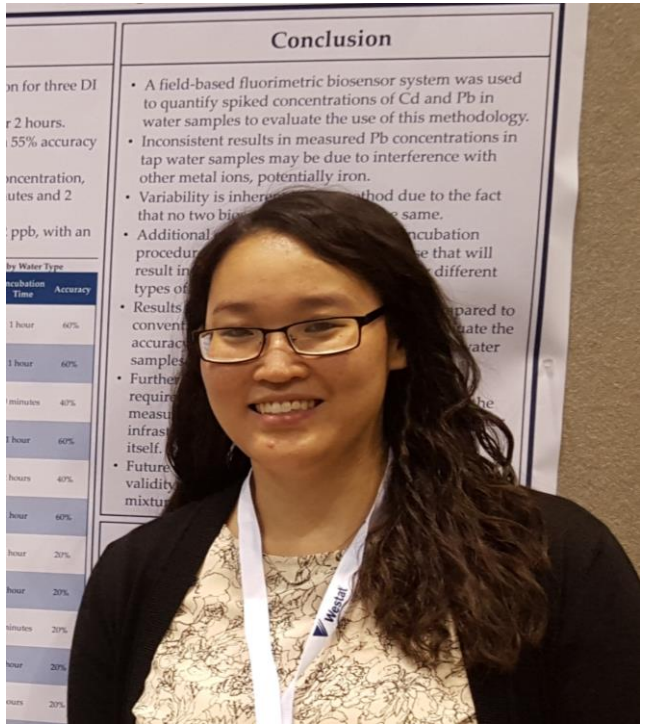
ISES 2021 Pre-Conference Workshop

August 6, 2021

Welcome!

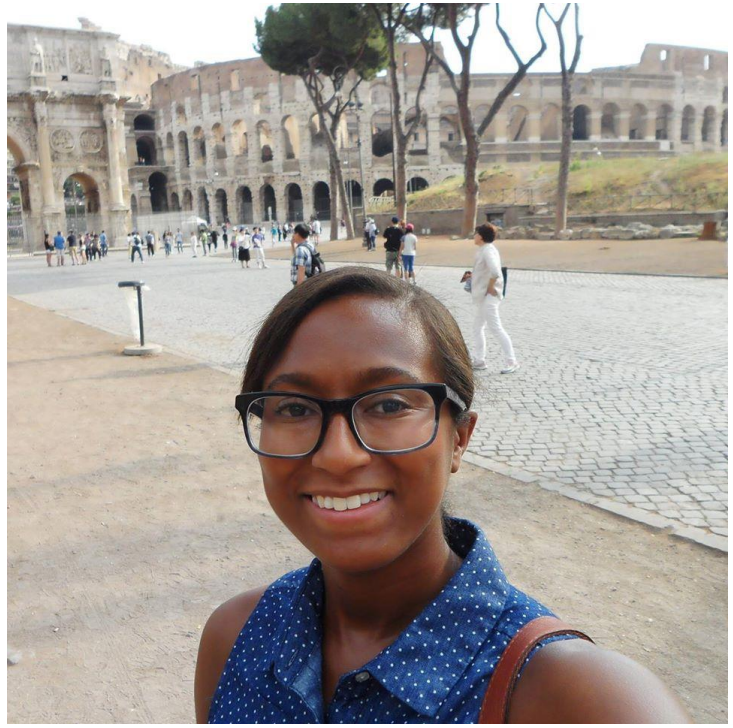
Alexandra Larsen, Ph.D.

US EPA | ORD | CPHEA | CPAD | TEAB-D



Alison Krajewski, Ph.D.

US EPA | ORD | CPHEA | HEEAD | IHAB



Lauren Wyatt, Ph.D.

US EPA | ORD | CPHEA | PHITD | CRB

Workshop Format

3x 50-minute sessions with 10-minute breaks in between.

Code examples in RStudio throughout.

All presentation materials are on GitHub:

https://github.com/USEPA/data-viz-ggplot2

Please feel free to ask questions!

Topics for Today

Section 1: What makes a “good” figure?

Section 2: The Basics (i.e. data formatting and ggplots)

Section 3: Customization (i.e. scales, colors, theme and facets)

Section 4: Complex plots (i.e. maps, examples)

Section 1

What makes a “good” figure?

Characteristics of Effective Plots

Graph type is appropriate for the data/results;

Scales are correctly formatted;

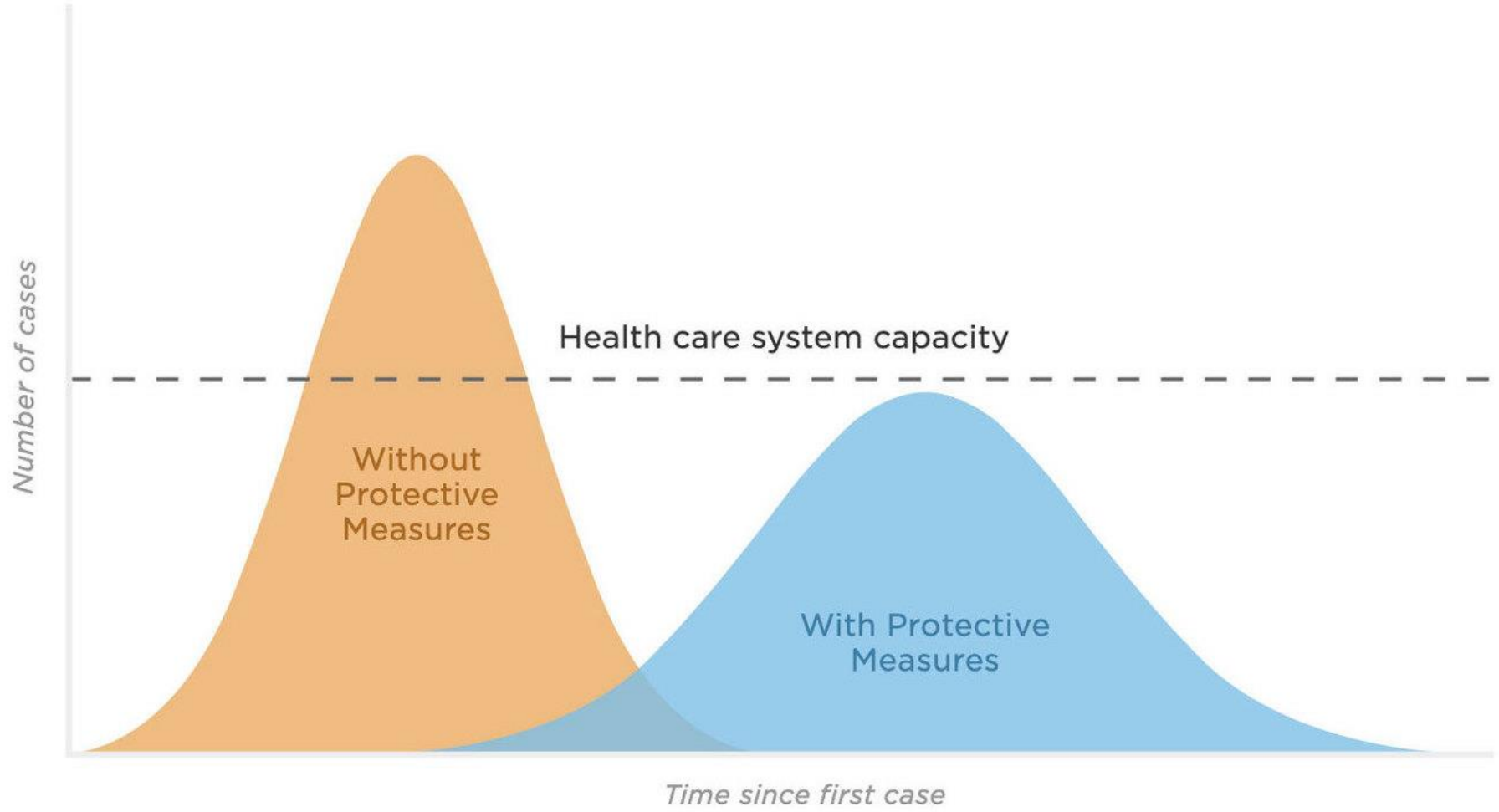
Message is clear enough to understand in a few minutes;

Formatting choices deliver the message instead of distract from it;

Facilitates informed decision-making.

lnes, Jason. “Effective Visualizations for Credible, Data-Driven Decision Making.” Rstudio Blog. Apr. 2020.

‘Flatten the Curve’



ge Credit: National Public Radio (NPR), Centers for Disease Control (CDC).

From the Literature:

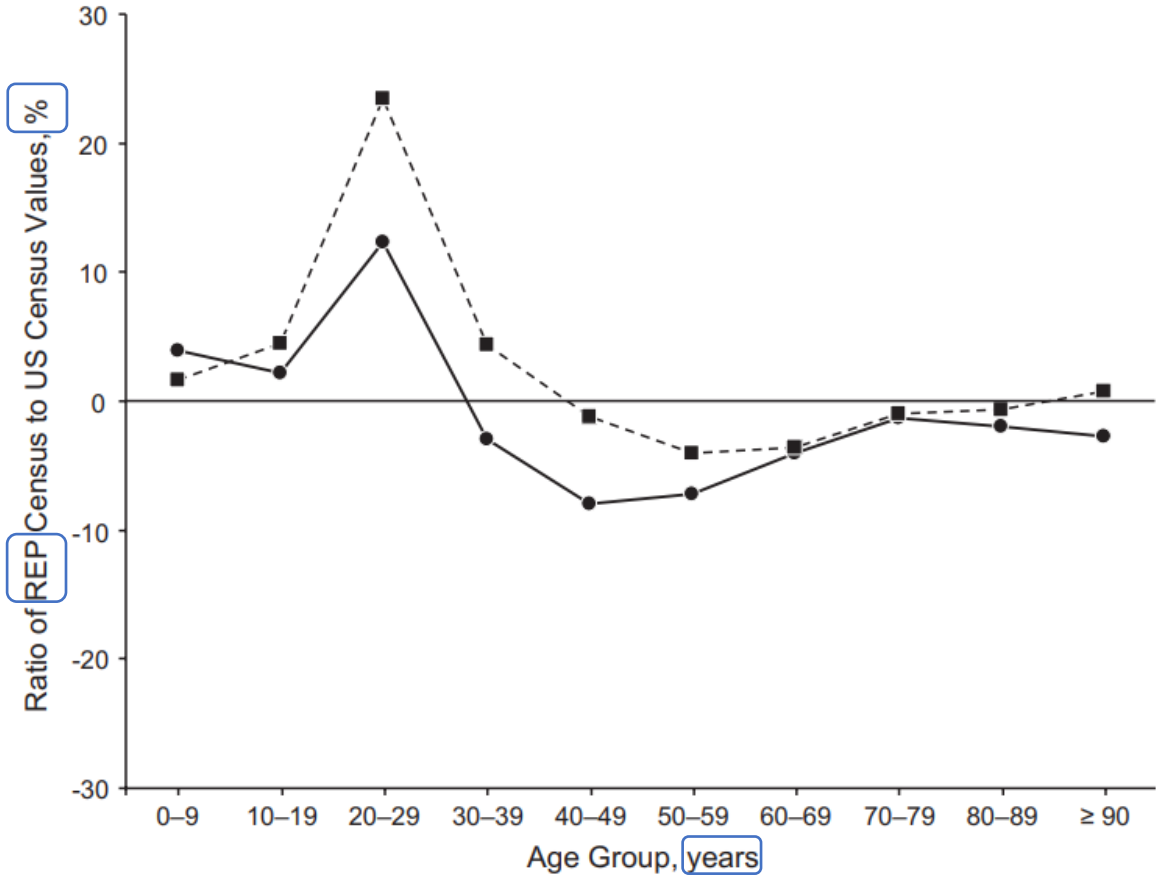


Figure 3. Age- and sex-specific capture rate by the Rochester Epidemiology Project (REP) medical records linkage system compared with US Census data (median capture rate in 1970, 1980, 1990, and 2000). Data from men (solid line, circle points) and women (dashed line, square points) are shown separately. The 0% line corresponds to perfect agreement between the system and the US Census. Values plotted above the 0% line indicate that the REP counted more persons than the US Census; values plotted below the 0% line signify that the REP counted fewer persons than the US Census.

St Sauver, Jennifer L.; Grossardt, Brandon R.; Yawn, Barbara P.; et al. Use of a Medical Records Linkage System to Enumerate a Dynamic Population Over Time: The Rochester Epidemiology Project. AMERICAN JOURNAL OF EPIDEMIOLOGY Volume: 173 Issue: 9 Pages: 1059-1068 Published: MAY 1 2011

ggplot2

Open-source R package for making scientific graphics.

"**gg**plot" = **G**rammar of **G**raphics (Leland Wilkinson).

Links variables in any data set to graphic components.

Allows for more flexibility and customization than built-in plots.

Lin Penderson, Thomas. "Plotting Anything with GGPlot2." celebRation Workshop. 2020.

Section 2

The Basics: data formatting and ggplots.

Data Formatting

Importing and Exporting Data Sets in R

R works with several data formats (e.g. .xlsx, .txt); **.csv** is easy and user-friendly.

To import .csv files, use **`read.csv()`**:

- Include the name and location of the .csv file, whether there are headers, etc.
- Assign imported file to a variable name; it becomes a 'data.frame()' object.

To export .csv files, use **write.csv()**:

- Include the data frame object, the file name, the name of the destination, whether to include headers, etc.

For other data formats, the import/export functions in R typically follow the format,

`{read/write}{_/.}{extension type}`,

But there are several formats, many of which are included in the examples…

Data Manipulation with `tidyr`

`Tidyr` is a library for data cleaning.

Includes intuitive functions for sub-setting, pivoting, etc.

Both `ggplot2` and `tidyr` use the **pipe operator**:

```
f(object, args) == object %>% f(args)
```

Allows for streamlined code that is easy to read.

The pipe operator is "`%>%`" in `tidyr` and "`+`" in `ggplot2`.

Long vs. Wide Data

Often need to convert between long and wide data.

Long (narrow, stacked):

One columns contain the values, and the other contains the description. Going from wide to long lengthens the data, increases the number of rows.

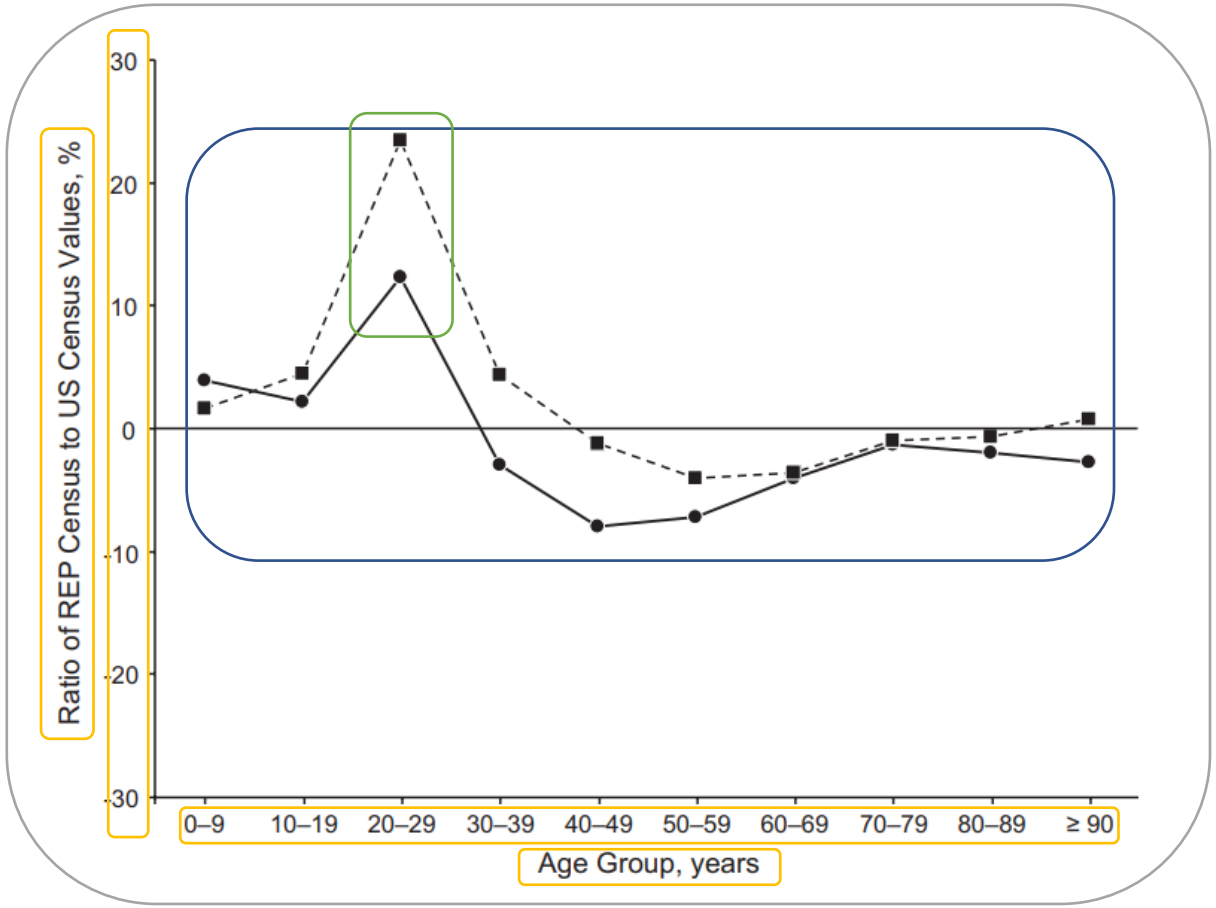
Wide:

Each variable in the data has its' own column. Going from long to wide shortens the data. *Very* useful format for plotting.

Go to Example…

ggplots

Basic Structure




```
ggplot(data = dat.csv) +
  geom_point(aes(x, y, shape, lty))+
  geom_hline(y = 0) +
  theme_bw() +
  ylim(c(-30, 30)) +
  ggtitle(null)
  xlab("Age Group, years") +
  ylab("Ratio of REP Census…") +
  scale_line_manual(guide = F) +
  scale_shape_manual(guide = F)
```

Geometries
(Plot type)

Themes
(background, style)

Labels, Axes
(breaks, limits etc.)

Scales, Legends
(Controls the aesthetics)

Creating a ggplot

Starts with a call to **`ggplot()`**, add graphic components in layers.

- E.g. plot type, scales, theme, etc.

In each component, specify data set and map aesthetics to the data.

- Aesthetic mapping: links variables in data to graphic component (e.g. x, y, color, shape, etc.)
- Data sets must be a data.frame() type, not numeric(), matrix(), etc.

ggplot2 Reference. https://ggplot2.tidyverse.org/reference/

Plot Types and Geometries (geoms)

`geom_{plot-type}()`

- Controls the points, lines, bars, polygons that go into making each type of plot
- Commonly used:
  - geom_point (scatterplot)
  - geom_histogram (histogram)
  - geom_boxplot (box plot)
  - geom_map (maps)
- Comprehensive list: ggplot2 Reference page

Go to Example…

ggplot2 Reference. https://ggplot2.tidyverse.org/reference/

Saving Figures

`ggsave()`

- Defaults to saving the last plot created; call immediately after making your figure.
- Can specify figure dimensions, resolution (dpi), format.
- Several output formats (png, jpg, tiff, etc.)

`grid.arrange()`, **`multiplot()`**, **`ggarrange()`**

- Combine multiple plots into one figure (alternative to facet)

Go to example…

Section 3

Customization: Scales, colors, theme, and facets.

Scales

Scales

- What are they?
  - Controls the mapping from data to aesthetics
  - Everything inside the aes() will have scales
  - Each scale is a function from a region in data space (domain of scale) to a region in aesthetic space (range of scale)

Wickham H. ggplot2: Elegant Graphic for Data Analysis. New York: Springer-Verlag; 2009.

Modifying Scales

- Scales are made up of three pieces separated by an underscore (_)
- Scale + name of aesthetic (e.g. colour, shape, or x/y) + the name of scale (e.g. discrete, continuous, brewer)
  - Examples:
    - scale_x_continuous()
    - scale_color_discrete()

Labels

- x and y axes labels will default to the variable name
- Modify the scales to change the label
  - scale_x_continuous(name="Label Name")
  - ylab("Label Name")
- Labels can include superscript, subscript, and mathematical expressions

Breaks

- Breaks control which values appear as tick marks on axes and keys on legends
- Breaks can be set on continuous or categorical scale
  - Used for labels, colors

Legends

- Legends can display multiple aesthetics (color, shape), from multiple layers
- Symbols displayed in legend varies based on the geom() used in layer
- Legends have more details that can be manipulated
  - Vertically or horizontally
  - Columns or rows
  - Size

Go to code and examples in markdown

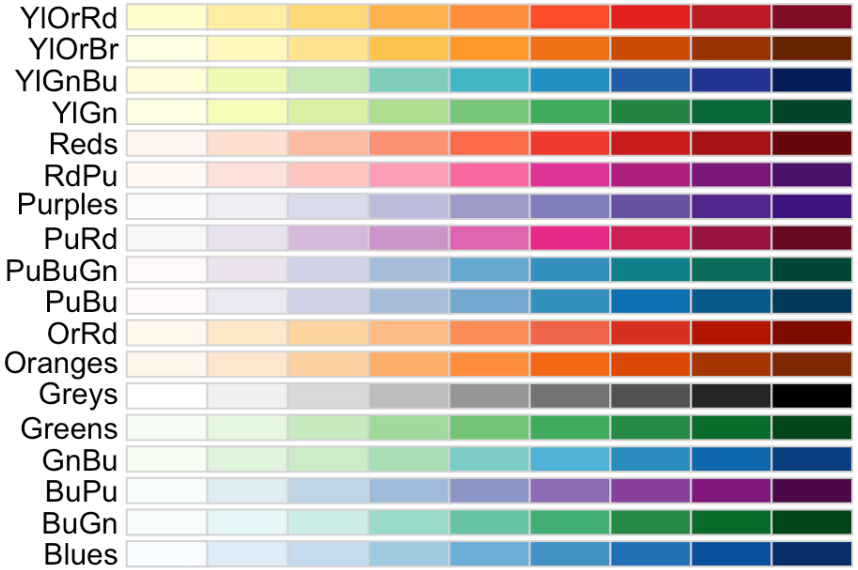
Colors

Colors

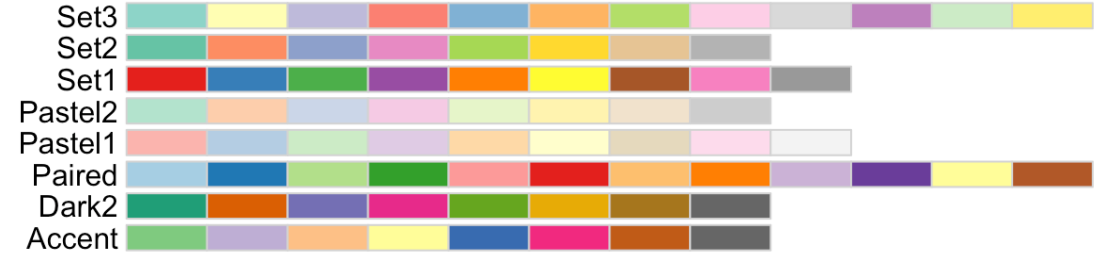
- Continuous
  - Note: for continuous color scales, keep the color scheme constant, and use a gradient scale
  - scale_colour_gradient() or scale_fill_gradient()
  - scale_color_distiller() or scale_fill_distiller()
- Discrete
  - Note: for discrete color scales, keep the color scheme color blind friendly
  - scale_colour_brewer()
    - Uses “ColorBrewer” colours (https://colorbrewer2.org)
  - scale_colour_grey()
    - Helpful for when needing grey-scale figures
  - scale_colour_manual()

ColorBrewer

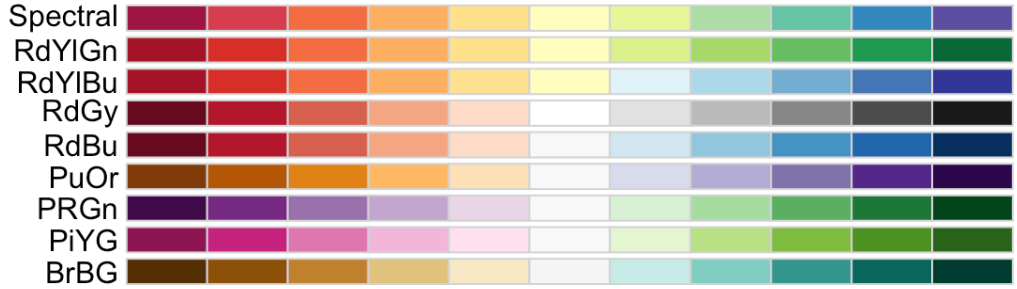
Sequential



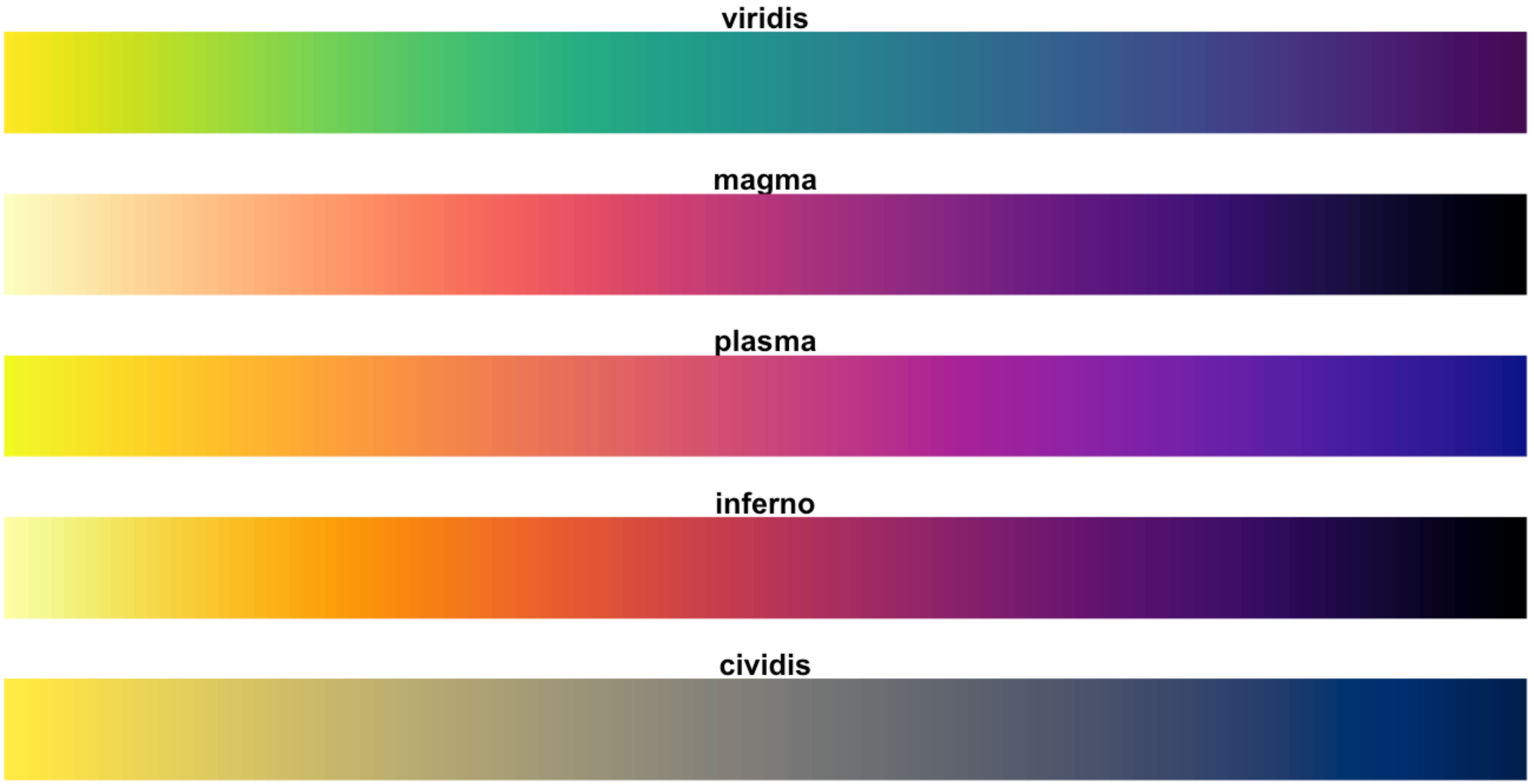
Qualitative



Divergent



Viridis



Creating Color Palettes

- Use breaks to create color palette
- Can use pre-existing color palettes
    - Specify in scale_color_brewer()
- Create a color palette with HEX codes
    - First, create a set of values
    - Second, specify with scale_color_manual

Go to code and examples in markdown

Theme

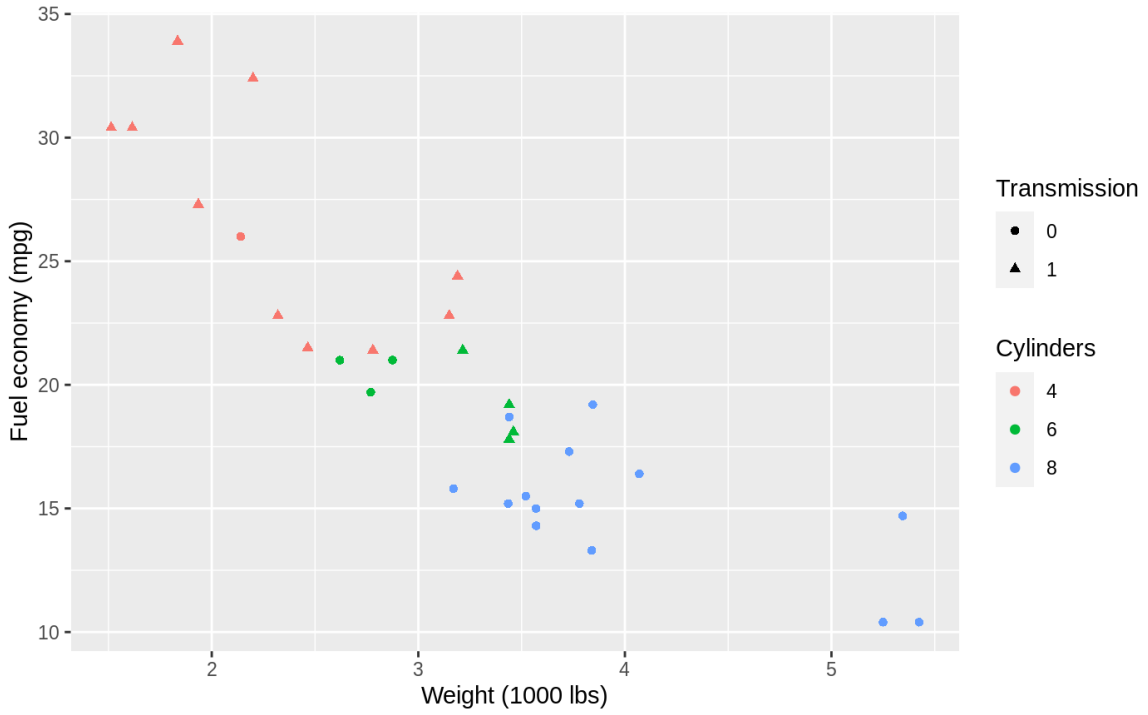
Theme

- What is a theme?
- Pre-existing themes
- Naming convention
- Customizing plot example

Theme

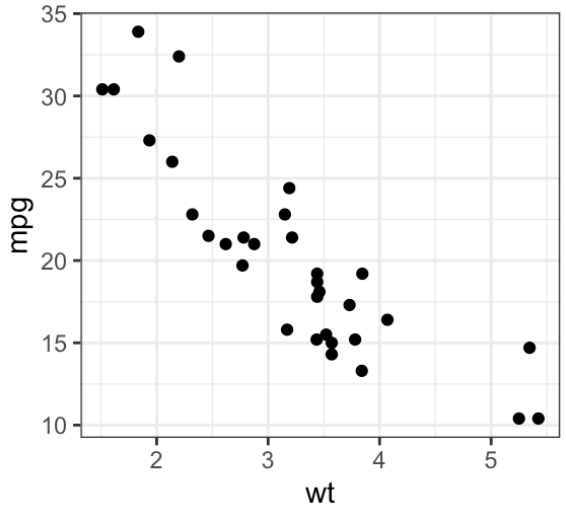
Customize non-data part of plots

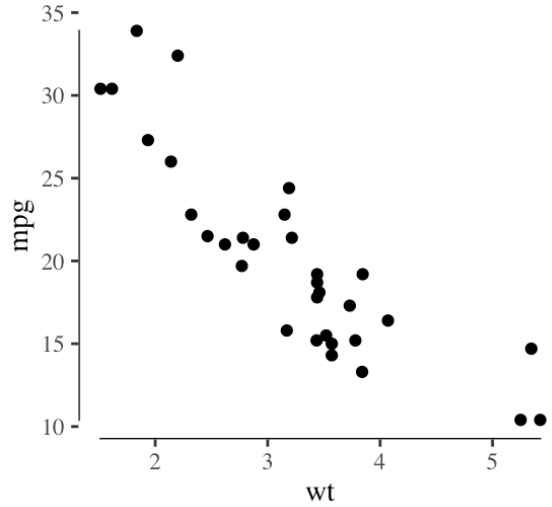
- Titles, labels, fonts, background, gridlines, legends
- Data exploration → Polished figure w/ focused message

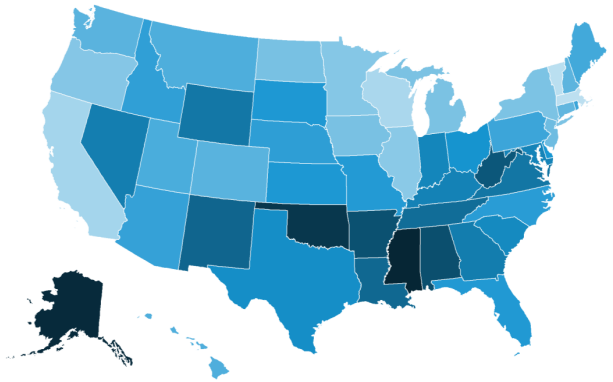


Use theme to focus attention to data

- Choosing – decluttering
- Maps – remove plot borders



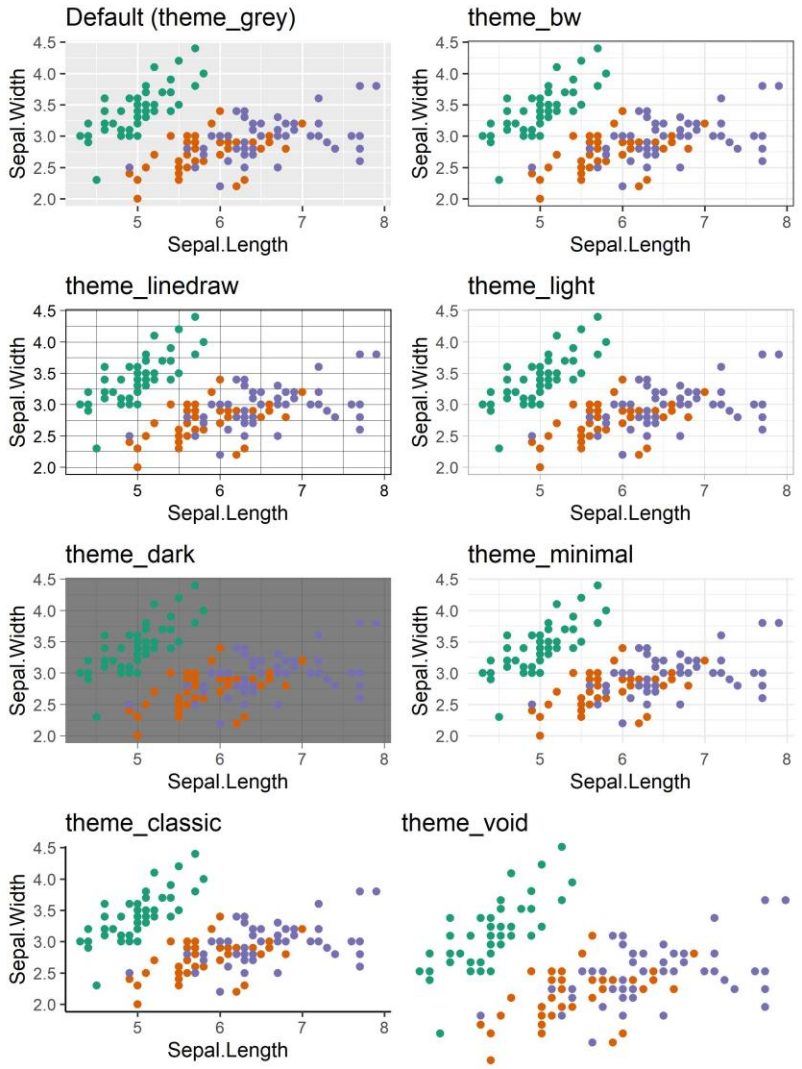




Pre-existing themes

Details

`theme_gray`	The signature ggplot2 theme with a grey background and white gridlines, designed to put the data forward yet make comparisons easy.
`theme_bw`	The classic dark-on-light ggplot2 theme. May work better for presentations displayed with a projector.
`theme_linedraw`	A theme with only black lines of various widths on white backgrounds, reminiscent of a line drawing. Serves a purpose similar to `theme_bw`. Note that this theme has some very thin lines (<< 1 pt) which some journals may refuse.
`theme_light`	A theme similar to `theme_linedraw` but with light grey lines and axes, to direct more attention towards the data.
`theme_dark`	The dark cousin of `theme_light`, with similar line sizes but a dark background. Useful to make thin coloured lines pop out.
`theme_minimal`	A minimalistic theme with no background annotations.
`theme_classic`	A classic-looking theme, with x and y axis lines and no gridlines.
`theme_void`	A completely empty theme.



- Great place to start, can customize further with theme()

Custom theme – theme()

- Main components
  - Line elements
    - axis lines, minor and major grid lines, plot panel border, axis ticks, etc.
  - Text elements
    - plot title, axis titles, legend title and text, axis tick mark labels, etc.
  - Rectangle elements
    - plot background, panel background, legend background, etc.

- Functions
  - element_line(color, size, linetype)
  - element_text(face, color, size, hjust, vjust, angle)
  - element_rect(fill, color, size, linetype)

Custom theme

- Naming convention
  - General
    - theme(
      axis.text = element_text(size = text_size))

Change text size for all text

  - More specific
    - theme(
      axis.text.x = element_text(size = text_size))

Change text size for only x-axis

Theme example with scatterplot

- Data
  - Daily ozone measurements in 2 cities
  - Time series
  - Source: Los Angeles Ozone Pollution Data, 1976 (package: mlbench)

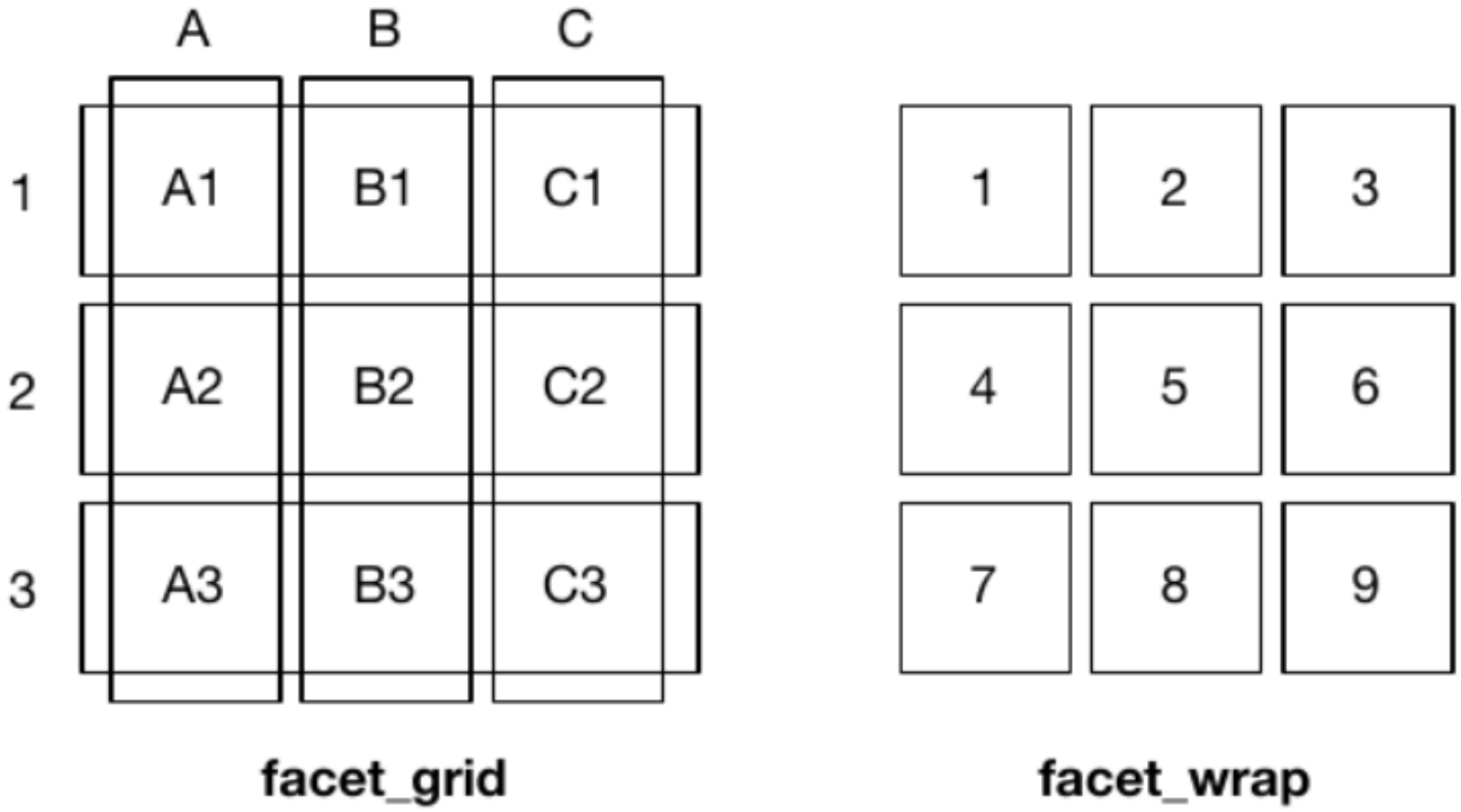
- Improvements
  - Gridlines
  - Text size
  - Rotate axis labels
  - Spacing between panels
  - Legend position

Facets

Facets

- What is it?
  - Facets generate small groupings with a different subset of the data
  - Powerful tool for exploratory data analyses
    - Readily compare patterns in different parts of data to see differences or similarities
    - Panel layout may carry meaning
  - Three types
    - facet_null(): a single plot
    - facet_wrap(): wraps a ribbon of panels
    - facet_grid(): produces a grid of panels defined by variables forming the rows and columns

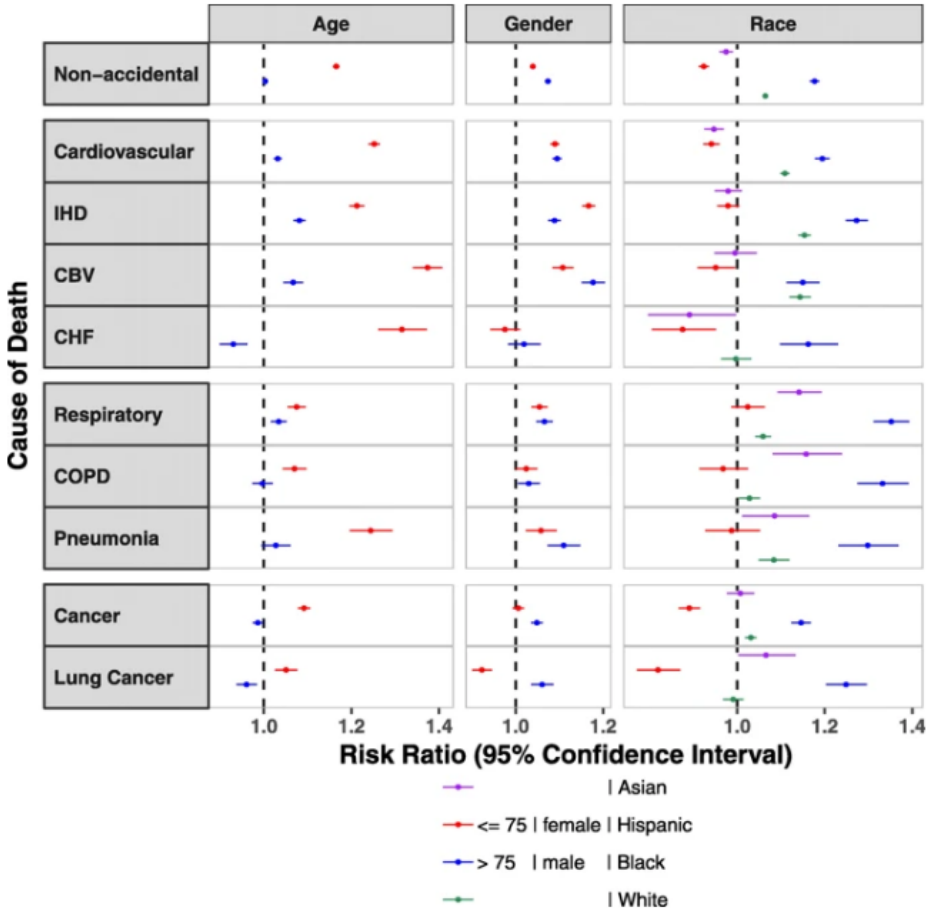
Facet Grid vs. Facet Wrap



Wickham H. ggplozt2: Elegant Graphic for Data Analysis. New York: Springer-Verlag; 2009.

Example from the Literature

Fig. 2



Modification of the SES-adjusted Association between $PM_{2.5}$ and Cause-specific Mortality by Age, Sex, and Race. For each cause of death, we examined effect modification using interaction terms for age, sex and race respectively in the SES-adjusted models. Results are expressed as the risk ratio and 95% CIs per 10 $\mu g/m^3$ increase in 12-month average $PM_{2.5}$. Abbreviations: IHD (Ischemic heart disease), CBV (Cerebrovascular disease), CHF (Congestive heart failure), COPD (Chronic Obstructive Pulmonary disease), SES (Socio-Economic Status), $PM_{2.5}$ (particles with aerodynamic diameters < 2.5 μm). Note: Each subgroup in the death-group box follows the same order defined in the figure legend

Wang B, Eum K, Kazzemiparkouhi F, Li C, Manjourides J, Pavlu V, and Suh H. The impact of long-term PM2.5 exposure on specific causes of death: exposure-response curves and effect modification among 53 million US Medicare beneficiaries. Environmental Health 2020; 19: 20. doi: https://doi.org/10.1186/s12940-020-00575-0

Facet Grid

- Lays out figures in a grid defined by . ~ option
  - .~a spreads that values of variable a across columns, which allows for comparisons of the y-axis because the vertical scales are aligned
  - b~. Spreads the values of variable b down rows, which allows for comparisons of the x-axis because the horizontal scale are aligned
  - a~b spreads variable a across columns and variable b down rows
    - Usually put the variable with the highest number of levels in the columns

Facet Wrap

- Control wrap with: ncol, nrow, as.table, and dir
  - ncol and nrow controls the number of columns and rows in the arrangement
  - as.table controls the layout to be like a table, with highest values at the bottom right (as.table=TRUE) or the highest values at the top right (as.table=FALSE)
  - dir controls the direction of the wrap (horizontal or vertical)

Scales in Facets

- Position of scales can be the same in all panels (fixed) or vary between panels (free)
  - scales = “fixed”: the x and y axes are fixed across the panels
  - scales = “free_x”: the x axis is free, but the y axis is fixed
  - scales = “free_y”: the y axis is free, but the x axis is fixed
  - scales = “free”: x and y axes vary across the panel

Go to code and examples in markdown

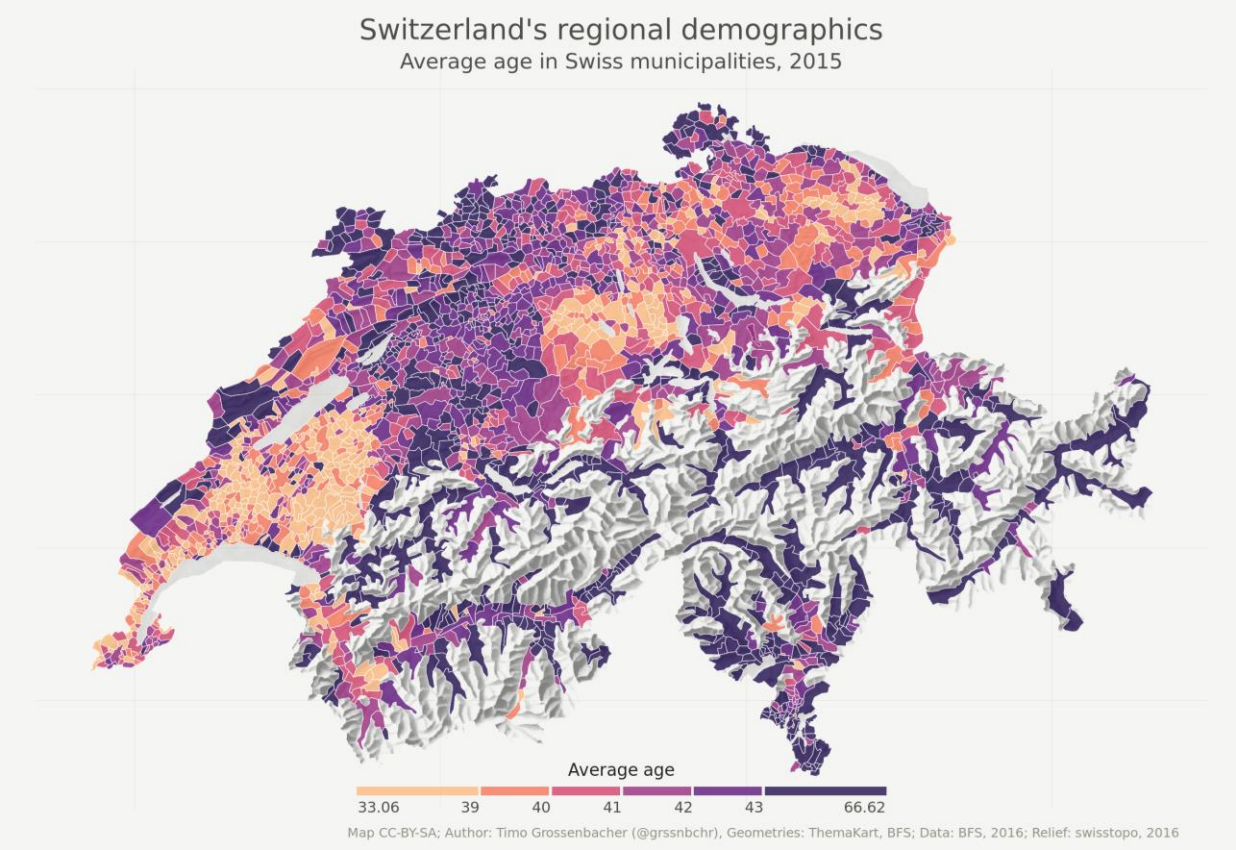
Section 4

Complex plots: Maps, examples from the literature.

Mapping

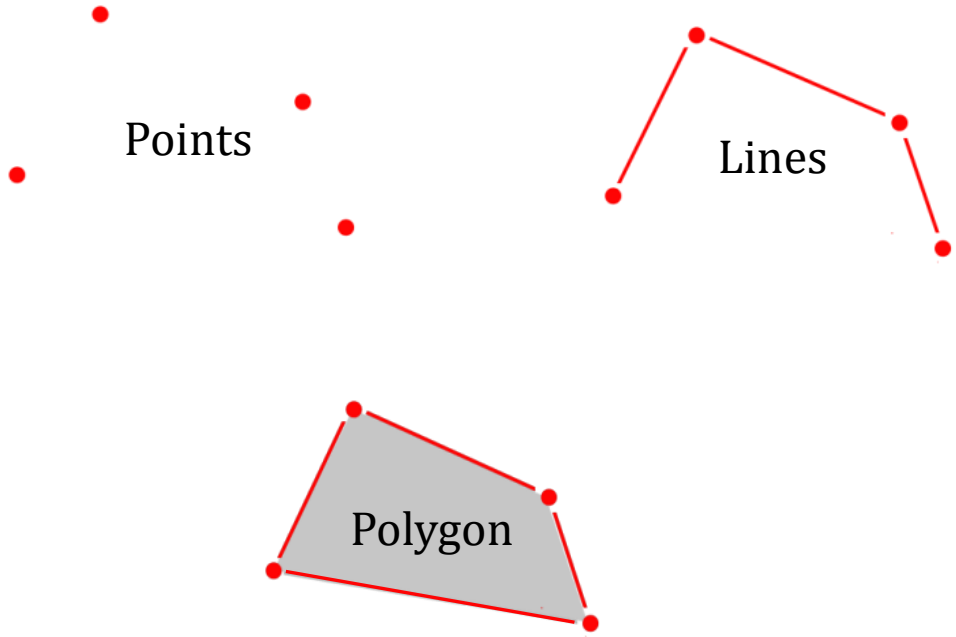
Mapping

- Fundamentals
- Layout and formatting
- Color palettes
- Create and customize examples
  - Continuous
  - Categorical

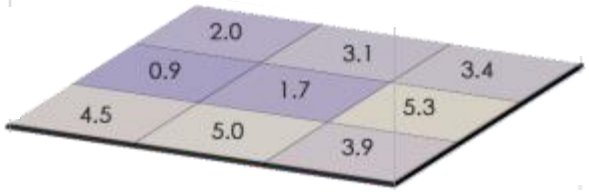


Fundamentals – spatial data types

- Vectors
  - Points (cities, landmarks)
  - Lines (roads, rivers)
  - Polygons (country borders)



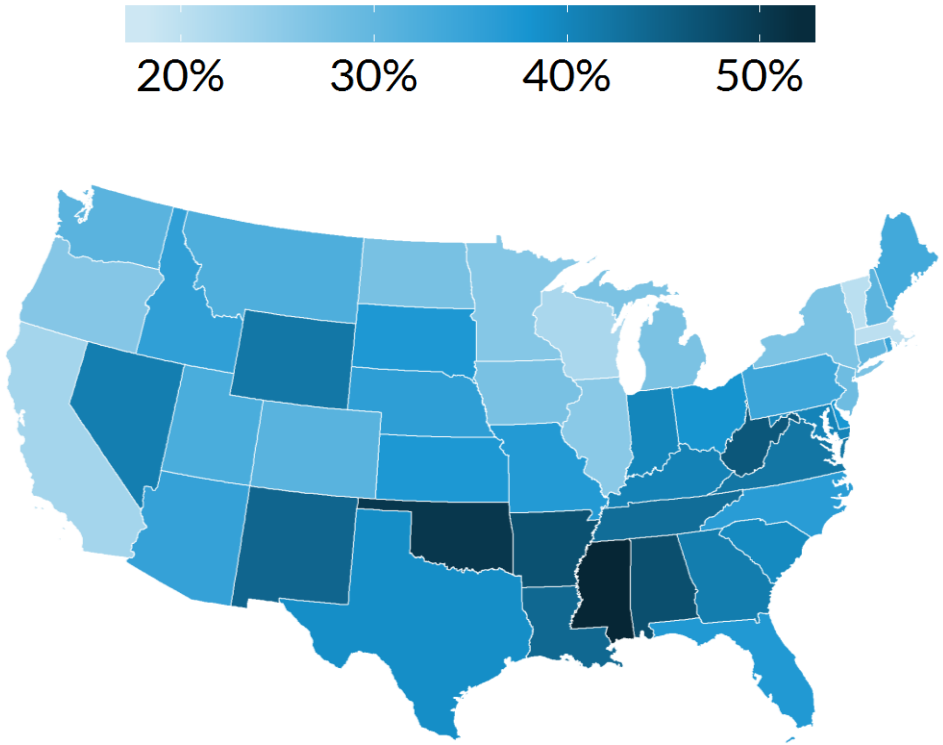
- Raster



Fundamentals – data prep → plot

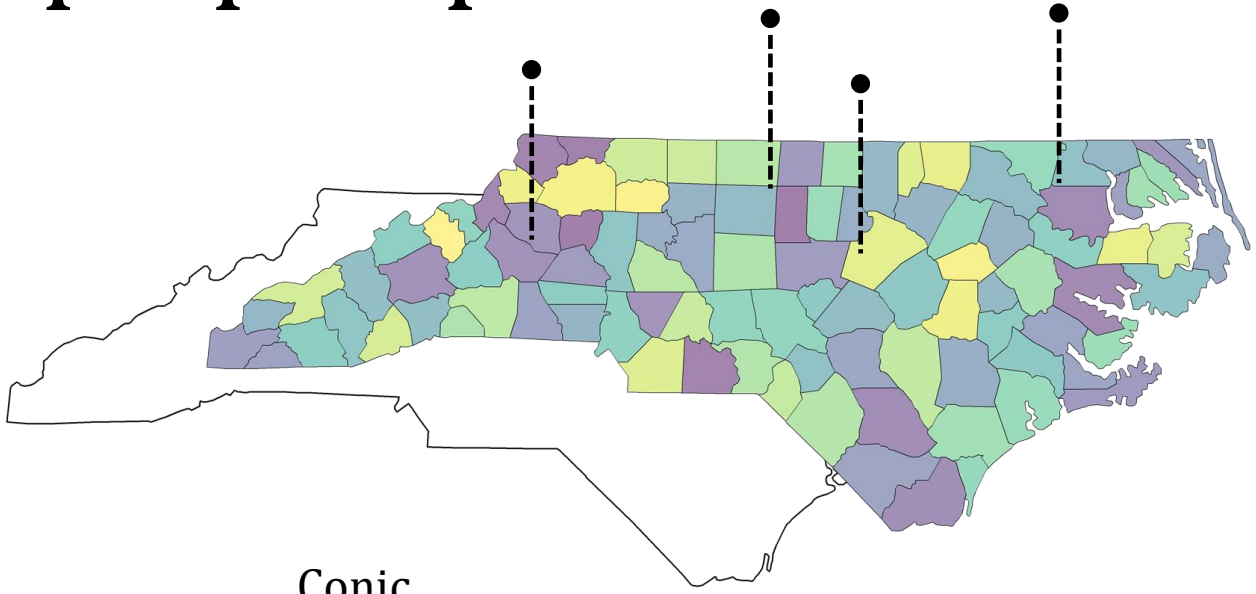
sf package (spatial features)

- Load data
  - Shapefiles
  - Data to display
- Join data of interest to shapefile



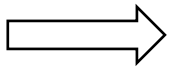
Fundamentals – data prep → plot

- Geometry order is important!



- Consider map projection

Cylindric



Conic

Choropleth example

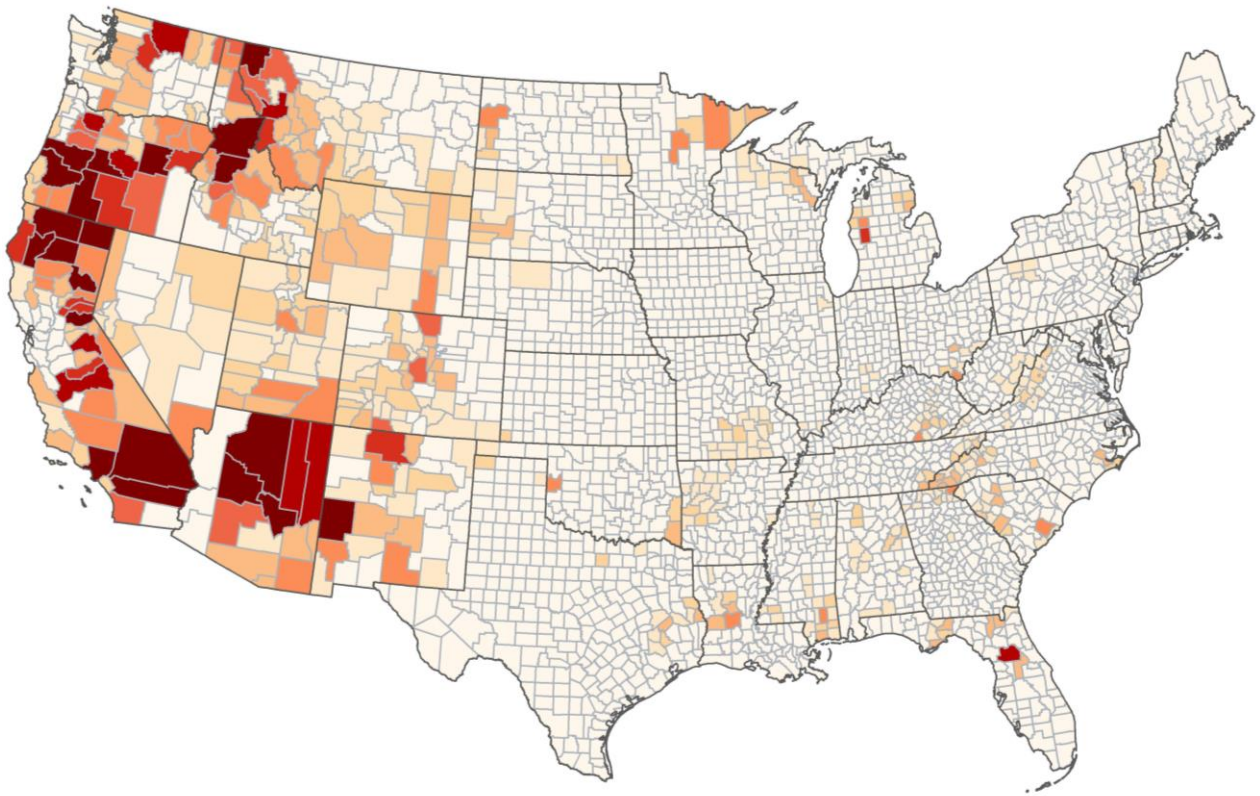
US counties

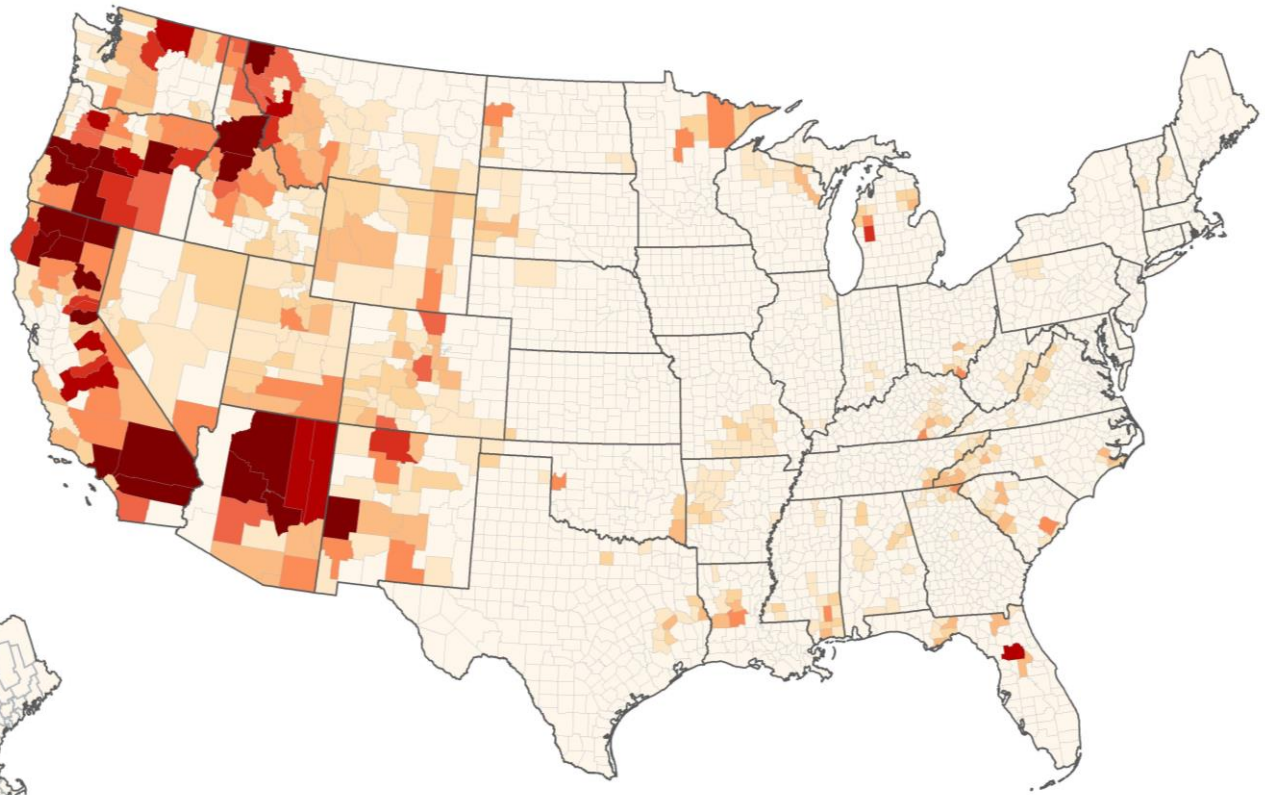
- Data
  - Spatial wildfire occurrence data for the United States
  - County-level counts of fires (2008 and 2009)
  - Source: USDA Forest Service

- Prepare data
  - Load shapefiles for state and county boundaries
    - Some packages have shapefiles for certain geographies (maps)
    - Manually load, st_read()
  - Combine with data of interest

- Improvements
  - Adjust focus to map, background adjustment
  - Projection
  - Color palette
  - Line colors and thickness
  - Facet

Reducing county line size





References

References

- Pederson TL. ggplot2 Workshop [Internet]. Copenhagen: GitHub; 2020 [updated 2020 March 25; cited 2020]. Available from: https://github.com/thomasp85/ggplot2_workshop.
- Peng RD and Dominici F. Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health. New York: Springer-Verlag; 2008. Available from: https://www.springer.com/gp/book/9780387781662
- R Studio. Data Visualizations with ggplot2 Cheat Sheet [Internet]. Boston: R Studio; 2020 [updated 2015 March; cited 2020]. Available from: https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
- Scherer C. A ggplot2 Tutorial for Beautiful Plotting in R [Internet]. 2019 [updated 2019 November 1; cited 2020]. Available from: https://cedricscherer.netlify.app/2019/08/05/a-ggplot2-tutorial-for-beautiful-plotting-in-r/
- Wickham H. ggplot2: Elegant Graphic for Data Analysis. New York: Springer-Verlag; 2009. Available from: https://ggplot2-book.org/
- Wilkinson L. The Grammar of Graphics. 2nd ed. New York: Springer-Verlag; 2005. Available from: https://www.springer.com/gp/book/9780387245447
- ZevRoss. Beautiful plotting in R: A ggplot2 cheatsheet [Internet]. Ithaca: ZevRoss; 2014 [updated 2016 January 20; cited 2020]. Available from: http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/

References for Data

- United States Environmental Protection Agency (US EPA). National Air Quality: Status and Trends of Key Air Pollutants, Air Quality Trends. Available from: https://www.epa.gov/air-trends. [Accessed 2020].
- Institute for Health Metrics and Evaluation (IHME). United States Hypertension Estimates by County 2001-2009. 2013. Seattle, United States of America: Institute for Health Metrics and Evaluation (IHME). Available from: http://ghdx.healthdata.org/record/ihme-data/united-states-hypertension-estimates-county-2001-2009. [Accessed 2020].
- U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. Spatial wildfire occurrence data for the United States, 1992-2011 [FPA_FOD_20130422] (1st Edition) Data publication contains GIS data Author(s): Short, Karen C. Publication Year:2013. Available from: https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0009. [Accessed 2020].
- National Cancer Institute (NCI) and Centers for Disease Control and Prevention (CDC). State Cancer Profiles, Incidence Rate Tables. 2018. https://statecancerprofiles.cancer.gov/incidencerates/index.php. [Accessed 2020].
- Centers for Disease Control and Prevention (CDC). National Environmental Public Health Tracking. 2018. https://www.cdc.gov/nceh/tracking/. [Accessed 2020].
- Rdocumentation for "Epi"Package.

  https://rdocumentation.org/packages/Epi/versions/2.44

References for Examples

- St Sauver JL, Grossardt BR, Yawn BP, Melton LJ 3rd, Rocca WA. Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. Am J Epidemiol. 2011 May 1;173(9):1059-68. doi: 10.1093/aje/kwq482. Epub 2011 Mar 23.
- Liu JC , Wilson A, Mickley LJ, Dominici F, Ebisu K, Wang Y, Sulprizio MP, Peng RD, Yue X, Son JY, Anderson GB, Bell ML. Wildfire-specific Fine Particulate Matter and Risk of Hospital Admissions in Urban and Rural Counties. Epidemiology. 2017 Jan;28(1):77-85. doi: 10.1097/EDE.0000000000000556.
- Sanders NJ, Barreca AI, Neidell MJ. Estimating Causal Effects of Particulate Matter Regulation on Mortality. Epidemiology. 2020 Mar;31(2):160-167. doi: 10.1097/EDE.0000000000001153.
- Pun VC, Kazemiparkouhi F, Manjourides J, Suh HH. Long-Term PM2.5 Exposure and Respiratory, Cancer, and Cardiovascular Mortality in Older US Adults. Am J Epidemiol. 2017 Oct 15;186(8):961-969. doi: 10.1093/aje/kwx166.
- Sidney S, Quesenberry CP Jr, Jaffe MG, Sorel M, Go AS, Rana JS. Heterogeneity in national U.S. mortality trends within heart disease subgroups, 2000-2015. BMC Cardiovasc Disord. 2017 Jul 18;17(1):192. doi: 10.1186/s12872-017-0630-2.
- Wang B, Eum KD, Kazemiparkouhi F, Li C, Manjourides J, Pavlu V, Suh H. The impact of long-term PM2.5 exposure on specific causes of death: exposure-response curves and effect modification among 53 million U.S. Medicare beneficiaries. Environ Health. 2020 Feb 17;19(1):20. doi: 10.1186/s12940-020-00575-0.
- Correia AW, Pope CA 3rd, Dockery DW, Wang Y, Ezzati M, Dominici F. Effect of air pollution control on life expectancy in the United States: an analysis of 545 U.S. counties for the period from 2000 to 2007. Epidemiology. 2013 Jan;24(1):23-31. doi: 10.1097/EDE.0b013e3182770237.