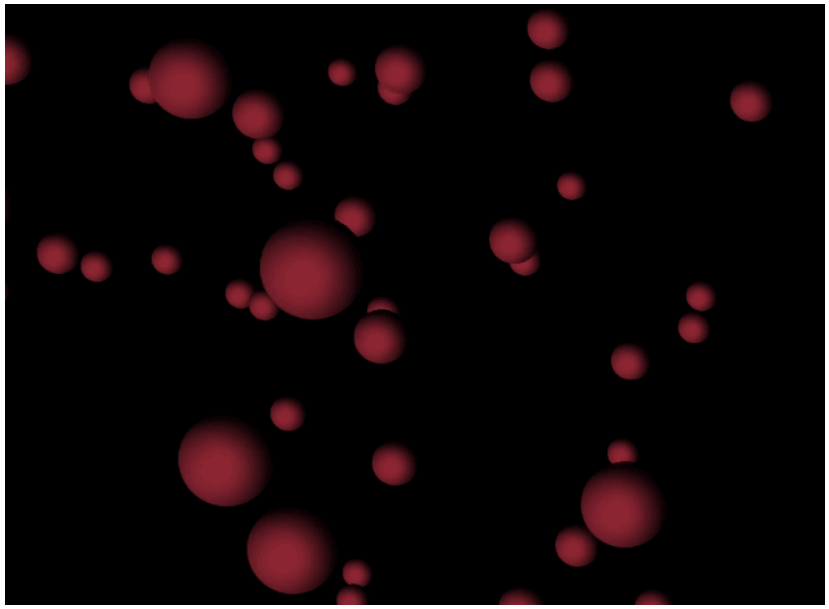
GR5702
Exploratory Data Analysis and Visualization

Prof. Joyce Robbins



http://flowingdata.com/2017/01/19/embedding-js-data-driven-environments-for-virtual-reality/

Today's Agenda (1/19/ 17)

- Final Project
- R

(5 minute break at 6:15pm)

Final Project

- 1 to 4 people
- Step #1: choose a topic that interests you
- Step #2: find a dataset that hasn't been overanalyzed
- Step #3: post your idea on Piazza to find team members
- May be more EDA or presentation-oriented
- Final form can be a document or web-based
- Will be graded on

Final Project

- The following will be valued:
  - Ability of techniques to produce data insights
  - Appropriateness of visualization techniques
  - Clarity and coherency
  - Reproducibility
  - Explanation / Documentation (with a clear audience in mind)
  - Overall quality
- More details to follow...

Explanation / Documentation (a.k.a. writing)

Choose a genre:

- industry report
- research brief
- academic journal article
- news article
- blog post

Health Insurance Coverage in the United States: 2015

Current Population Reports

By Jessica C. Barnett and Marina S. Vornovitsky

Issued September 2016
P60-257(RV)

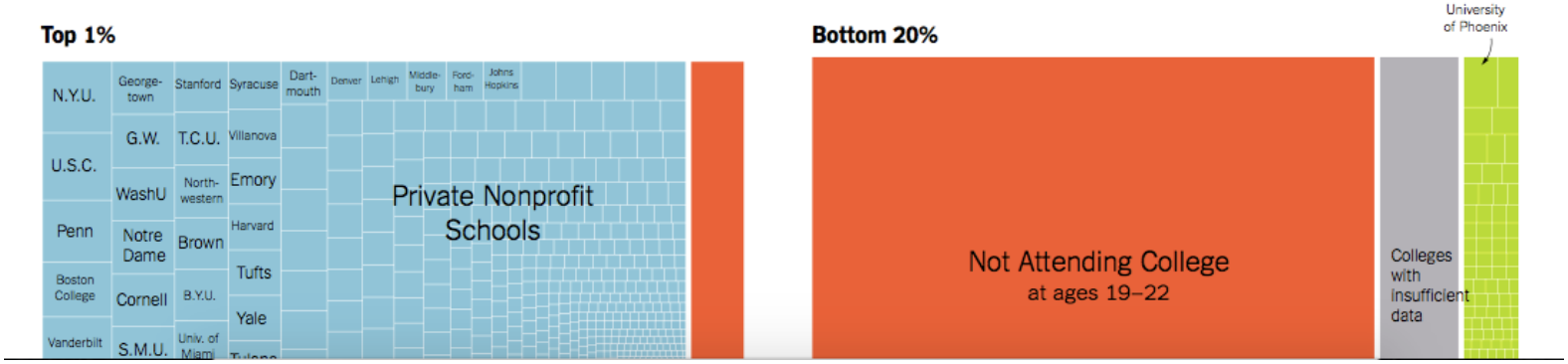
United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Some Colleges Have More Students From the Top 1 Percent Than the Bottom 60. Find Yours.

JAN. 18, 2017

Where the top 1% and the bottom 20% go to college



Geophysical Research Letters

AN AGU JOURNAL

Explore this journal >

Research Letter

Was Venus the first habitable world of our solar system?

M. J. Way, Anthony D. Del Genio, Nancy Y. Kiang, Linda E. Sohl, David H. Grinspoon, Igor Aleinov, Maxwell Kelley, Thomas Clune

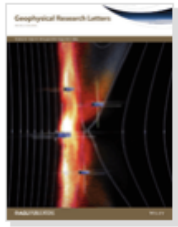
First published: 28 August 2016 Full publication history

DOI: 10.1002/2016GL069790 View/save citation

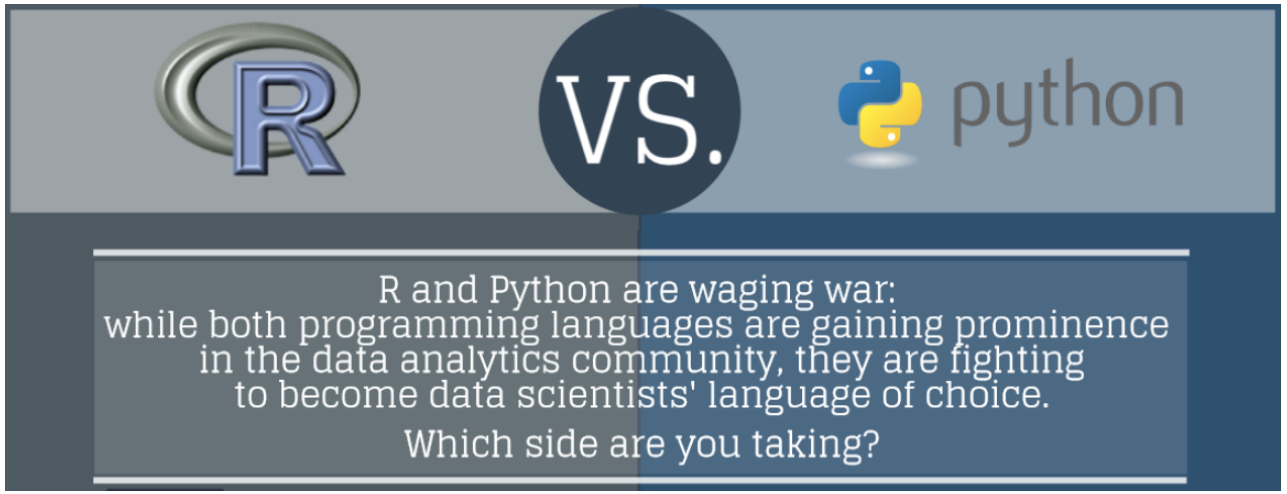
Cited by: 0 articles Citation tools

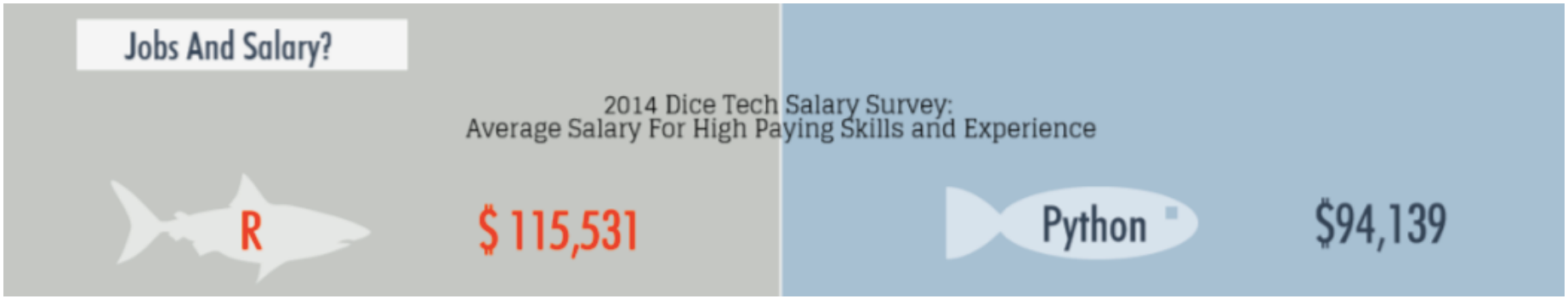
Am score 883

Funding Information

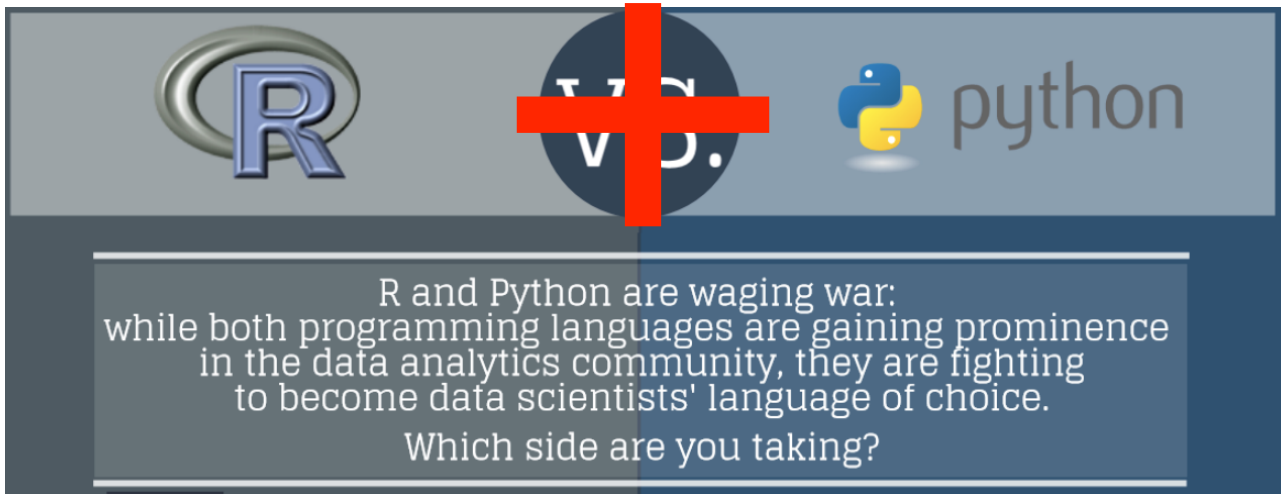


View issue TOC
Volume 43, Issue 16
28 August 2016
Pages 8376–8383





https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis



https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

A picture says more than a thousand words

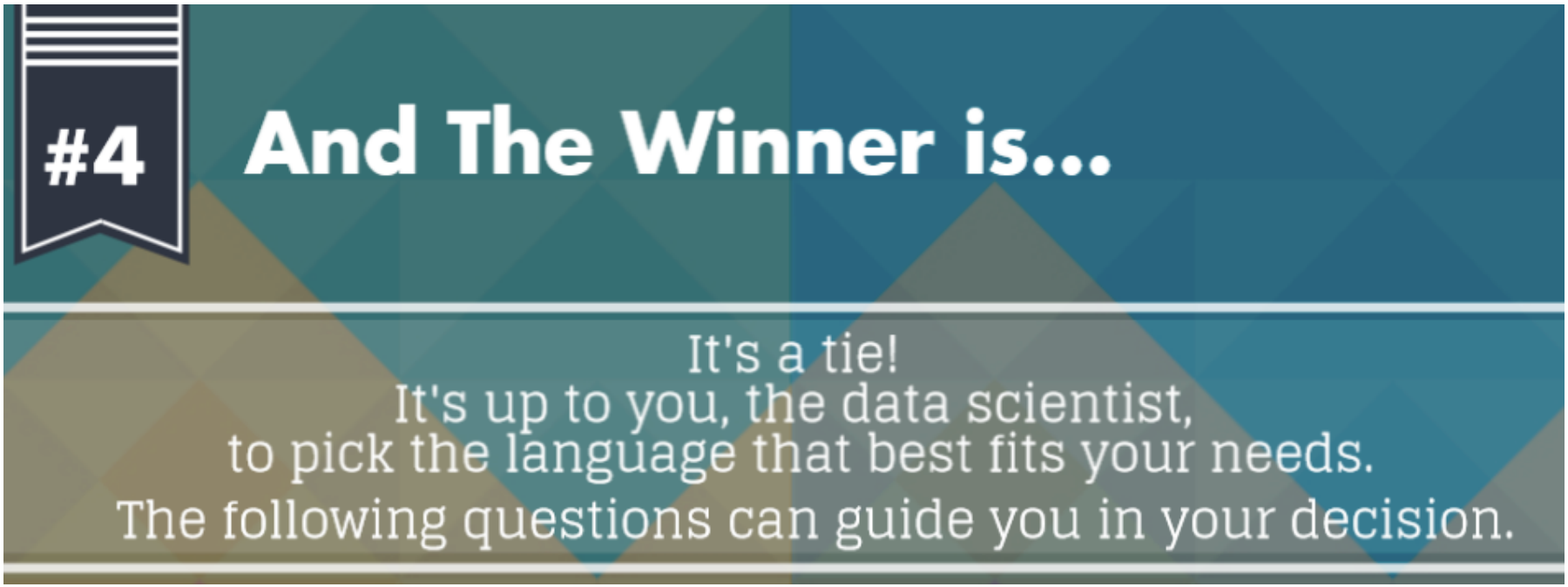
Visualized data can be understood more efficiently and effectively than the raw numbers alone.

R + visualization
= perfect match

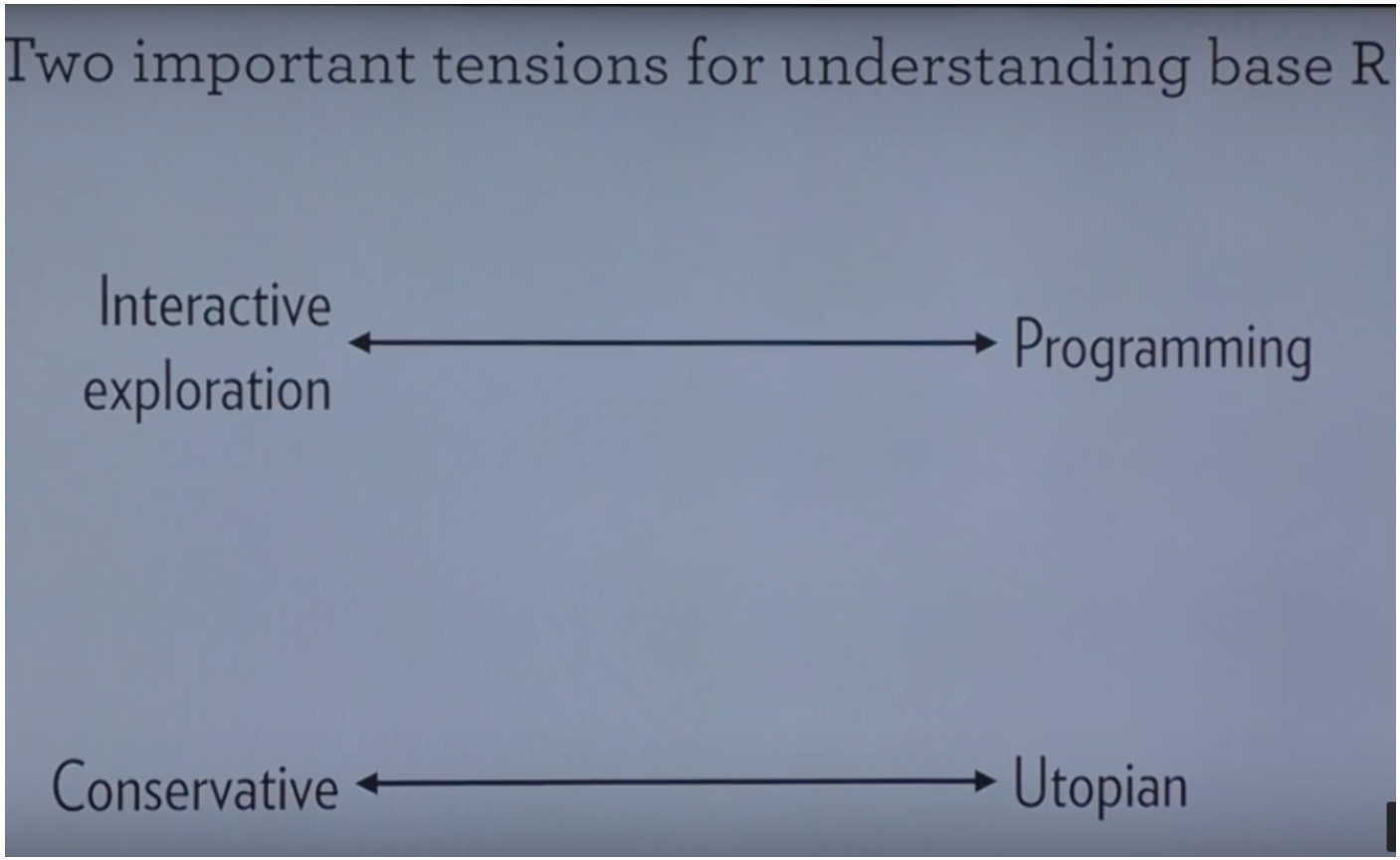
ggplot2 To make pretty graphs, including the opportunity to use grammar of graphics to create layered, customizable plots

Moreover, in comparison to R

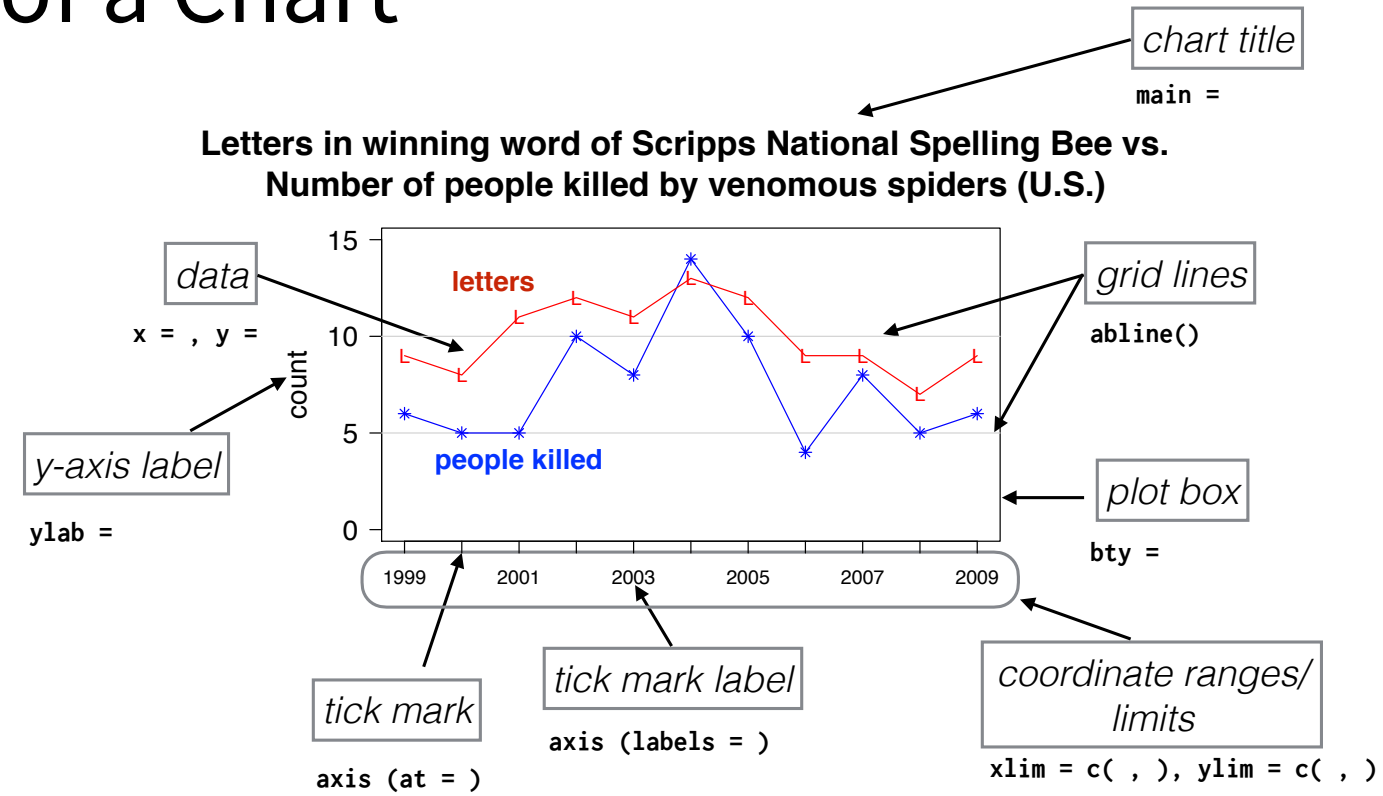
"Visualizations in Python are usually more convoluted, and the results are not nearly as pleasing to the eye or as informative."


#4
And The Winner is...
It's a tie!
It's up to you, the data scientist,
to pick the language that best fits your needs.
The following questions can guide you in your decision.

base R vs. tidyverse



Parts of a Chart



CREATE A NEW PLOT

Bar charts **barplot**(*height*, …)
- *bar labels* `names.arg =`
- *border* `border =`
- *fill color* `col =`
- *horizontal* `horiz = TRUE`

Box plots **boxplot**(*x*, …)
- *horizontal* `horizontal = TRUE`
- *box labels* `names =`

Dot plots **dotchart**(*x*, …)
- *dot labels* `labels =`

Histograms **hist**(*x*, …)
- *breakpts* `breaks =`

Line charts **plot**(*x*, `type = "l"`)
- *line type* `lty =` `"blank"|0` `"solid"|1` `"dashed"|2` `"dotted"|3`
- *line width* `lwd =`

Scatterplots **plot**(*x*, …)
- *symbol* `pch =`

REMOVE	
axis labels	`ann = FALSE`
axis, tickmarks, and labels	`xaxt = "n"` `yaxt = "n"`
plot box	`bty = "n"`

NOTE: Many of the parameters here can be also be set in `par()`. *See R help for more options.*

ADJUST	
allow plotting out of plot region	`xpd = TRUE`
aspect ratio	`asp =`
axis limits	`xlim =, ylim =`
axis lines to match axis limits	`xaxs = "i"` , `yaxs = "i"` *(internal axis calculation)*

ADD TEXT

location

axis labels	`xlab =, ylab =`
subtitle	`sub =`
title	`main =`

size

(magnification factor)

all elements	`cex =`
axis labels	`cex.lab =`
subtitle	`cex.sub =`
tick mark labels	`cex.axis =`
title	`cex.main =`

style

font face	`font = 1` *(plain)* 2 (*bold*) 3 (*italic*) 4 (*bold italic*)
font family	`family = "serif"` `"sans" "mono"`

position

text direction	`las = 1` (horizontal)
justification	`adj = 0 .5 1` (left, center, right)

ADD TO AN EXISTING PLOT

Add new plot `[`*any plot function*`]`
`(…, add = TRUE)`
ex. `barplot(x, add = TRUE)`

Axes **`axis (`***side,…* **`)`**
location `side = 1 2 3 4` (bottom, left, top, right)
tick mark:
labels `labels =`
location `at =`
remove `tick = FALSE`
rotate text `las = 1` (horizontal)

Axis labels **`mtext (`***text,…* **`)`**
location `side = 1 2 3 4` (bottom, left, top, right)
lines to skip `line =` *(from plot region, default = 0)*
position `at =` *x or y-coord* (depending on side)
justification `adj = 0 .5 1` (left, center, right)

Lines **`lines (`***x,…* **`)`**
line style `lty =`
line width `lwd =`
color `col =`

Points **`points (`***x,…* **`)`**
symbol `pch =`

color `col =`
fill color `bg =` `(pch: 21-25 only)`

Text **`text (`***x, y, text,…* **`)`**
position (rel. to x,y) `pos = 1 2 3 4` (below, left, above, right) (default=center)

Title **`title (`***main,…* **`)`**
axis labels `xlab =, ylab =`
subtitle `sub =`
title `main =`