hello ggplot2!

Dr. Jennifer (Jenny) Bryan
Department of Statistics and Michael Smith Laboratories
University of British Columbia

jenny@stat.ubc.ca
@JennyBryan
https://github.com/jennybc
http://www.stat.ubc.ca/~jenny/

thanks to ...

Vancouver R Users Group
Tavis Rudd and Tilman Holschuh -- admin

Rob Balshaw and the
BC Centres for Disease Control -- host

Casey Shannon, Nick Fishbane -- helpers + content

please see this GitHub repository for all references, examples worked with live coding, etc.

https://github.com/jennybc/ggplot2-tutorial

these slides just remind me to discuss some Big Ideas by putting them in a Big Font

stackoverflow is your friend

use tags!

stackoverflow Questions Tags Tour Users

Search

suppress legend [ggplot2] search

21 results relevance newest votes active

ggplot2 is an actively maintained open-source chart-drawing library for R, written by Hadley Wickham, based upon the principles of "Grammar of Graphics". It partially replaces R's basic plot and the lattice package, while providing a clean, powerful, orthogonal and fun API.

learn more... | top users | synonyms (2)

14 votes

A: ggplot legend issue w/ geom_point() and geom_text()

or, if you need to specify the size of text inside the aes, then **legend** = FALSE **suppress** drawing the **legends** of the geom: p <- ggplot(data = df, aes(x = x, y = y, size = count)) + geom_point() p + geom_text(aes(label = label, size = 150, vjust = 2), show_guide = FALSE) ...

answered nov 19 '10 by kohske

11 votes

A: removing a layer legend in ggplot

)(this, ...) : "**legend**" argument in geom_XXX and stat_XXX is deprecated. Use show_guide = TRUE or show_guide = FALSE for display or **suppress** the guide display. I would recommend upgrading **ggplot2**. ... Depending on the version of **ggplot2** you are using you get this problem. Using **ggplot2** vs 0.9.0 on R2.14.1 I get this graph: which does not include the **legend** for the vline. In this version ...

answered mar 26 '12 by Paul Hiemstra

stackoverflow is your friend

use tags!

stackoverflow Questions Tags Tour Users

Search

are for loops evil [r] search

6 results relevance newest votes active

R is a free, open-source programming language and software environment for statistical computing, bioinformatics and graphics. Questions should have a minimal example, see tinyurl.com/m3fryge. For statistical questions please use stats.stackexchange.com.

learn more... | top users | synonyms (1)

5 votes

A: For loops in R and computational speed

of information about R's for **loops** on the main Stackoverflow site. For example, the question Speed up the **Loop** Operation in **R** has at least two excellent answers which I found very helpful. Also, the **R** Inferno ... , paricularly in a double for **loop**. That's why it's interesting that innocent-looking things like brackets are actually function calls.) The first place you will be told to look when trying to extend **R** ...

answered jun 23 by Flounderer

4 votes 1 answer

Q: Ranged/Filtered Cross Join with R data.table

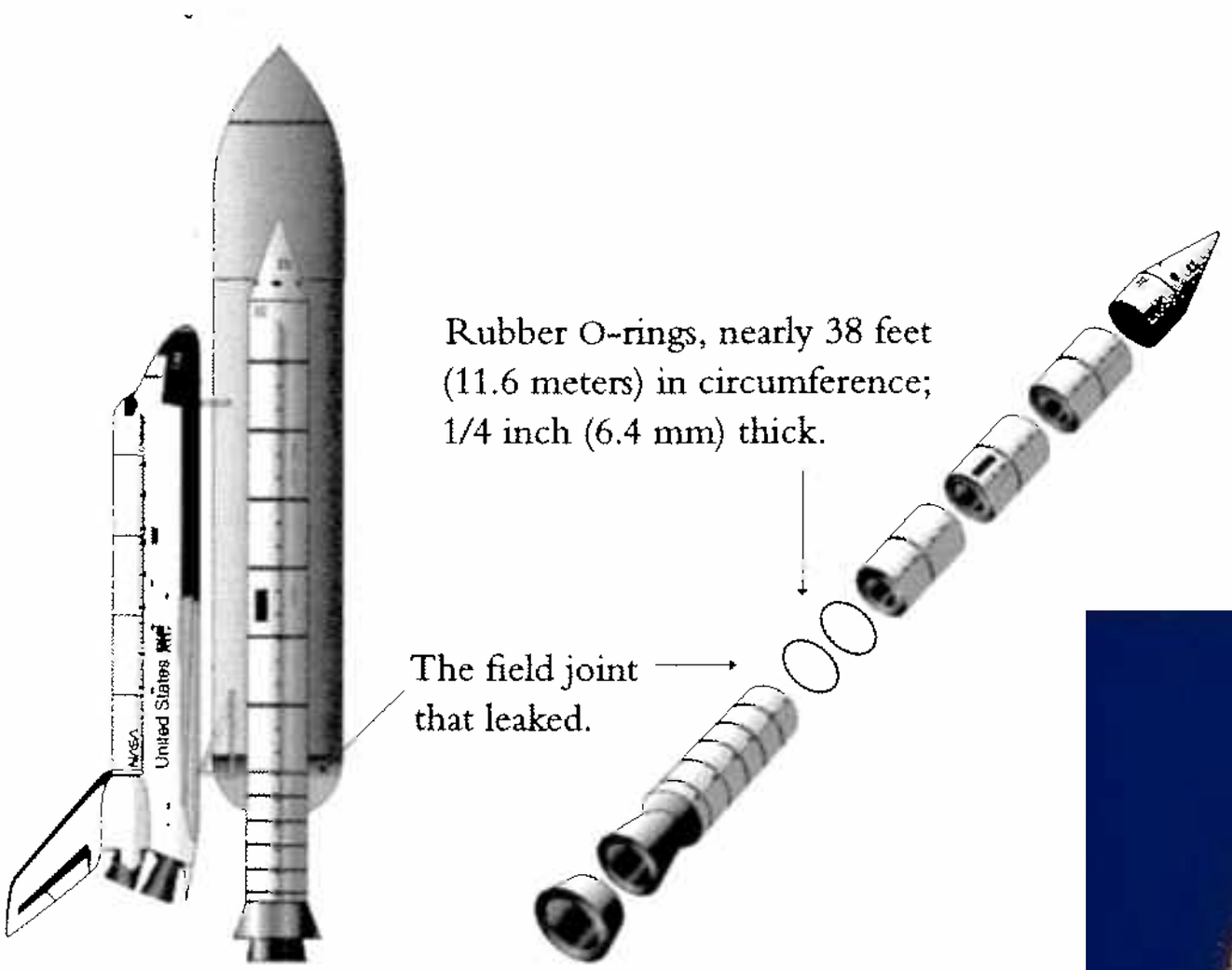
it to be but at least is reasonable in terms of memory consumption (I will let it run overnight on my real scenario ~ 1 Million rows). I tried changing the data table keys (using the dates instead of id's); it did not have any impact. As expected, explicitly writing the **loop** in **R** (manualIter) crawls. ... suggest a high performing approach avoiding the full cross join? See test example below doing the job with the **evil** full cross join. library(data.table) # Test data. dt1 <- data.table(id1=1:10, D=2 ...

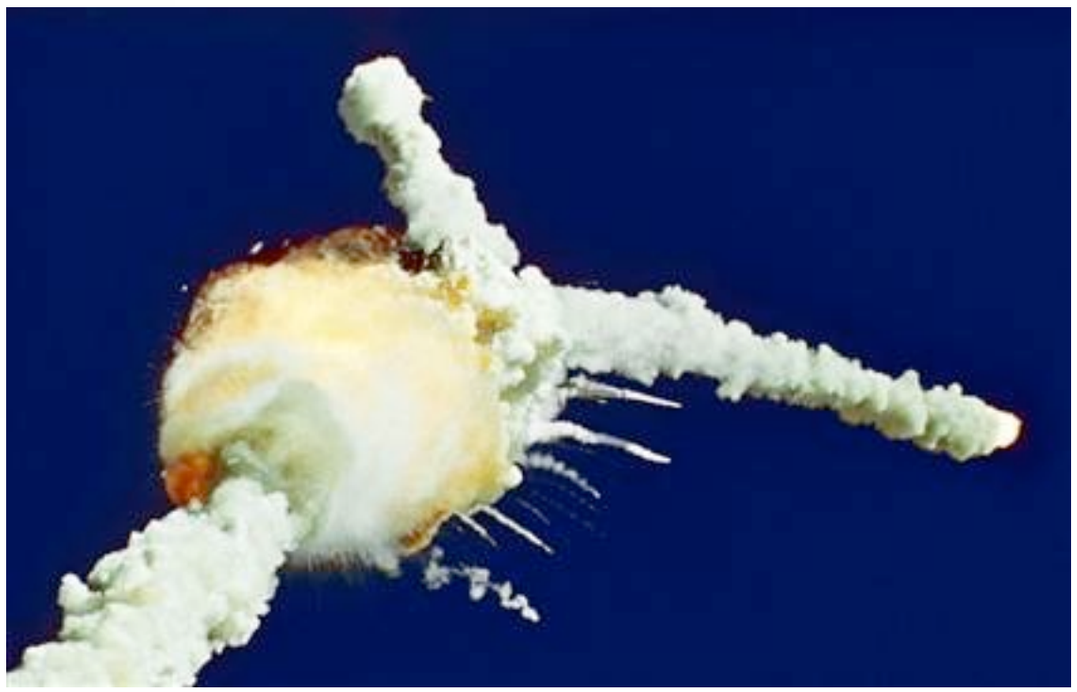
r data.table

asked feb 25 by Patrick

“A picture is worth a thousand words”

1986 Challenger space shuttle disaster
Favorite example of Edward Tufte





http://msnbcmedia1.msn.com/j/msnbc/Components/Photos/050709/050609_columbia_hmed_6p.hmedium.jpg

TEMPERATURE CONCERN ON SRM JOINTS

27 JAN 1986

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36°--66°
61A LH CENTER FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

BLOW BY HISTORY

SRM-15 WORST BLOW-BY
- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
- 2 CASE JOINTS (30-40°)

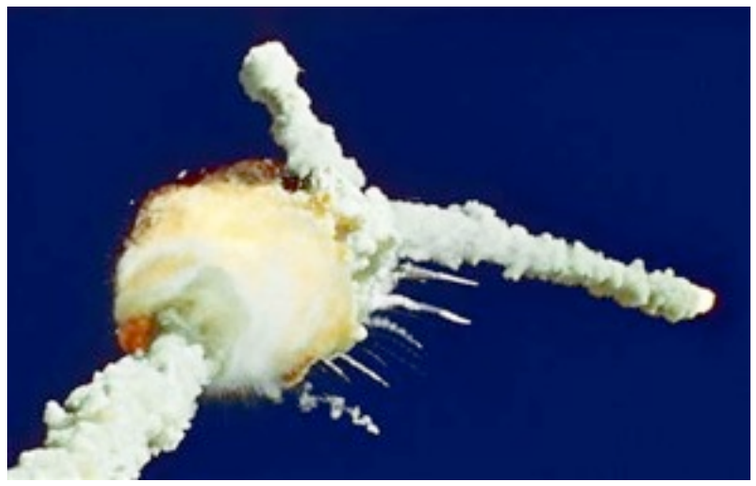
SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

MOTOR	O-RING
DM-1	47
DM-2	52
QM-3	48
QM-4	51
SRM-15	53
SRM-22	75
SRM-25	29 27

“A picture is worth a thousand words”

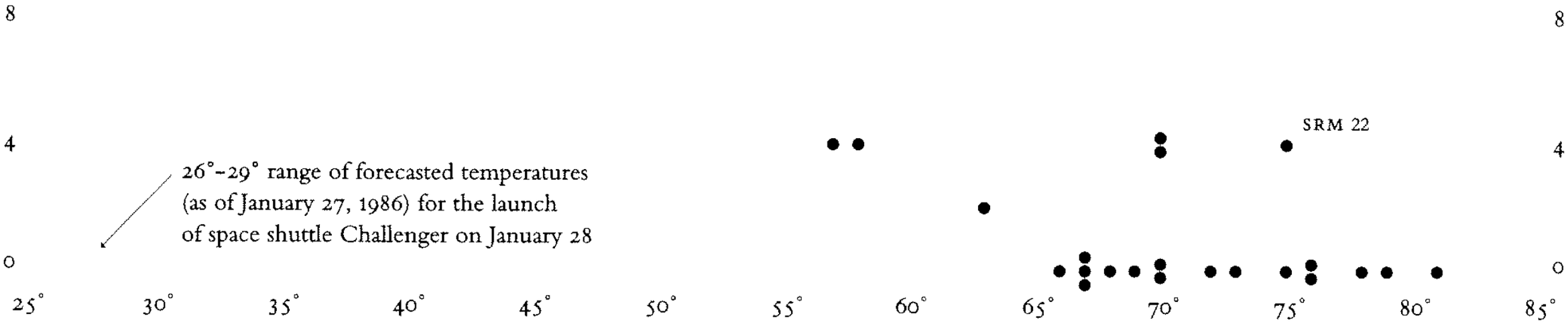


O-ring damage
index, each launch

12

SRM 15

12



Temperature (°F) of field joints at time of launch

“A picture is worth a thousand words”

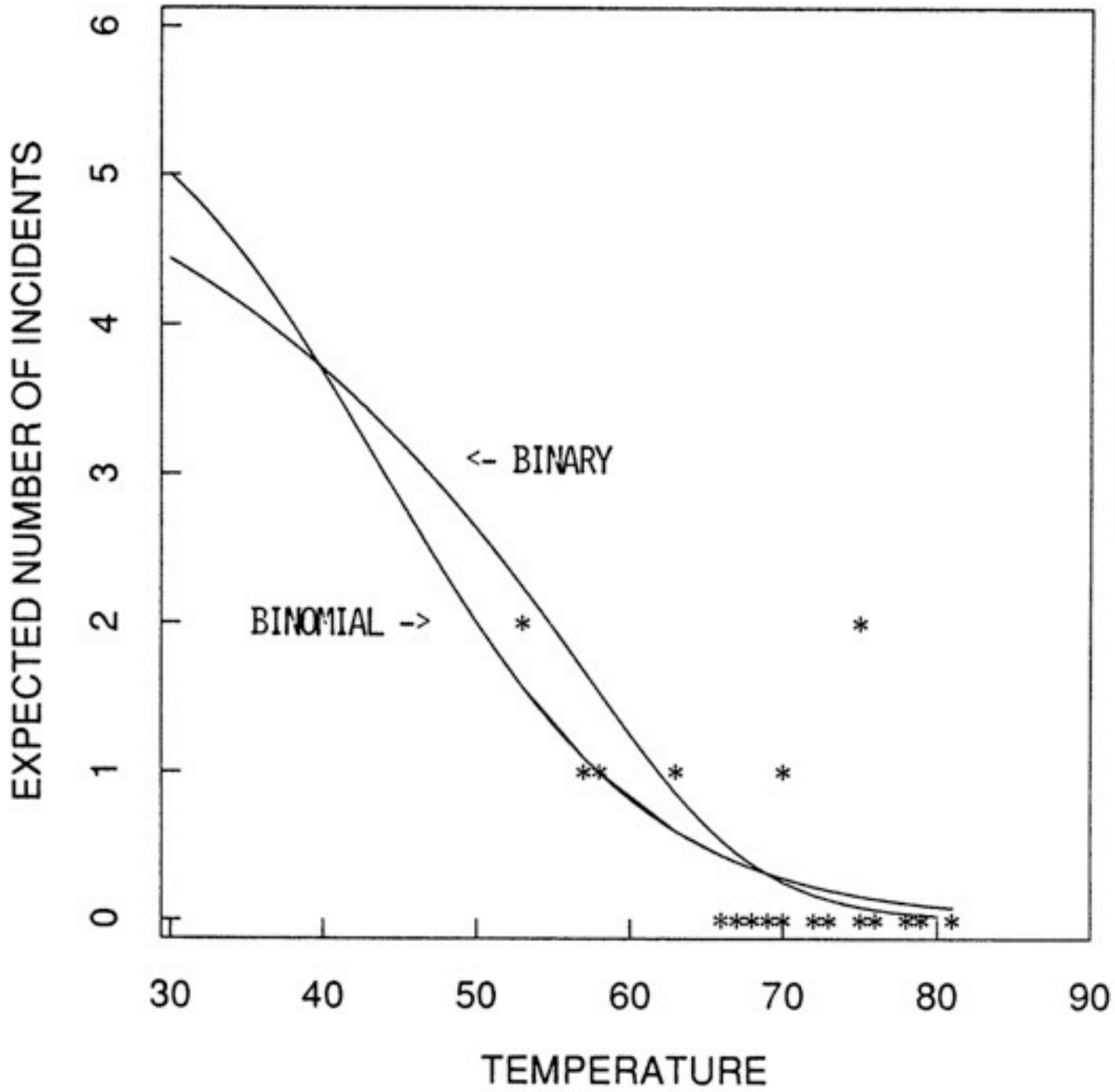
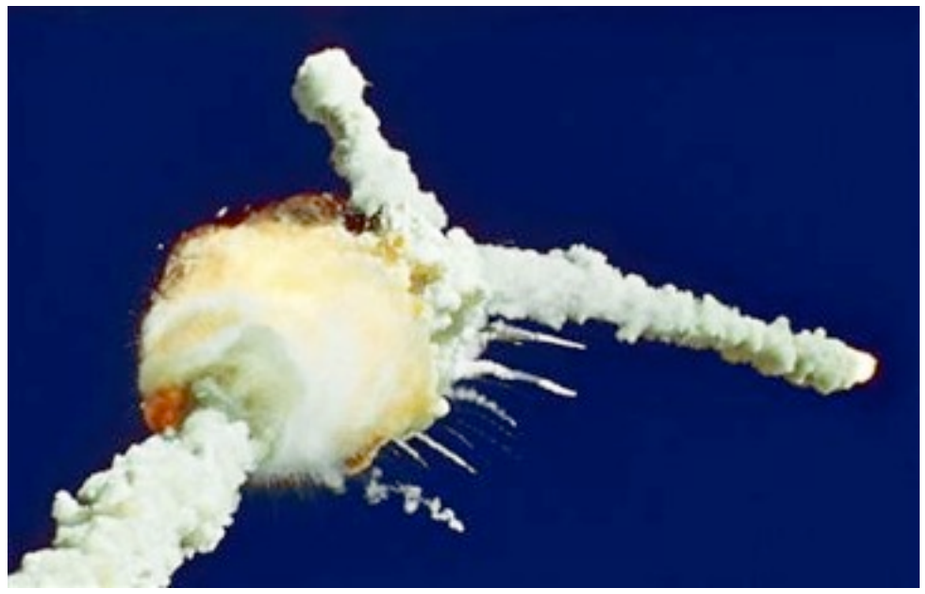


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.



Siddhartha R. Dalal; Edward B. Fowlkes; Bruce Hoadley. Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. JASA, Vol. 84, No. 408 (Dec., 1989), pp. 945-957. Access via JSTOR.

Edward Tufte
http://www.edwardtufte.com

BOOK:
Visual Explanations: Images and Quantities, Evidence and Narrative

Ch. 5 deals with the Challenger disaster
That chapter is available for $7 as a downloadable booklet:
http://www.edwardtufte.com/tufte/books_textb

“A picture is worth a thousand words”

Always, always, always plot the data.

Replace (or complement) ‘typical’ tables of data or statistical results with figures that are more compelling and accessible.

Whenever possible, generate figures that overlay / juxtapose observed data and analytical results, e.g. the ‘fit’.

base or traditional graphics

vs

`lattice` package

ships with R, but must load

```
library(lattice)
```

vs

`ggplot2` package

must be installed and loaded

```
install.packages("ggplot2", dependencies = TRUE)
library(ggplot2)
```

Two main goals for statistical graphics

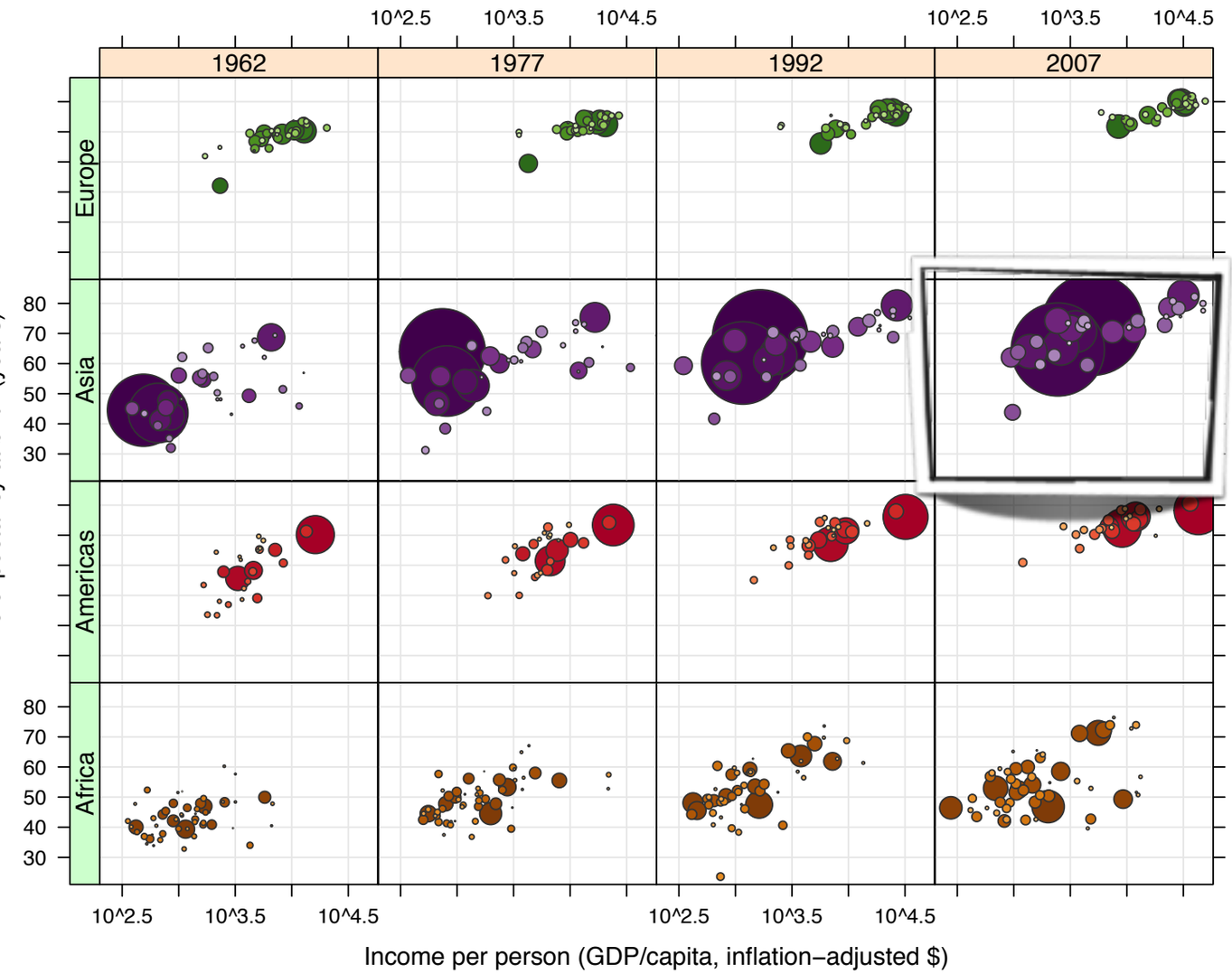
- To facilitate comparisons.
- To identify trends.

lattice and ggplot2 achieve these goals with less fuss

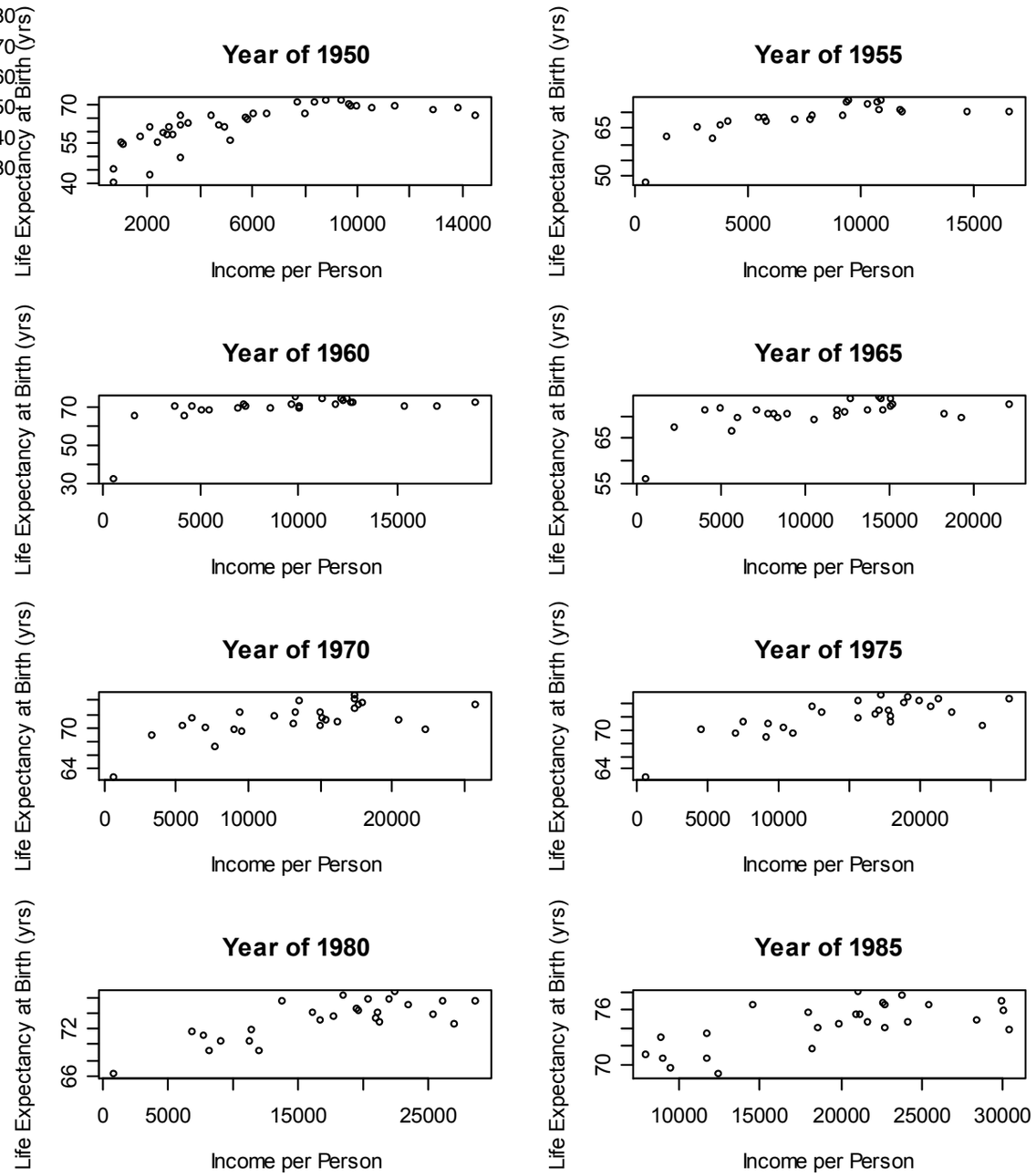
lattice

"multi-panel conditioning"

```
lifeExp ~ gdpPercap | continent * year
```



Assignment 1: Best Set of Graphs

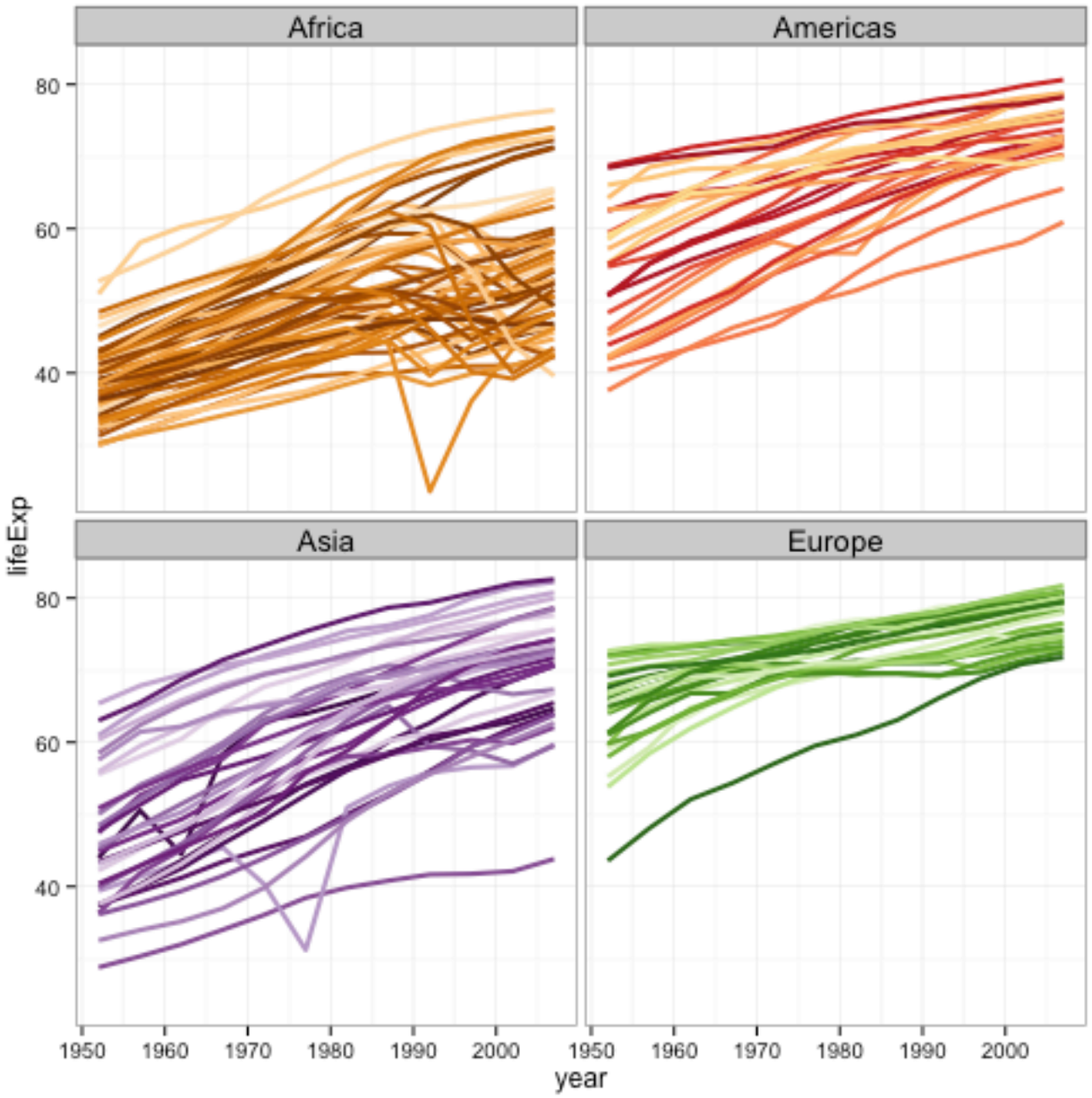


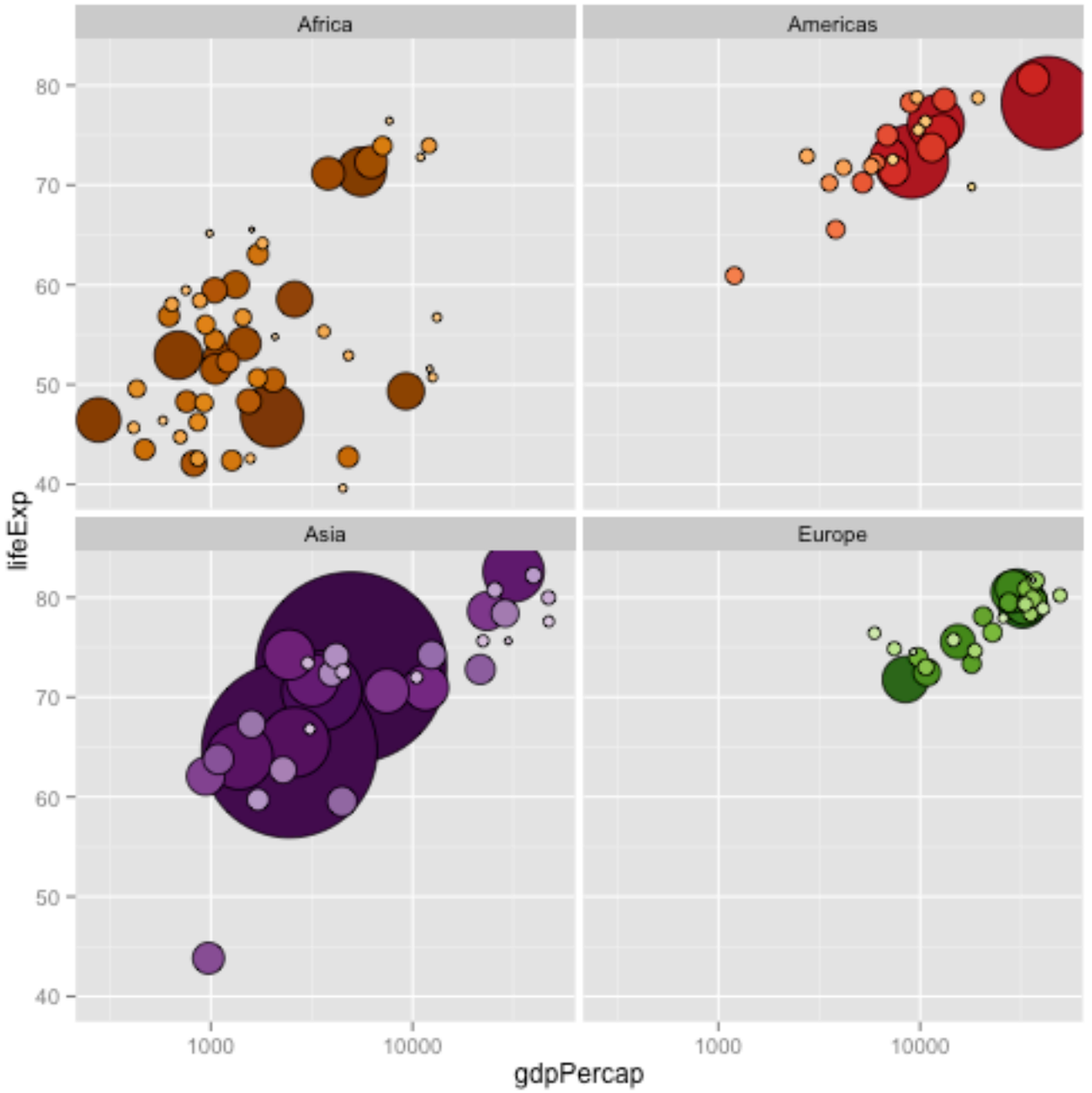
base

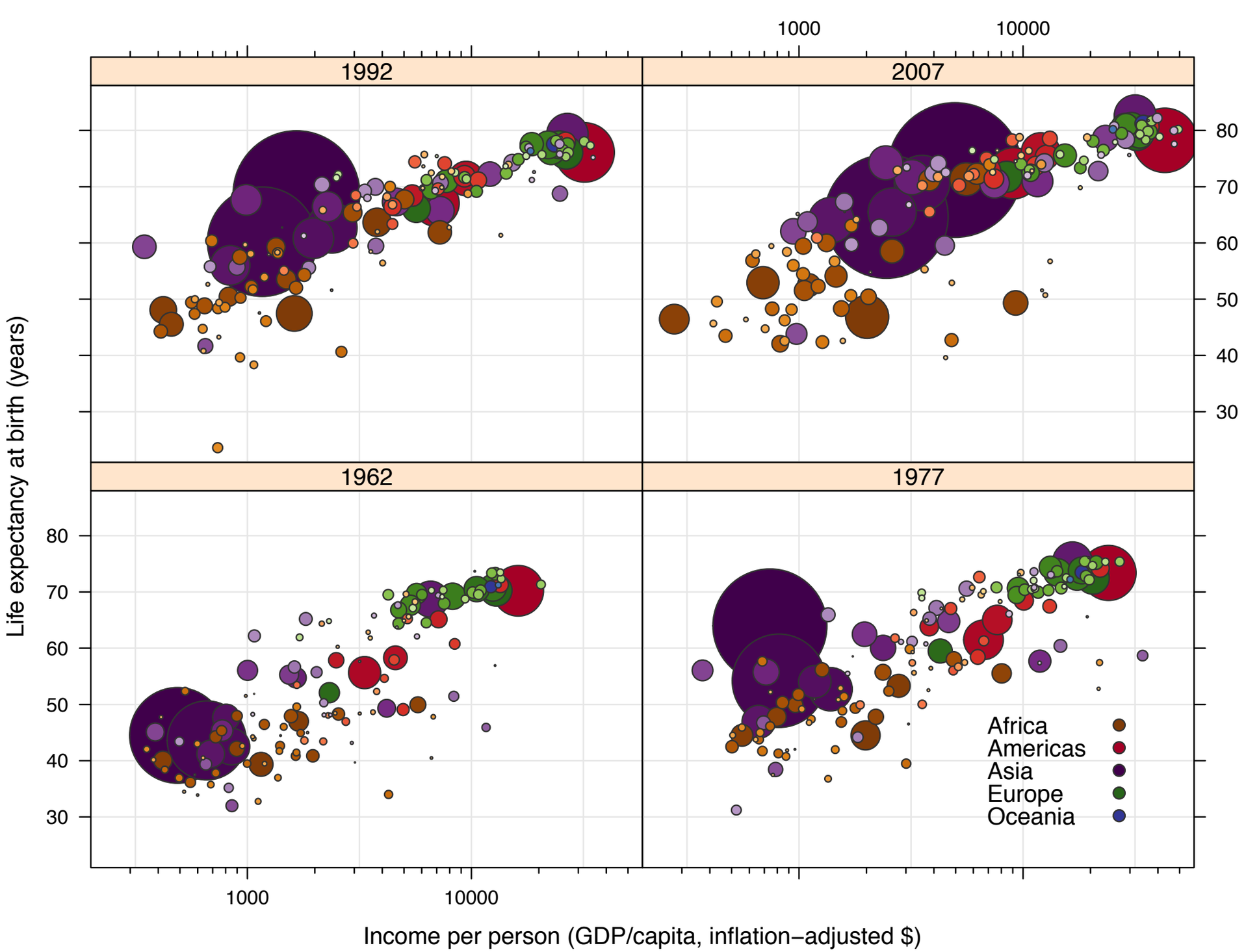
ggplot2

"facetting"

```
ggplot(...) + ... +
  facet_wrap(~ continent)
```







lattice

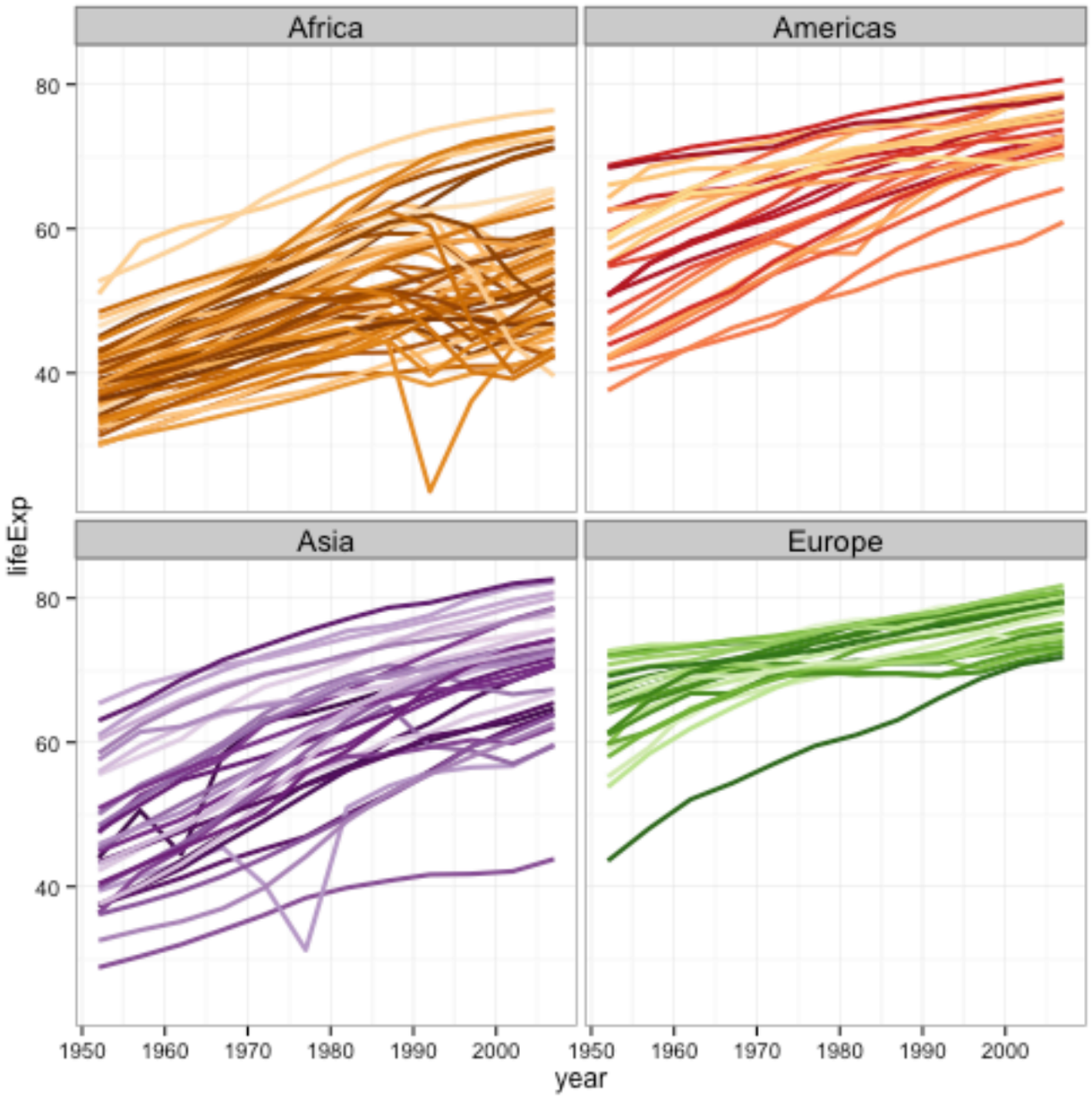
"groups and superposition"

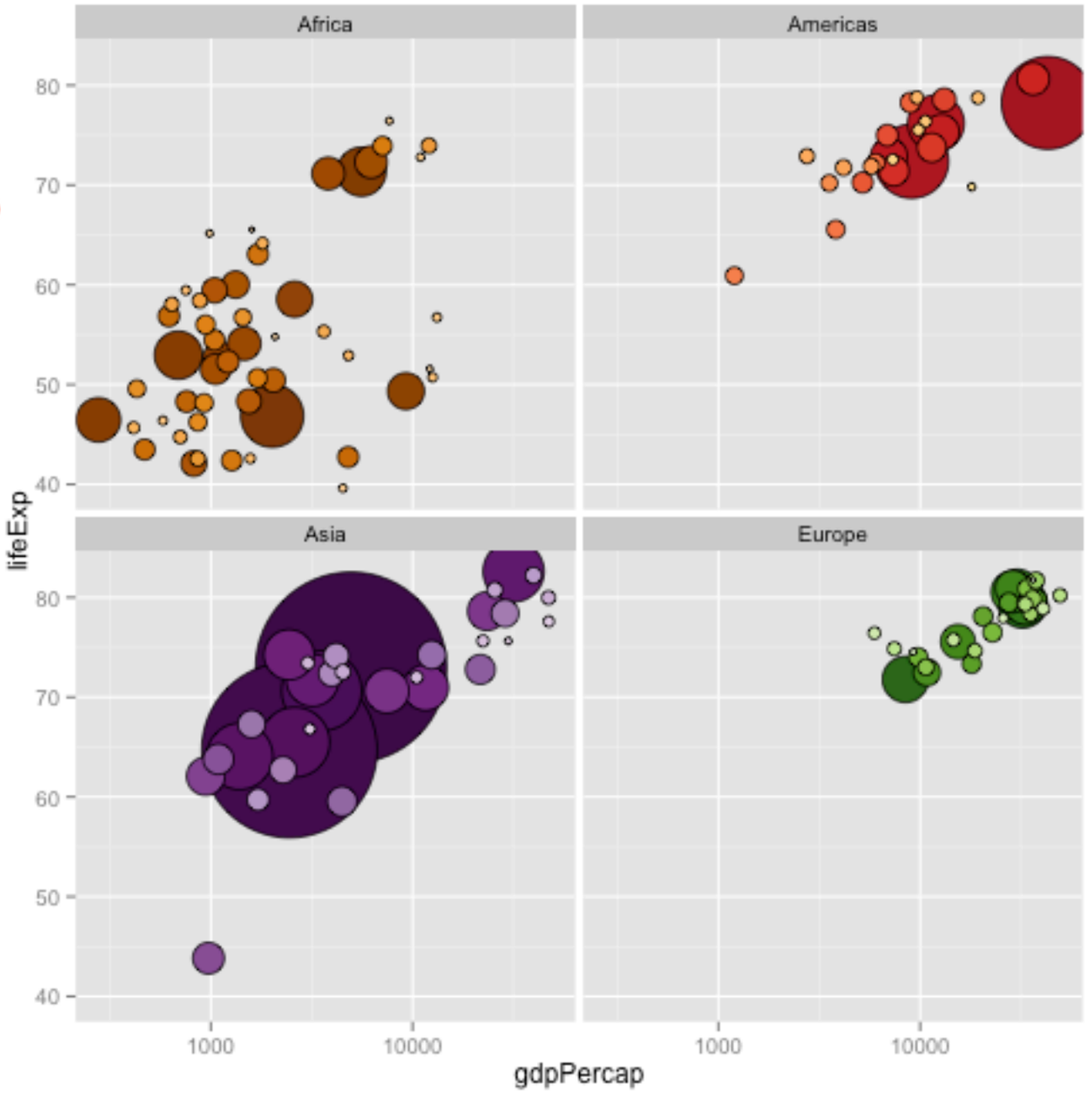
```
lifeExp ~ gdpPercap | year, group = country
```

ggplot2

"aesthetic mapping"

```
ggplot(...) + ... +
  aes(fill = country)
```

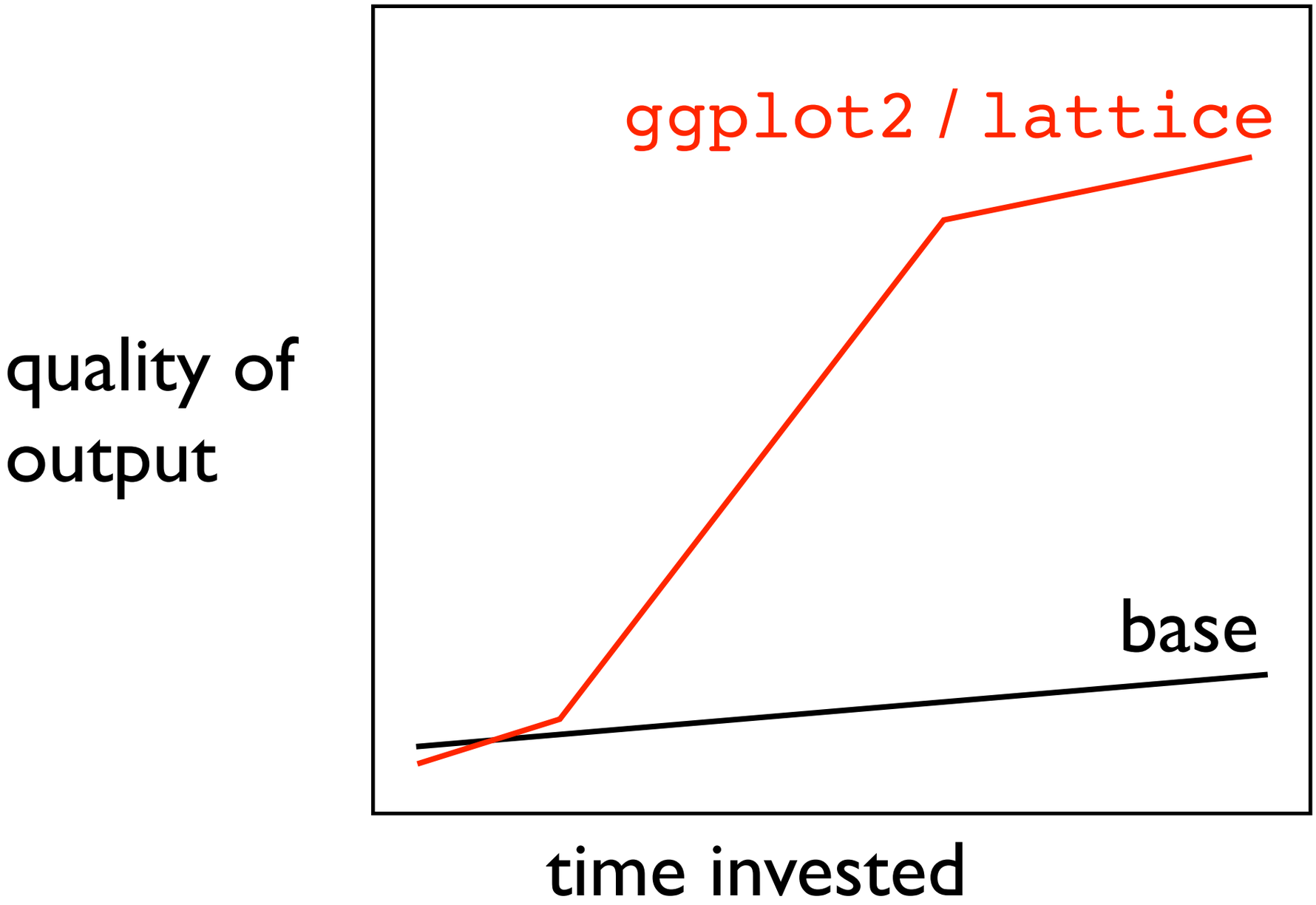




TO DO:

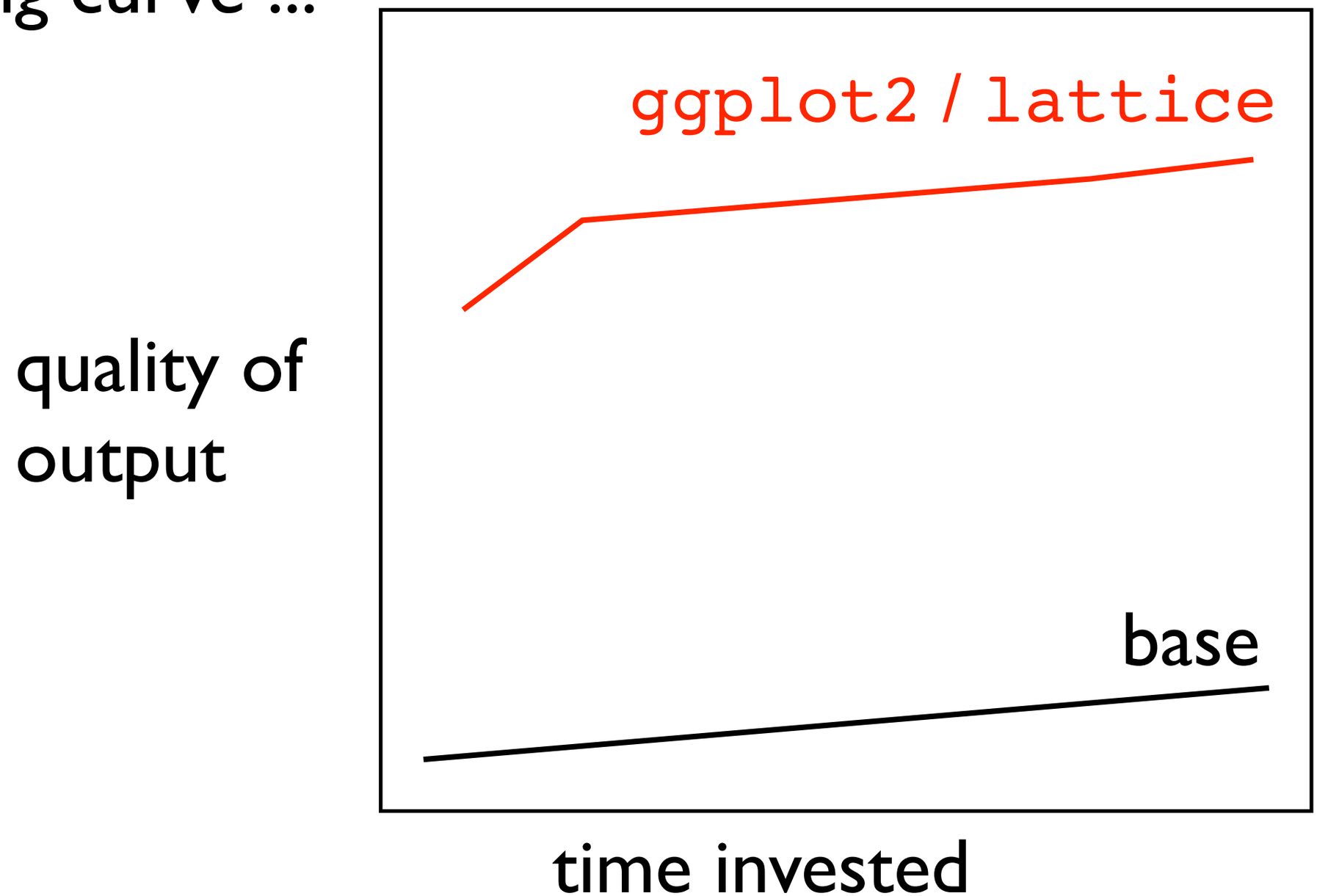
add similar eye candy for overlaying, e.g. a smooth fit

week one ....



* figure is totally fabricated but, I claim, still true

after you've climbed the steepest part of the learning curve ...



* figure is totally fabricated but, I claim, still true

use data.frames

use factors

be the boss of your factors

keep your data tidy

reshape your data

if you are struggling with a plot,

ask yourself:

am I breaking one or more of these “rules”?

often that is the real, hidden reason for struggle

use data.frames

use factors

be the boss of your factors

keep your data tidy

reshape your data

master `read.table()`

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", row.names, col.names,
           as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)
```

master `reorder()`

reorder.default {stats} R Documentation

Reorder Levels of a Factor

Description

`reorder` is a generic function. The `"default"` method treats its first argument as a categorical variable, and reorders its levels based on the values of a second variable, usually numeric.

Usage

```
reorder(x, ...)

## Default S3 method:
reorder(x, X, FUN = mean, ...,
        order = is.ordered(x))
```

from Wickham's *Tidy Data*

In **tidy data**:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

messy

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

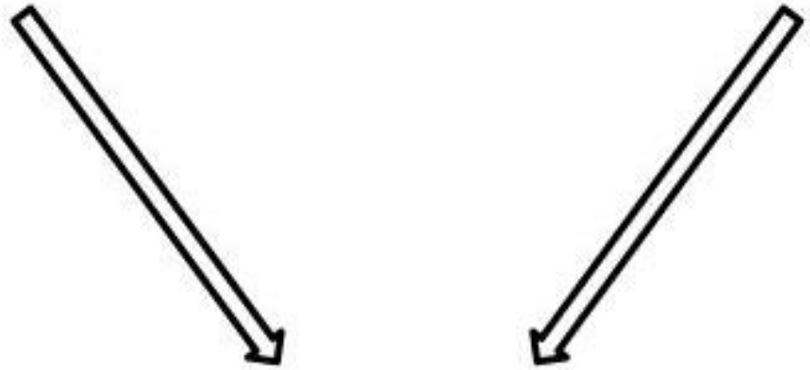
tidy

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

from White et al's *Nine simple ways* ...

	Habitat		
Species	X	Y	Z
A	0	3	0
B	1	0	2

Species	HabitatX	HabitatY	HabitatZ
A	0	3	0
B	1	0	2



Species	Habitat	Abundance
A	Y	3
B	X	1
B	Z	2

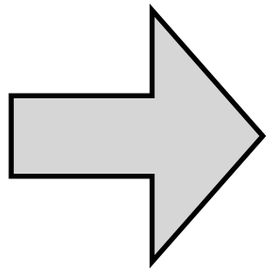
reshape your data

data has a tendency to get shorter and wider, but tall and thin often better for analysis + visualization

from Wickham's *Tidy Data*
see also reshape2

melt

row	a	b	c
a	1	4	7
b	2	5	8
c	3	6	9

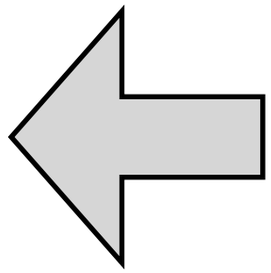


row	column	value
a	a	1
b	a	2
c	a	3
a	b	4
b	b	5
c	b	6
a	c	7
b	c	8
c	c	9

from Wickham's *Tidy Data*
see also reshape2

cast

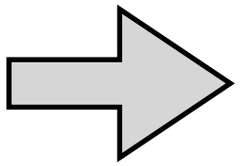
row	a	b	c
a	1	4	7
b	2	5	8
c	3	6	9

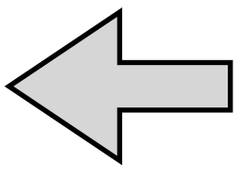


row	column	value
a	a	1
b	a	2
c	a	3
a	b	4
b	b	5
c	b	6
a	c	7
b	c	8
c	c	9

row	a	b	c
a	1	4	7
b	2	5	8
c	3	6	9

melt





cast

row	column	value
a	a	1
b	a	2
c	a	3
a	b	4
b	b	5
c	b	6
a	c	7
b	c	8
c	c	9

typical usage pattern:

melt to facilitate analysis and visualization

cast to make compact tables that are nicer for eyeballs

in addition to:

`reshape2`

see also:

`plyr`

`dplyr`

ggplot2

we will not use qplot() function

no training wheels

you're here ...
I assume you want to ride this bike

data, in data.frame form

aesthetic: map variables into properties people can perceive visually ... position, color, line type?

geom: specifics of what people see ... points? lines?

scale: map data values into "computer" values

stat: summarization/transformation of data

facet: juxtapose related mini-plots of data subsets

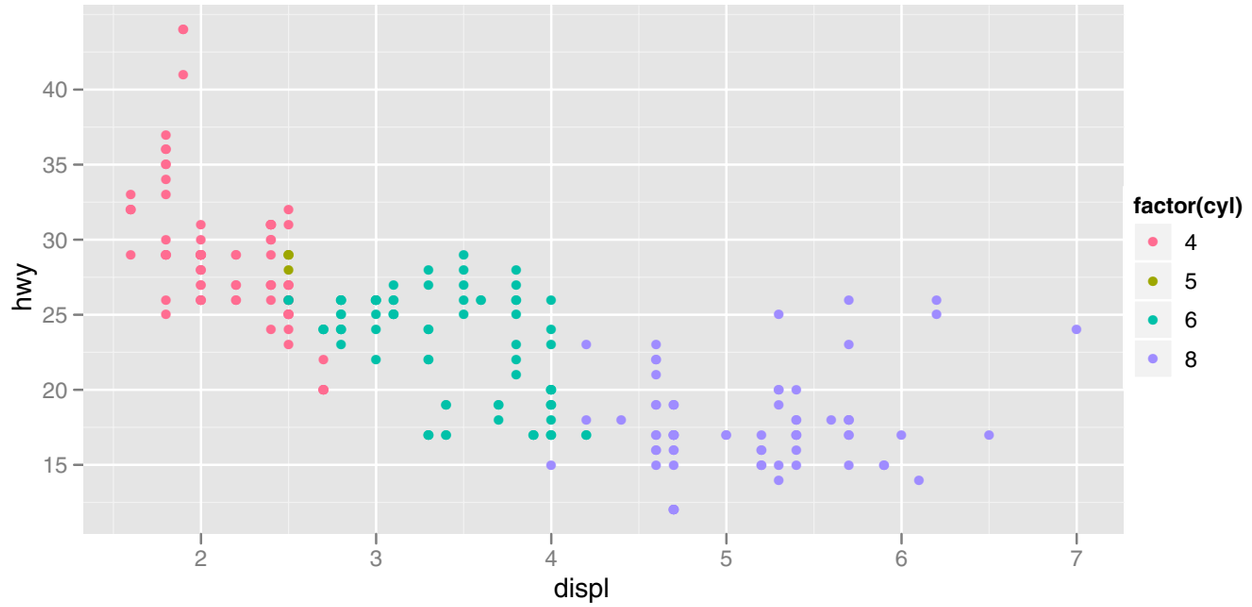


Fig. 3.1: A scatterplot of engine displacement in litres (displ) vs. average highway miles per gallon (hwy). Points are coloured according to number of cylinders. This plot summarises the most important factor governing fuel economy: engine size.

manufacturer	model	disp	year	cyl	cty	hwy	class
audi	a4	1.8	1999	4	18	29	compact
audi	a4	1.8	1999	4	21	29	compact
audi	a4	2.0	2008	4	20	31	compact
audi	a4	2.0	2008	4	21	30	compact
audi	a4	2.8	1999	6	16	26	compact
audi	a4	2.8	1999	6	18	26	compact
audi	a4	3.1	2008	6	18	27	compact
audi	a4 quattro	1.8	1999	4	18	26	compact
audi	a4 quattro	1.8	1999	4	16	25	compact
audi	a4 quattro	2.0	2008	4	20	28	compact

x	y	colour
1.8	29	4
2.0	31	4
2.0	30	4
2.8	26	6
2.8	26	6
3.1	27	6
1.8	26	4
1.8	25	4
2.0	28	4

x	y	colour	size	shape
0.037	0.531	#FF6C91	1	19
0.074	0.594	#FF6C91	1	19
0.074	0.562	#FF6C91	1	19
0.222	0.438	#00C1A9	1	19
0.222	0.438	#00C1A9	1	19
0.278	0.469	#00C1A9	1	19
0.037	0.438	#FF6C91	1	19
0.037	0.406	#FF6C91	1	19
0.074	0.500	#FF6C91	1	19

mapping data to aesthetics

scaling: data units → "computer" units

base graphics cause a figure to exist as a “side effect”

ggplot2 (and lattice) construct the figure as an R object

obviously you’ll need to print it to see it

this tutorial consisted largely of live coding ... see the repo for indicative content

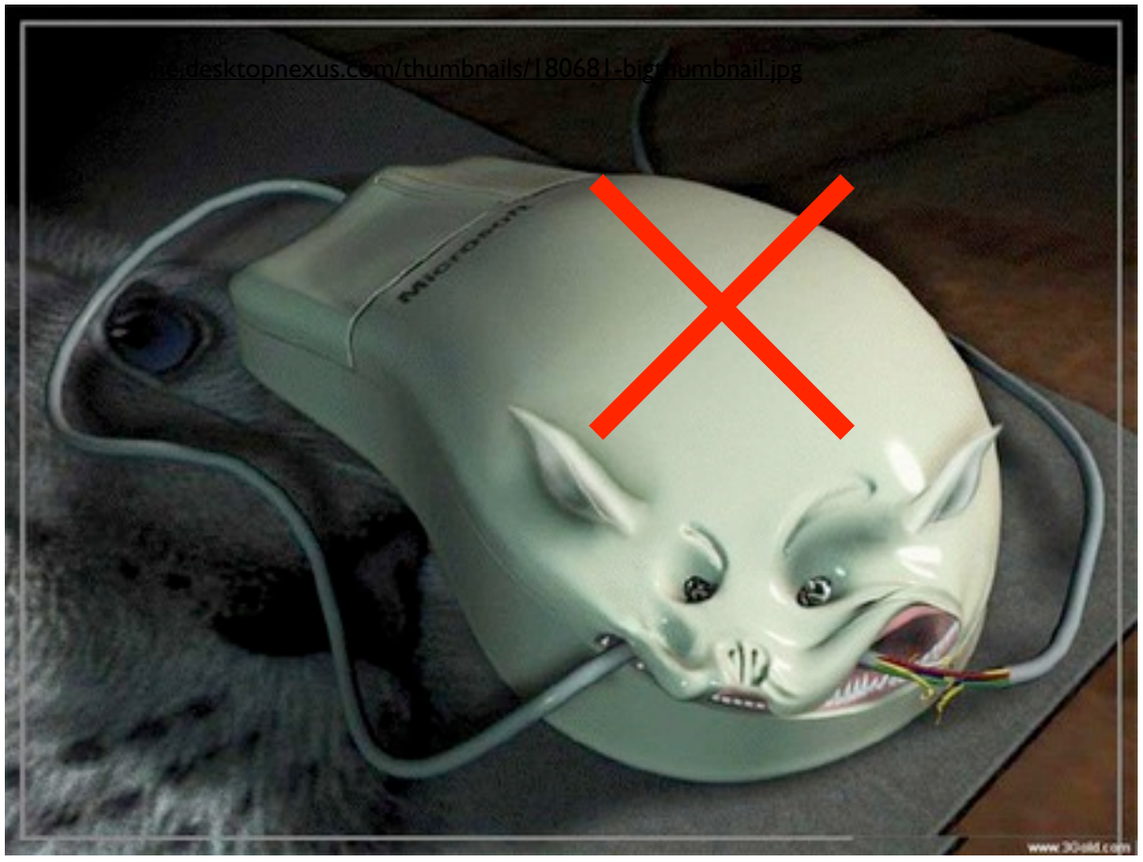
https://github.com/jennybc/ggplot2-tutorial

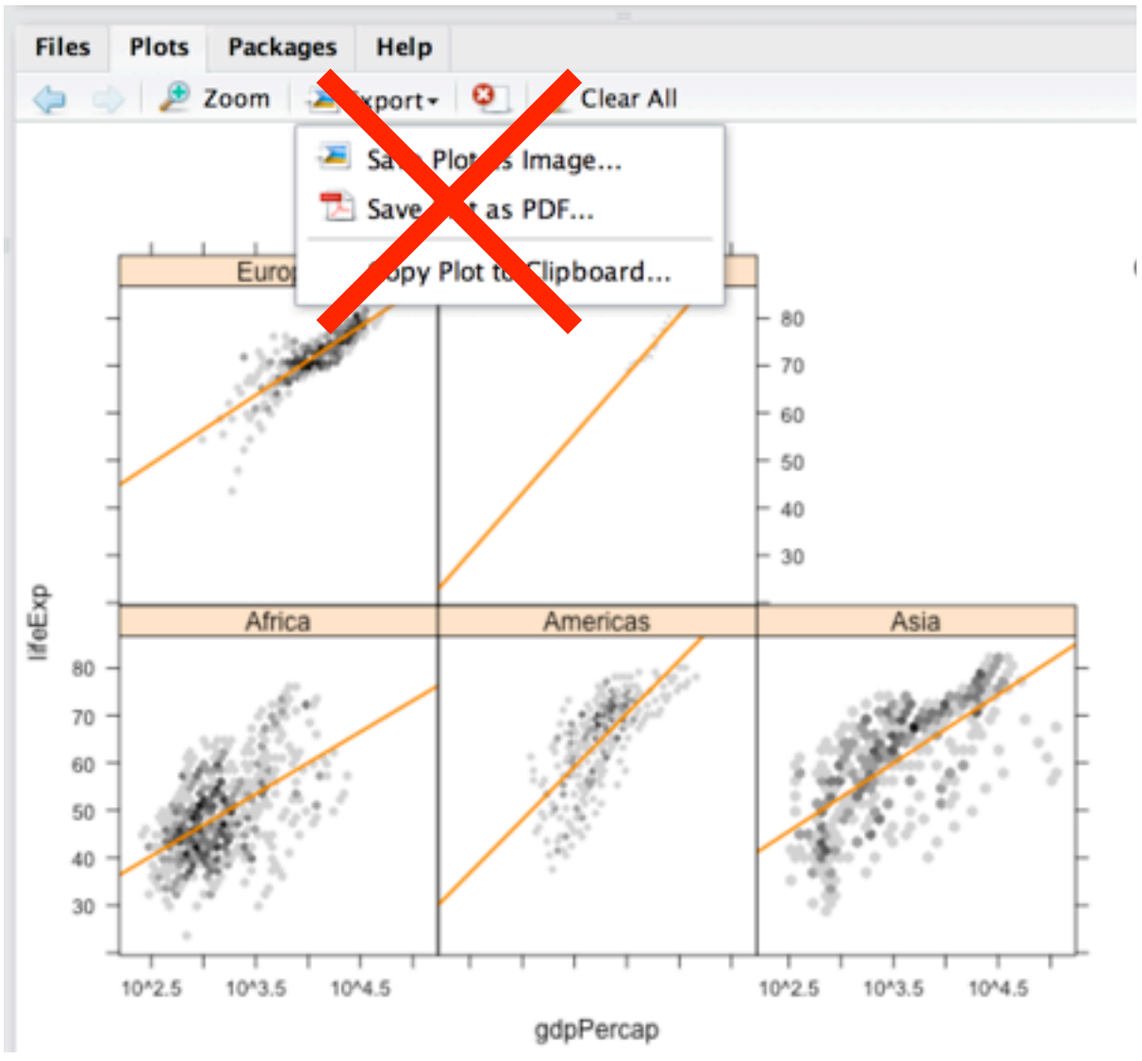
saving figures to file

do not save figures mouse-y style

not self-documenting

not reproducible





most correct method:

```
pdf("awesome_figure.pdf")
plot(1:10)
dev.off()
```

postscript(), svg(), png(), tiff(), ...

fine for everyday use:

```
plot(1:10)
dev.print(pdf,"awesome_figure.pdf")
```

postscript(), svg(), png(), tiff(), ...

next slide from here:

Data Visualization with R & ggplot2

Karthik Ram

September 2, 2013

- If the plot is on your screen

```
ggsave("~/path/to/figure/filename.png")
```

- If your plot is assigned to an object

```
ggsave(plot1, file = "~/path/to/figure/filename.png")
```

- Specify a size

```
ggsave(file = "/path/to/figure/filename.png", width = 6,
height =4)
```

- or any format (pdf, png, eps, svg, jpg)

```
ggsave(file = "/path/to/figure/filename.eps")
ggsave(file = "/path/to/figure/filename.jpg")
ggsave(file = "/path/to/figure/filename.pdf")
```

```
p <- ggplot(...) + ...
p #delete or comment this out if non-interactive
ggsave(p, file = “path/to/figure/filename.png”)
```

Use this workflow if the script might be run non-interactively.

Why? If you do not specify the plot explicitly, the default is to draw the last interactively drawn plot. That won’t exist in a non-interactive session and your plot files will be blank.

This can be frustrating. Ask me how I know.

```
p <- ggplot(...) + ...
ggsave(p, "filename.png", scale = 0.8)
```

Adjust the "`scale`" parameter to get multiple versions of a plot destined for different targets, e.g., for use in a presentation vs. a poster. vs a manuscript.

`scale < 1` makes the various plot elements bigger relative to the plotting area

`scale > 1` makes them smaller

YMMV but try `scale = 0.8` for posters/slides