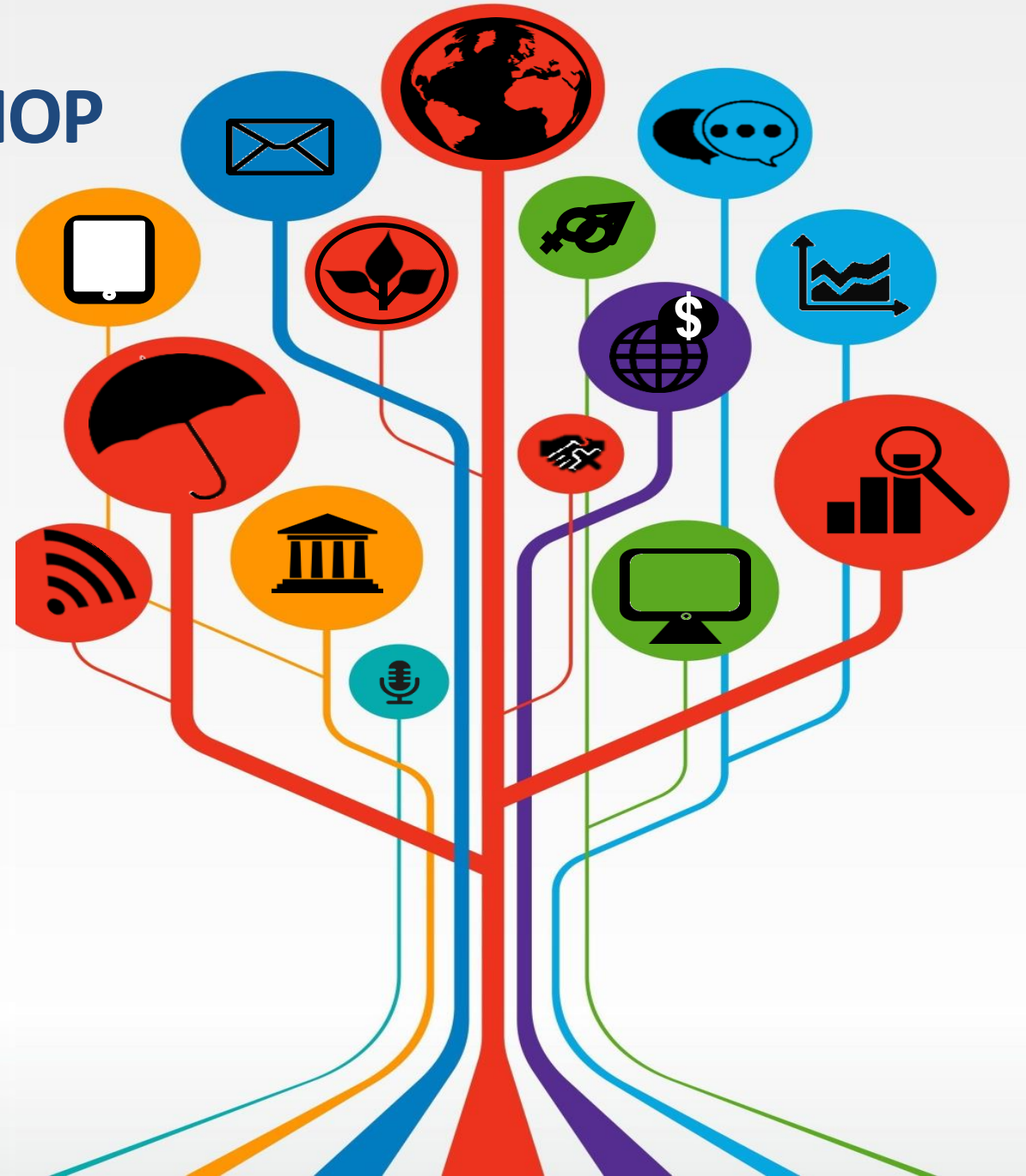


FIELD COORDINATOR WORKSHOP

Manage Successful Impact Evaluations

18 - 22 JUNE 2018
WASHINGTON, DC



WORLD BANK GROUP

Data Visualization

Stata Track 2

Prepared by DIME Analytics

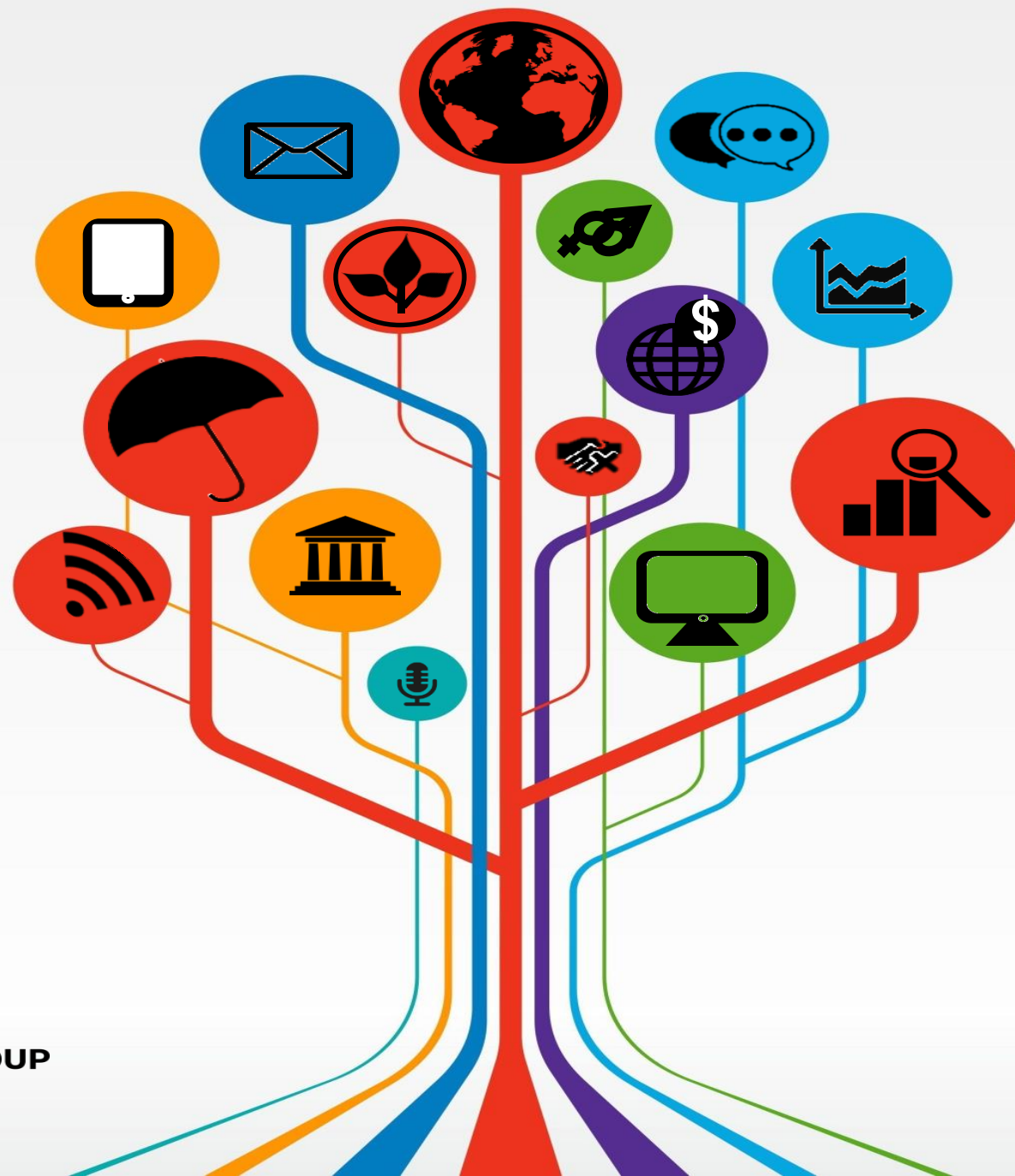
DIMEAnalytics_Internal_Use_Only@worldbank.org

Presented by Benjamin Daniels and Sakina Shibuya

bdaniels@worldbank.org / sshibuya@worldbank.org

<http://www.worldbank.org/en/research/dime>

June 21, 2018



Tables give all the details

- What's happening in this regression table? What's important?

TABLE 3—MEASURES OF ACCESS TO CARE JUST BEFORE 65 AND ESTIMATED DISCONTINUITIES AT 65

	1997–2003 NHIS				1992–2003 NHIS			
	Delayed care last year		Did not get care last year		Saw doctor last year		Hospital stay last year	
	Age 63–64 (1)	RD at 65 (2)	Age 63–64 (3)	RD at 65 (4)	Age 63–64 (5)	RD at 65 (6)	Age 63–64 (7)	RD at 65 (8)
Overall sample	7.2	–1.8 (0.4)	4.9	–1.3 (0.3)	84.8	1.3 (0.7)	11.8	1.2 (0.4)
<i>Classified by ethnicity and education:</i>								
<i>White non-Hispanic:</i>								
High school dropout	11.6	–1.5 (1.1)	7.9	–0.2 (1.0)	81.7	3.1 (1.3)	14.4	1.6 (1.3)
High school graduate	7.1	0.3 (2.8)	5.5	–1.3 (2.8)	85.1	–0.4 (1.5)	12.0	0.3 (0.7)
At least some college	6.0	–1.5 (0.4)	3.7	–1.4 (0.3)	87.6	0.0 (1.3)	9.8	2.1 (0.7)
<i>Minority:</i>								
High school dropout	13.6	–5.3 (1.0)	11.7	–4.2 (0.9)	80.2	5.0 (2.2)	14.5	0.0 (1.4)
High school graduate	4.3	–3.8 (3.2)	1.2	1.5 (3.7)	84.8	1.9 (2.7)	11.4	1.8 (1.4)
At least some college	5.4	–0.6 (1.1)	4.8	–0.2 (0.8)	85.0	3.7 (3.9)	9.5	0.7 (2.0)
<i>Classified by ethnicity only:</i>								
White non-Hispanic	6.9	–1.6 (0.4)	4.4	–1.2 (0.3)	85.3	0.6 (0.8)	11.6	1.3 (0.5)
Black non-Hispanic (all)	7.3	–1.9 (1.1)	6.4	–0.3 (1.1)	84.2	3.6 (1.9)	14.4	0.5 (1.1)
Hispanic (all)	11.1	–4.9 (0.8)	9.3	–3.8 (0.7)	79.4	8.2 (0.8)	11.8	1.0 (1.6)

Note: Entries in odd numbered columns are mean of variable in column heading among people ages 63–64. Entries in even numbered columns are estimated regression discontinuities at age 65, from models that include linear control for age interacted with dummy for age 65 or older (columns 2 and 4) or quadratic control for age, interacted with dummy for age 65 and older (columns 6 and 8). Other controls in models include indicators for gender, race/ethnicity, education, region, and sample year. Sample in columns 1–4 is pooled 1997–2003 NHIS. Sample in columns 5–8 is pooled 1992–2003 NHIS. Samples for regression models include people ages 55–75 only. Standard errors (in parentheses) are clustered by quarter of age.

But figures *tell the story*

- This is the data that generates those estimates.
- You can see exactly what is happening very quickly!

Even more importantly:
You don't have to look for it.
The eye is drawn to the story!

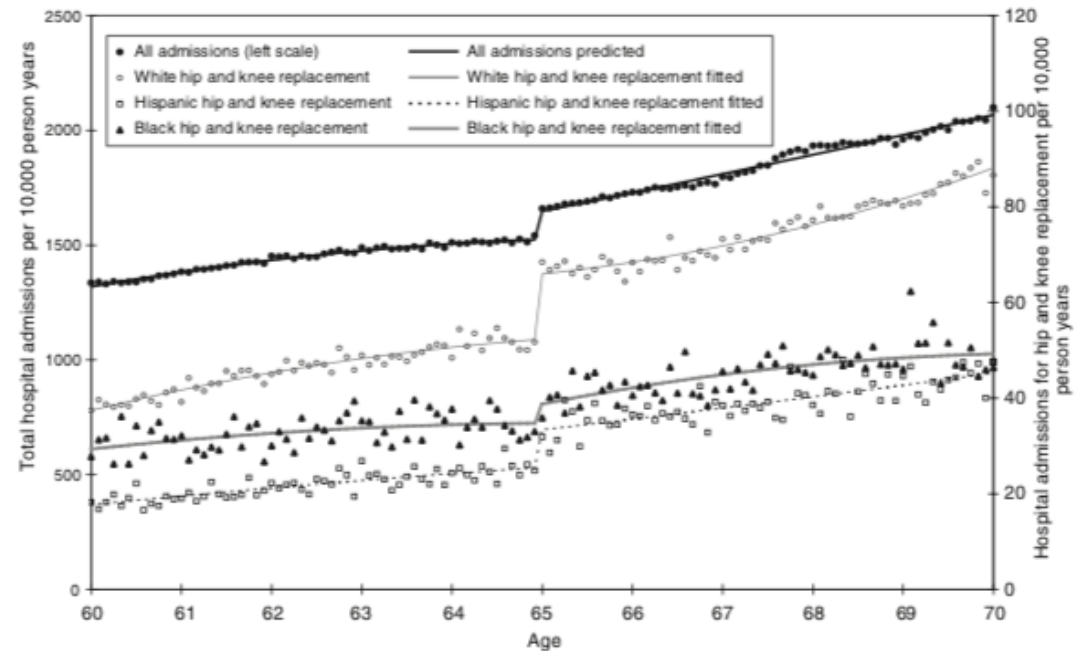
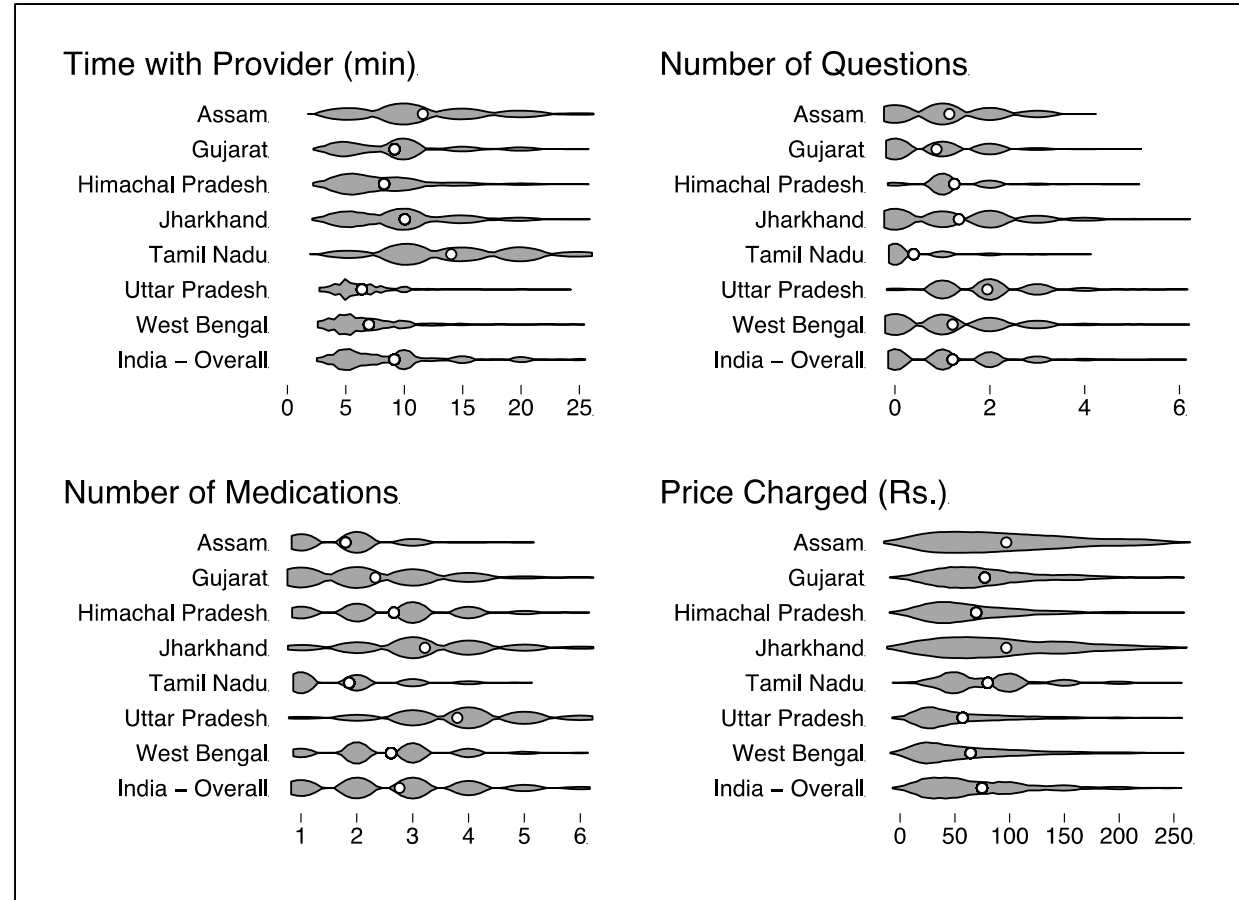
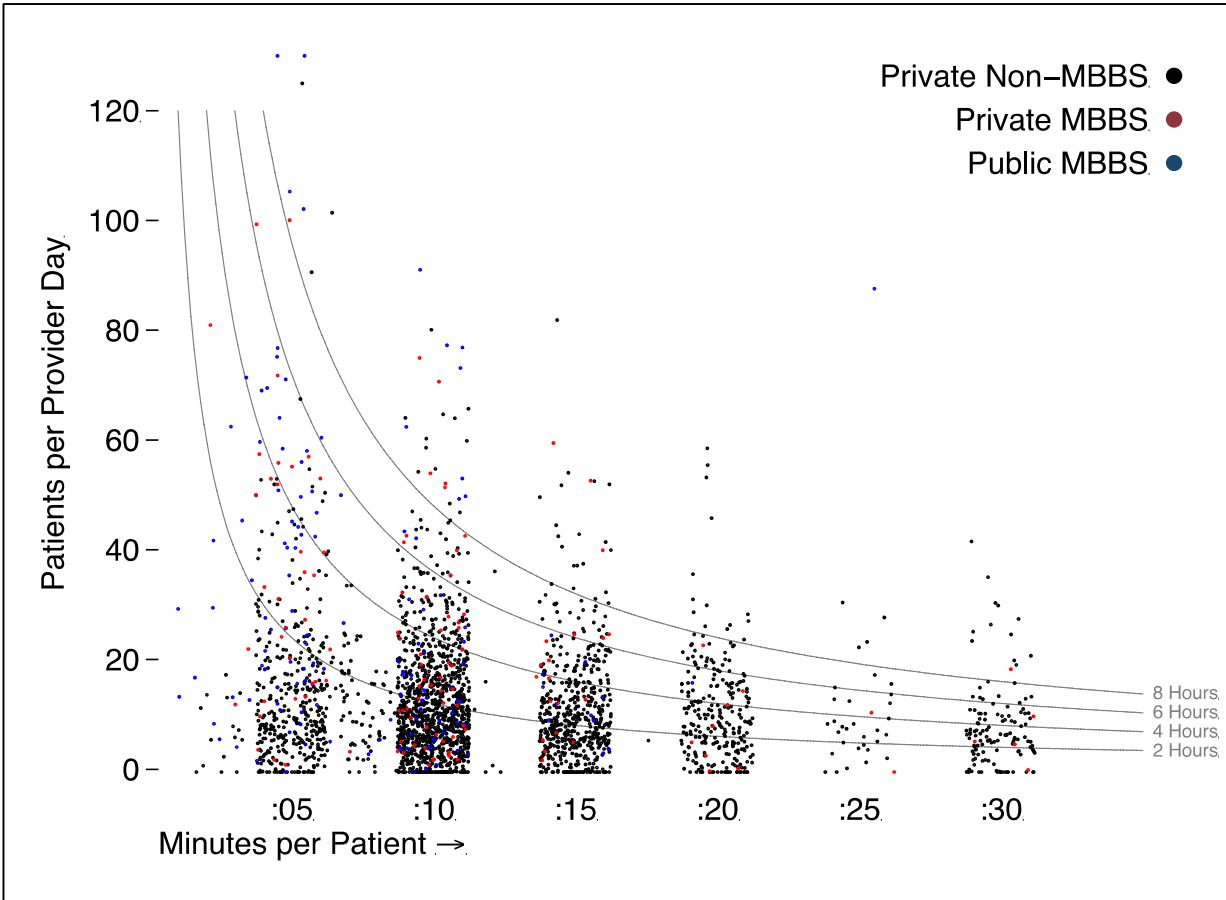
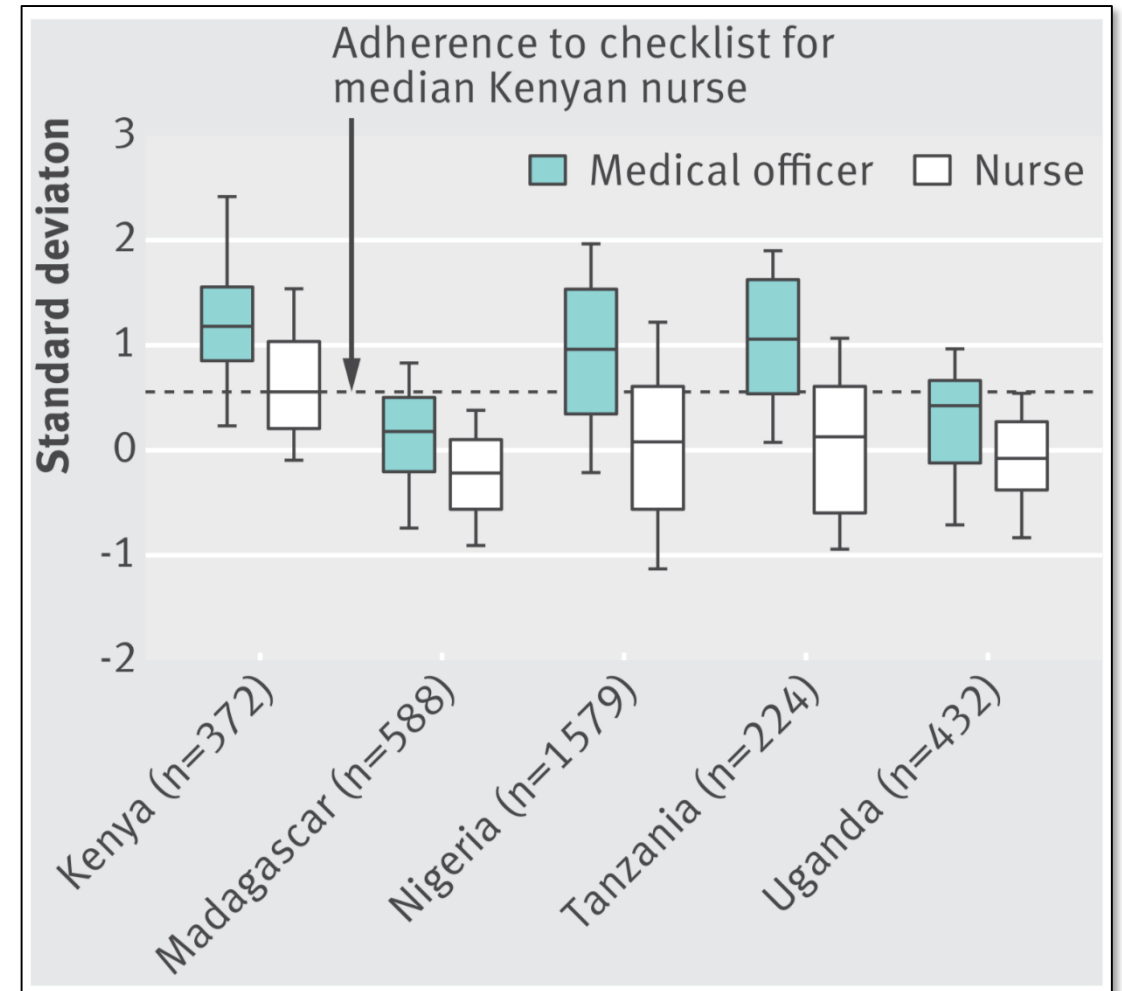
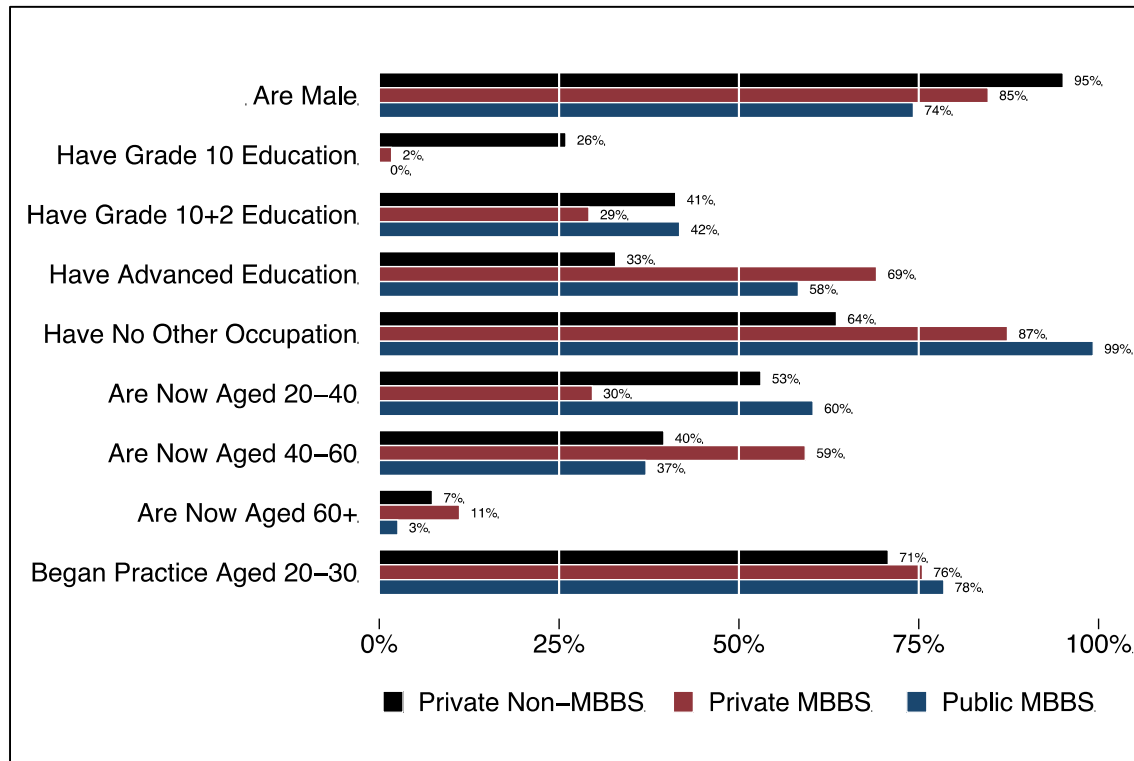


FIGURE 3. HOSPITAL ADMISSION RATES BY RACE/ETHNICITY

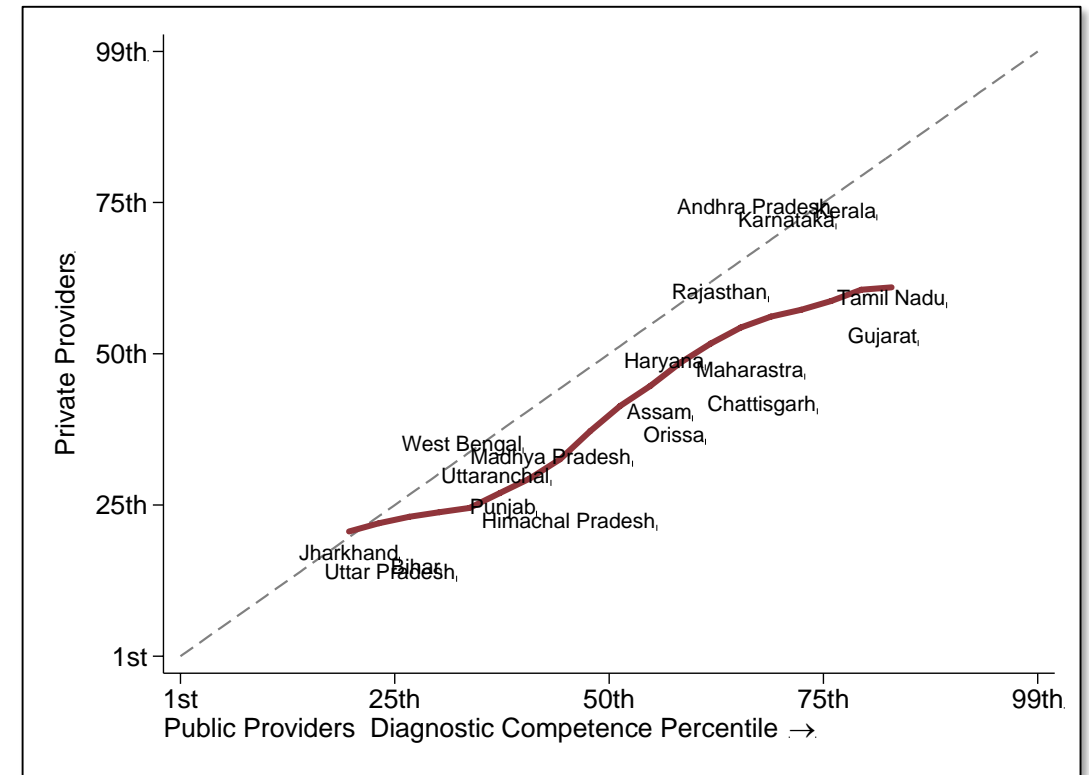
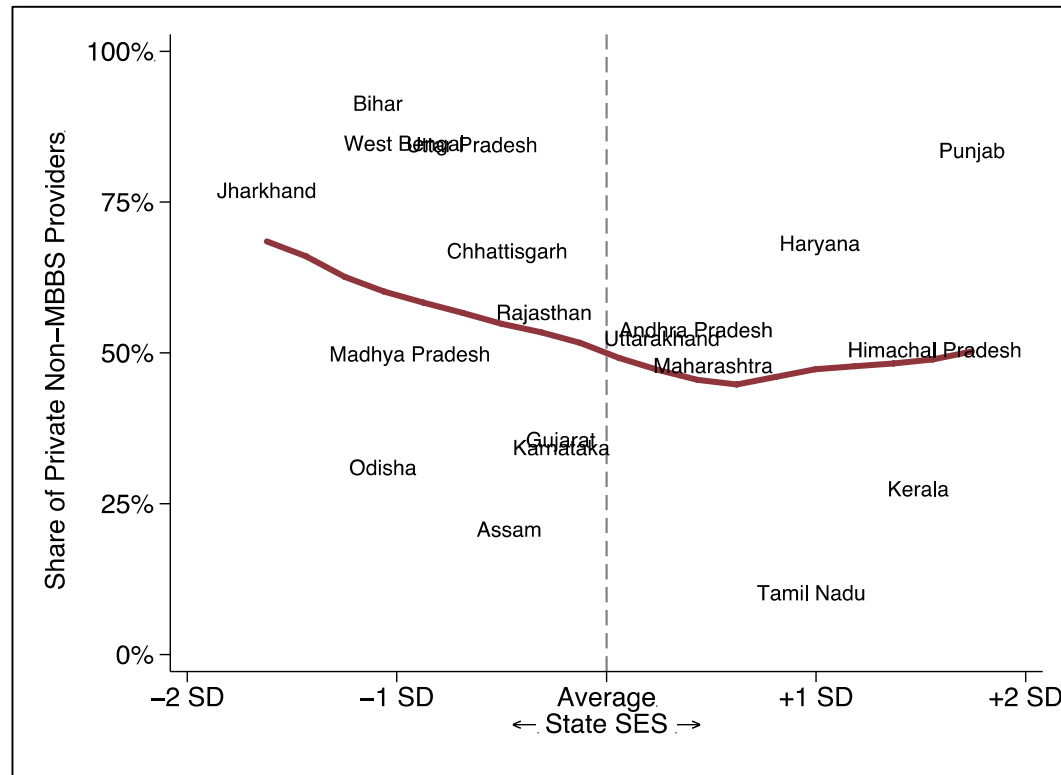
Examples: Examining distributions



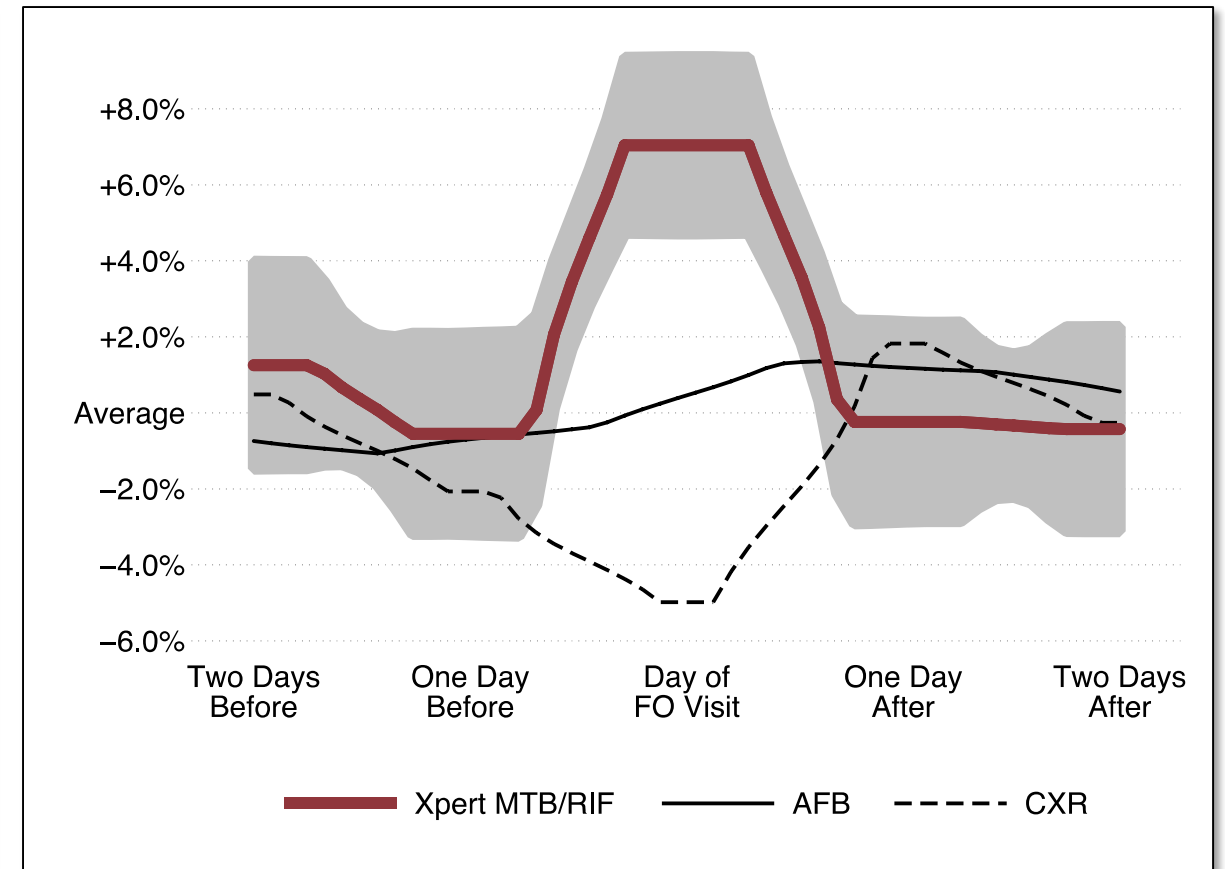
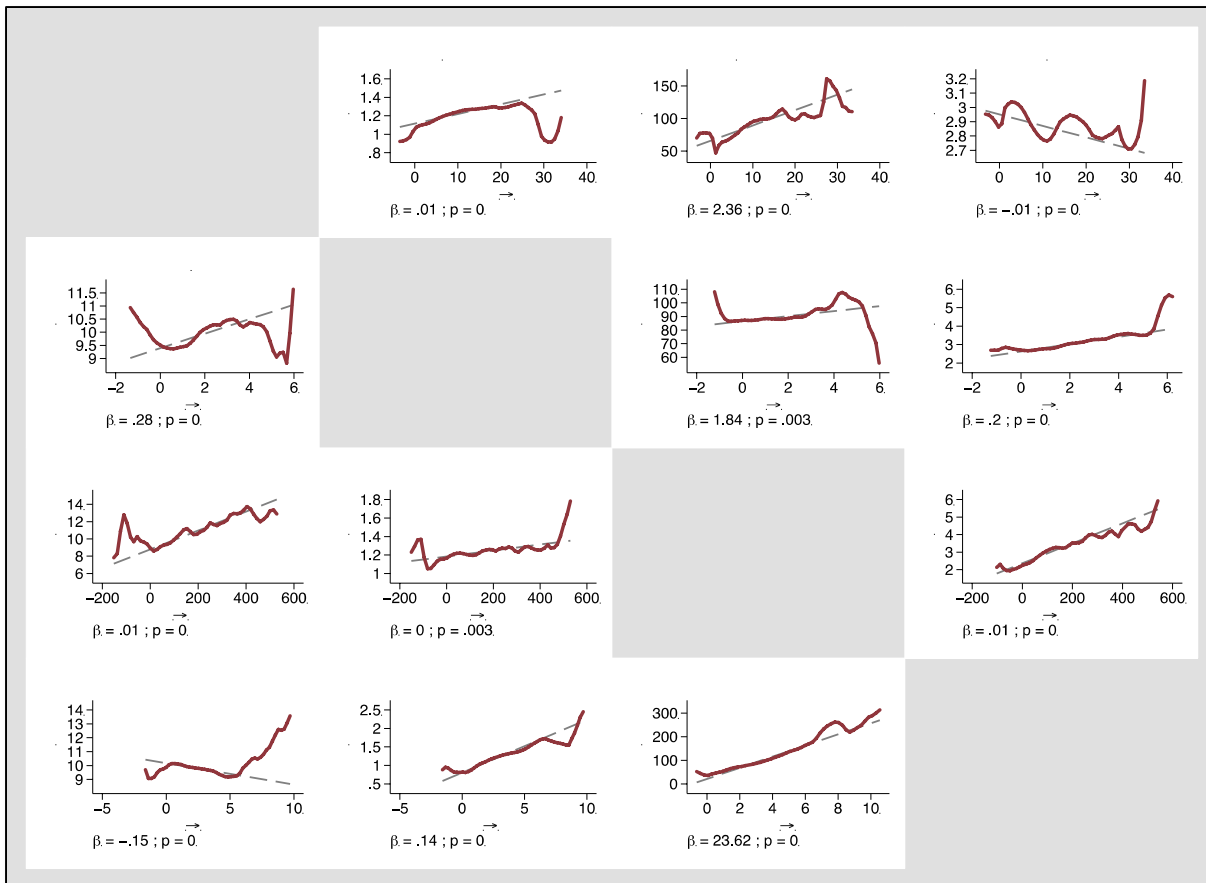
Examples: comparing means



Some examples: comparing correlations

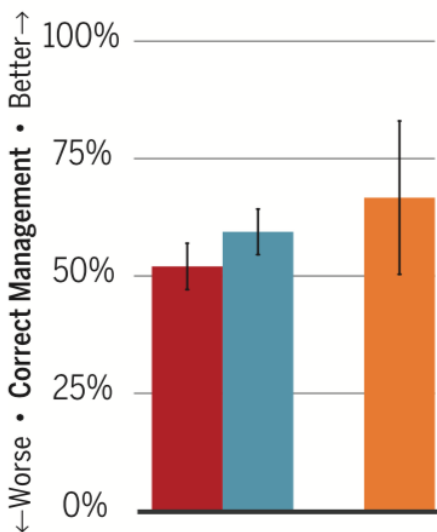


Examples: Searching for patterns

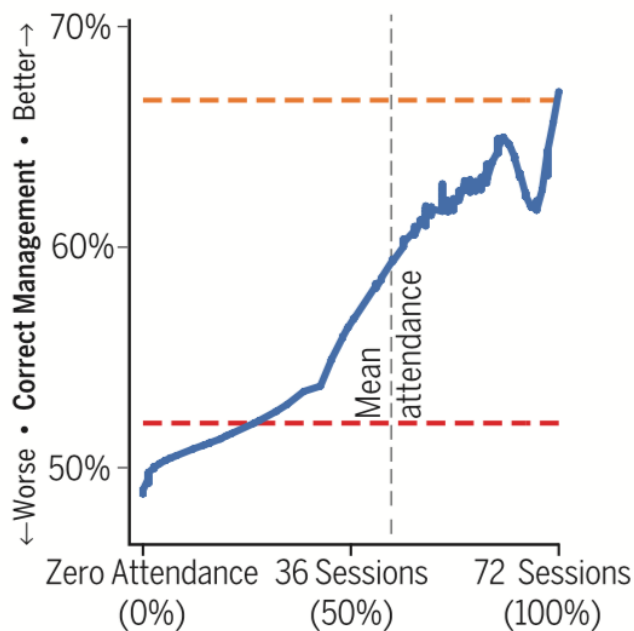


Examples: telling a story about treatment takeuptake

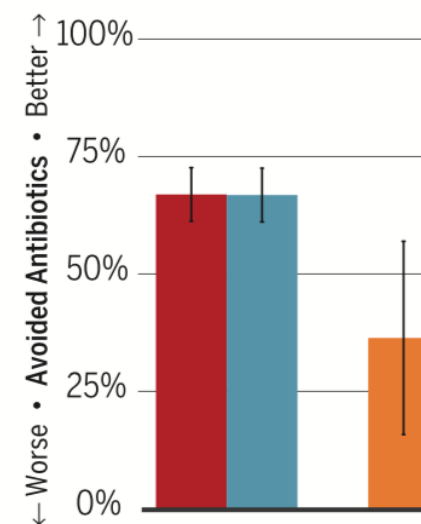
Despite 56% mean attendance, trained informal providers correctly managed more cases, closing half the gap with the public sector.



Providers who completed the full training course correctly managed cases as often as public-sector doctors.



However, training had no impact on the avoidance of unnecessary antibiotics.



Control: Untrained Informal

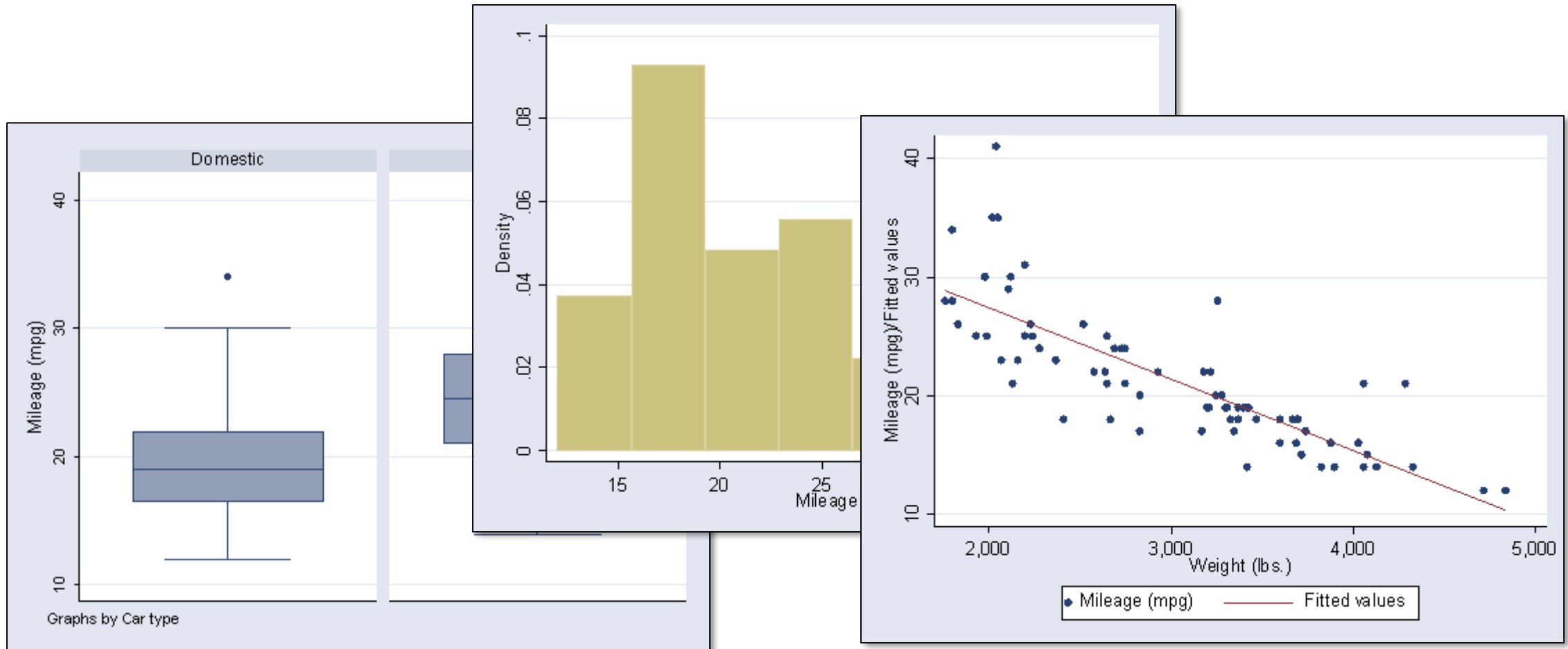
Trained Informal

Benchmark: Public Sector

Why do these charts look so good?

- They tell a story
- They use consistent visual (design) language
- They draw the eye to what is important

Stata default graphs are not very attractive



Stata has three core built-in graph functions

[*graph graphtype*]

- graphs which plot one or more variables *on one axis*

[*twoway graphtype*]

- graphs which plot two variables *together on an x,y axis*

[*histogram*] , [*kdensity*] , [*lowess*]

- Essential *distributional* commands

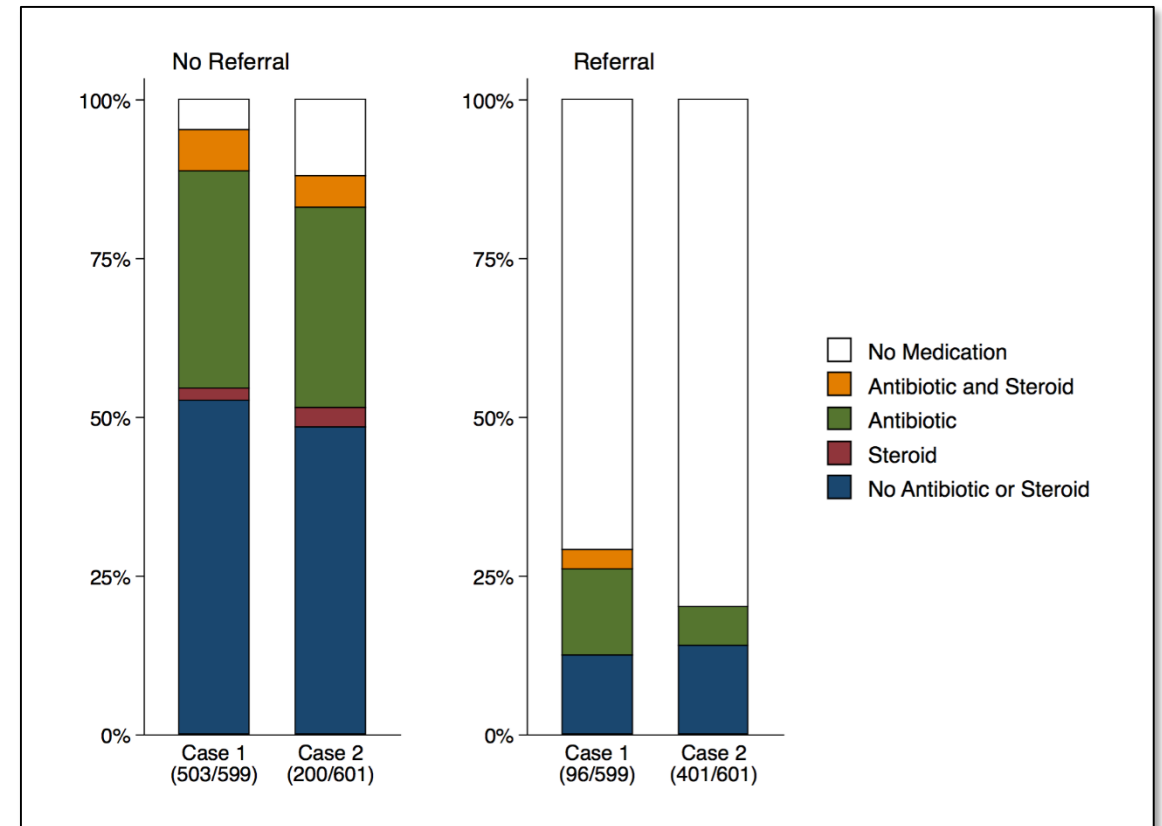
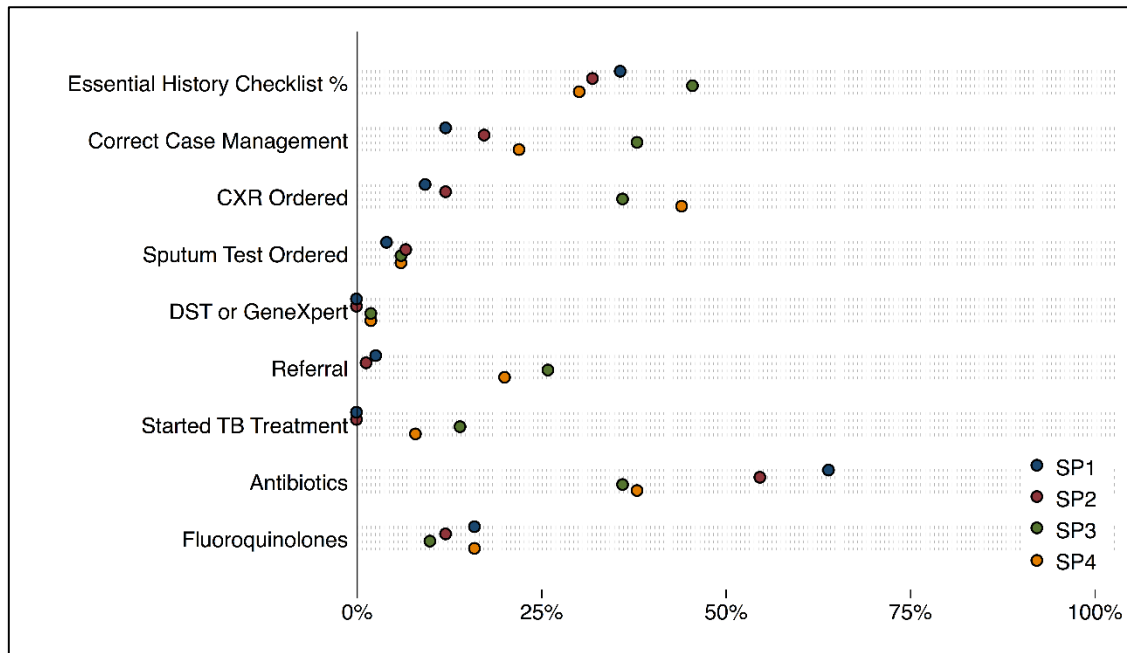
The other graph commands are implemented in terms of **graph**, which provides the following capabilities:

Command	Description
graph bar	bar charts
graph pie	pie charts
graph dot	dot charts
graph matrix	scatterplot matrices
graph twoway	twoway (y-x) graphs, including
graph twoway scatter	scatterplots
graph twoway line	line plots
graph twoway function	function plots
graph twoway histogram	histograms
graph twoway *	more

Smoothing and densities:

Command	Description
kdensity	kernel density estimation, univariate
lowess	lowess smoothing
lpoly	local polynomial smoothing

Oneway [graph] plots can be very informative



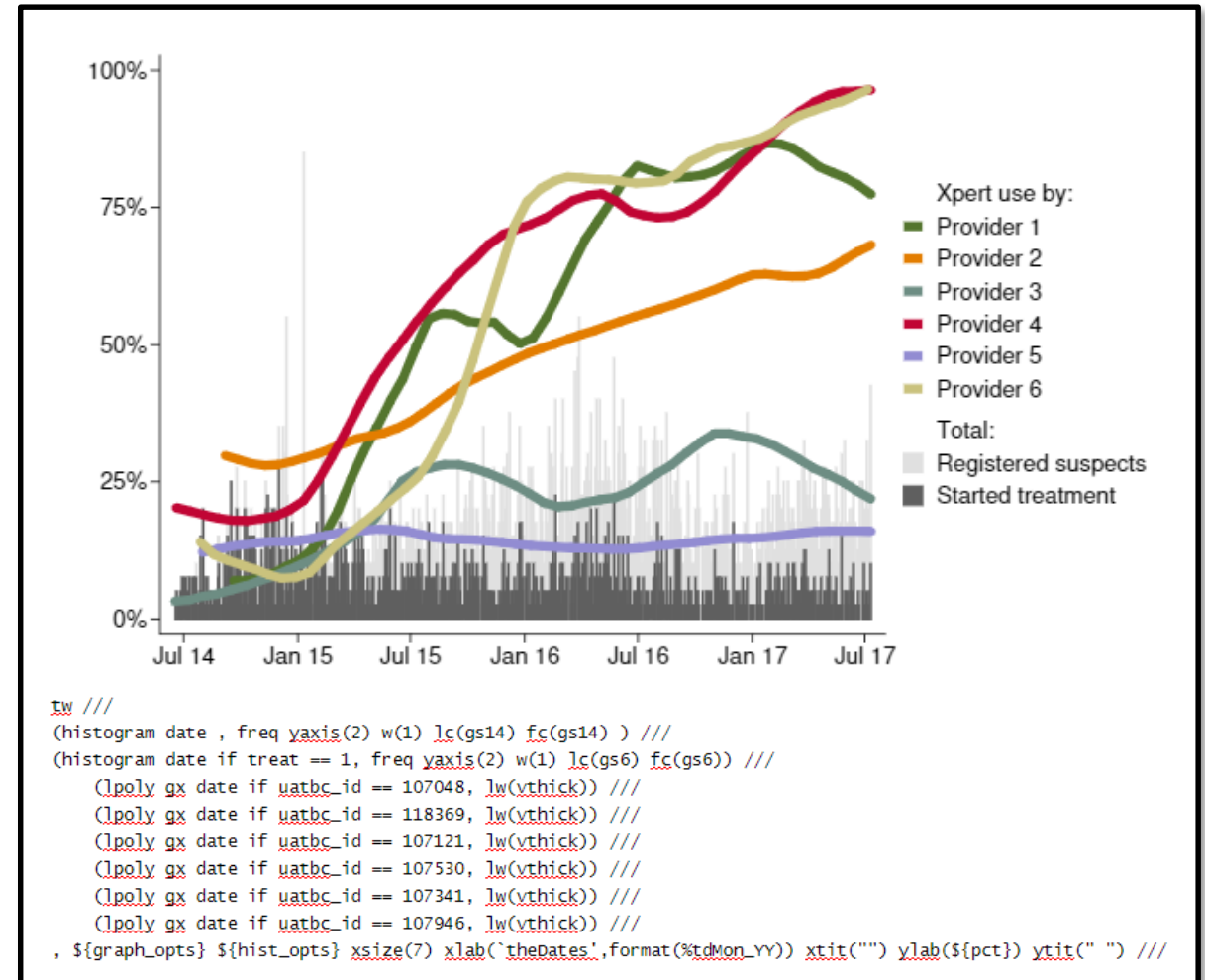
https://github.com/qutubproject/lancetid2015/blob/master/tables_figures.do (Figure S1)

https://github.com/qutubproject/lancetid2016/blob/master/tables_figures.do (Figure 2)

[*twoway*] graphs can be stacked up

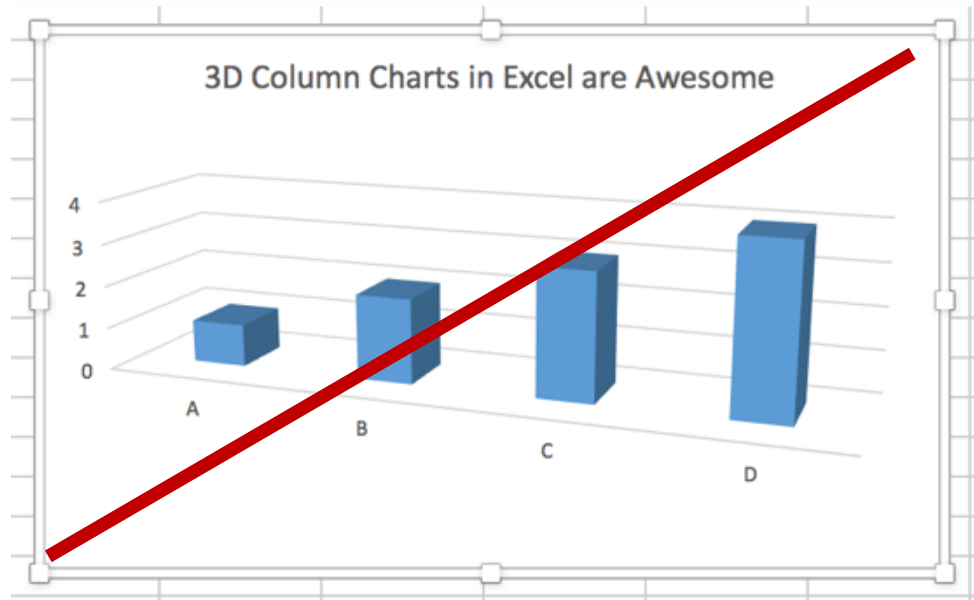
- The axes are abstract, so you do not need to use the same variables or the same units for each graph!
- Syntax:

```
[tw ///  
  (type var1 var2 , opts) ///  
  (type var3 var4 , opts) ///  
  , opts]
```
- Each can have its own if/in and options

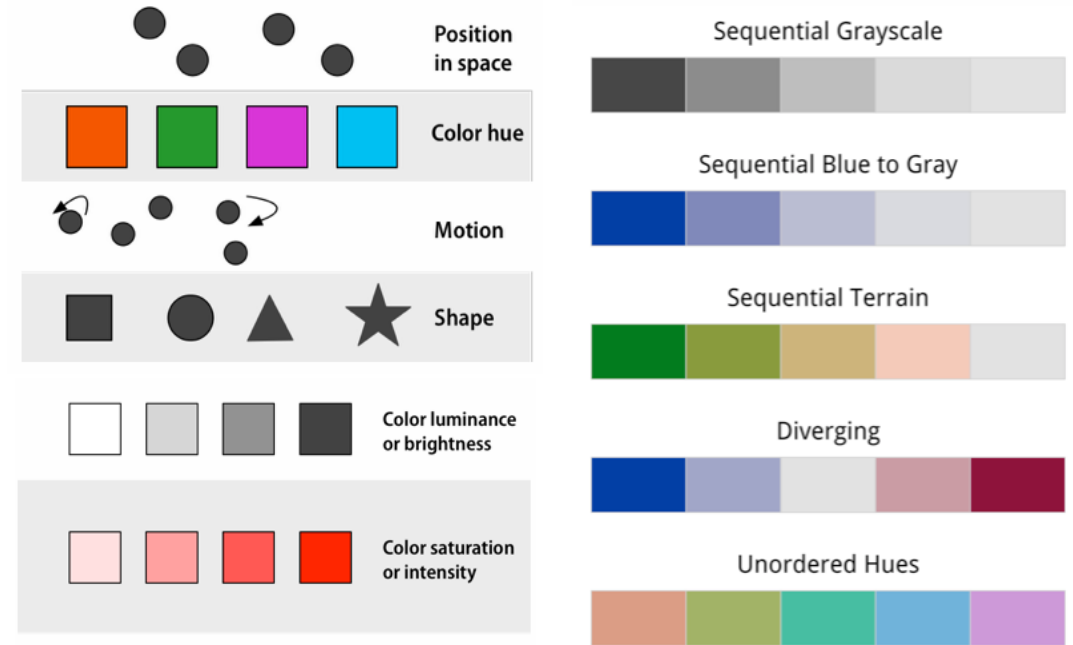


Charts show information across *dimensions*

Not these dimensions!



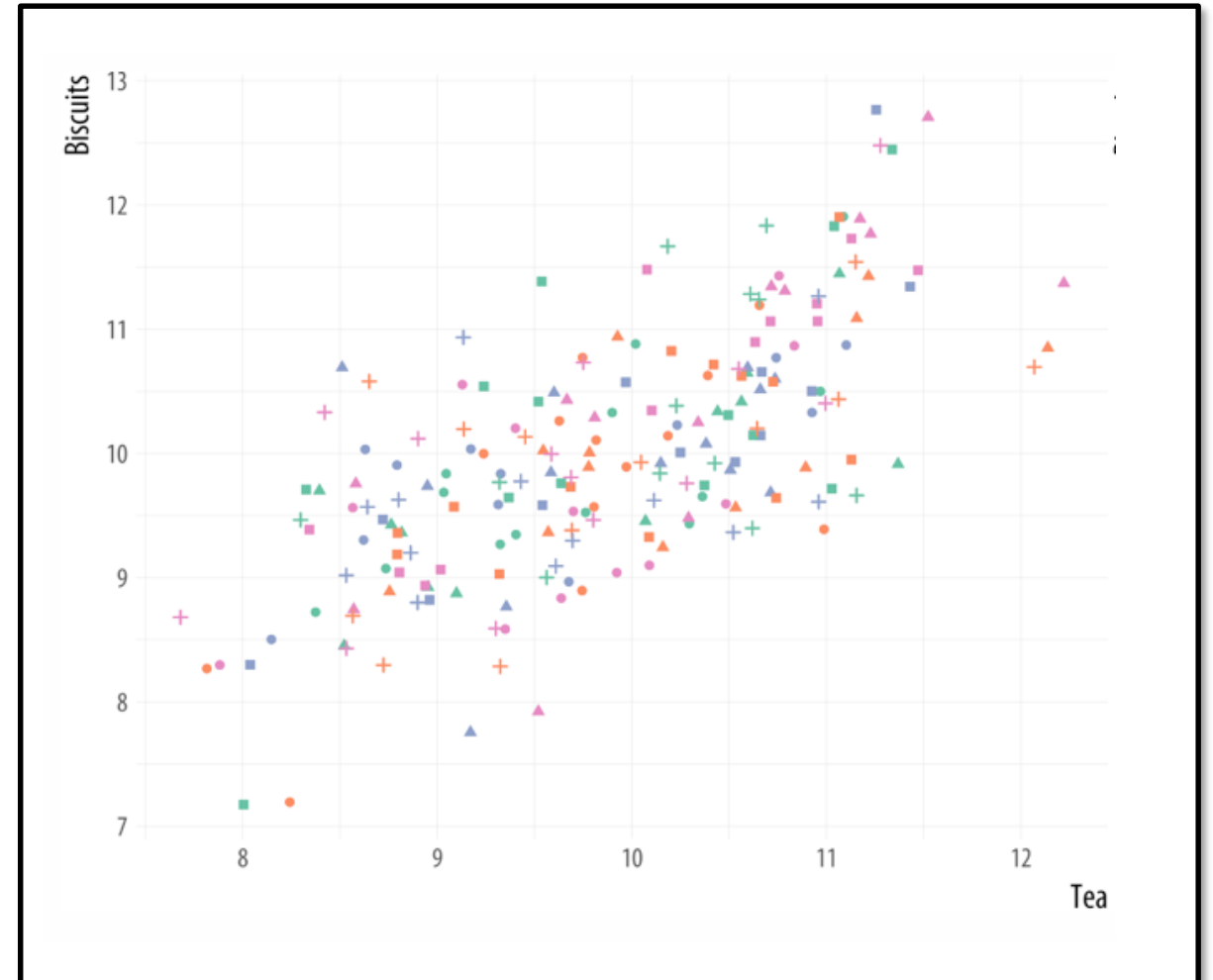
<http://socviz.co>



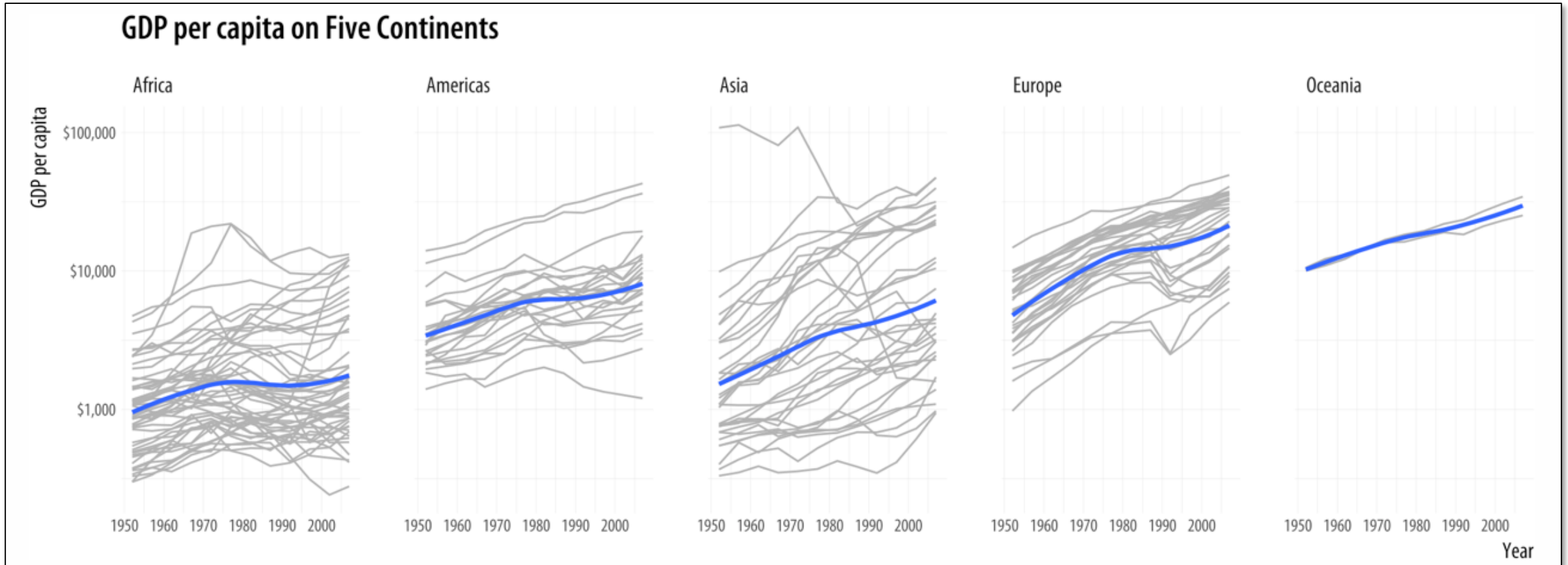
These dimensions!

Design with dimensions in mind

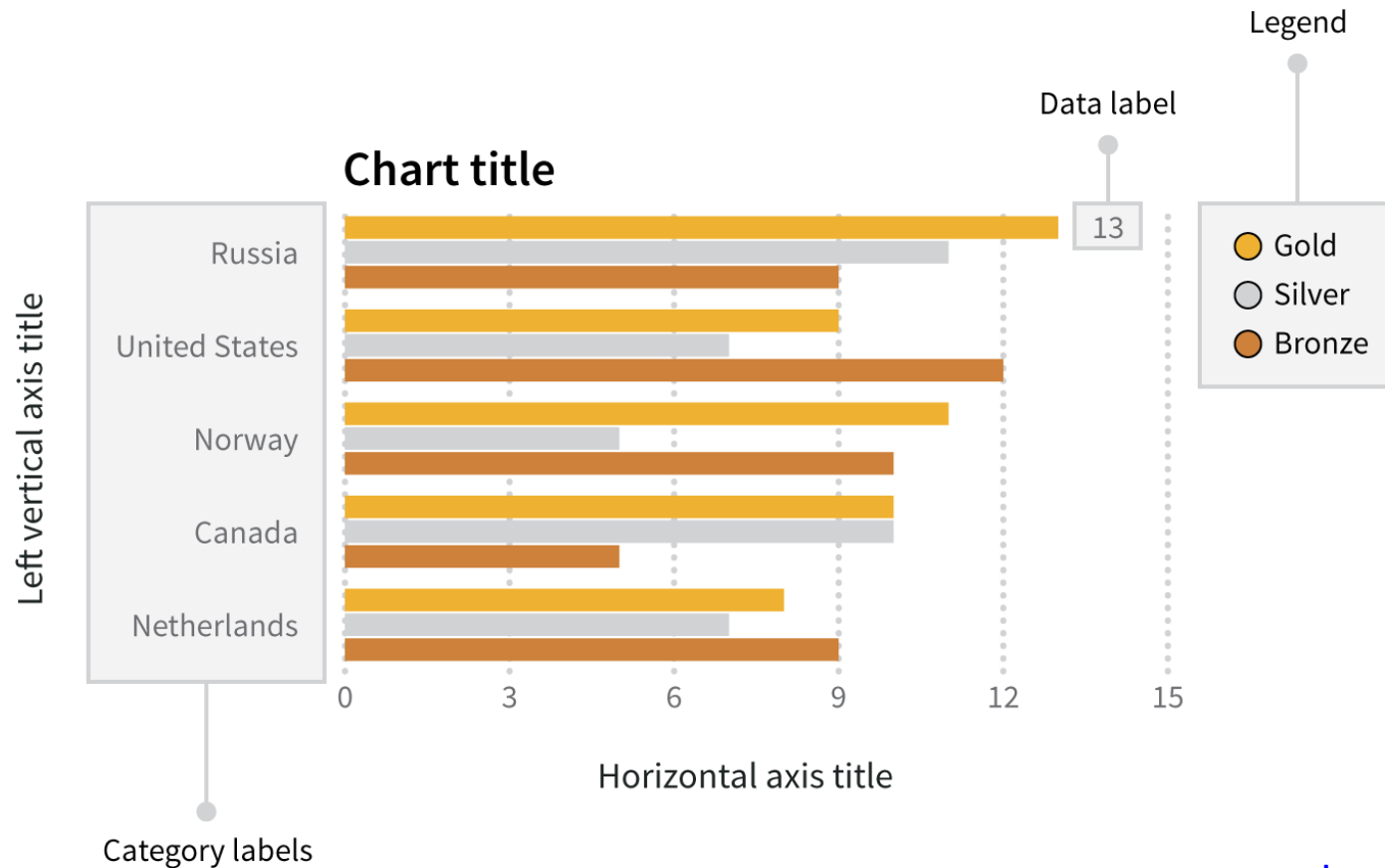
- A chart with a lot of information can blend together like TV static



Use design language to give charts meaning

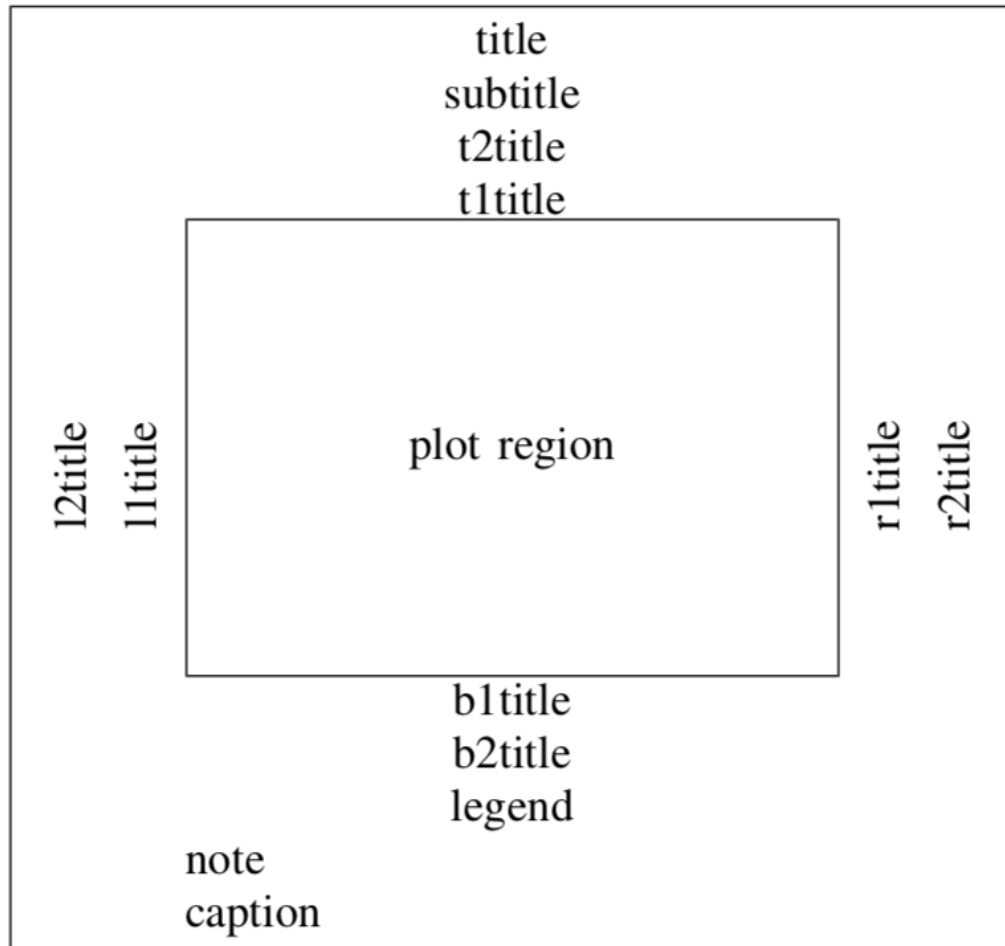


Anatomy of most graphs



<https://infoactive.co/data-design/ch14.html>

Components of Stata graphs

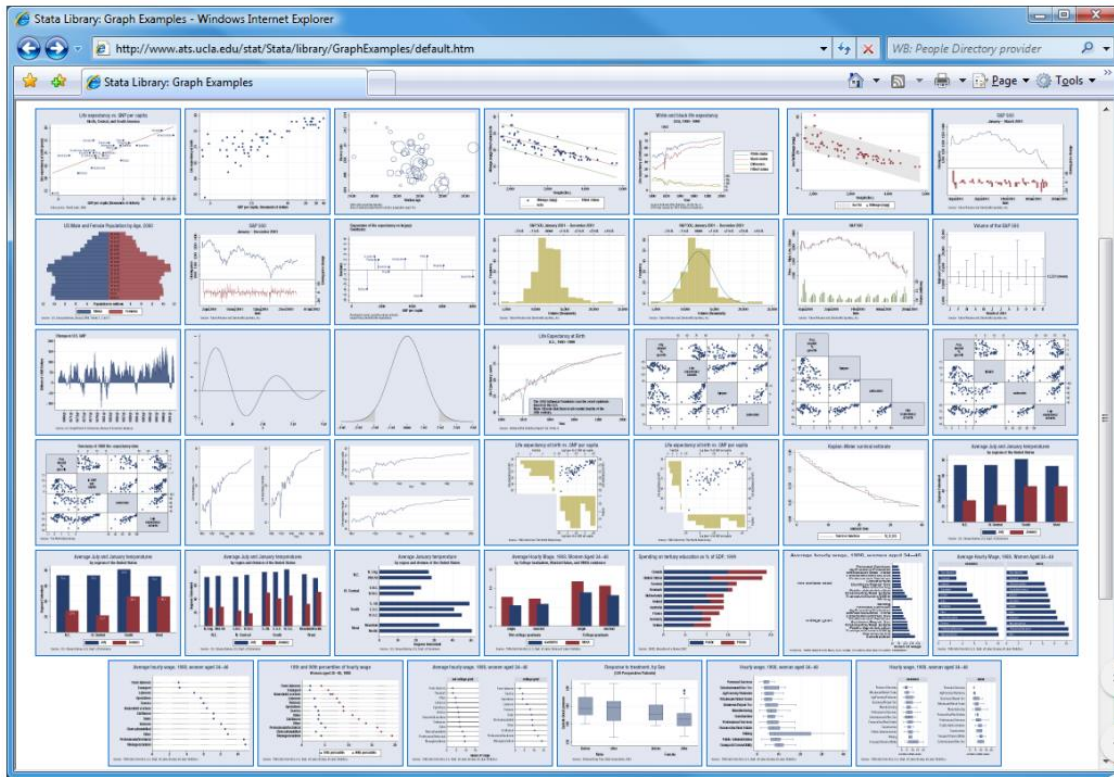


Interpretation of clock position()ring(k), $k > 0$ and ring(0)

	11	12	1	
10	10 or 11	12	1 or 2	2
9	9	0	3	3
8	7 or 8	6	4 or 5	4
	7	6	5	

Best place to start: [*h tw*]

Lots of graph types by default



Title

stata.com

graph twoway — Twoway graphs

Syntax

Menu

Description

Remarks and examples

Syntax

[graph] twoway plot [*if*] [*in*] [, *twoway_options*]

where the syntax of *plot* is

[(*c*) *plotype varlist* ..., *options*] [11]

plotype

Description

scatter

scatterplot

line

line plot

connected

connected-line plot

scatteri

scatter with immediate arguments

area

line plot with shading

bar

bar plot

spike

spike plot

dropline

dropline plot

dot

dot plot

rarea

range plot with area shading

rbar

range plot with bars

rspike

range plot with spikes

rcap

range plot with capped spikes

rcapsym

range plot with spikes capped with symbols

rscatter

range plot with markers

rline

range plot with lines

rconnected

range plot with lines and markers

Best place to start: [*h twoway_options*]

- Major graphing elements:
 - Lines
 - Shapes
 - Points
- Major styling elements:
 - Fill color
 - Outlines
 - Sizes
- Major graphing elements:
 - Lines
 - Labels

Title [stata.com](#)

twoway_options — Options for twoway graphs

Syntax Description Options Remarks and examples Also see

Syntax

The *twoway_options* allowed with all *twoway* graphs are

<i>twoway_options</i>	Description
<i>added_line_options</i>	draw lines at specified <i>y</i> or <i>x</i> values
<i>added_text_options</i>	display text at specified (<i>y,x</i>) value
<i>axis_options</i>	labels, ticks, grids, log scales
<i>title_options</i>	titles, subtitles, notes, captions
<i>legend_options</i>	legend explaining what means what
<i>scale(#)</i>	resize text, markers, line widths
<i>region_options</i>	outlining, shading, graph size
<i>aspect_option</i>	constrain aspect ratio of plot region
<i>scheme(schemename)</i>	overall look
<i>play(recordingname)</i>	play edits from <i>recordingname</i>
<i>by(varlist, ...)</i>	repeat for subgroups
<i>nodraw</i>	suppress display of graph
<i>name(name, ...)</i>	specify name for graph
<i>saving(filename, ...)</i>	save graph in file
<i>advanced_options</i>	difficult to explain

See [G-3] *added_line_options*, [G-3] *added_text_options*, [G-3] *axis_options*, [G-3] *title_options*, [G-3] *legend_options*, [G-3] *scale_option*, [G-3] *region_options*, [G-3] *aspect_option*, [G-3] *scheme_option*, [G-3] *by_option*, [G-3] *nodraw_option*, [G-3] *name_option*, [G-3] *saving_option*, [G-3] *advanced_options*.

Working with shapes

Showing different symbols by varying mymbol()

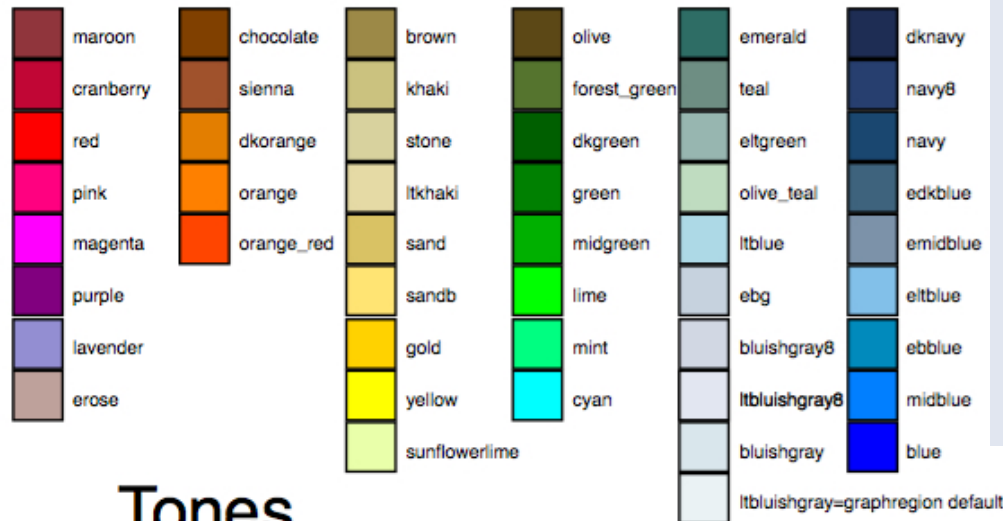
● D	○ Dh	● o	○ oh
◆ D	◇ Dh	◆ d	◇ dh
▲ T	△ Th	▲ t	△ th
■ S	□ Sh	■ s	□ sh
++	+ smplus	× X	× x
· p			

Showing different symbol sizes by varying msize()

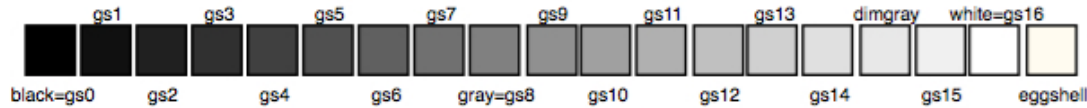
· vtiny	· tiny	· vsmall	• small
◆ medsmall	◆ medium	● medlarge	● large
● large	● huge	● vhuge	● ehuge

Working with colors

Colors



Tones



Showing different marker colors by varying mcolor()

◆black ◆gs0 ◆gs1 ◆gs2 ◆gs3 ◆gs4
◆gs5 ◆gs6 ◆gs7 ◆gs8 ◆gs9 ◆gs10

Showing different marker fill colors by varying mfcolor()

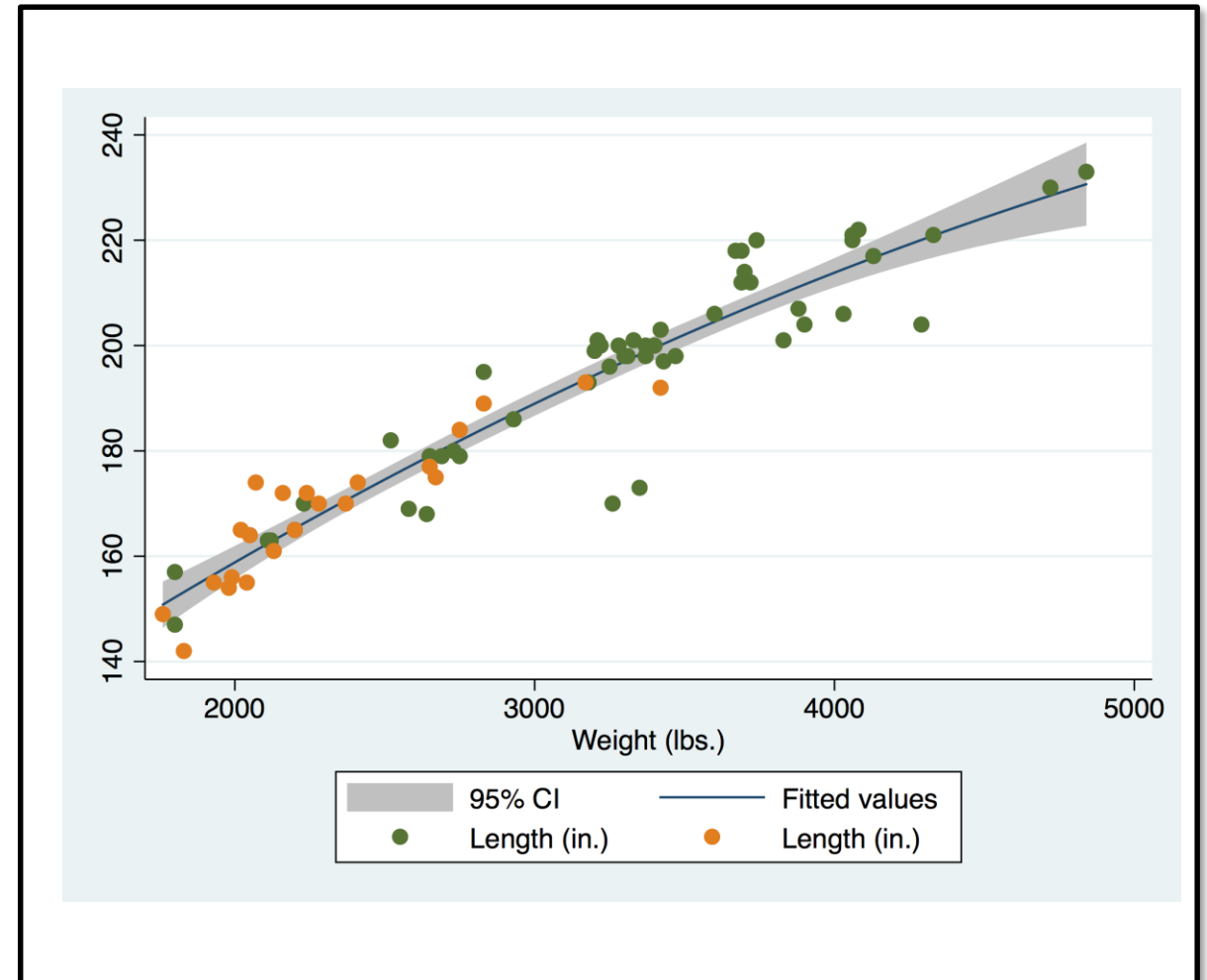
◆black ◆gs0 ◆gs1 ◆gs2 ◆gs3 ◆gs4

Showing different marker outline colors by varying mcolor()

●black ●gs0 ●gs1 ●gs2 ●gs3 ●gs4
●gs5 ●gs6 ●gs7 ●gs8 ●gs9 ●gs10
●gs11 ●gs12 ●gs13 ●gs14 ●gs15 ●gs16
●white ●blue ●bluishgray ●brown ●cranberry ●cyan
●dimgray ●dkgreen ●dknavy ●dkorange ●eggshell ●emerald
●forest_green ●gold ●gray ●green ●khaki ●lavender
●lime ●ltblue ●ltbluishgray ●ltkhaki ●magenta ●maroon
●midblue ●midgreen ●mint ●navy ●olive ●olive_teal
●orange ●orange_red ●pink ●purple ●red ●sand
●sandb ●sienna ●stone ●teal ●yellow ●ebg
●ebblue ●edkblue ●eltblue ●eltgreen ●emidblue ●erose

Every graph starts from the basics

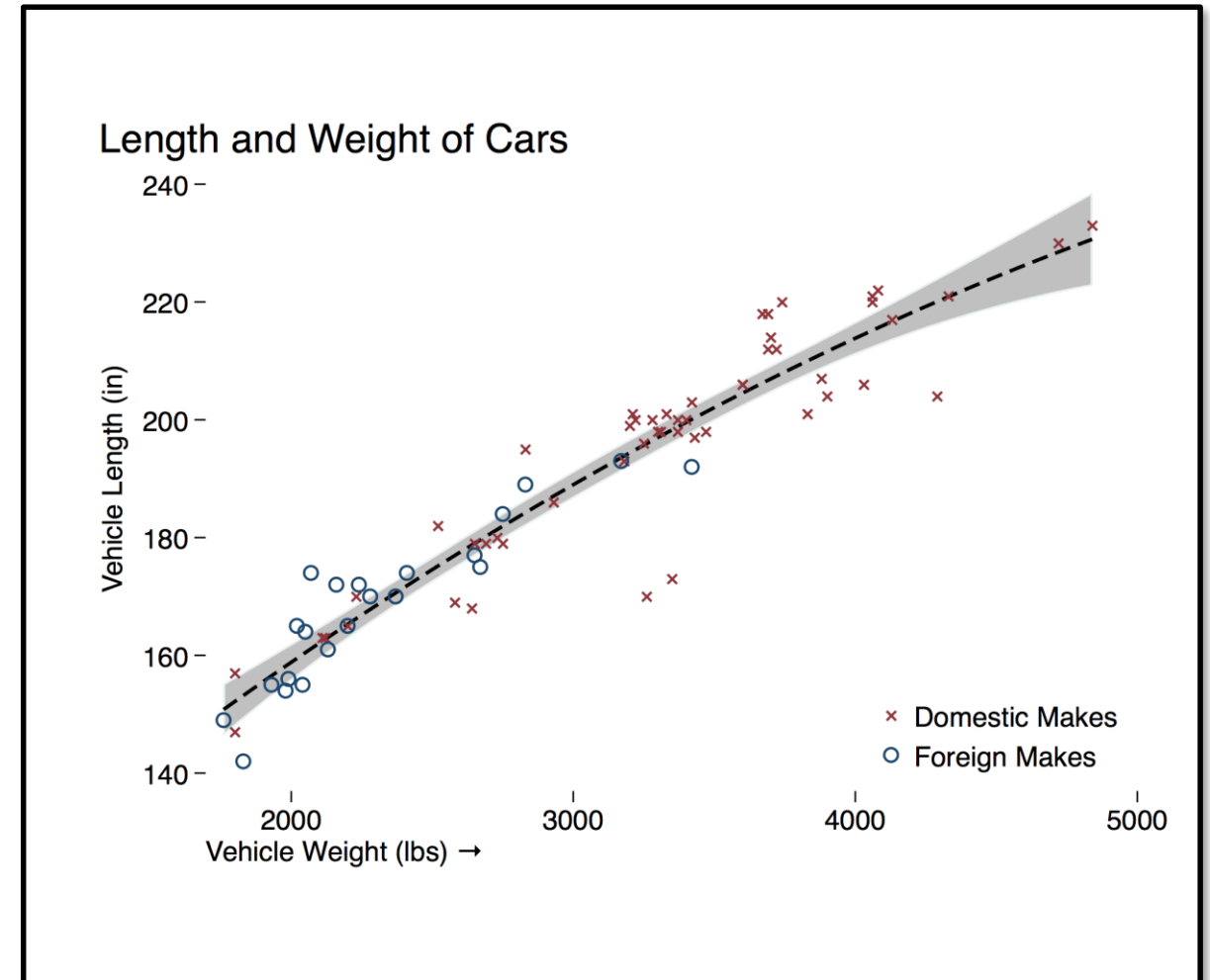
```
tw ///  
(qfitci length weight) ///  
(scatter length weight  
    if foreign == 0) ///  
(scatter length weight  
    if foreign == 1)
```



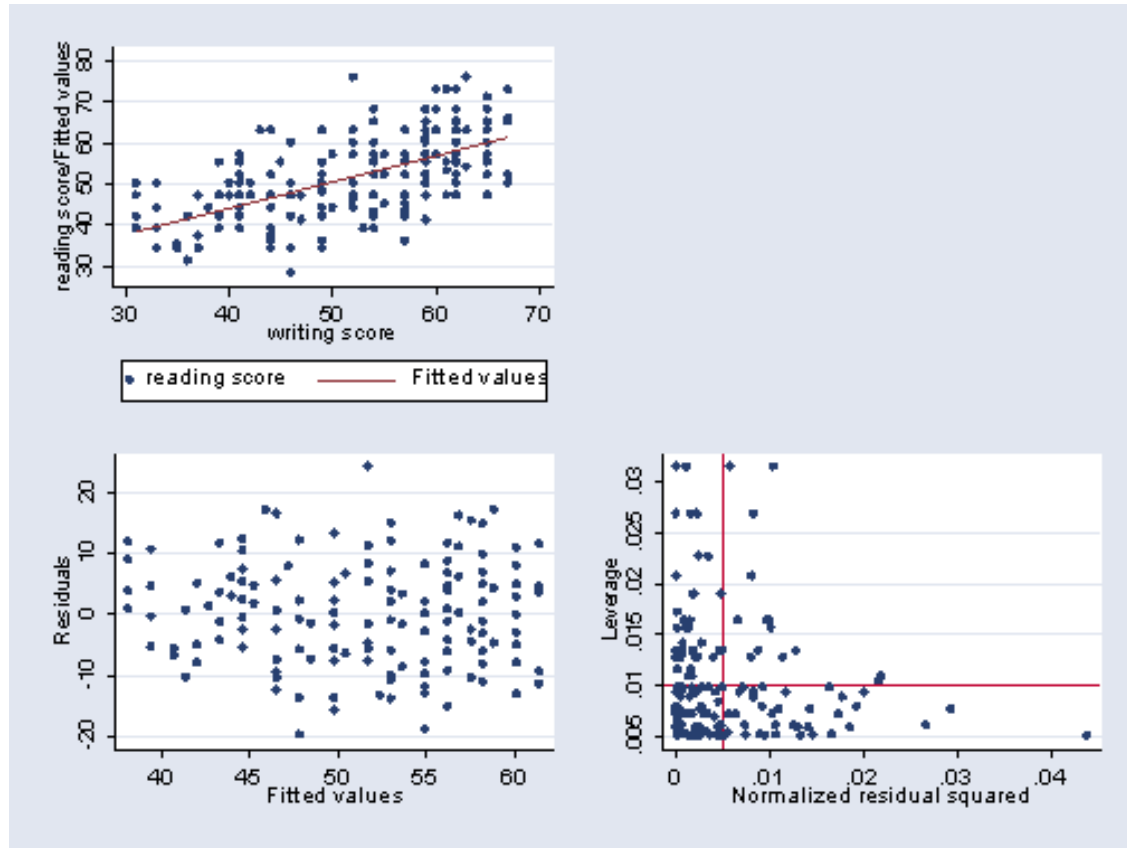
With styling, we have a pretty graph

- Options ... everywhere!

```
1 * Graph for slides
2
3 sysuse auto , clear
4
5 global graph_opts ///
6 title(, justification(left) color(black) span pos(11)) ///
7 graphregion(color(white) lc(white) lw(med) la(center)) ///
8 ylab(,angle(0) nogrid) xtit(,placement(left) justification(left)) ///
9 yscale(noline) xscale(noline) legend(region(lc(none) fc(none)))
10
11
12 tw ///
13 (qfitci length weight ///
14   , lp(dash) lc(black) lw(medthick) cip(rarea) alw(none)) ///
15 (scatter length weight if foreign = 0 ///
16   , ms(X) mc(maroon)) ///
17 (scatter length weight if foreign = 1 ///
18   , ms(O) mc(navy)) ///
19   , ${graph_opts} ///
20   xtitle("Vehicle Weight (lbs) {&rarr}") ///
21   ytitle("Vehicle Length (in)") ///
22   title("Length and Weight of Cars") ///
23   legend(order(3 "Domestic Makes" 4 "Foreign Makes") ///
24     ring(0) pos(5) c(1))
```



Graphs can be combined and exported



graph export ///

“filename” /// (.png or .eps)

, replace

With .png, specify “width(1000)” for higher resolution

.eps files can scale to any size on most modern software (but hard to preview on older systems)

DIME Resources (please contribute!)

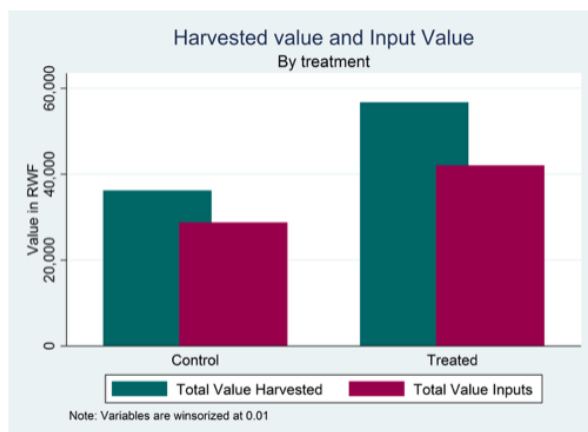
<https://worldbank.github.io/Stata-IE-Visual-Library/>

<https://worldbank.github.io/stata/>

IE Visual Library in Stata

Bar plots

Bar plot of two variables by treatment



```
# d;  
graph bar w_total_val_harvested_a w_total_val_inputs_a,  
    over(treated)  
    bargap(-30)  
    legend(label(1 "Total Value Harvested")  
           label(2 "Total Value Inputs"))  
    bar (1, color("0 102 102") )  
    bar (2, color("153 0 76") )  
    ytitle ("Value in RWF")  
    title ("Harvested value and Input Value")  
    subtitle ("By treatment")  
    note ("Note: Variables are winsorized at 0.01");  
# d cr
```

stata

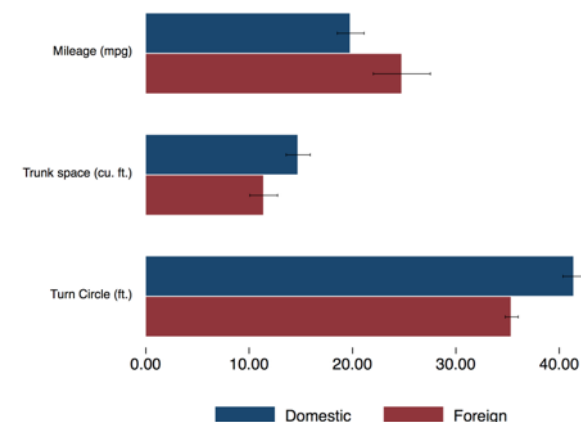
Stata Commands for Data Management
and Analysis

View the Project on GitHub
[worldbank/stata](https://github.com/worldbank/stata)

Commands for Data Analysis

betterBar

betterBar creates bar graphs for multiple variables with confidence intervals, setting `by()` and `over()` groups, adding labels and legends, and various styling commands.



```
wb_git_install betterBar  
sysuse auto , clear  
betterBar mpg trunk turn ///  
    , over(foreign) se ///  
    barlook(1 lw(thin) lc(white) fi(100))
```

Thank you!

