Some graphical tips for Stata users

Nicholas J. Cox
Department of Geography

Durham University

Pleased to join you from Britain

Here are some graphical tips, some old, some new(er).

You or your team might consider them when designing and coding graphics.

The talk mixes examples from official and community-contributed commands and details both large and small.

Ground rules

Examples won't include full Stata code.

Code to be posted online (in due course) will be fully reproducible.

I use Stata 19.50 with various other commands installed, as will be explicit.

If you are using Stata 17 or an earlier version, watch out for scheme `stcolor`. You will need to use a different scheme.

SJ means *Stata Journal*.

SSC means Statistical Software Components archive.

We have not looked at our results until we have displayed them effectively.

John Wilder Tukey. 1977.
Exploratory Data Analysis.
Reading, MA: Addison-Wesley, p.56.

1915–2000

Small stuff

The devil is often in the details…

Scatter plots

The scatter plot is the workhorse of statistical graphics.

John M. Chambers. 1977. *Computational Methods for Data Analysis*. New York: John Wiley, p.221.

Its invention is attributed to Sir John Herschel (1792–1871).

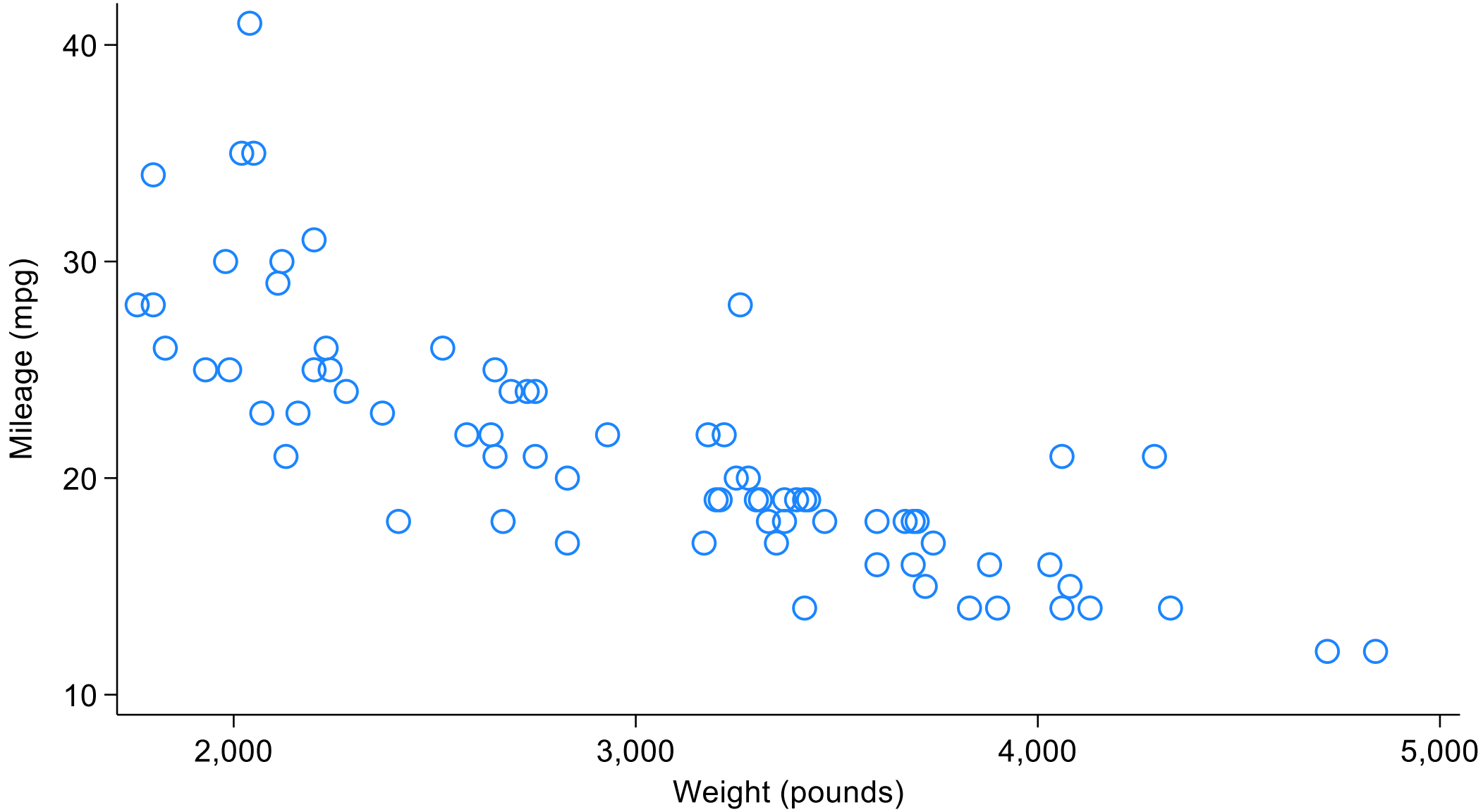
Use open markers

To me open or hollow circles oh or oh are the most suitable default marker symbols, as they tolerate overlap well and are easier on the eye.

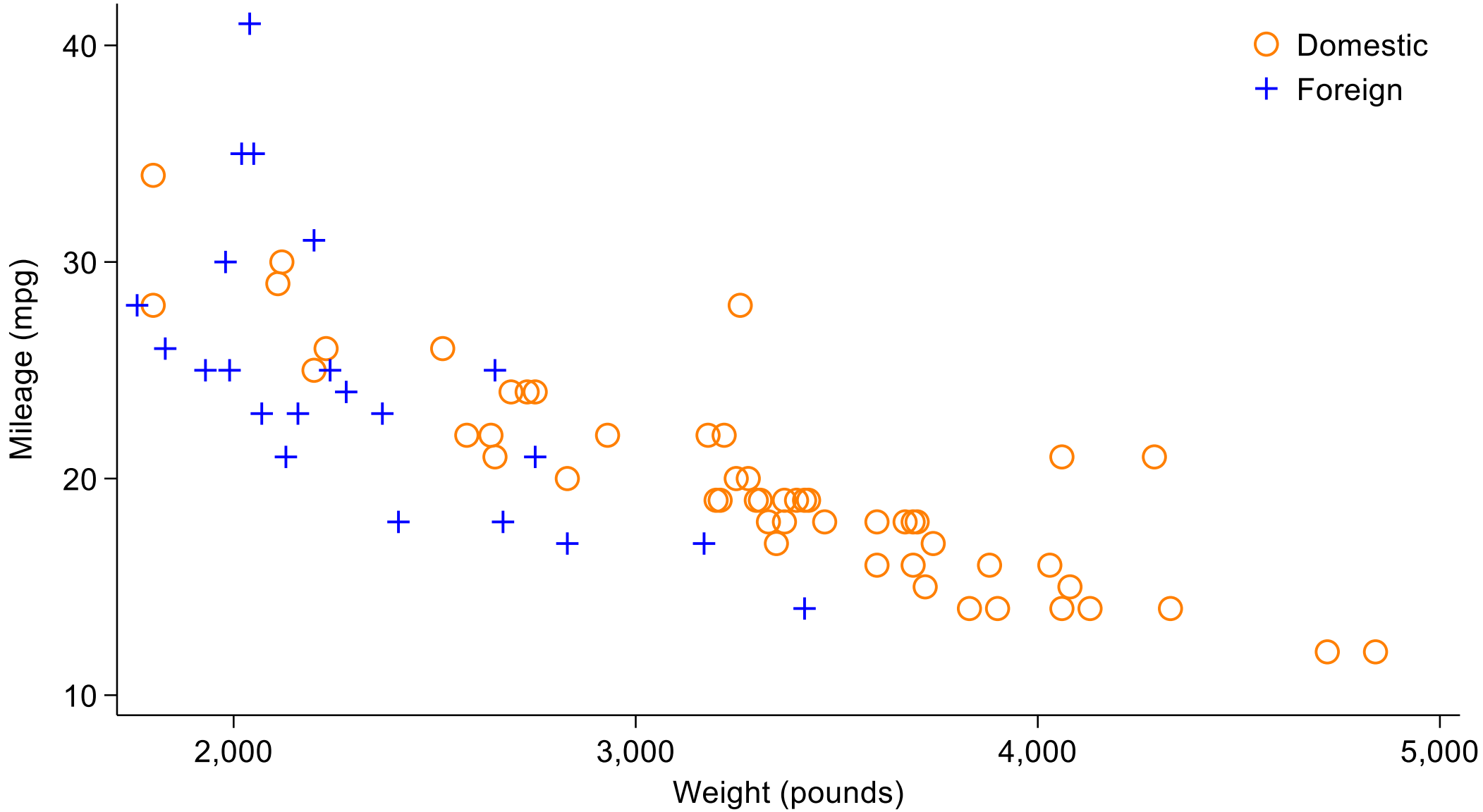
Other open markers, such as squares, diamonds and triangles, often work well too.

Plus signs + combine well with open markers, even when values are almost or exactly identical.

Resist the puzzling fashion that every marker is a solid circle!



auto data


Domestic
Foreign
Mileage (mpg)
40
30
20
10
2,000
3,000
4,000
5,000
Weight (pounds)
auto data

Use marker labels as markers

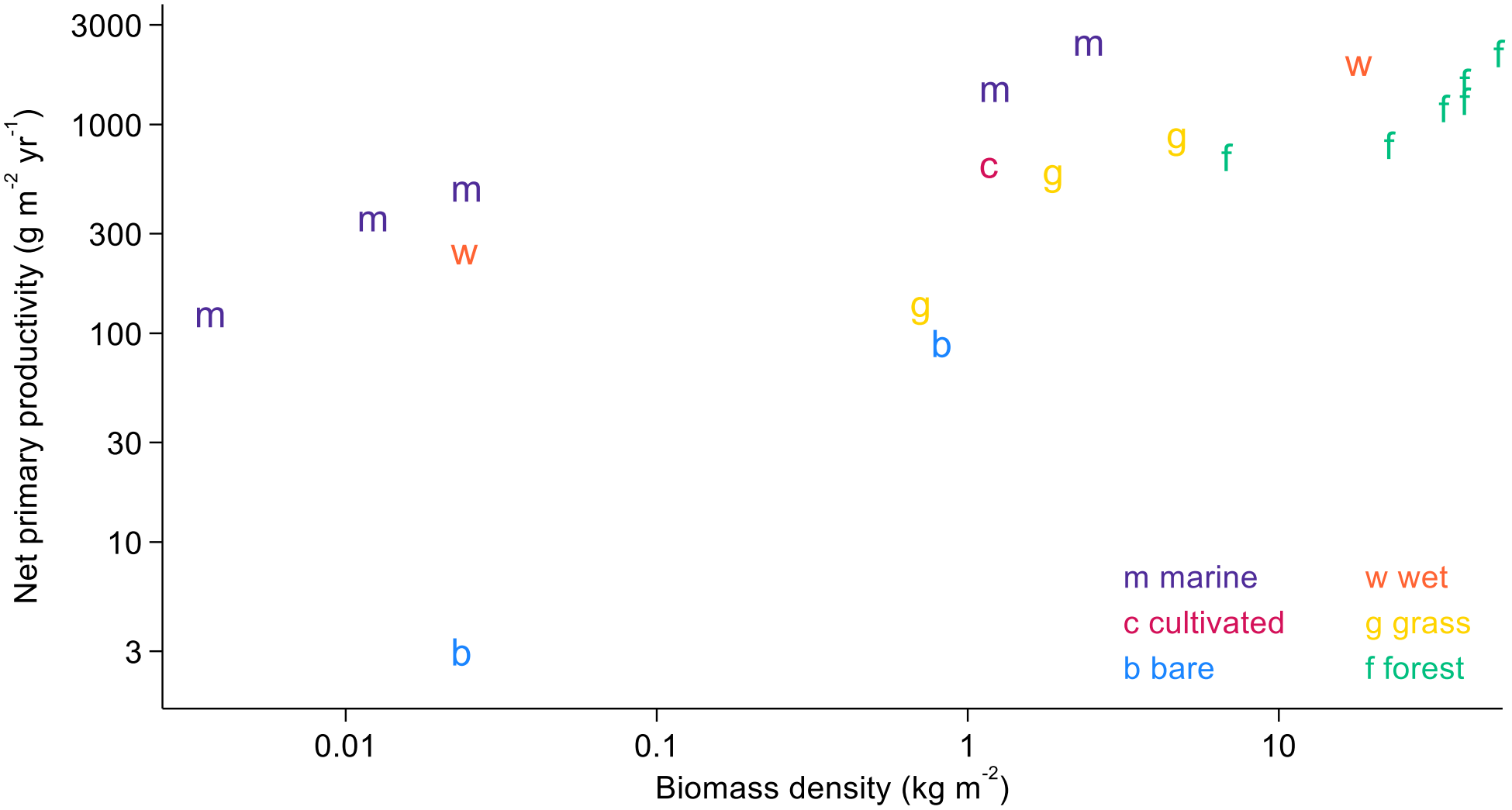
You can use marker labels as markers with

`mlabel(`*`varname`*`) mlabpos(0)`

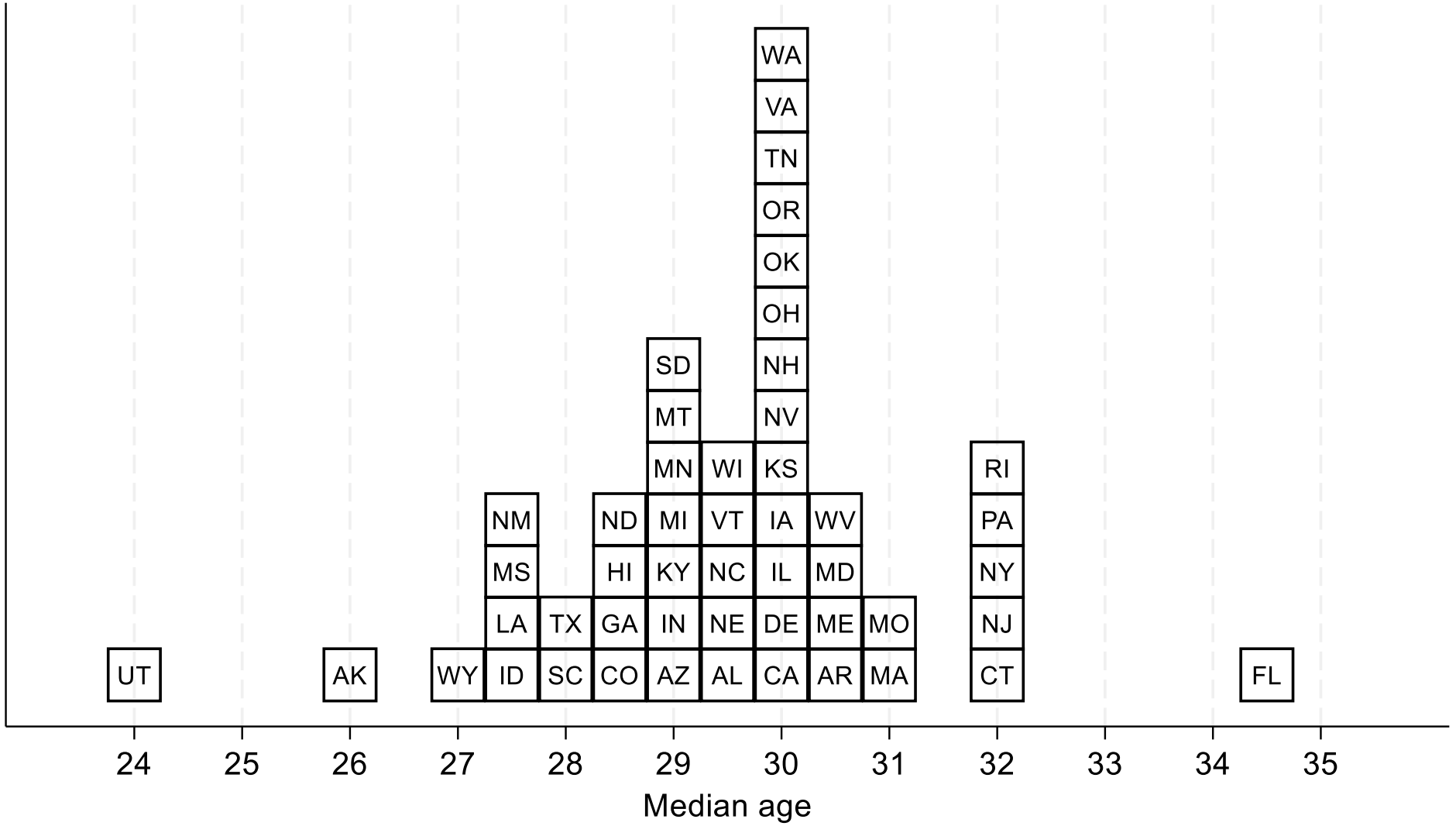
where variable *`varname`* can usefully contain (say) single characters, such as a-z or 0-9, or two- or three-letter abbreviations (TLAs).

See *SJ* 5: 604–606 (2005).

Compare using marker labels ***and*** markers together.



Robert H. Whittaker. 1975. *Communities and Ecosystems.*
New York: Macmillan, p.224



US Census 1980: half-year bins

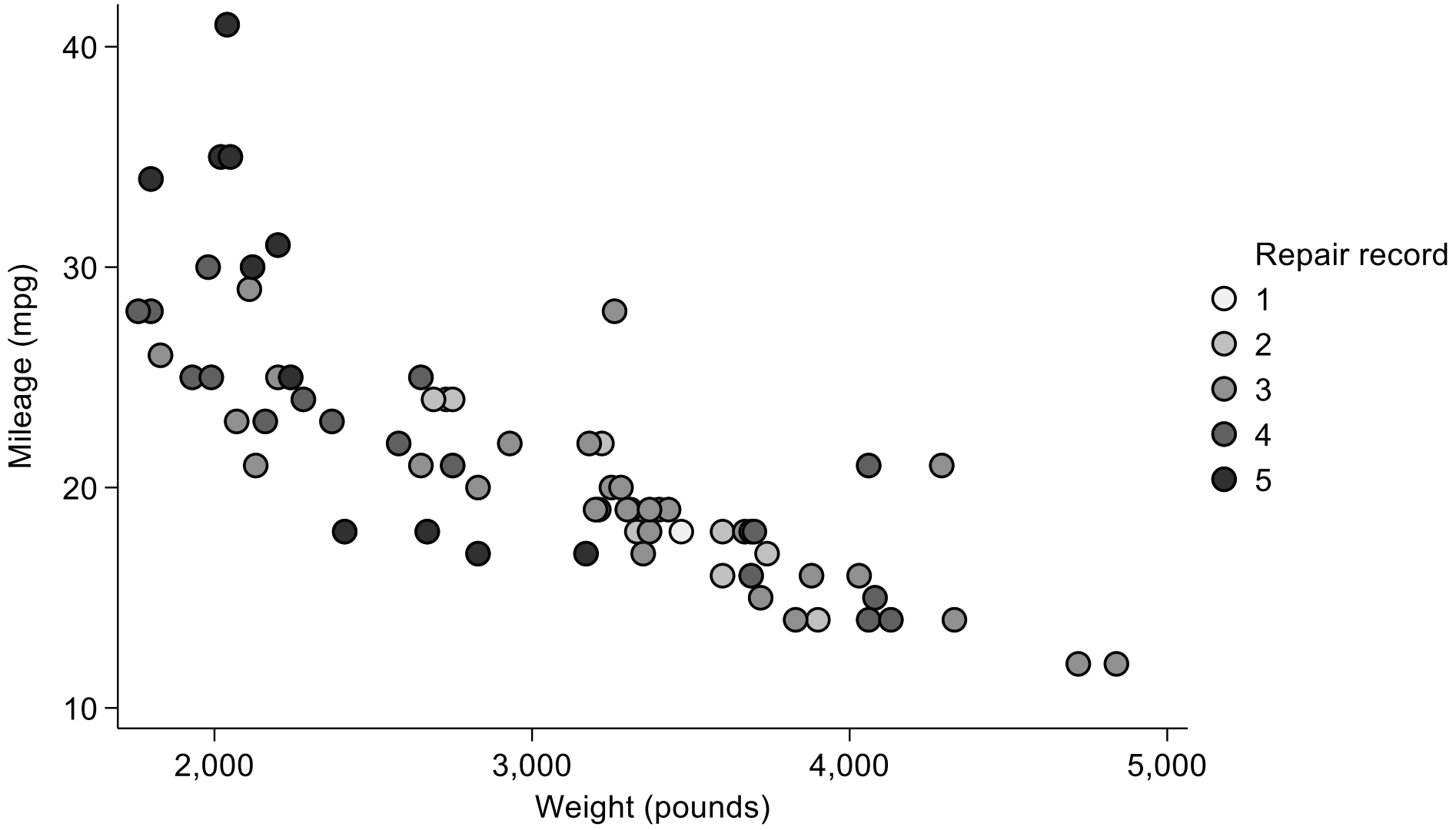
Go grey

Grey is good for markers and lines in a backdrop:
more examples later.

You can usually publish graphs with grey when a journal
won't support full colour.

Ordered sequences are clear when classifying data points.
Use both `mfcolor()` and `mlcolor()`.
Darker outline colours enhance visibility.

See *SJ* 5: 604–606 (2005).


40
30
20
10
Mileage (mpg)
2,000
3,000
4,000
5,000
Weight (pounds)
Repair record
1
2
3
4
5

auto data

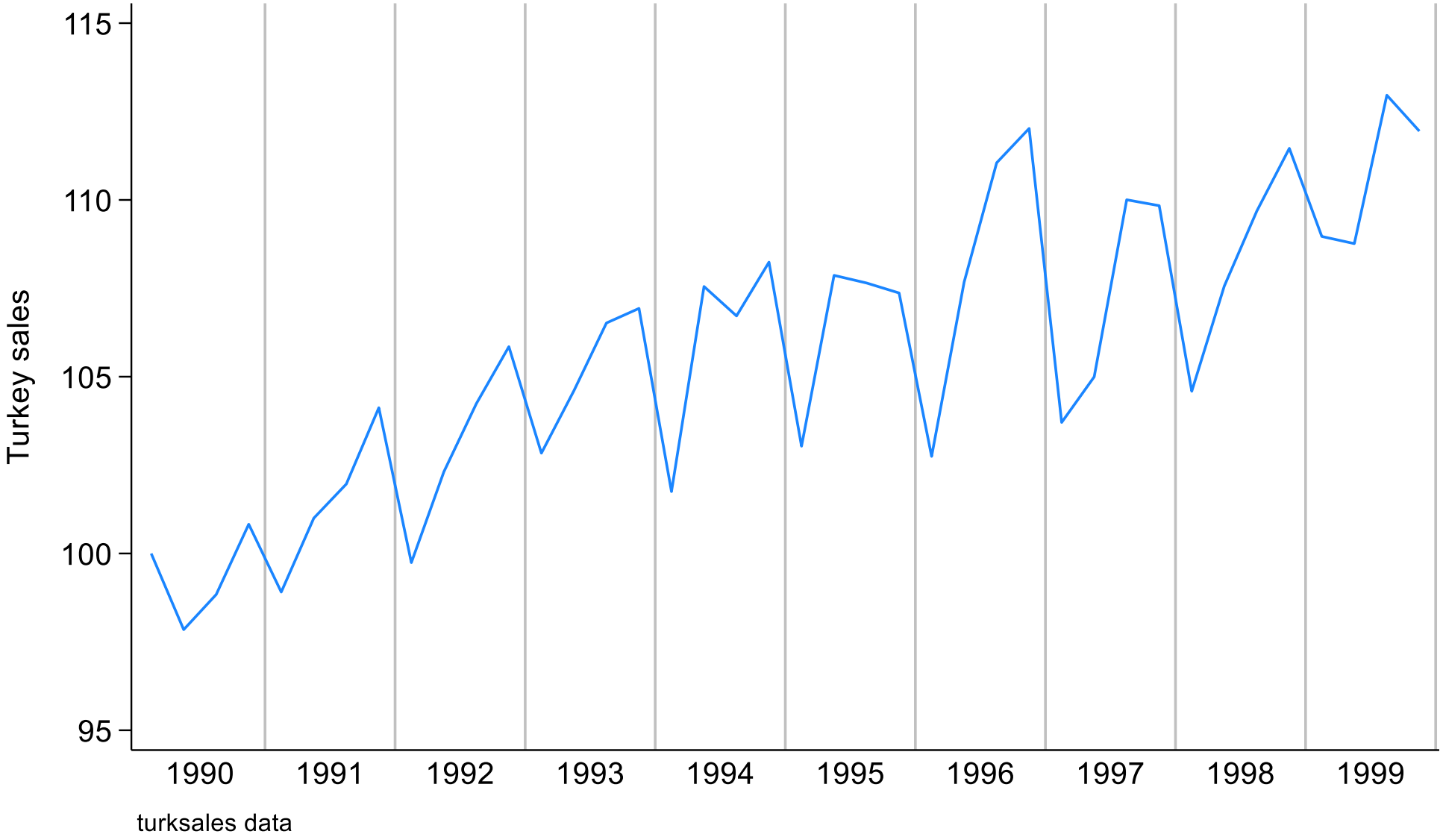
Labels without ticks can make sense

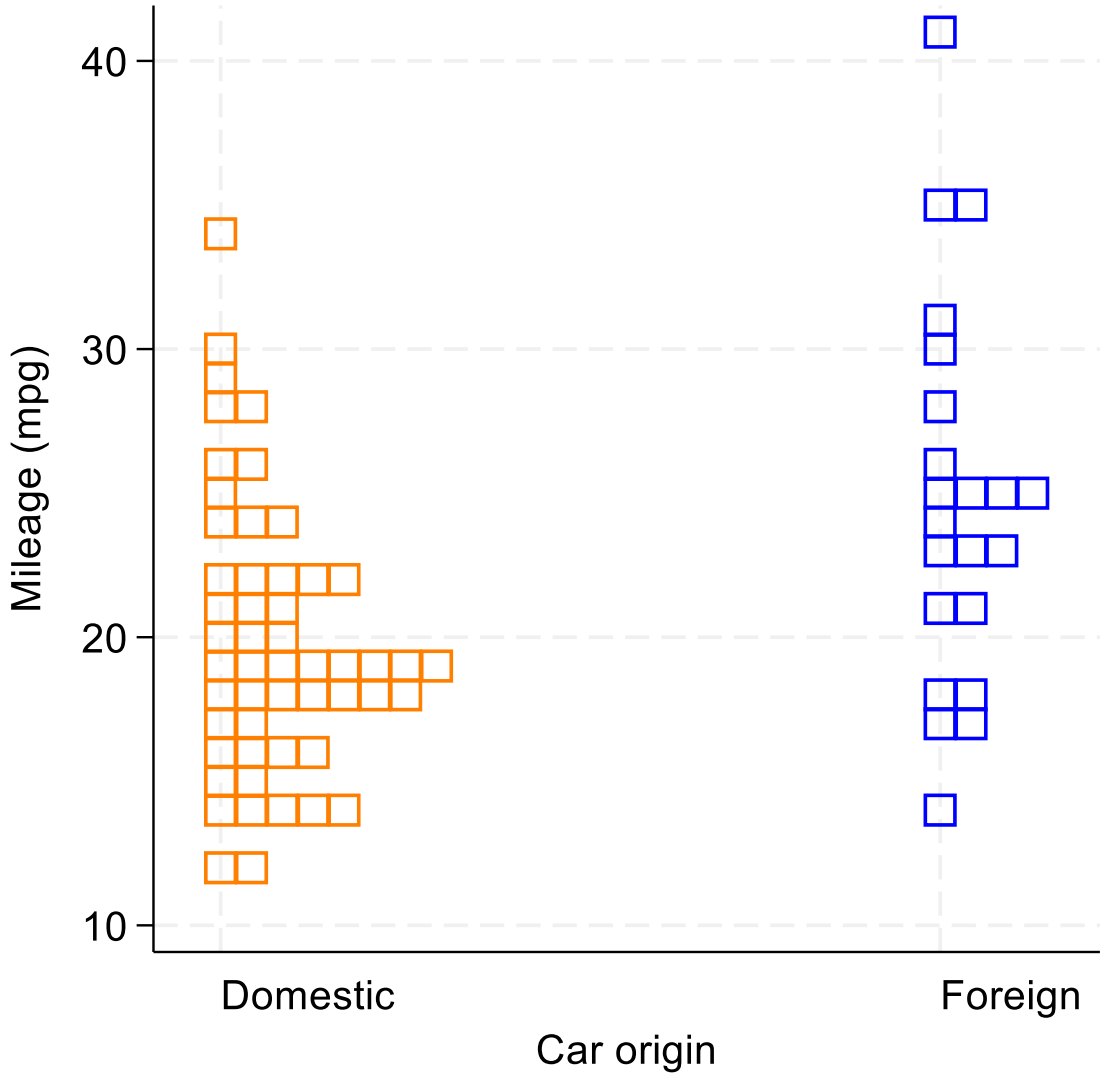
Labels without ticks are good for

◊ times placed at interval midpoints

◊ category names on one axis of a `twoway` graph.

See *SJ* 7: 590–592 (2007)
and more generally on ticks *SJ* 19: 741–747 (2019).


Turkey sales
115
110
105
100
95
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
turksales data



auto data

Use ticks or pipes in marginal rug plots

Precisely which values occur in the data?

`levelsof` can yield a list.

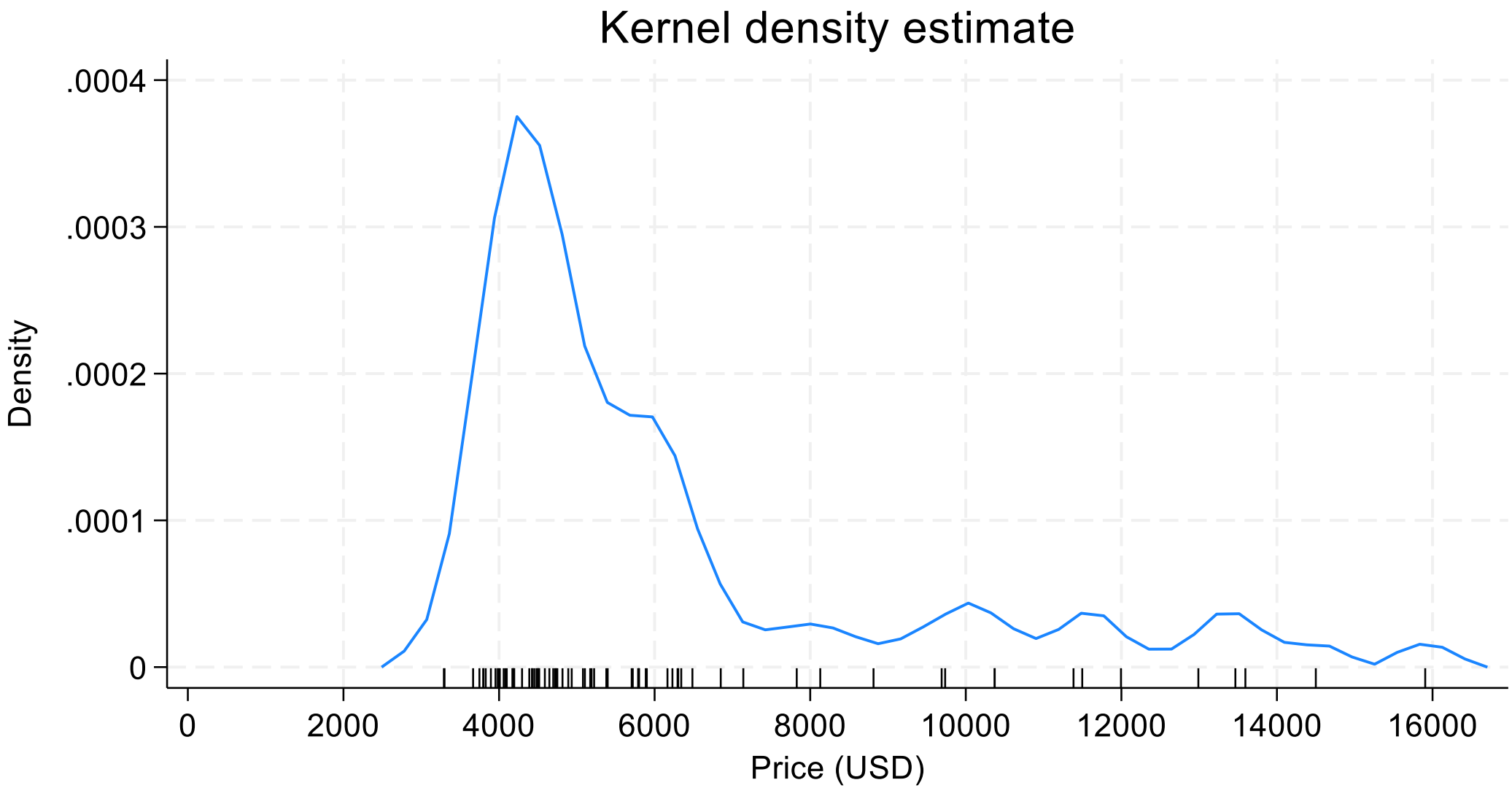
Kernel density estimates can be easier to interpret with a sight of distinct data values.

Much more on rugs at *SJ* 25: 491–497 (2025).

Rugs go back at least to

David Brunt. 1917. *The Combination of Observations*. London: Cambridge University Press.

Can you give an earlier example?



biweight kernel; width USD 800

auto data

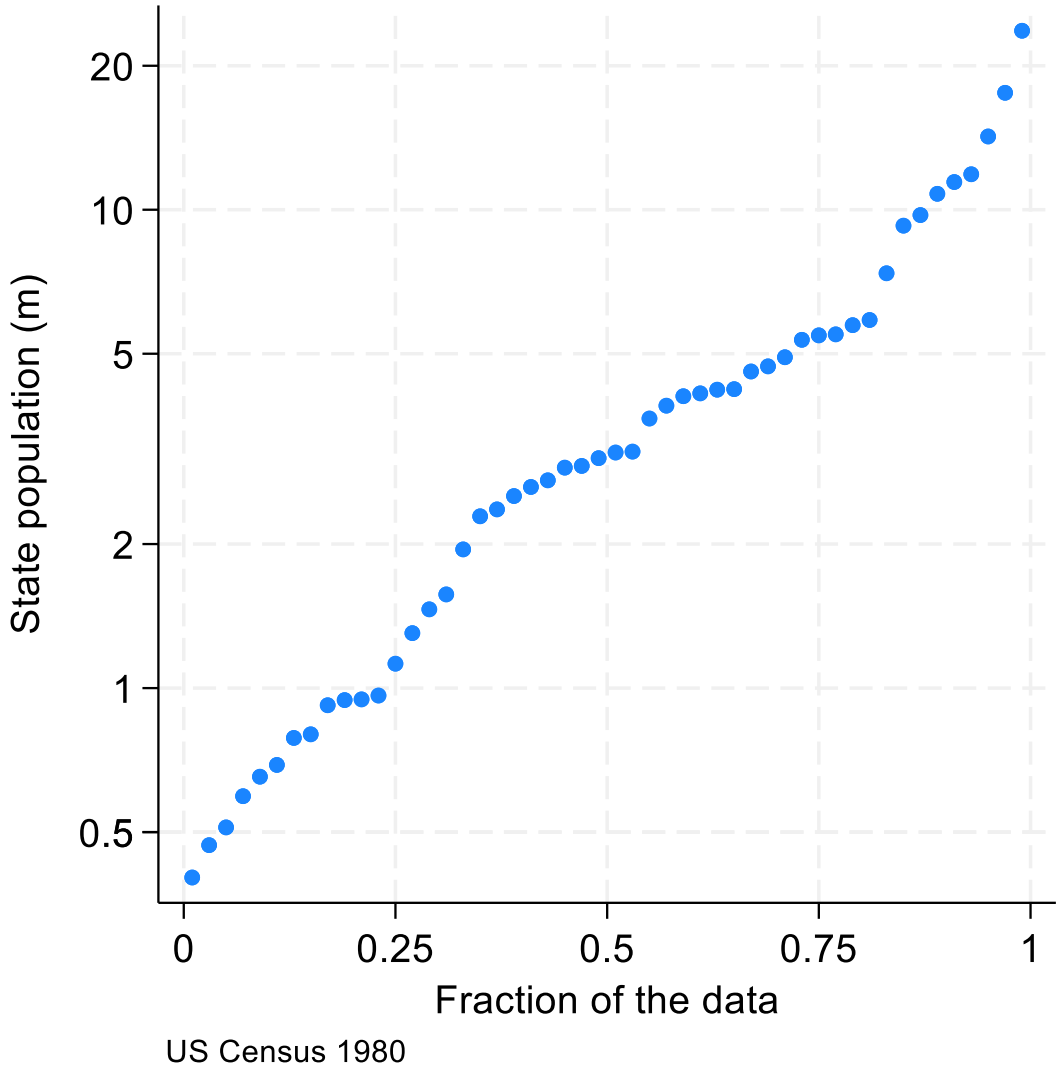
Use nice labels for logarithmic scales

With logarithmic scales, multiples of 1, 3, 10 or 1, 2, 5, 10 look good as labels.

log base 10 of 3 is 0.477.

log base 10 of 2 is 0.301; of 5 is 0.699.

See `niceloglabels` and *SJ* 18: 262–286 (2018)
(and updates 2020, 2025)



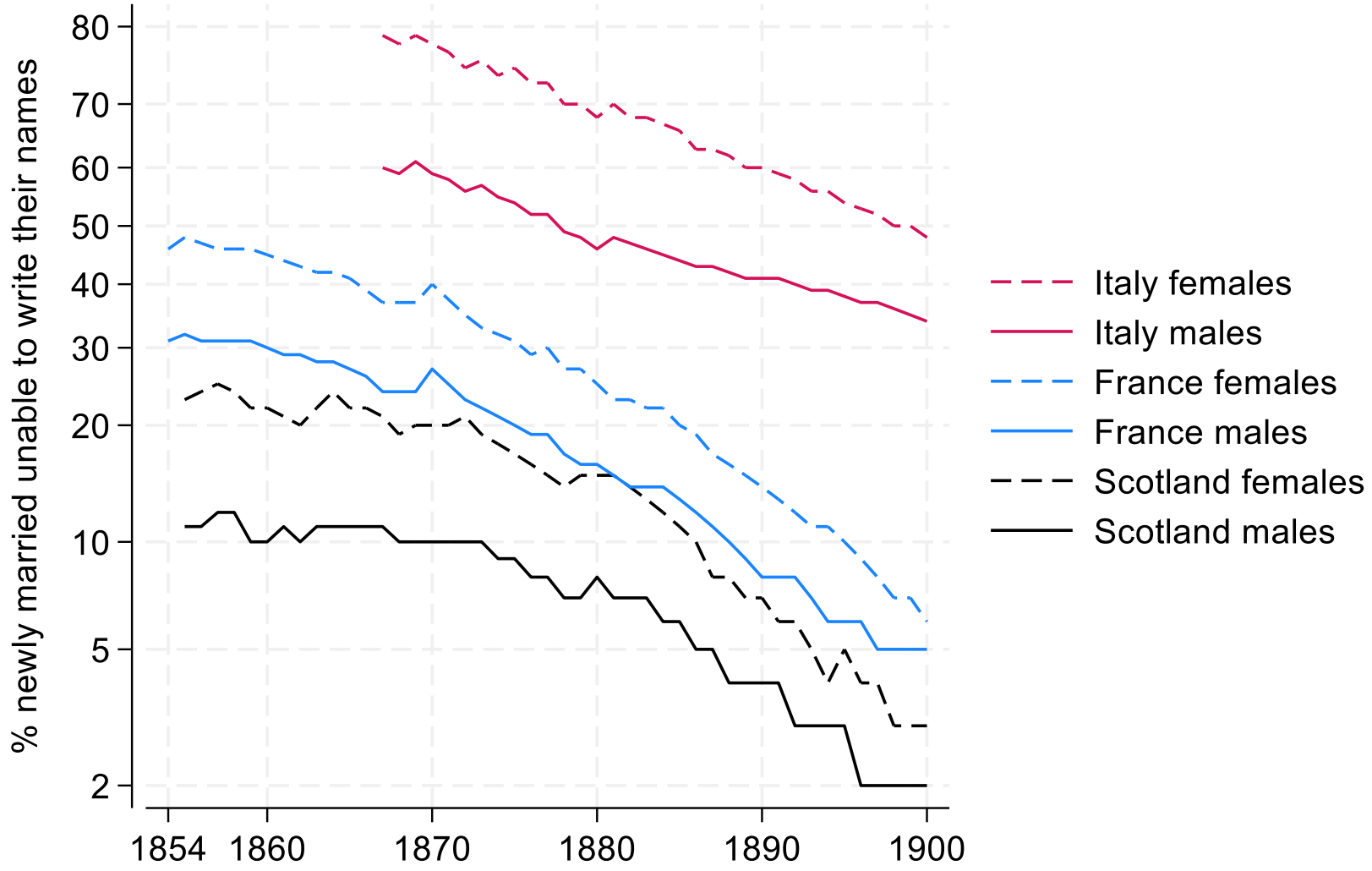
US Census 1980

Consider other nonlinear scales

Although Stata privileges only linear and logarithmic scales, any nonlinear scale may be helpful for graphs (square root, reciprocal, logit, etc.) so long as you can do the calculations and fix the axis labels.

See *SJ* 8: 142–145 (2008) and *SJ* 22: 975–995 (2022), especially the `mylabels` command.

This example uses a logit scale for illiteracy. Illiteracy has bounds 0 and 100% and something like a downward S-curve over time is often seen. Logit scale stretches the tails relative to the middle.



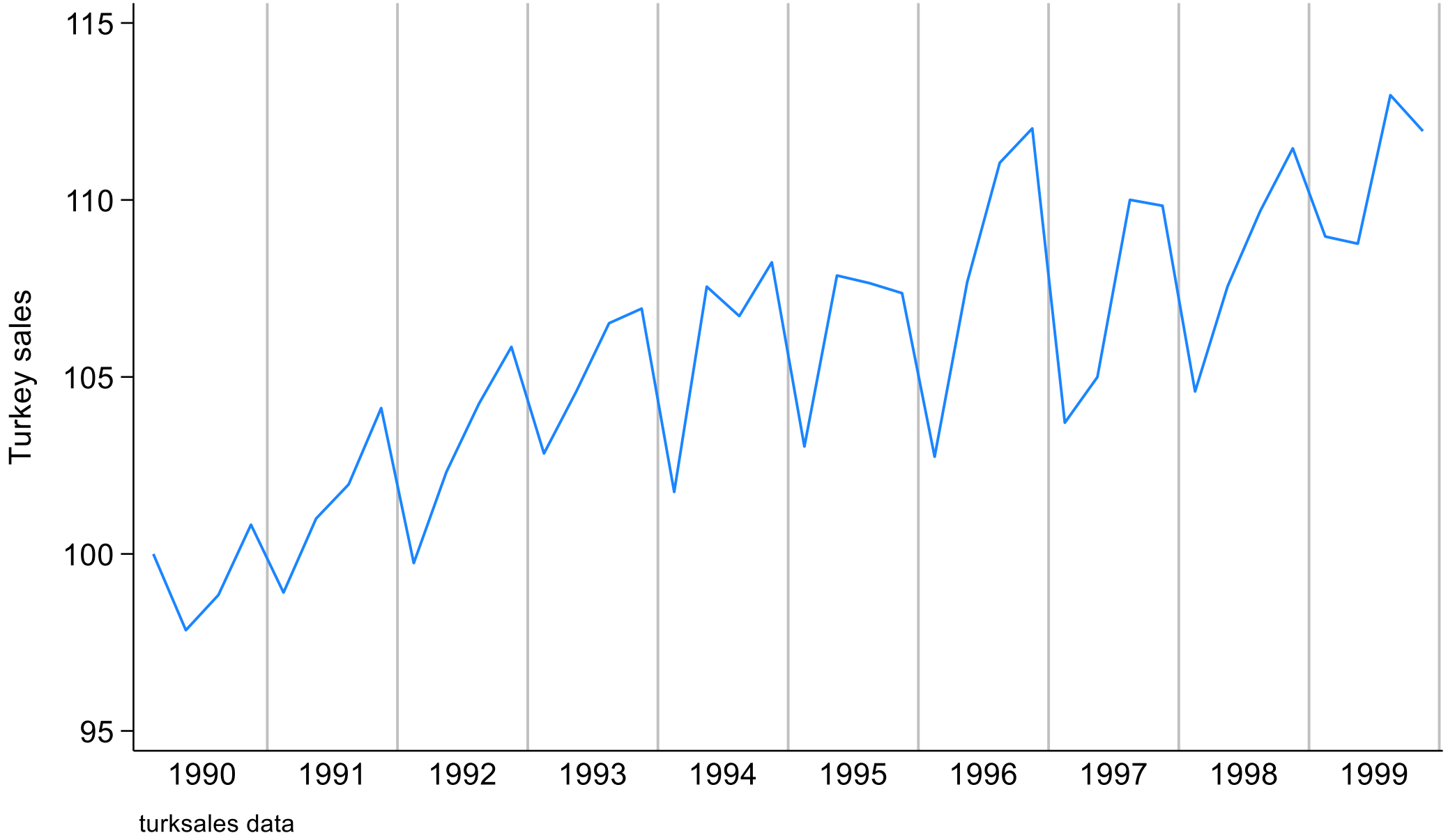
Carlo M. Cipolla. 1969. *Literacy and Development in the West.* Harmondsworth: Penguin, pp.121-125

Omit needless axis titles

Standard examples of needless titles are `Time` or `Date` or `Year`.

Who needs such titles, really?

Your teachers (should have) told you to explain each axis, showing definitions and units of measurement. They were right, except that sometimes a rule may be broken without damage.


Turkey sales
115
110
105
100
95
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
turksales data

Lose the legend! Kill the key!

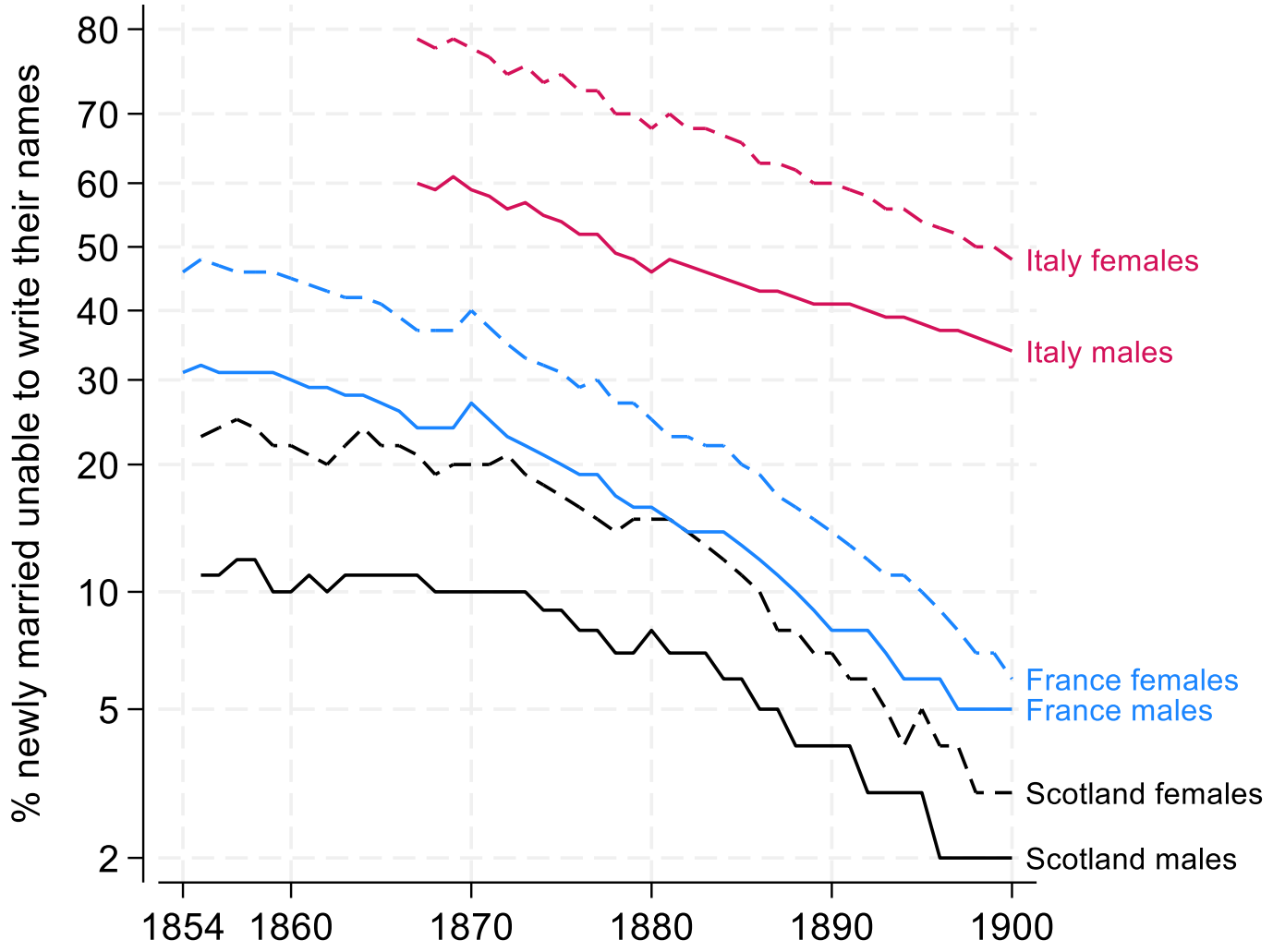
Legends are at best necessary evils: they oblige mental back and forth between legend items and data display and they can take up much valuable real estate.

Obnoxious examples are common with multiple time series, especially if a graph is tangled spaghetti.

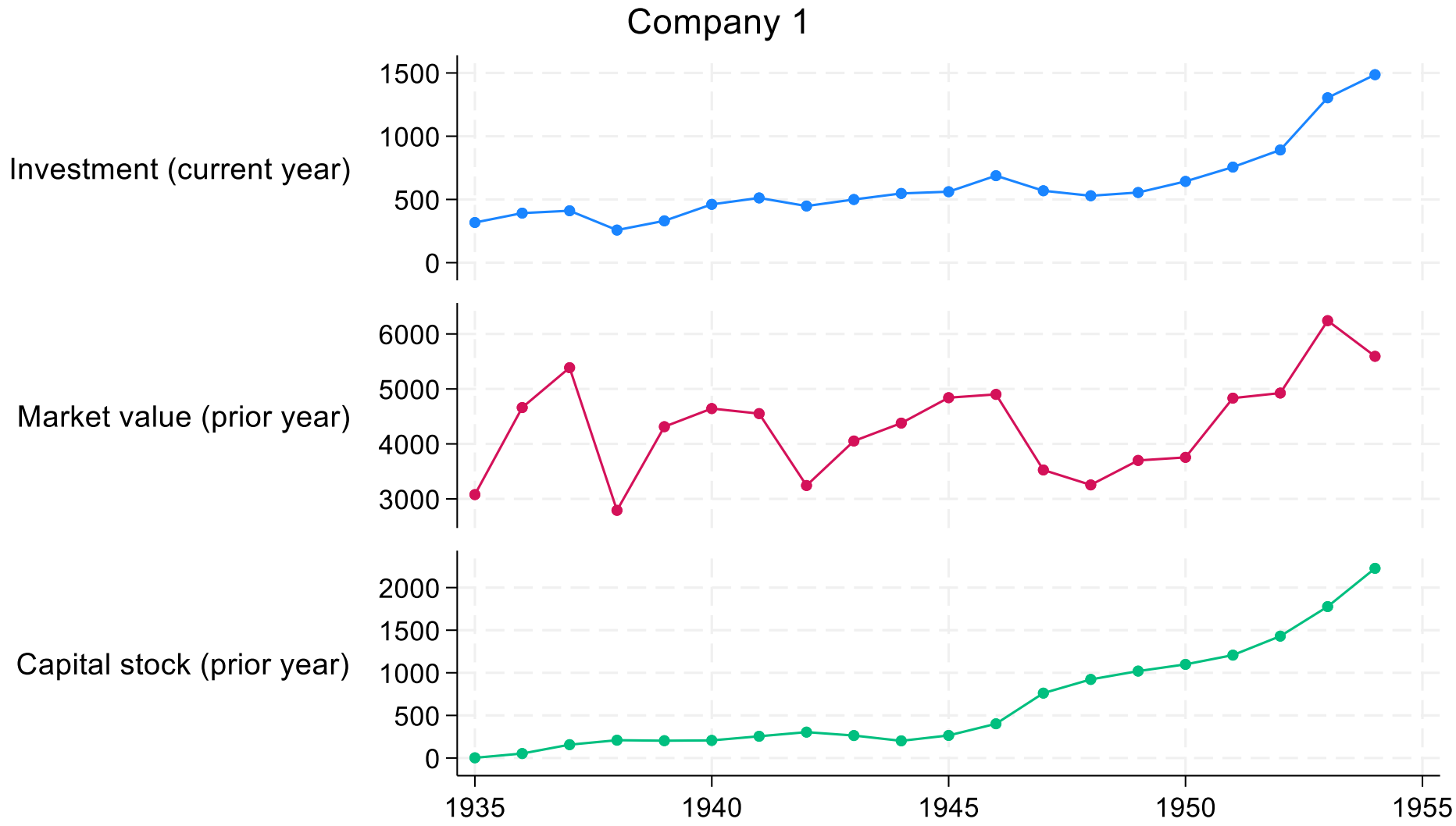
Better solutions may be direct labelling in the data region; axis labels; or panel subtitles.

You may need a reshaped data layout. See e.g. *SJ* 20: 1016–1027 (2020)

Matching text and lines or markers in colour is harder work, but usually worth the extra effort.



Carlo M. Cipolla. 1969. *Literacy and Development in the West.* Harmondsworth: Penguin, pp.121-125



Grunfeld data

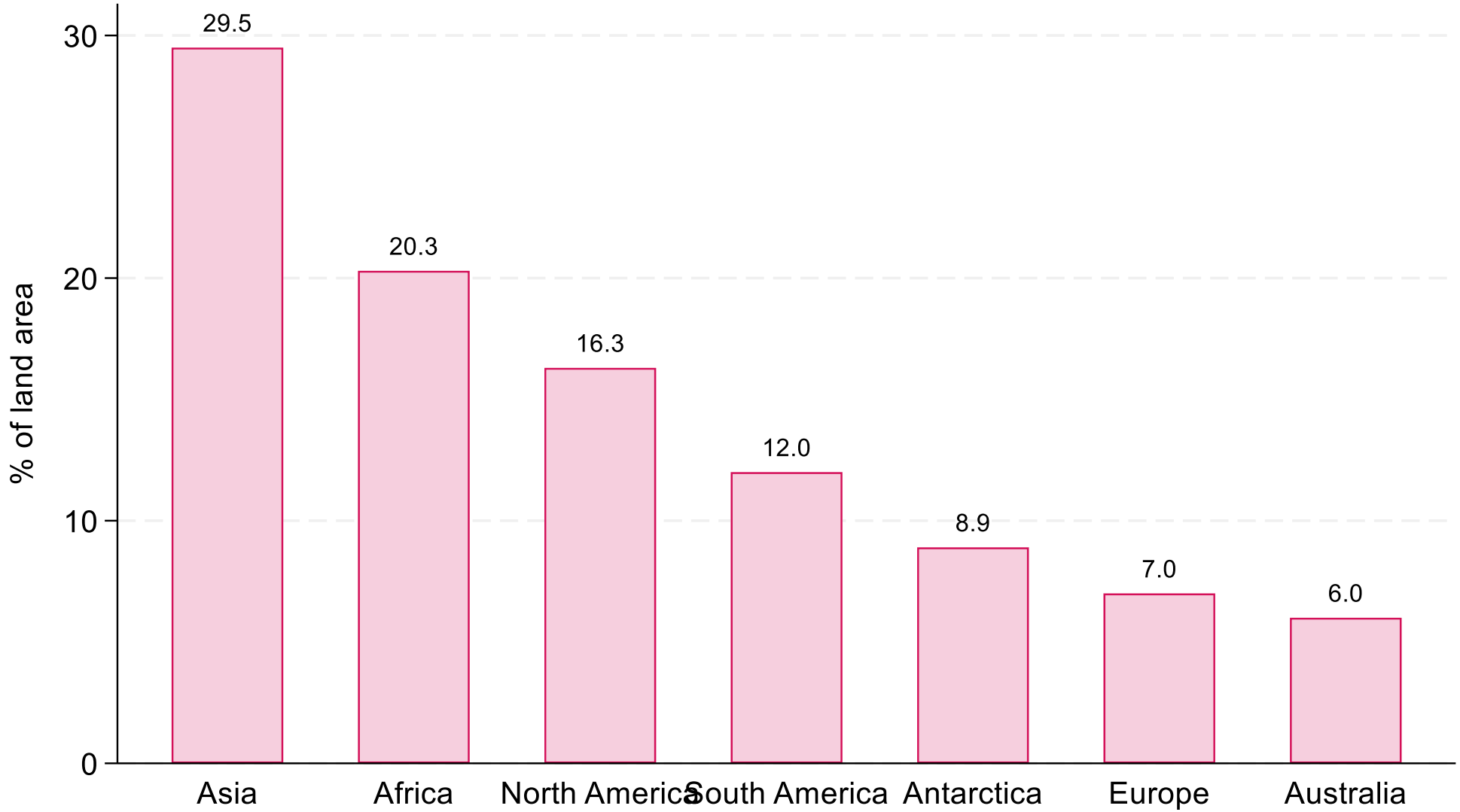
Horizontal is good

Most of us find it easier to read good-sized full horizontal text, not vertical or angled or over-reduced text or cryptic abbreviations.

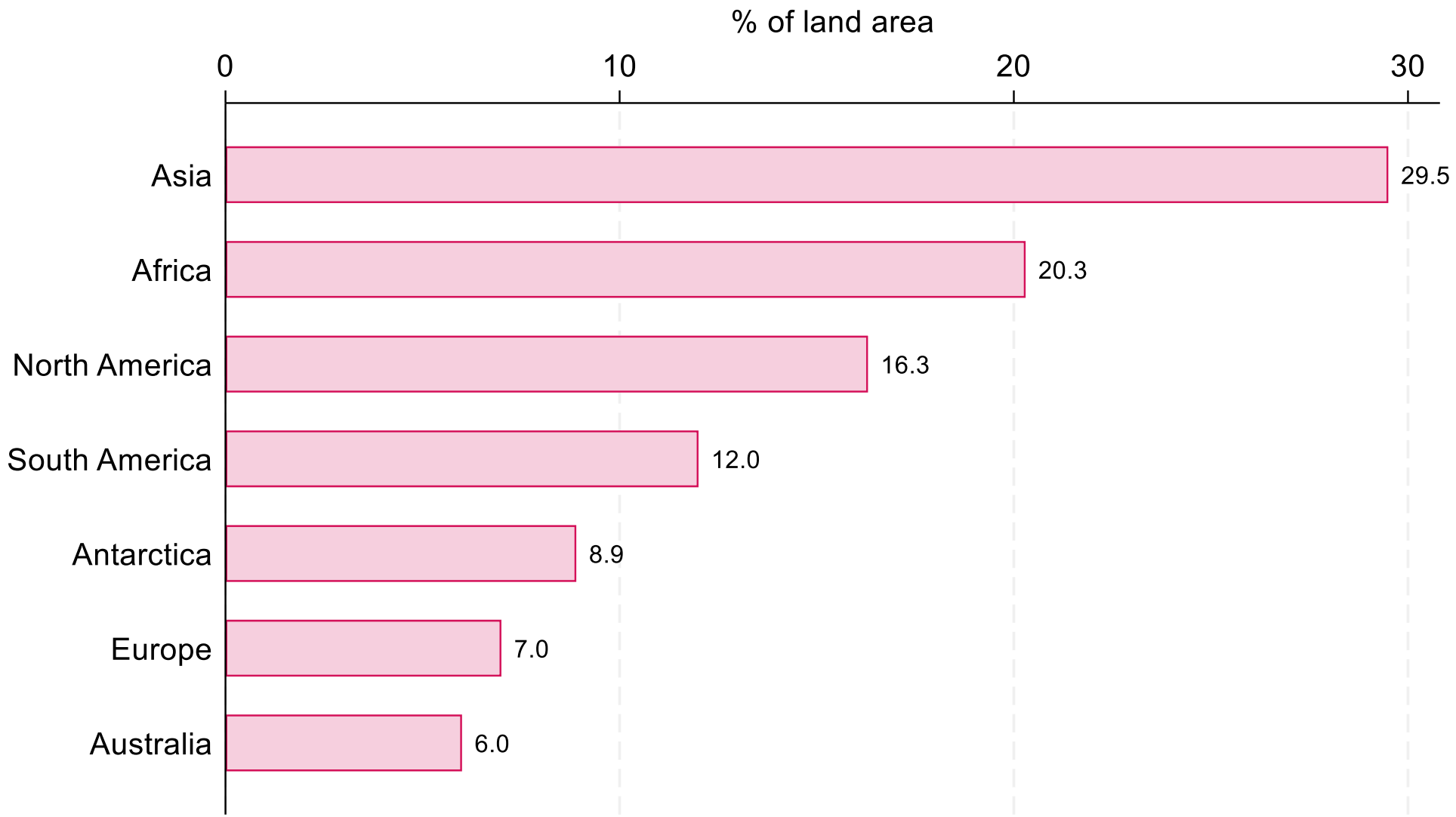
The principle that graphics must be readable may need to override a convention that responses or outcomes go on the vertical axis.

Faced with a simple problem — not enough space for category labels with a bar or column chart — the simplest solution is to go horizontal.

Incidentally, when graphs have table flavour, consider whether the horizontal axis would be better at the top. See *SJ* 12: 549–561 (2012).



2011. *World Political Reference Atlas*. Riga: Jana seta, p.11



2011. *World Political Reference Atlas*. Riga: Jana seta, p.11

Use `lpoly` for scatter plot smoothing

Stata offers several methods that are possible scatter plot smoothers.

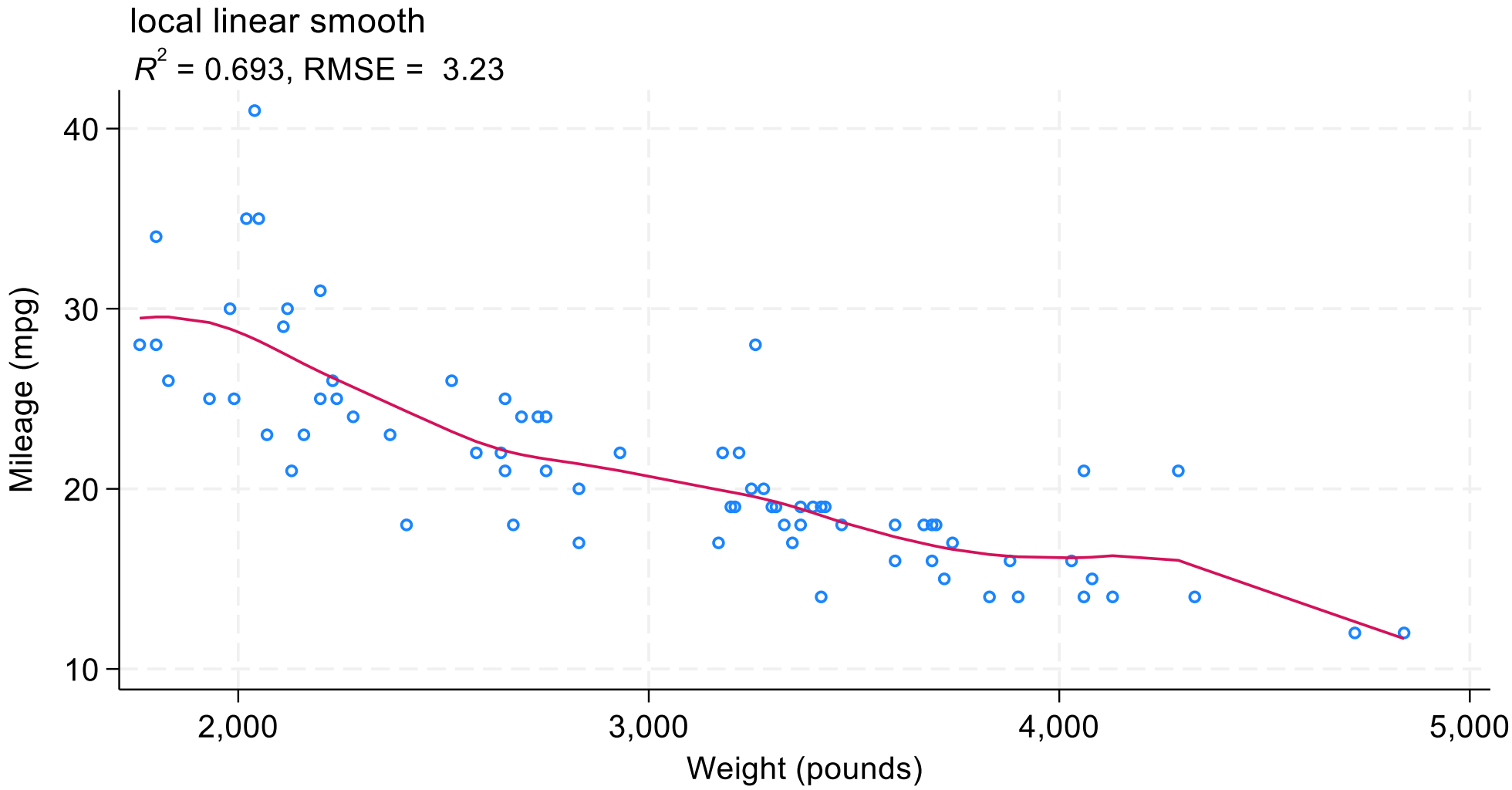
My affections have shifted over time but `lpoly` is my current favourite. It is highly flexible, fairly easy to explain, and associated with confidence interval machinery.

You may need to fight its defaults. See `localp` from SSC.

Still, it's a folk theorem that reasonable smoothers agree.

Generic advice: Always explain quite how and how much you smoothed.

If you don't, some readers will be curious or even furious.



kernel = biweight, degree = 1, bandwidth = 600

auto data

Convenient commands

Some personal favourites…

`graph dot`

`graph dot` is a greatly under-rated command. It encourages focus on comparison of values and use of nonlinear scales. It allows zeros to be omitted if they are irrelevant.

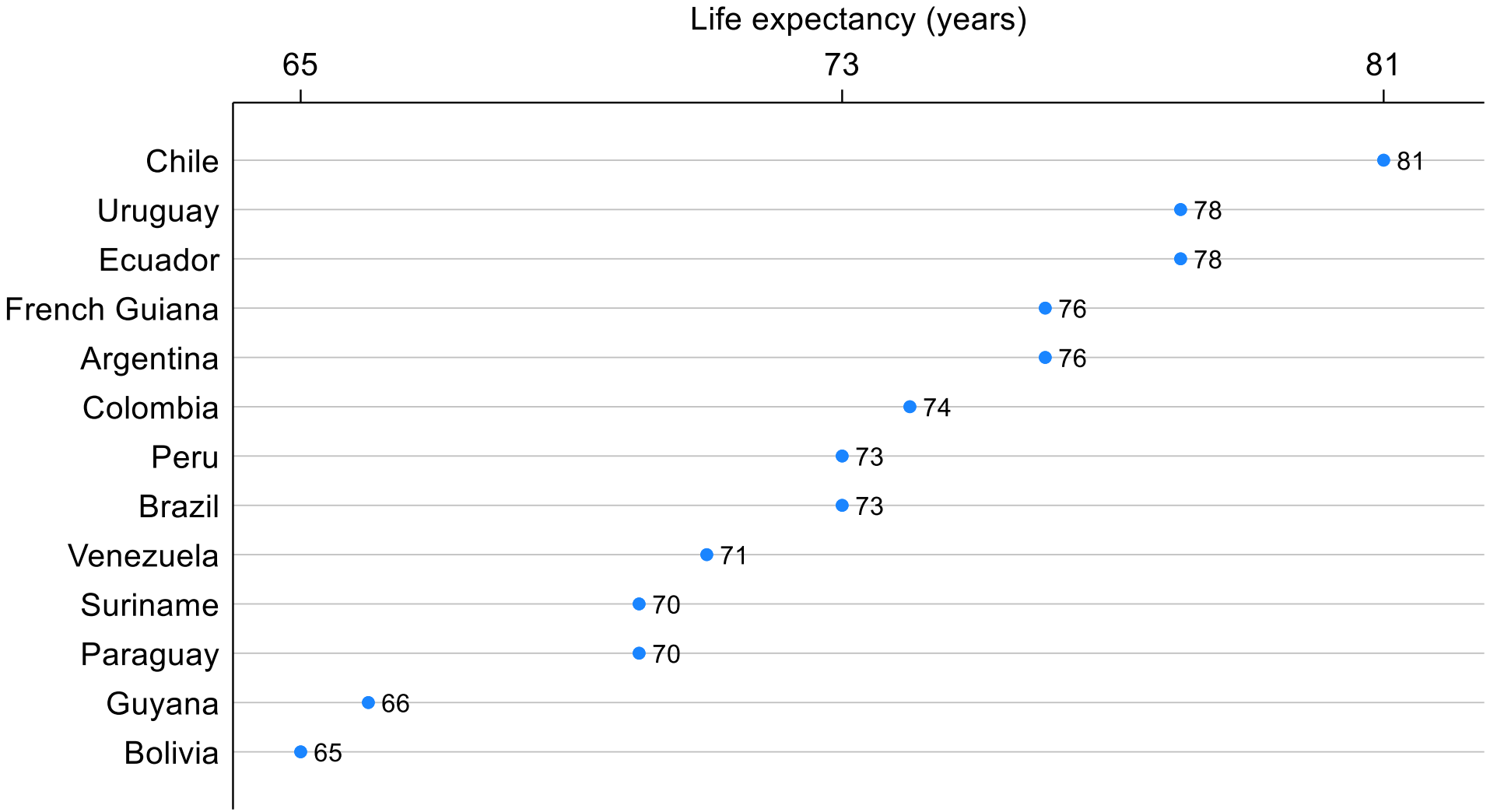
Many bar charts would work as well or better presented as dot charts.

The term *Cleveland dot chart* honours the advocacy of William S. Cleveland — e.g. *American Statistician* 38: 270–280 (1984) — but the idea can be found in George W. Snedecor's text from 1937.

Detail: I have found that the default dotted line grid can degrade on porting to other software.

I recommend options such as

```
linetype(line)
lines(lw(vthin) lc(gs12))
```

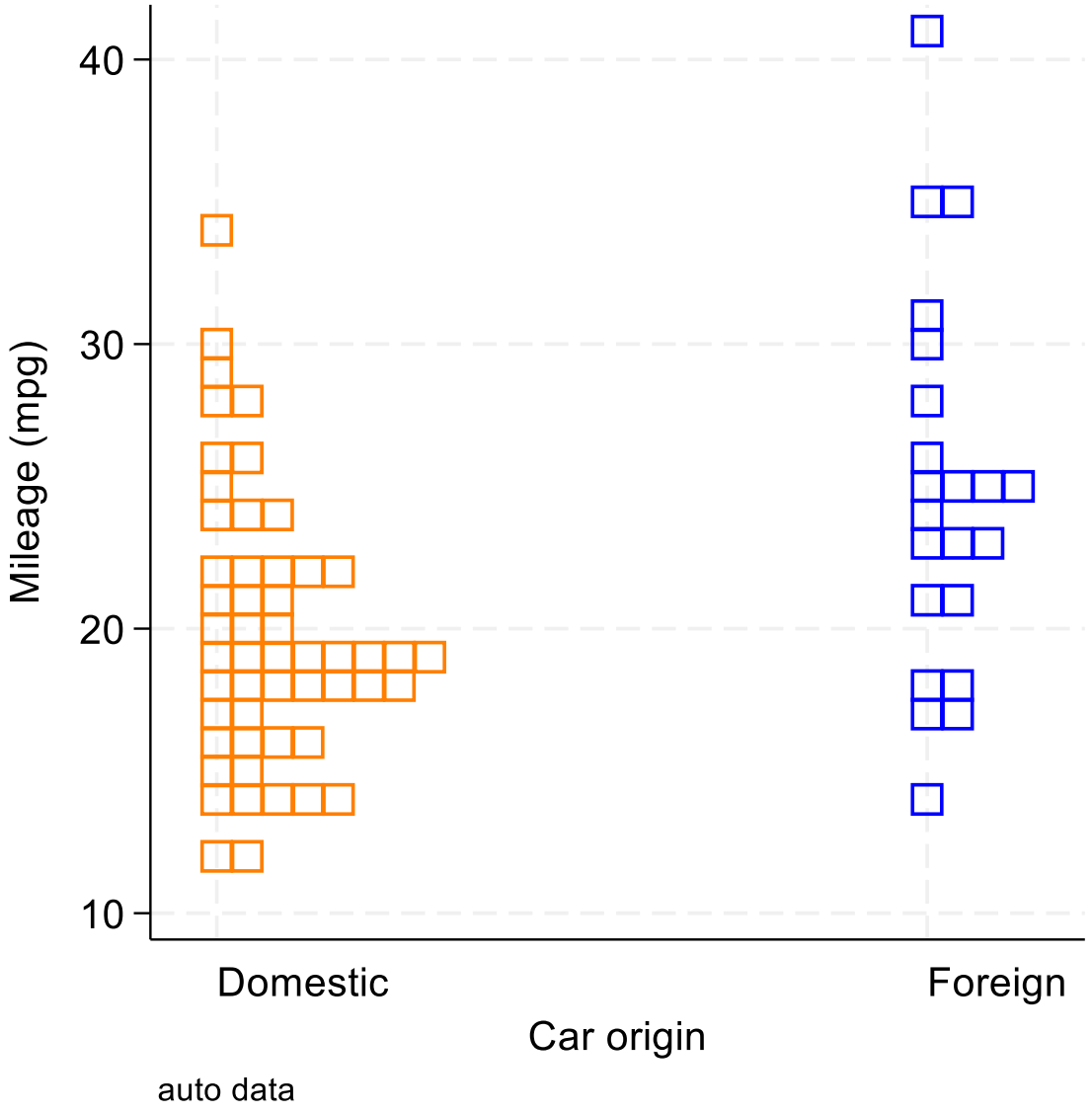


Population Reference Bureau. 2023.
World Population Data Sheet. p.11

`stripplot` for strip plots (and much more)

`stripplot` from SSC has morphed into (in effect) a superset of official command `dotplot`.

The focus is univariate distribution displays in which each data point is a separate marker.



auto data

Beyond box plots

Box plots are often poor choices. They may omit helpful detail or other pertinent summaries. They can be too easy to misinterpret, especially in the tails.

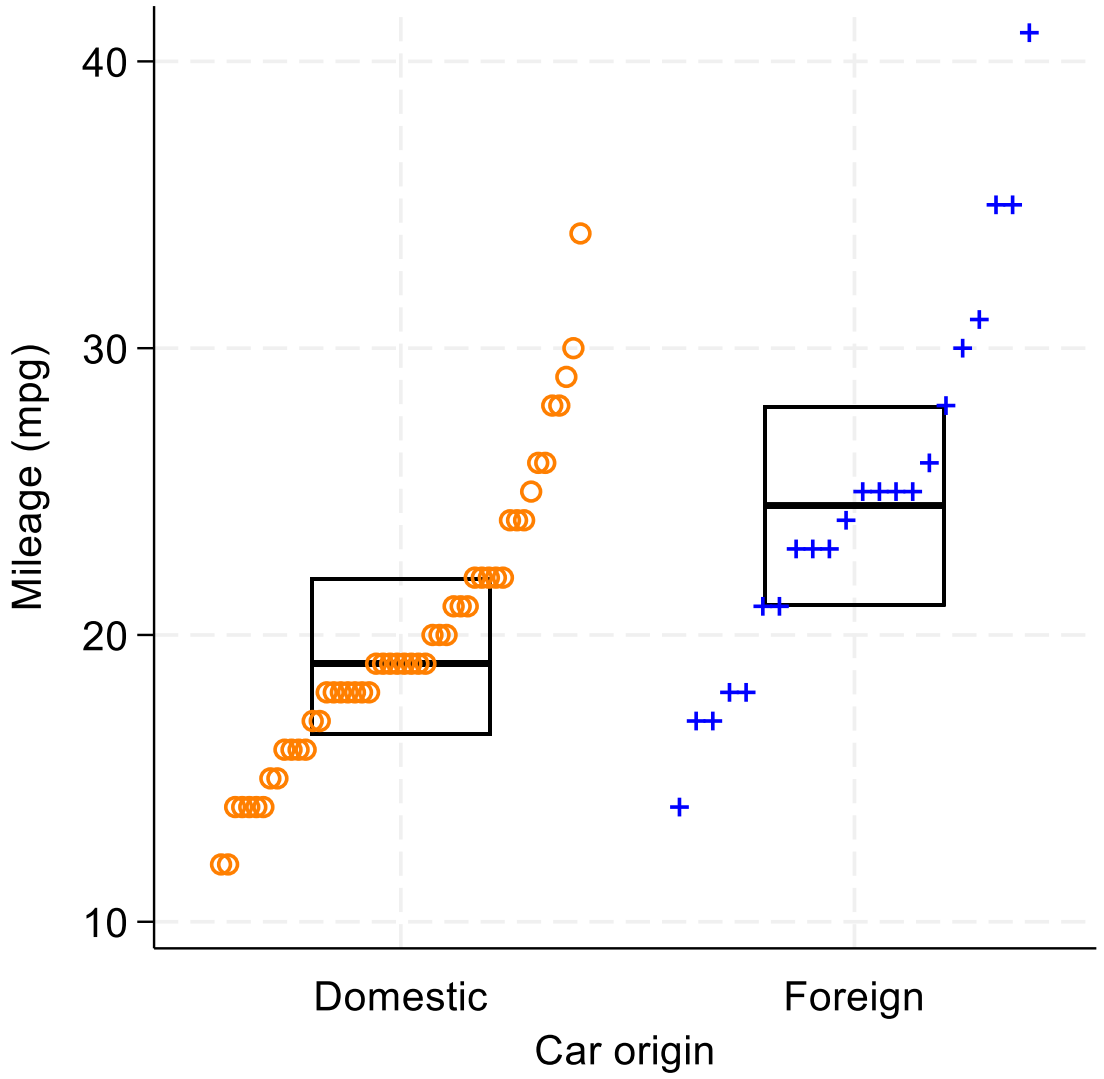
Tukey's rule of plotting individually data points if they are more than 1.5 IQR from the nearer quartile has lost whatever good rationale it had.

One alternative is a quantile-box plot that combines a quantile plot and a box plot. The term and the idea come from Emanuel Parzen (1929–2016).

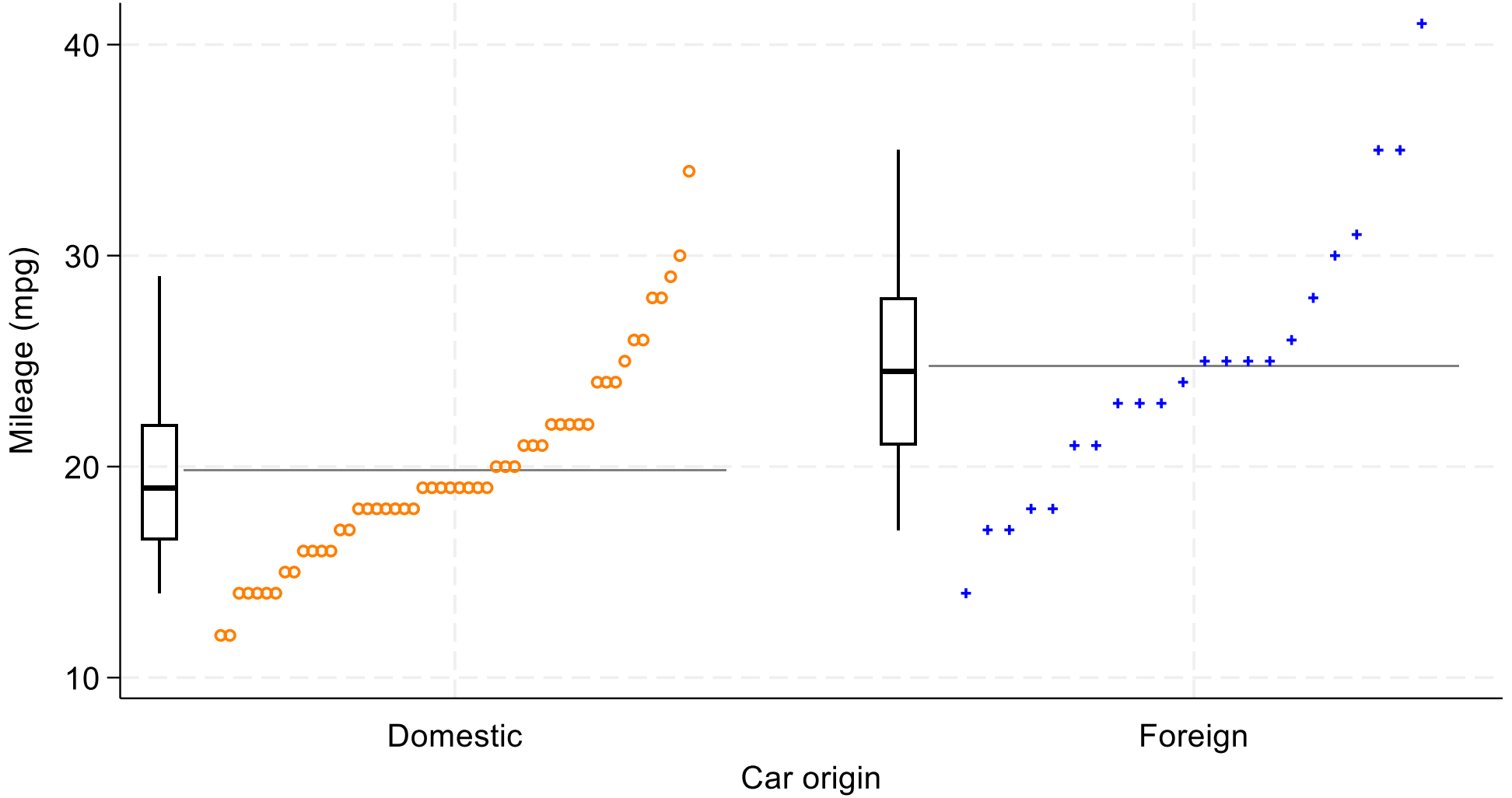
One design makes literal the simple point that one half of the data points belong *inside* the box, and so the other half belong *outside* the box.

Detail: Box plots under different names have a long history, including notably their use by geographers from 1933 on and going back at least as far as A.L. Bowley's recommendation circa 1897 to use various quantiles in distribution display.

See (again) `stripplot` from SSC.



auto data



auto data

whiskers to 5 and 95% points; longer lines show means

I suggest that these displays are easier to think about than the puzzlingly fashionable superimposition of jittered dots on box plots.

See also `qbplot` from SSC for a pedagogic or propaganda version.

A paper in progress might even appear in *SJ* 26(3).

As Nero Wolfe said in a different context

Tradition should be respected but not sanctified.

Rex Stout. 1975. *A Family Affair*. New York: Viking Press. Ch.3

Tom Sawyer, or rather Mark Twain, was there earlier.

Often, the less there is to justify a traditional custom, the harder it is to get rid of it.

Mark Twain. 1876. *The Adventures of Tom Sawyer*. Hartford, CT: American Publishing Company. Ch.5.

`qplot` (and `qqplotg`) for quantile plots

The main idea of quantile plots is just to plot ordered values against a version of cumulative probability. Using plotting positions such as (rank – 1/2) / sample size allows easy transformations. Normal quantile plots are just the most familiar examples for many people.

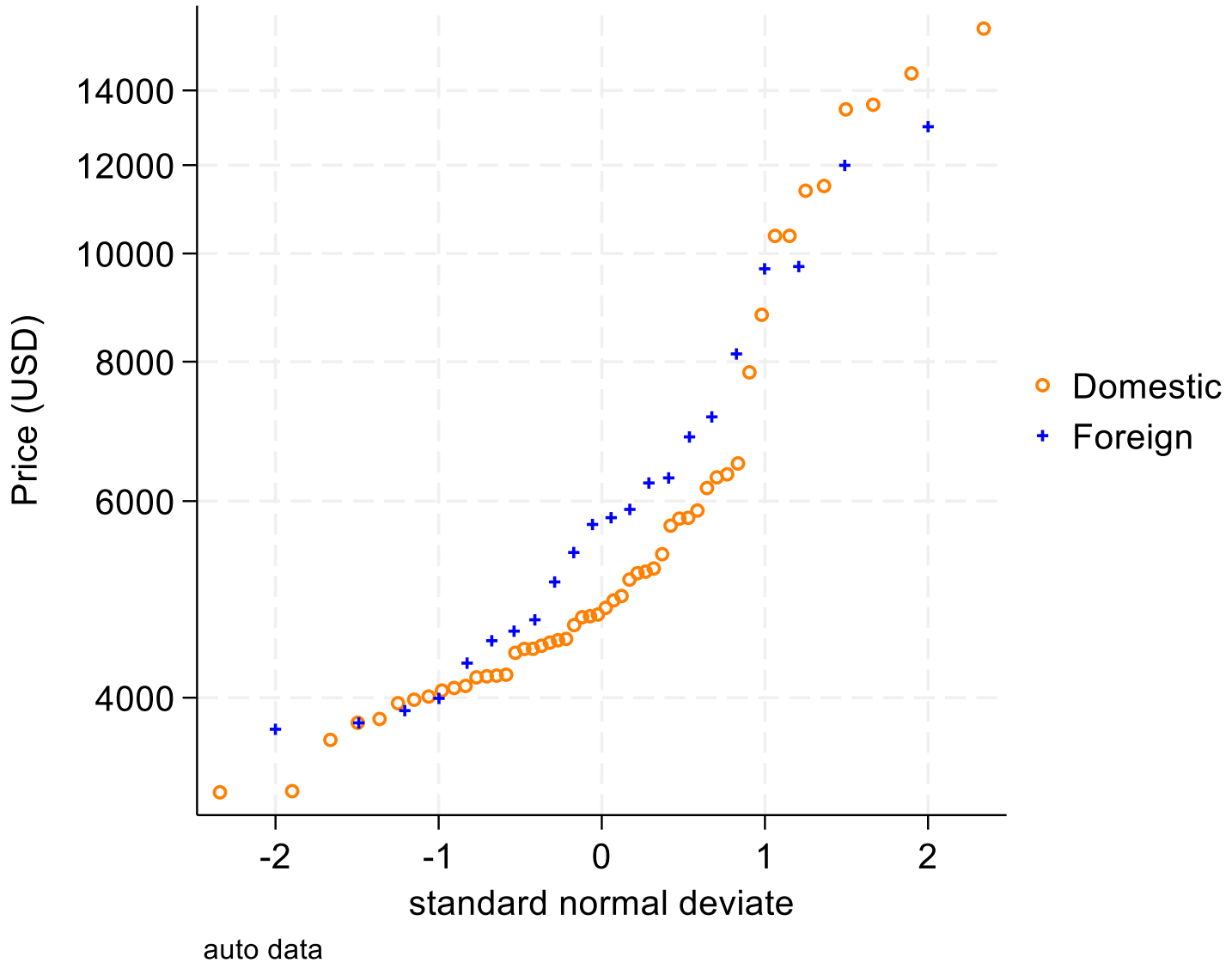
`qplot` and `qqplotg` from the *Stata Journal* are in effect supersets of official commands `quantile` and `qqplot`.

See

SJ 24: 514–534 (2024)

SJ 7: 275–279 (2007)

SJ 5: 442–460 (2005)


14000
12000
10000
8000
6000
4000
Price (USD)
-2
-1
0
1
2
standard normal deviate
Domestic
Foreign
auto data

`fabplot` for front-and-back plots

Front-and-back plots (my name, but not my idea) display each subset in turn in front with the other subsets as backdrop. The idea combines superimposed and juxtaposed plotting, with (sometimes) the benefits of both styles.

See *SJ* 21: 539–554 (2021)

SJ 19: 989–1008 (2019)

SJ 10: 670–681 (2010)

SJ 9: 499–503 (2009).

Some people struggle with multiple markers or line patterns or colours and a legend for each item. That battle is often lost before it starts.

The following example with 20 countries as panels is suggested to be of a size where different colours, marker symbols or line patterns would not work well. Any of those would require a legend, which itself would take up space.


Paraguay
Bolivia
Uruguay
Argentina
Venezuela
Brazil
Nicaragua
Peru
Ecuador
Colombia
Panama
Mexico
Dominican Republic
Cuba
Honduras
Chile
Guatemala
El Salvador
Costa Rica
Haiti
Greenhouse gas emissions (tonnes per head)
30
10
3
1
1990
2000
2010
2020

Florence Nightingale's data on deaths in Crimean War

Florence Nightingale (1820–1910) had a strongly statistical approach to medical and other questions among other outstanding contributions.

She reported on deaths of British soldiers in the Crimean War, recording numbers due to zymotic disease (loosely, infections), wounds and injuries, and all other causes.

Her circular graph is much repeated and lauded. How does a front-and-back compare?

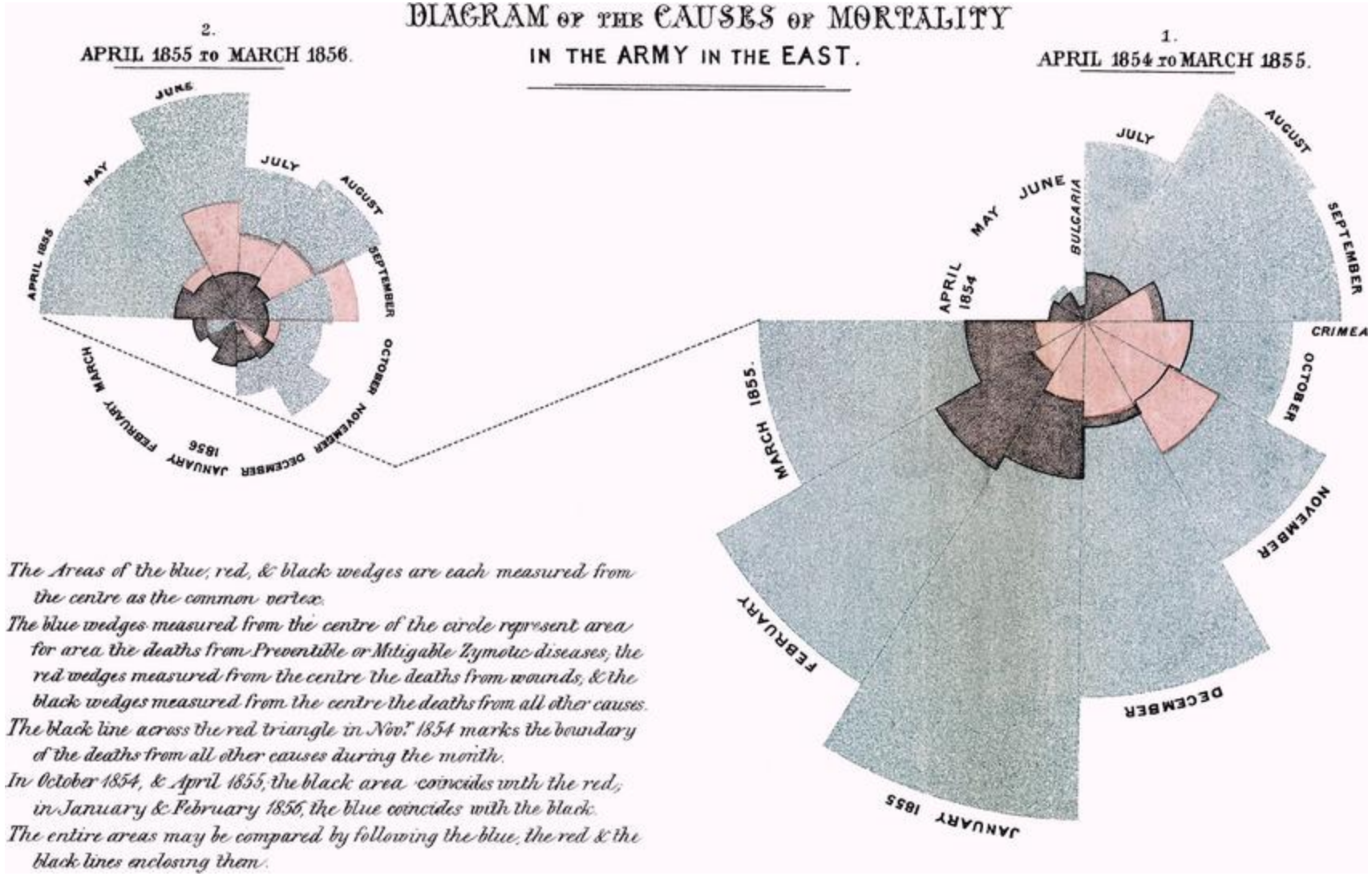
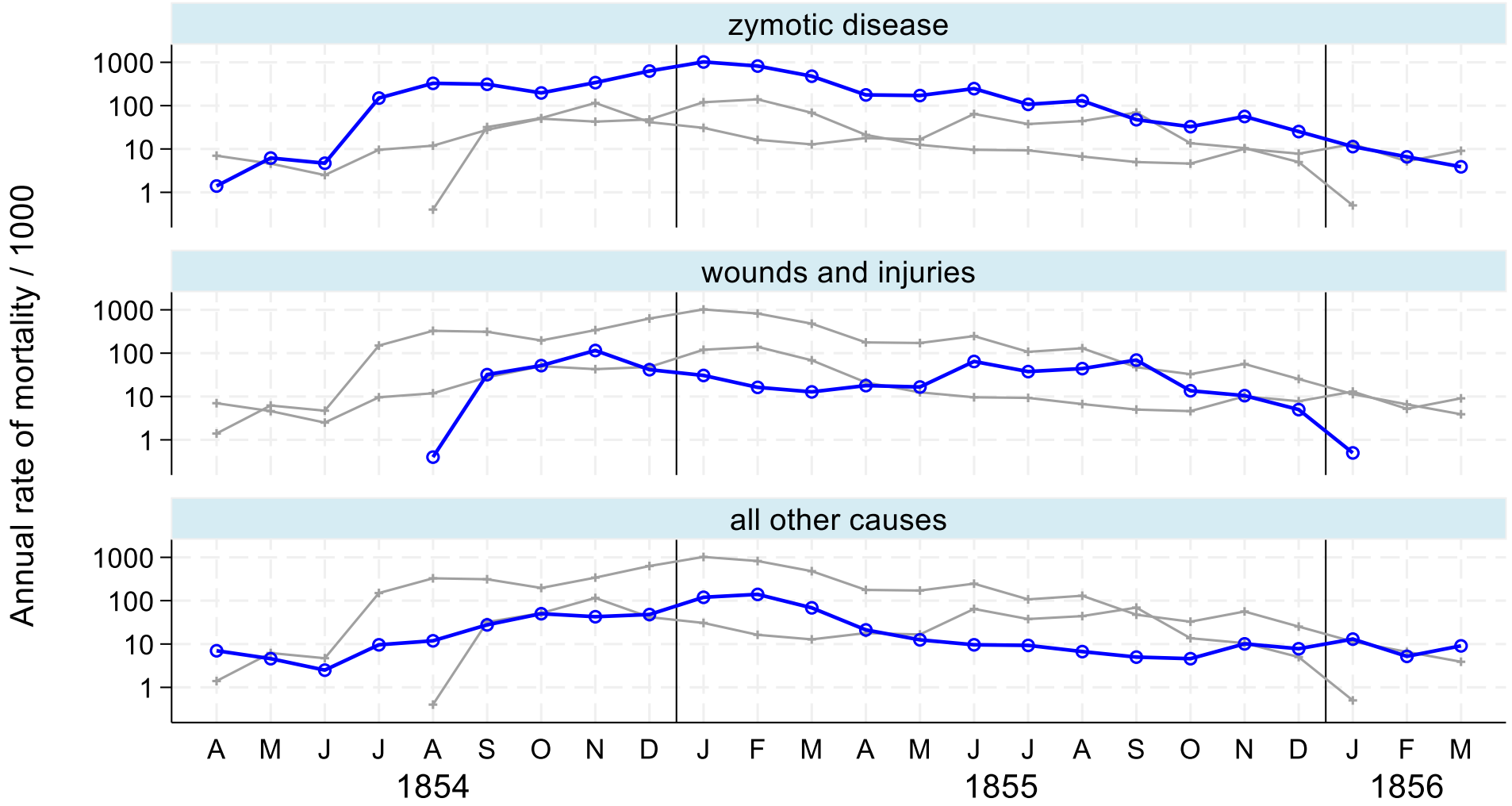


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.
2.
APRIL 1855 TO MARCH 1856.
1.
APRIL 1854 TO MARCH 1855.
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRUARY
MARCH
APRIL 1855
1856
MAY
APRIL 1854
MARCH 1855.
BULGARIA
CRIMEA
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

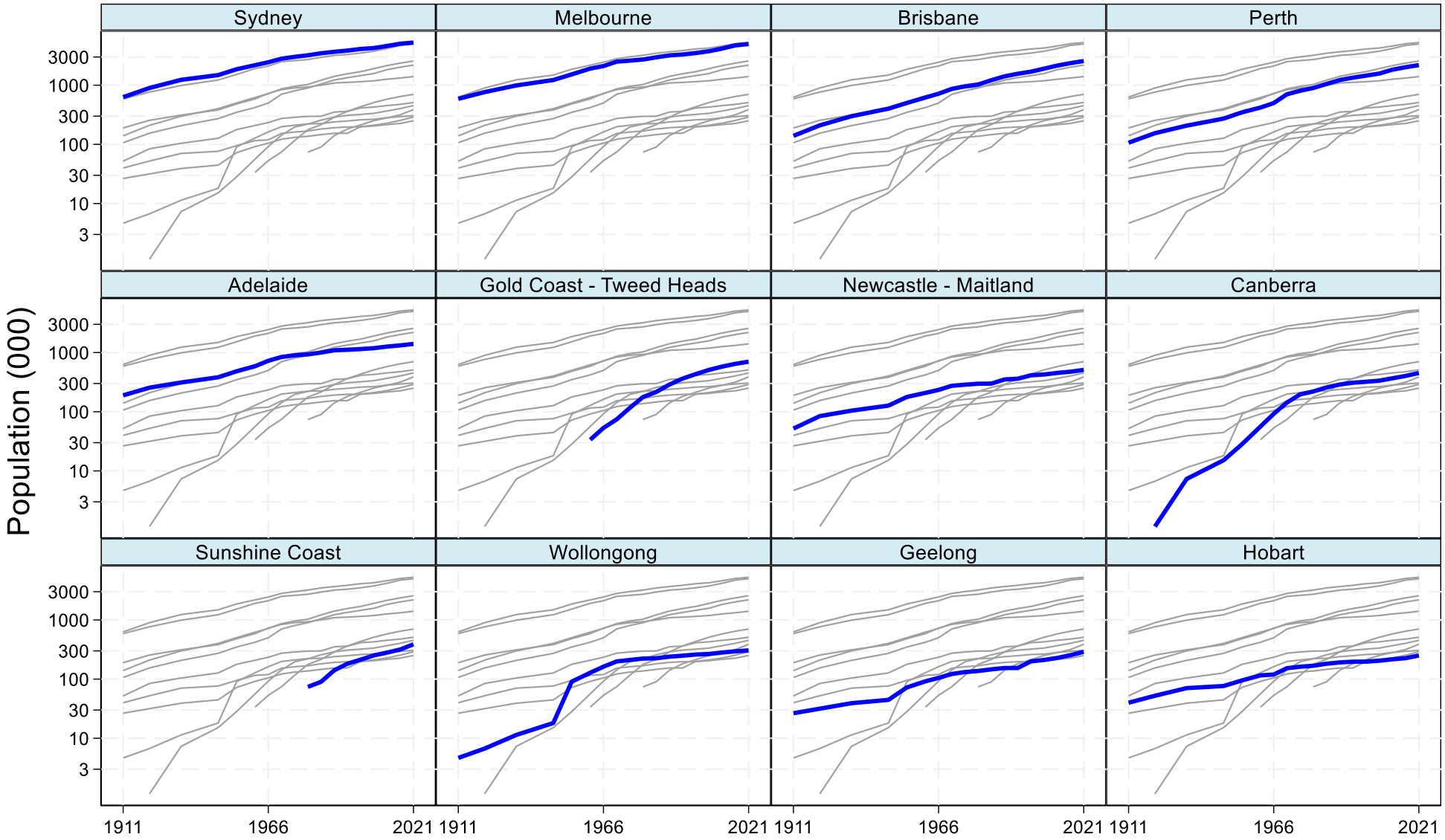


Florence Nightingale's data
https://understandinguncertainty.org/node/214

Australian city populations

My suggestions:

- ◊ Logarithmic scale is needed to shrink and stretch the support
- ◊ Name order (Adelaide first!) is not needed or helpful
- ◊ Each city should be plotted allowing comparison with others


Sydney
Melbourne
Brisbane
Perth
Adelaide
Gold Coast - Tweed Heads
Newcastle - Maitland
Canberra
Sunshine Coast
Wollongong
Geelong
Hobart
Population (000)
3000
1000
300
100
30
10
3
1911
1966
2021

`tabplot` for multiway bar charts

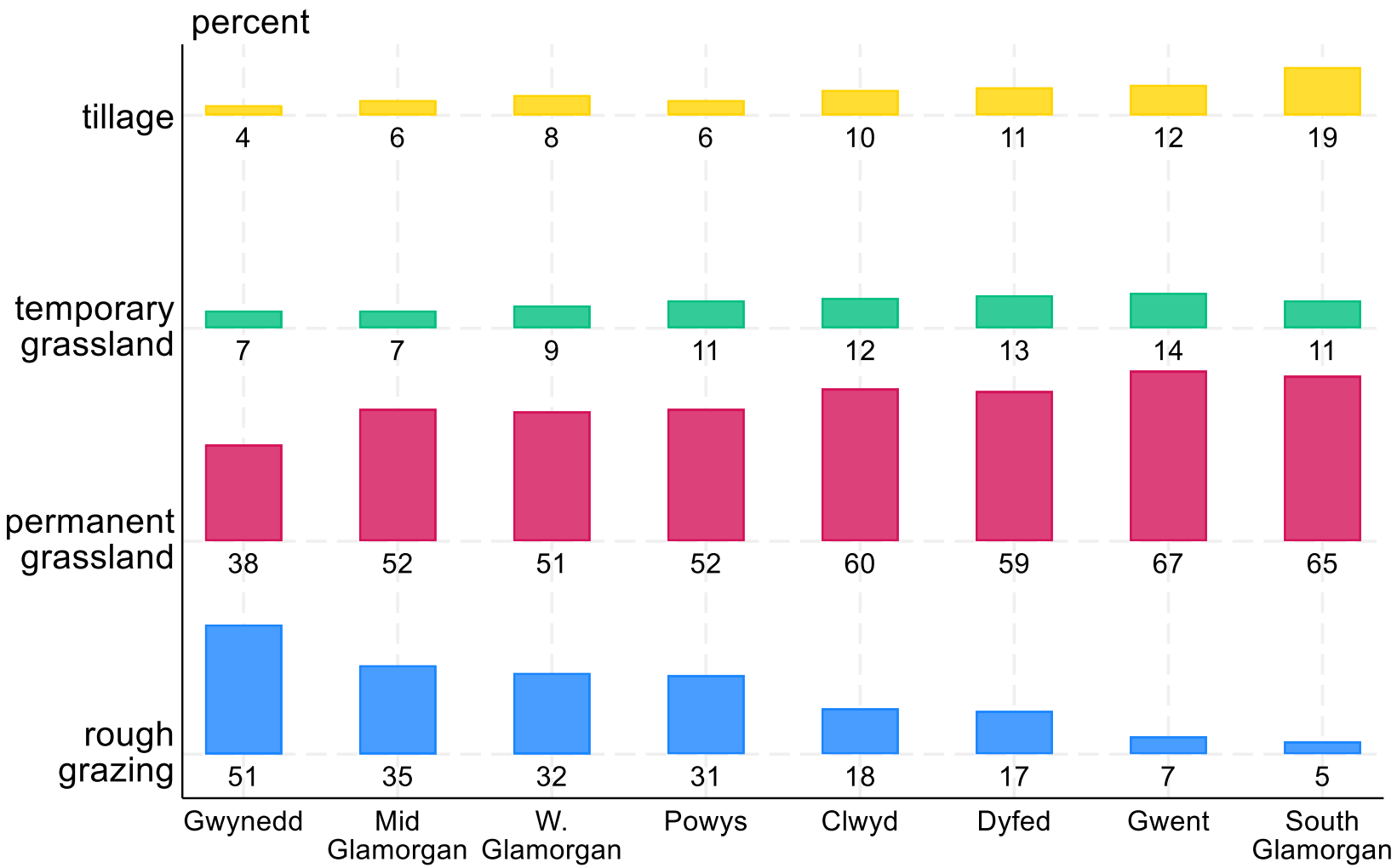
Stacked or divided bars are easy to understand in principle, but often not so effective in practice.

Pulling the bars apart in a table-like display lets zeros and small amounts or counts be visible as such and allows annotation, thus hybridizing graph and table ideas.

See *SJ* 16: 491–510 (2016) and later Software Updates.

Usually a legend can be omitted, as the axis labels explain categories or variables.

The earliest example I know comes from Charles Booth's team studying labour and life in London in 1889.



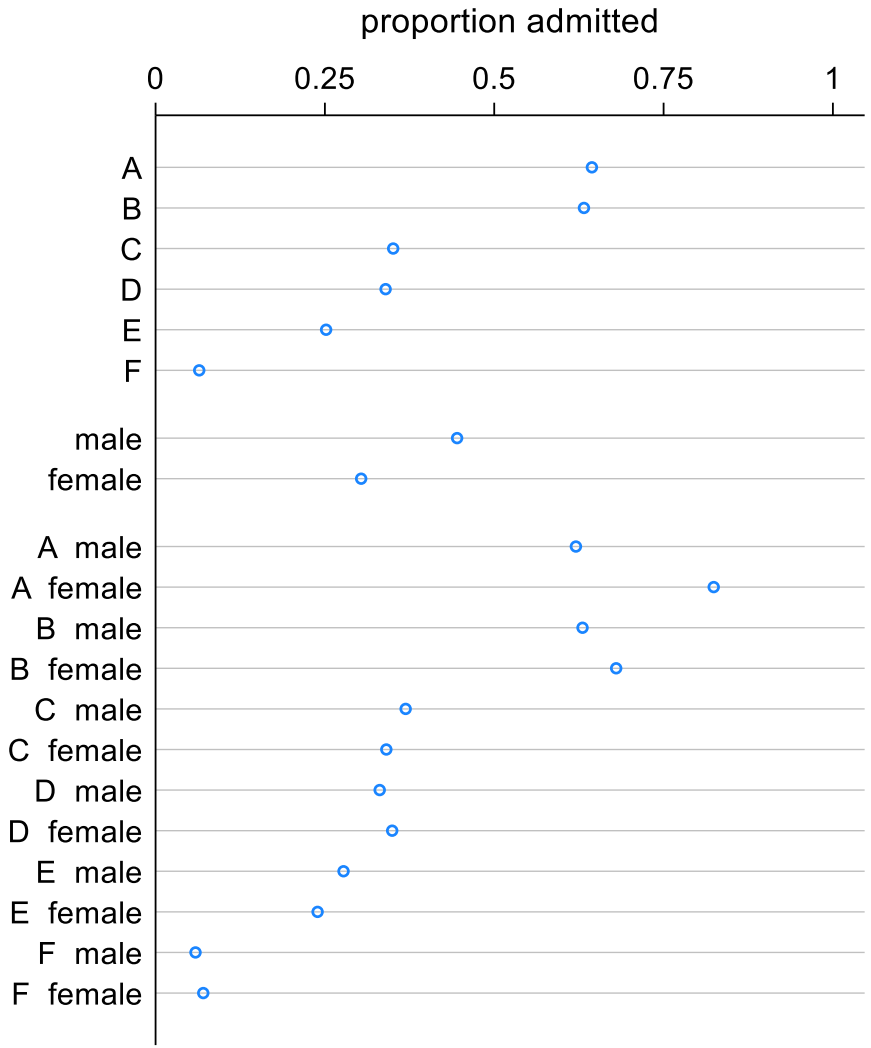
Rural land use 1978
John W. Aitchison. 1981. Triangles, tetrahedra and taxonomy. *Area* 13: 137-143

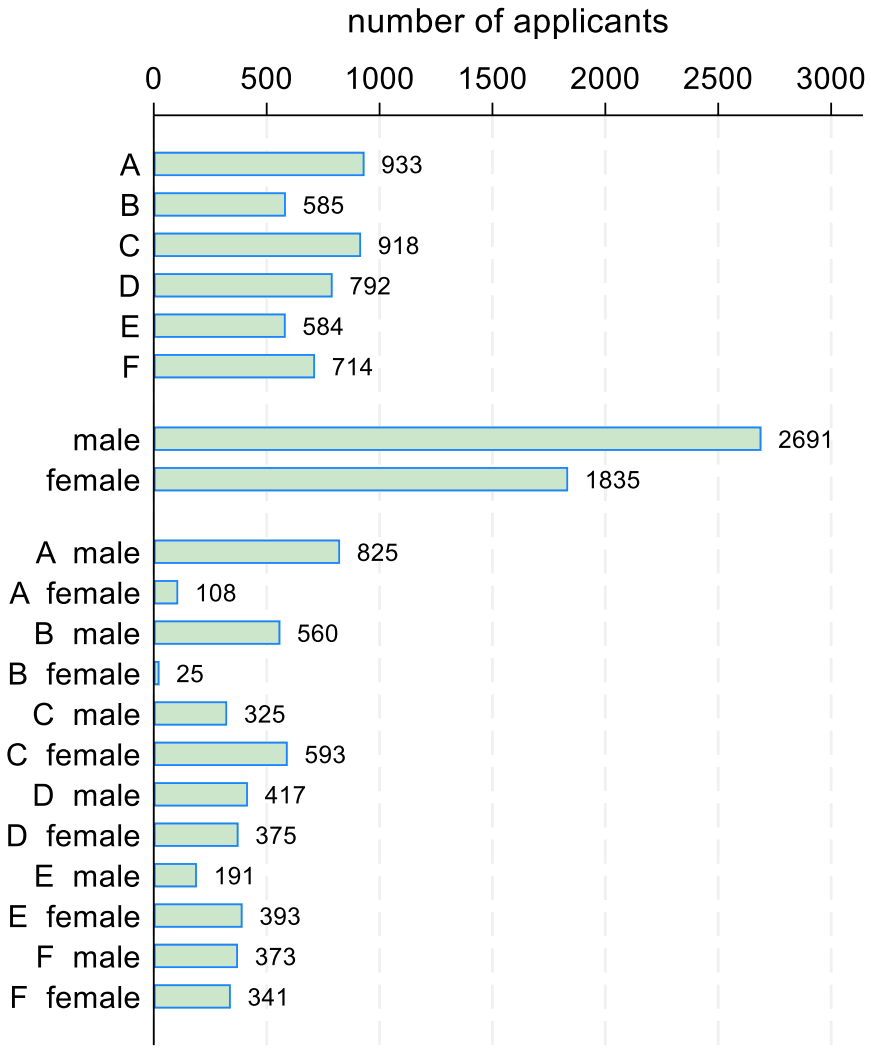
`designplot` for summaries by categorical predictors

`designplot` is a rewriting and generalization of official command `grmeanby`.

Its forte is showing results at two or more levels of aggregation or amalgamation. It is thus available to explore and explain amalgamation paradoxes such as Simpson's (Yule's, Pearson's).

See *SJ* 14: 975–990 (2014) and updates.





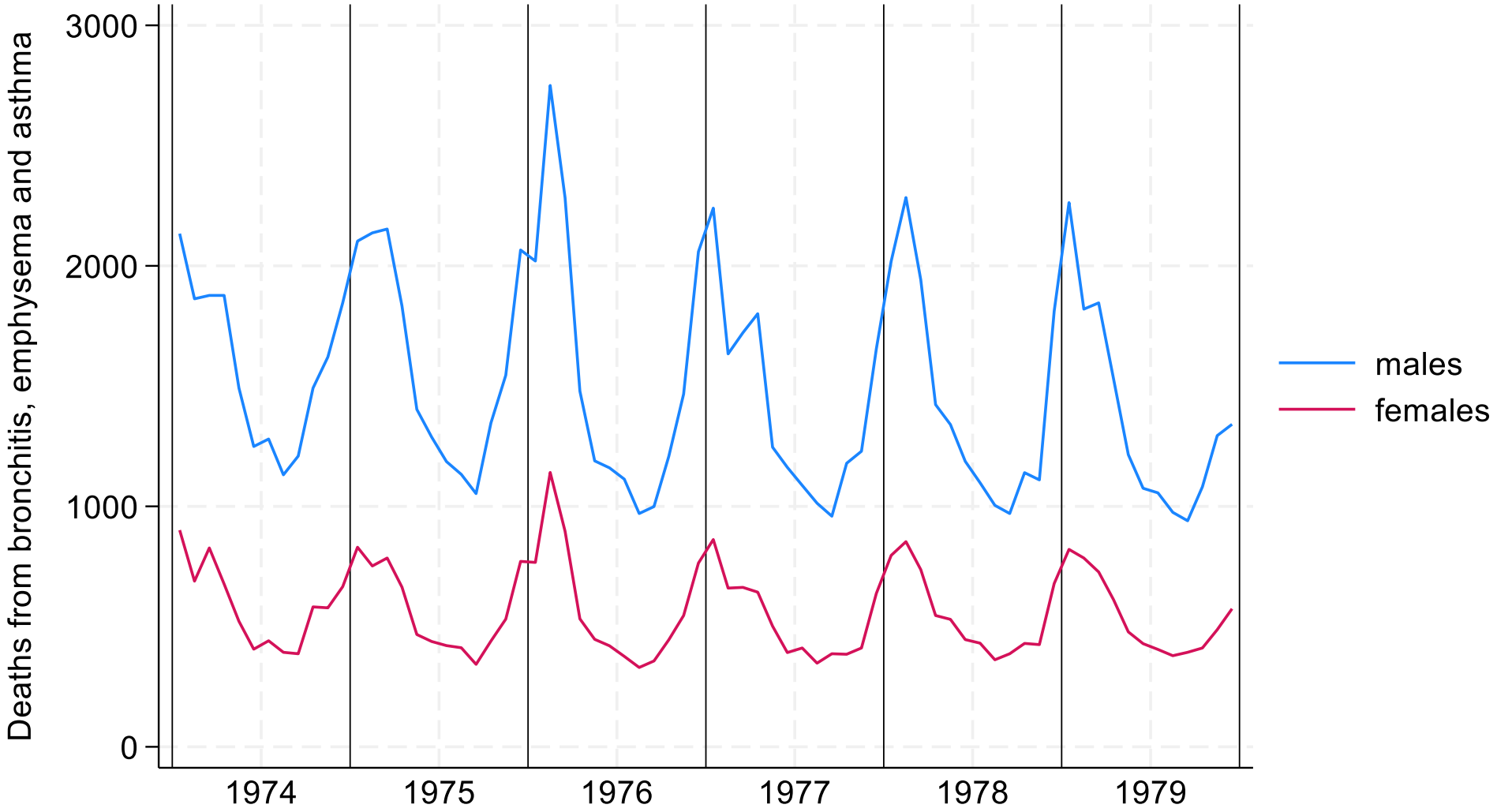
Berkeley admissions data

`cycleplot` for cycle plots

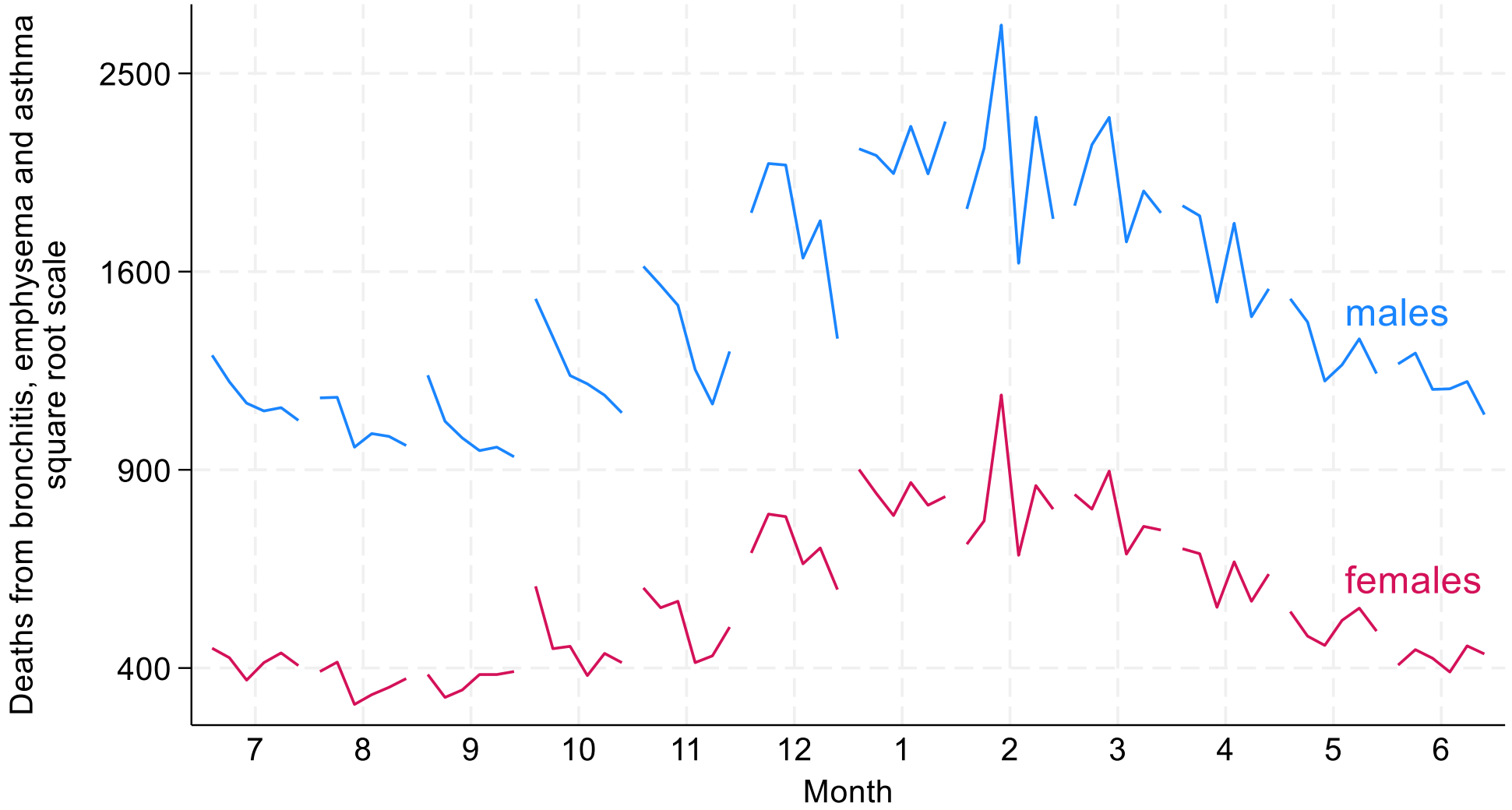
Data mixing longer term and shorter term variations in time are often helpfully shuffled to group Januarys or Mondays or whatever. That cuts down on truthful but unhelpful roller-coaster displays.

The idea was pushed vigorously and successfully by Bell Labs statisticians from the late 1970s on but can also be found in much older literature.

The example shows first a conventional line plot (month within year) and then a cycle plot (year within month).



Peter J. Diggle. 1990. *Time Series: A Biostatistical Introduction.* Oxford: Oxford University Press, p.238



Peter J. Diggle. 1990. *Time Series: A Biostatistical Introduction.* Oxford: Oxford University Press, p.238

Species of origin

See *SJ* 15: 574—587 (2015) for a riff on the simple idea that you may be better off moving the origin.

Stata strategy

Four simple points follow…

Stata technique: four simple tips quickly

When a command gets complicated, copy it to the Do-file Editor, space it out, and revise it as the equivalent of a do-file.

Use `name()` to keep graphs within a session.

Use `graph save` or `save()` to keep graphs beyond a session. `.png` can work well (and is required for Statalist).

The Graph Editor is especially useful for polishing graphs for presentation or publication.

Broader themes

Always think about the big picture too…

Divide and conquer (1 of 3)

Dividing a challenging problem into smaller or easier sub-problems may often be a good way to make progress.

Data reduction first

- ◊ Reduction of a dataset to whatever results you want to plot should often precede trying to plot those results.
- ◊ Official reduction commands include `collapse`, `contract`, `statsby`.
- ◊ Community-contributed commands include `corrci` (*SJ*), `cisets`, `pctilesets`, `quantilesets`, `momentsets`, `lmomentsets` (all SSC) .

Divide and conquer (2 of 3)

Long layout beats wide layout

- With panel or longitudinal data in particular, a long layout is often more flexible than a wide layout
- You often need to reach for `reshape long` (or occasionally `stack`)
- Once you have a long layout, a `by()` option call can help greatly: see e.g. *SJ* 20: 1016–1027 (2020)

The term *layout* (Clyde B. Schechter) is warmly recommended rather than the overloaded *format*.

Divide and conquer (3 of 3)

A good command often just does one thing very well

◊ Highly versatile commands can be too complicated to use, especially just occasionally

◊ Complicated commands are more challenging for programmers to write, to document, to maintain and to extend

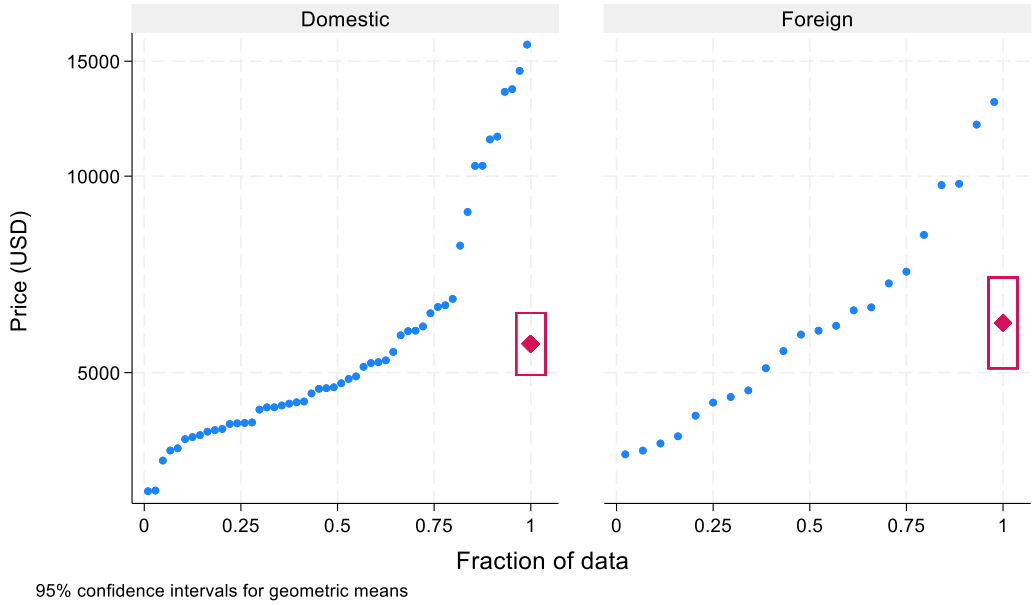
Personal story: `ciplot` was posted on SSC in 2003, for confidence interval plots. Its limitations quickly became evident.

Trying to extend it while not complicating code too much did not work.

Instead `cisets` was posted on SSC to calculate the intervals, separating out questions of how to plot them.

Here `cisets` is used to get the confidence intervals for geometric mean price and `qplot` to show quantile plots with extra point estimates and confidence intervals.

Incidentally, summarizing prices by their geometric mean goes back to Galileo at least.



Choose colours carefully

Too many colours can confuse, not clarify.

Red and blue and orange and blue are good contrasting pairs.

Never use red and green together (colour blindness!)

Rainbows are beautiful, but they belong in the sky.

Soften large patches of colour.

Black may seem boring but is a standard with good reason.

Transparency helps readability

Reducing and clarifying overlap on scatter plots with many data points is perhaps the most obvious application of transparency.

Showing probability densities as areas is another good example. The interpretation of area as representing probability ties in with any virtue it has psychologically.

Showing convex hulls is another. On a scatter plot a convex hull is the smallest convex polygon including a (sub)set of points. Here *include* implies that data points are either vertices (corners) or inside the polygon.

See this Statalist thread if interested in convex hulls.


Price (USD)
.0003
.0002
.0001
0
Density
0
5000
10000
15000
20000
auto data
Domestic
Foreign

Show differences

All graphical purposes need or imply comparisons. Comparisons often require focus on differences between values – possibly on a transformed scale.

If your interest is in differences, why not calculate and plot them directly?

Examples include

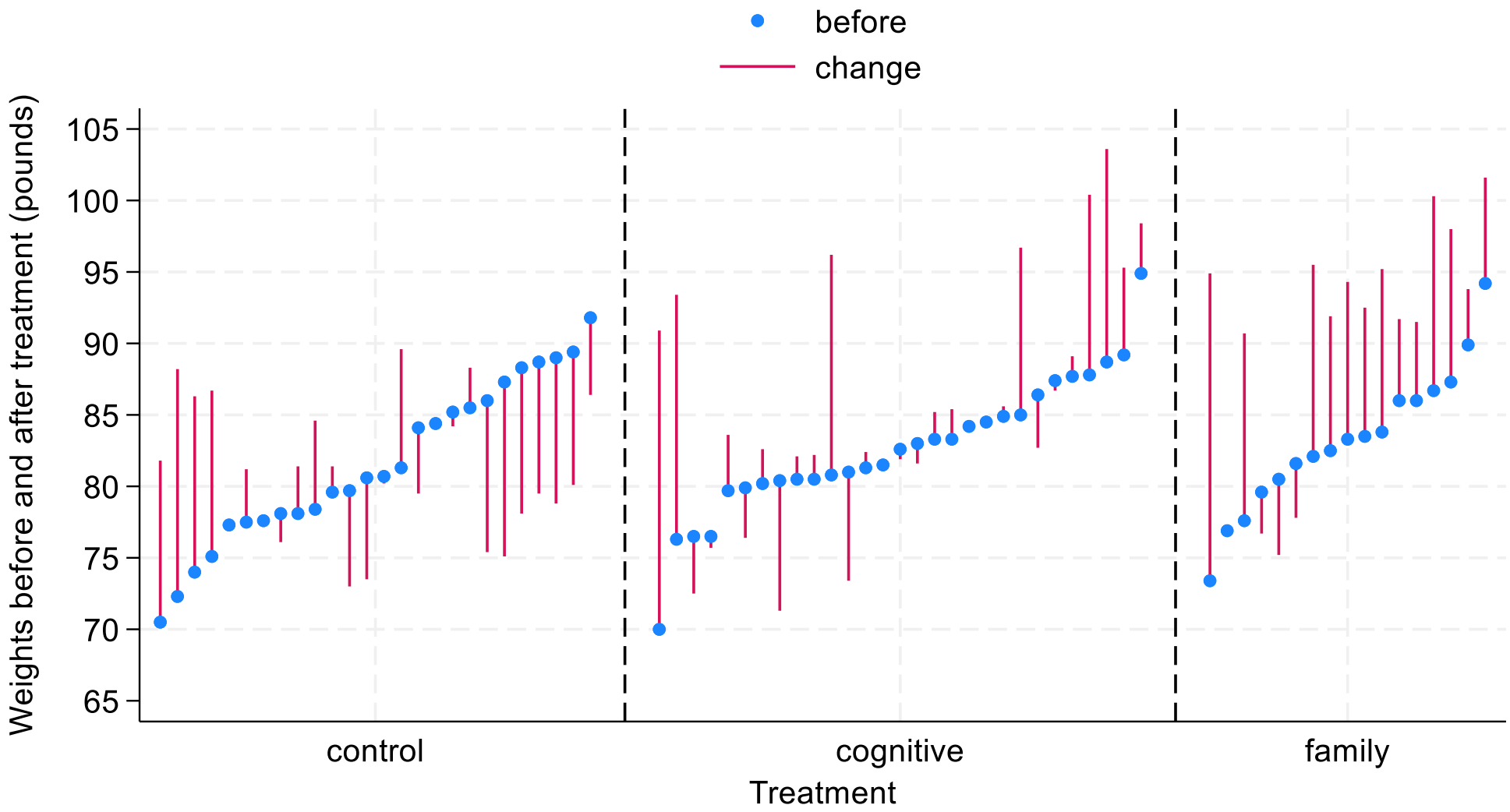
◊ residual plots

◊ plots of differences versus means (e.g. MA plots in genomics)

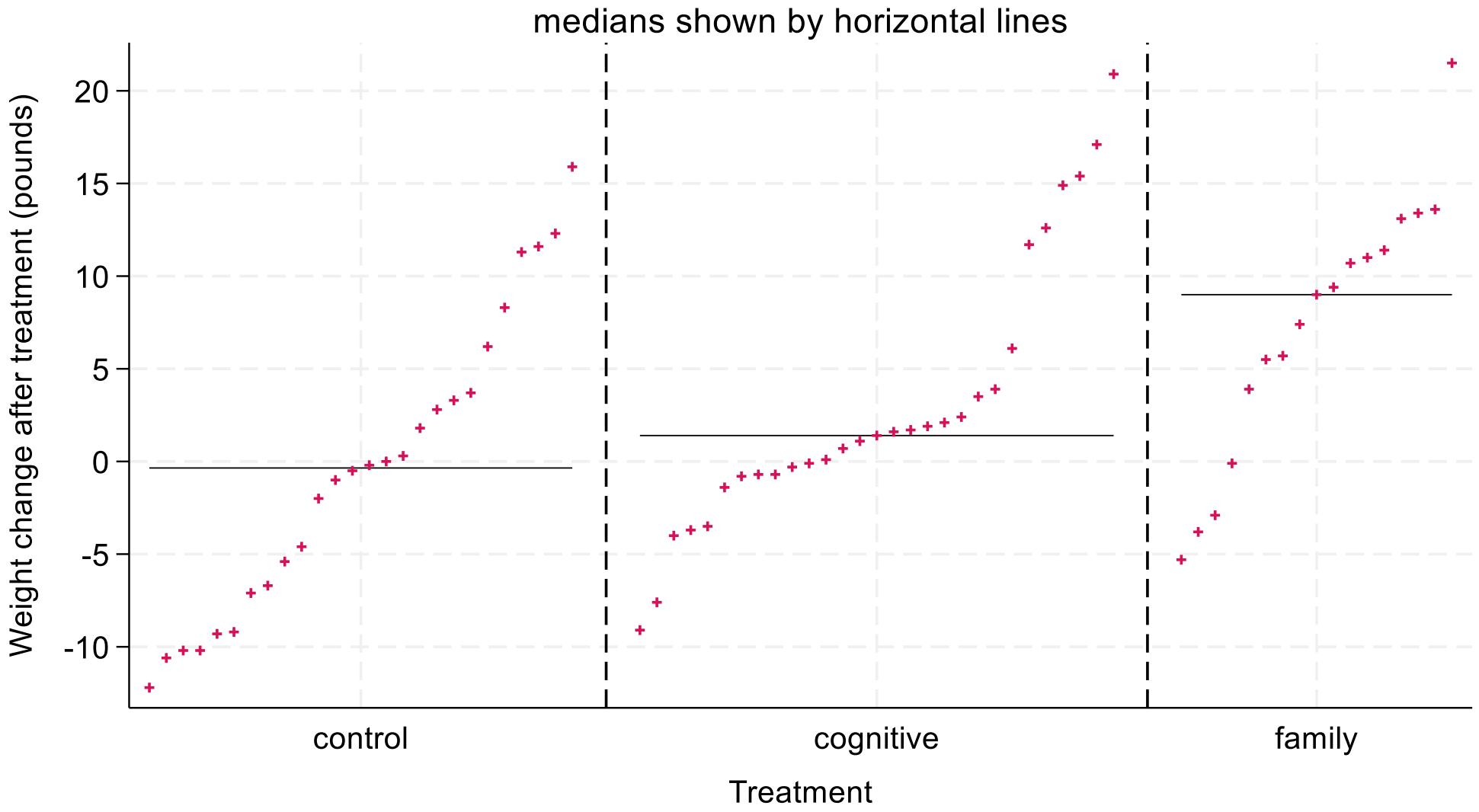
◊ plots of differences between corresponding quantiles.

The example following shows weight changes of anorexic girls under different treatments.

See *SJ* 9: 621–639 (2009) and *SJ* 24: 766–776 (2024).



David J. Hand *et al.* 1994. *A Handbook of Small Data Sets.* London: Chapman & Hall, p.229



David J. Hand *et al.* 1994. *A Handbook of Small Data Sets.* London: Chapman & Hall, p.229

Sorting helps

The order in which data arrive may be arbitrary or uninformative.

Alphabetical or identifier order may be a particularly dopey default for graphics.

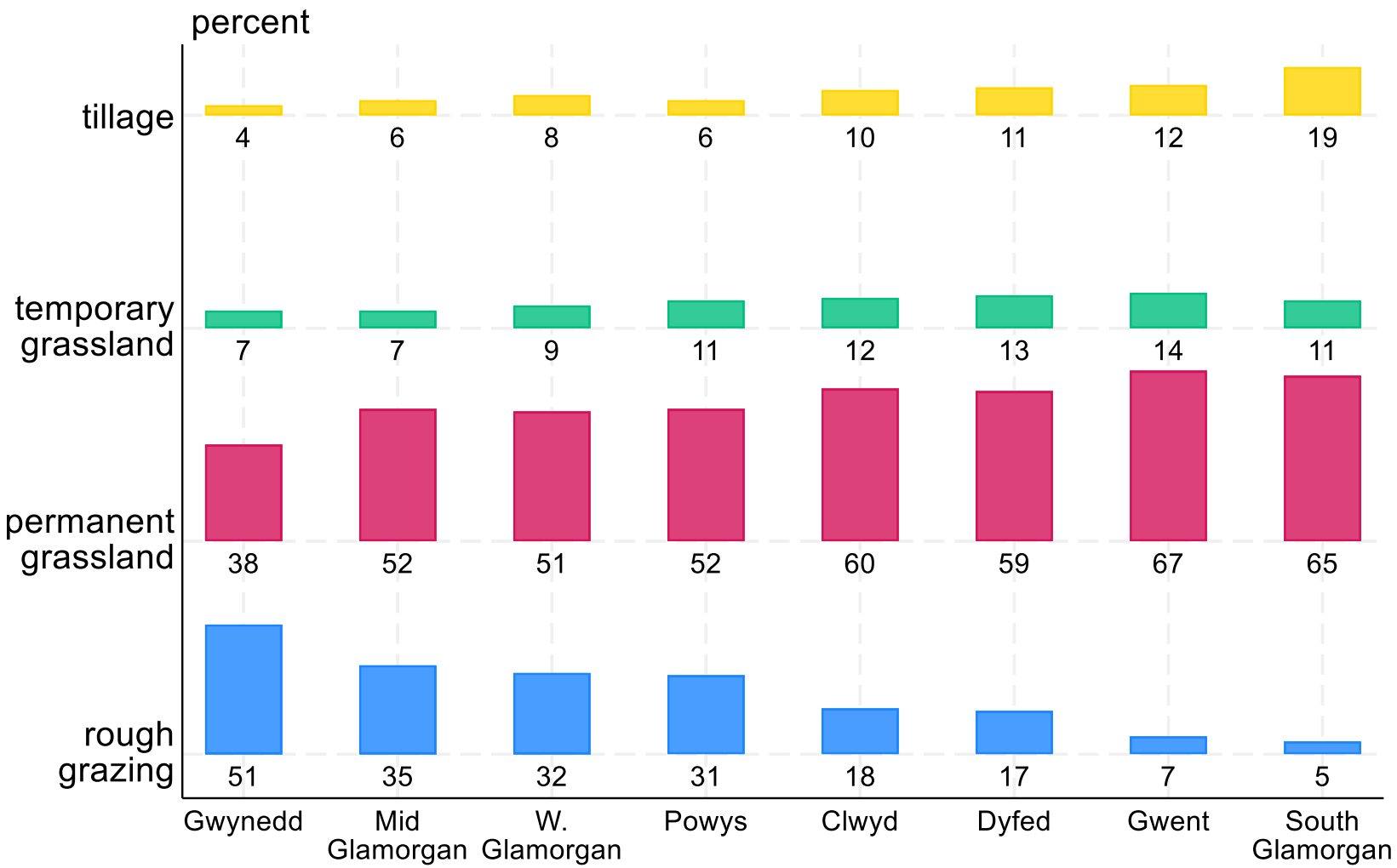
Sort instead on amount or count, latest value, or anything closer to the problem.

Among other methods, consider the `myaxis` command from *SJ* 21: 818–837 (2021).

In the original paper for the following example, both Welsh county names and land uses were ordered alphabetically.

Howard Wainer has rightly mocked this practice as Alabama first!

People in every country will suggest their own local variant: Adelaide, Aceh and so on.



Rural land use 1978
John W. Aitchison. 1981. Triangles, tetrahedra and taxonomy. *Area* 13: 137-143

Use small multiples

Small multiples (Edward R. Tufte's term) use the same idea repeatedly over multiple groups or variables.

Scatter plot matrices through `graph matrix` are one example. Any graphical method that supports an `over()` or a `by()` option could be another.

As a last resort, use `graph combine`.

Before that, consider reshaping data to allow use of `by()` or `over()`.

Show the data!

To improve learning from data,
credibility, and integrity,
show the data.

Edward R. Tufte. 2020. *Seeing with Fresh Eyes: Meaning, Space, Data, Truth*. Cheshire, CT: Graphics Press, p.101.

Graphs are for people

Why produce a graph?

◊ explorations

◊ indications

◊ conclusions

Who is the graph for?

◊ colleagues

◊ clients

◊ citizens

Graphics is – and yet is not – a spectator sport.

Given a novel or unfamiliar design, you need to play actively using data of interest to judge it fairly.

Does it convey what you already know, easily and effectively?

Does it tell you anything new and helpful?

The purpose of computing is insight, not pictures.

Lloyd Nicholas Trefethen. 2011.

Trefethen's Index Cards: Forty years of notes about People, Words and Mathematics.

Singapore: World Scientific, p.330 [3 March 1997]

Resources

The Graphics manual [G], naturally.

Michael N. Mitchell. 2022. *A Visual Guide to Stata Graphics*. College Station, TX: Stata Press. Especially useful as a reverse reference on official commands. Find a graph you like; see which command produced it.

Franz Buscha. 2025. *Graphs Everyone Should Know and How to Create Them in Stata*. College Station, TX: Stata Press. Wide ranging and friendly.

Stata Journal. The column *Speaking Stata* often discusses graphics. Many of the Tips in each issue are graphical.

The SSC archive also includes many user-written commands.

All such presentations are partial, if not prejudiced.

These were some of my tips for graphics. What are yours?

The next few slides gather together tips in summary form.

Some extras follow.

Reprise: small stuff

Use open markers
Use marker labels as markers
Go grey
Labels without ticks can make sense
Use ticks or pipes in marginal rug plots
Use nice labels for logarithmic scales
Consider other nonlinear scales
Omit needless axis titles
Lose the legend! Kill the key!
Horizontal is good
Use `lpoly` for scatter plot smoothing

Reprise: convenient commands

```
graph dot
stripplot
qplot
qqplotg
fabplot
tabplot
designplot
cycleplot
```

Reprise: Stata strategy

Do-file Editor

`name()`

`graph save` and `save()`

Graph Editor

Reprise: broader themes

Divide and conquer

Choose colours carefully

    Transparency helps readability

Sorting can help

Show differences

Use small multiples

Show the data!

Graphs are for people

Quotations on normal quantile plots

Yudi Pawitan (2001, 92)

The normal QQ-plot is a useful exploratory tool even for nonnormal data. The plot shows skewness, heavy-tailed or short-tailed behaviour, digit preference, or outliers and other unusual values.

Ramanathan Gnanadesikan (1977, 168; 1997, 193)

Although a normal probability plot does not provide a single-statistic-based formal test, as a graphical tool it conveys a great deal more information about the configuration of the observations than any single summary statistic is likely to do.

Michael Hills (1974, 28)

It can be useful to plot an observed distribution against the standard Gaussian even though there is no question of it being Gaussian in shape. The motive is that it is easier to study a distribution by comparing it with a standard shape than just by looking at it.

Pawitan, Y. 2001.
In All Likelihood: Statistical Modelling and Inference Using Likelihood.
Oxford: Oxford University Press.

Gnanadesikan, R. 1977.
Methods for Statistical Data Analysis of Multivariate Observations.
New York: John Wiley. (second edition 1997)

Hills, M. 1974.
Statistics for Comparative Studies.
London: Chapman and Hall.

The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them.

Ronald Aylmer Fisher. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, S.7. 1890–1962

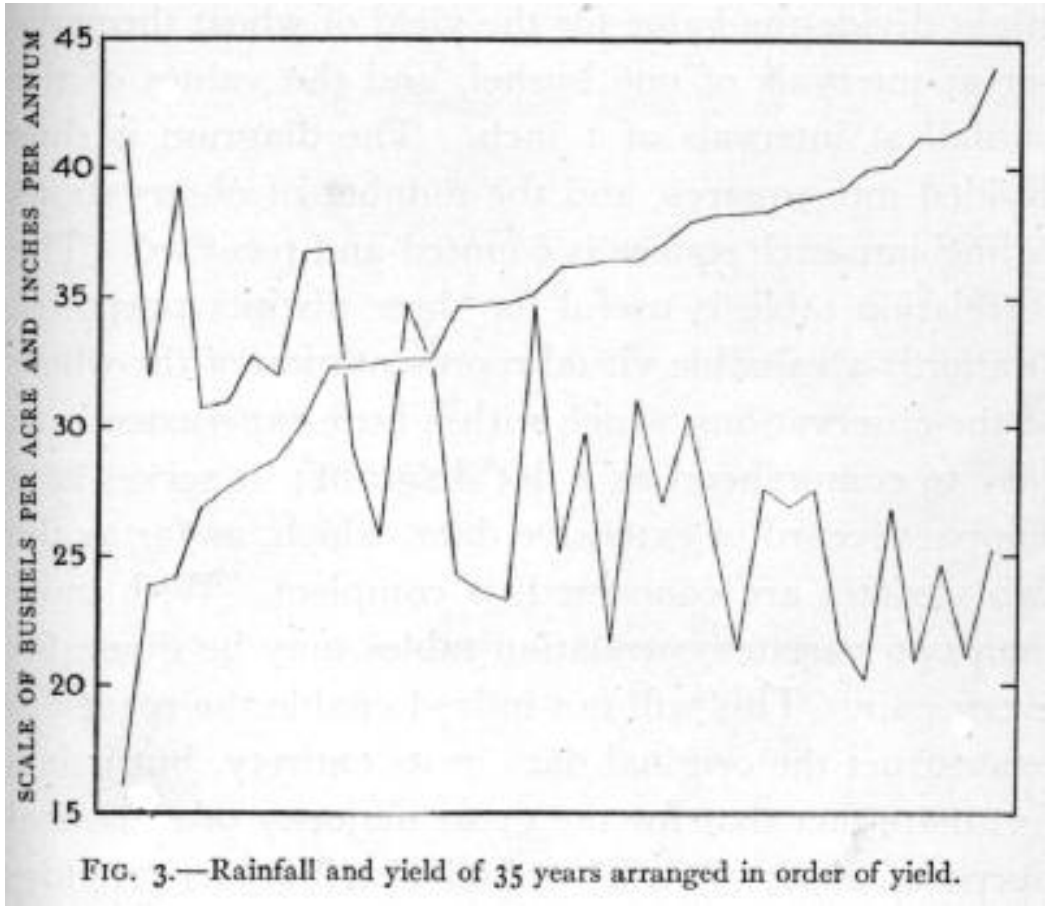


FIG. 3.—Rainfall and yield of 35 years arranged in order of yield.

Reader reactions to your graphs

Most wanted

Aha! With this design I can see structure in my data and details that I need to think about!

Least wanted

Huh? What is going on here? What am I supposed to see?

In between

Wow! How did you do that?

This was a presentation to the Stata Oceania meeting, 6 February 2026.