Data visualization

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

My goal in this lecture is to explain some basic principles of data visualization. My focus here is on simple graphs with not too many data points.

Displaying data well

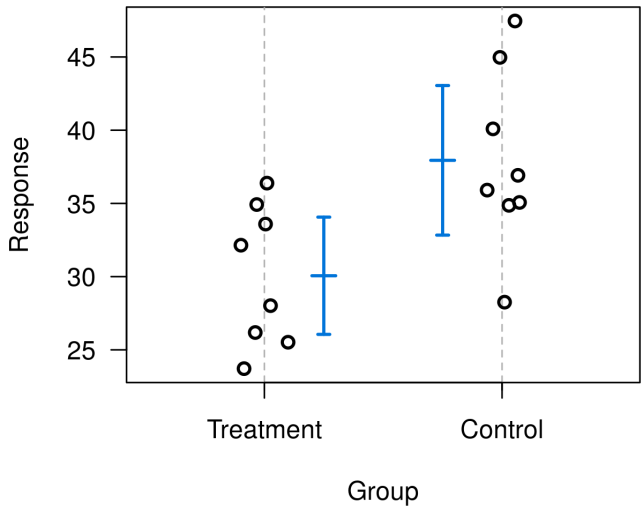
- Be accurate and clear.
- Let the data speak.
  - Show as much information as possible, taking care not to obscure the message.
- Science not sales.
  - Avoid unnecessary frills (esp. gratuitous 3d).
- In tables, every digit should be meaningful. Don't drop ending 0's.

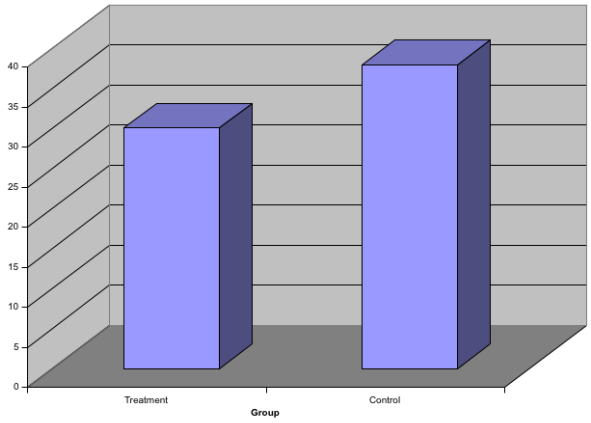
The key principles of data visualization for us are to be accurate and clear, and to let the data speak. We want to show as much information as possible, but to a point: we don't want to obscure the message.

For data scientists, this should be about science rather than sales. You want to focus on conveying the truth not persuading or arguing a particular angle. And you want to avoid unnecessary frills. That's not to say that good graphs can't be pretty, but they should be over-the-top with junk that gets in the way of presenting the data, like gratuitous 3-dimensional rendering.

In tables, every digit should be meaningful, but you shouldn't drop ending zeros as they indicate precision.

Show the data

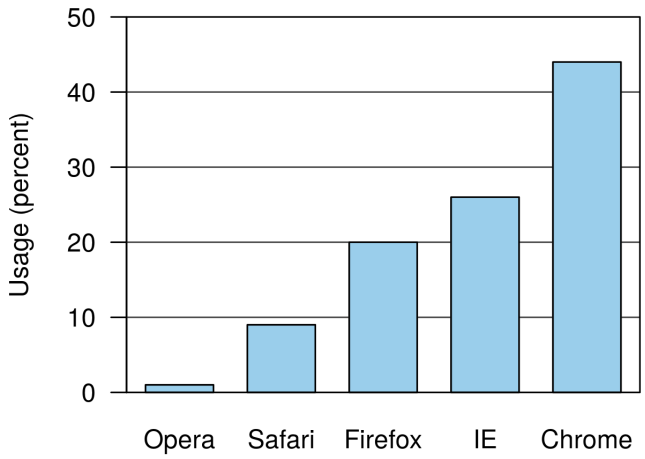


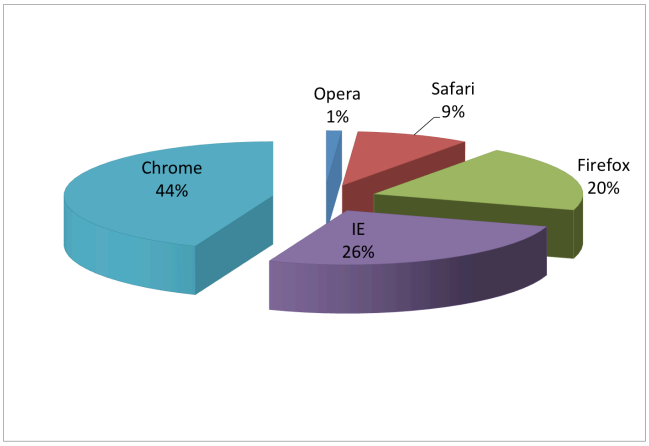


First, show the data. Particularly in a case like this where you have just a few data points in each of two groups, it would be a shame to not show the actual data. Bar plots with antennas give you just a couple of numbers, and 3-dimensional bars make it hard even to see them. In the figure on the right, the bars are sitting in front of the axis, and so you need to project back in order to figure out what the numbers are.

There's a lot of fancy stuff that you can do in Excel. But is it helping the reader to better understand the data, or is it making it harder?

Avoid pie charts





Pie charts are seldom a good solution. 3d pie charts are the worst.

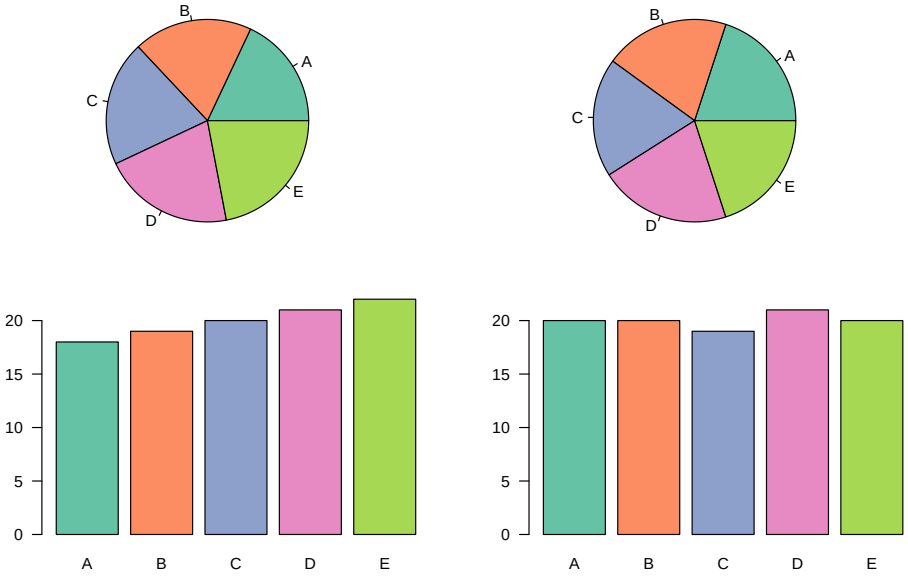
The problem is, humans are terrible at judging areas. That's why pie charts so often show the percentages next to the slices. To do so seems an admission of defeat.

I don't much like bar charts, but they're much more effective for the actual comparison of values, which is the point of such a chart.

Pie charts have one advantage: for numbers that add up to 100%, they make that part clear. But in all other ways, they are ineffective. And the more slices there are, the worse they get.

There can be a case for pie charts with two slices. But I would mostly avoid them, and would at least switch to a bar chart.

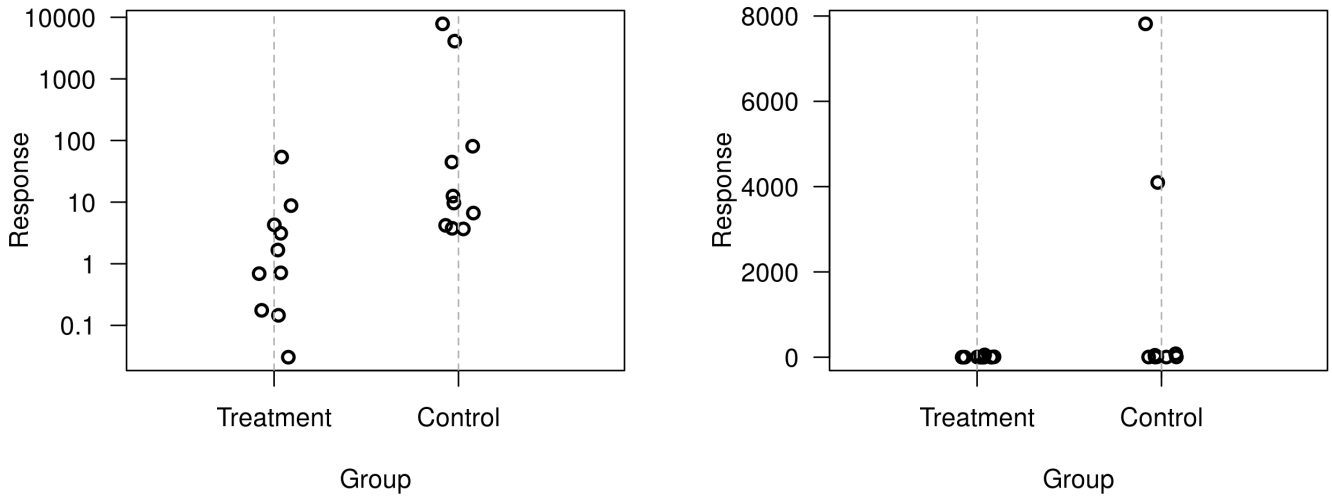
Avoid pie charts



via @MonaChalabi (bit.ly/pie_vs_barchart)

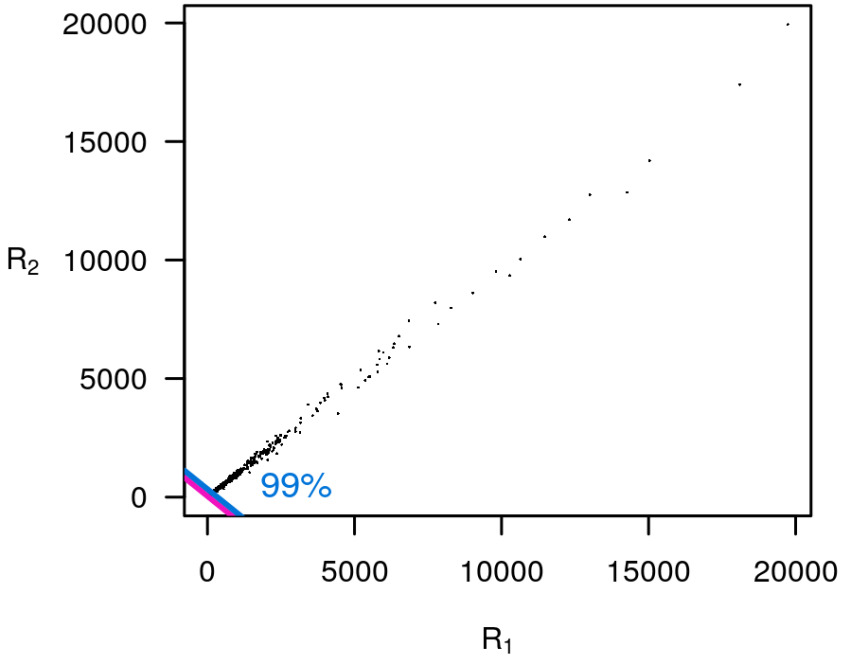
Here's another example. You really can't distinguish the values in the pie chart, but it's easy to compare them when viewed as a bar chart.

Consider logs



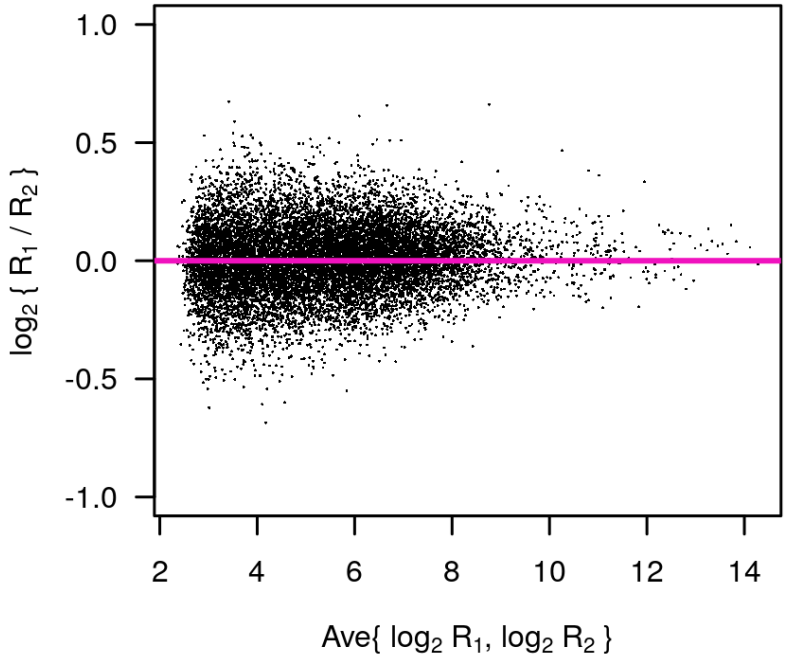
Always consider transforming the data by taking logs. This is particular important if the data span multiple orders of magnitude, which is particularly common for cytokines and gene expression measurements.

Consider logs



This is a plot of gene expression values from two replicate microarrays. The values are so skewed, that unless you take logs, the figure is dominated by a few very large points. 50% of the data are below the pink line, and 99% of the data are below the blue line. If we take logs, we get to see more of the data.

Consider differences

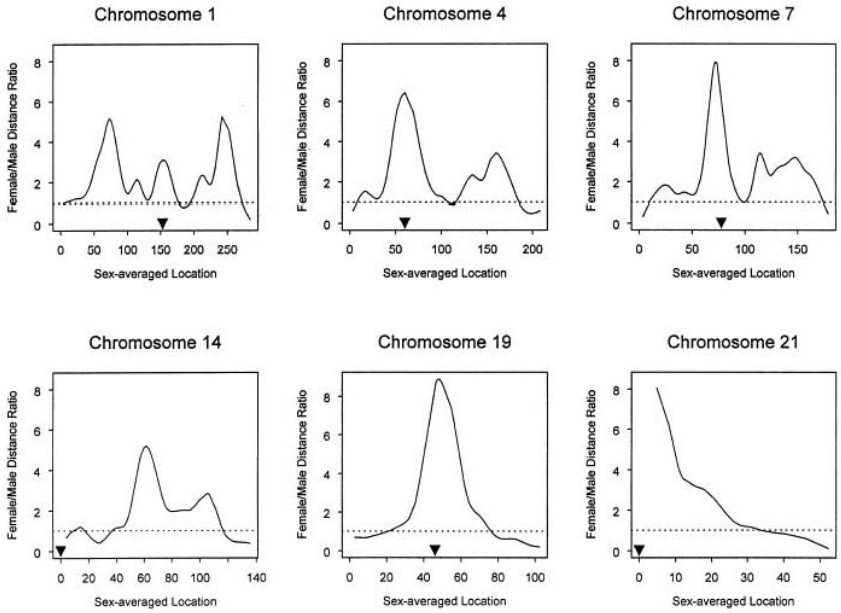


Also, if you're interested in the differences between the two arrays, it's best to subtract the values and look at the differences directly. The differences on the log scale are the same as the log ratios.

This prevents us all from having to rotate our heads 45 degrees to look for points off the diagonal.

This is sometimes called an "MA plot," though I'm not sure why "M" and "A." I'd call it a difference-vs-average plot. Note how it also makes more complete use of the space.

Another "take logs" example



Broman et al., Am J Hum Genet 63:861-869, 1998, Fig. 1

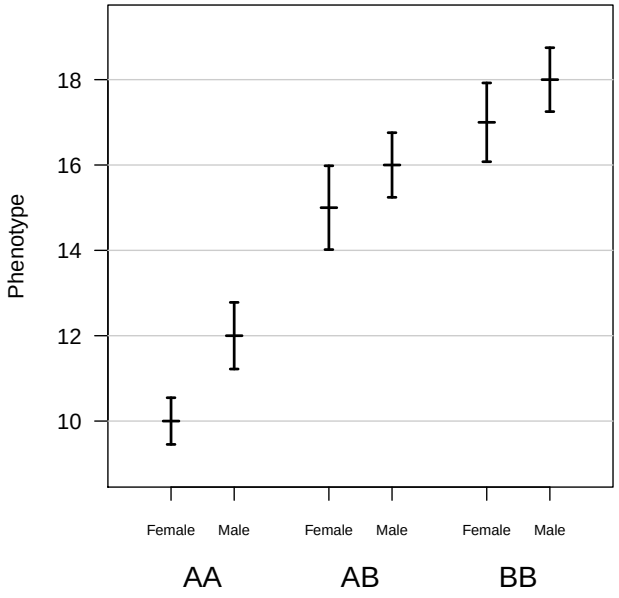
Ratios are another case where it can be important to take logs.

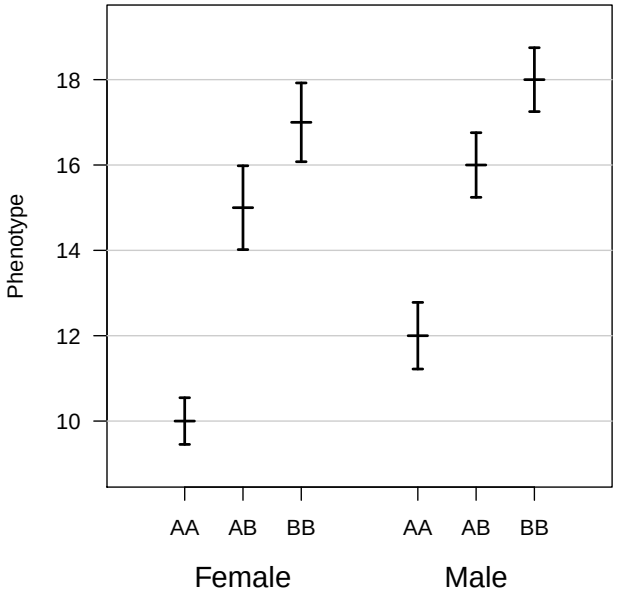
I've shown this figure before; it's a bit of an embarrassment to me now. This is a plot of a female/male ratio. When female > male, the values span from 1 to infinity, whereas if male > female, the values are smashed between 0 and 1.

If I'd taken logs, there'd be a nice symmetry between the ratios > 1 and those < 1.

Ease comparisons

(things to be compared should be adjacent)





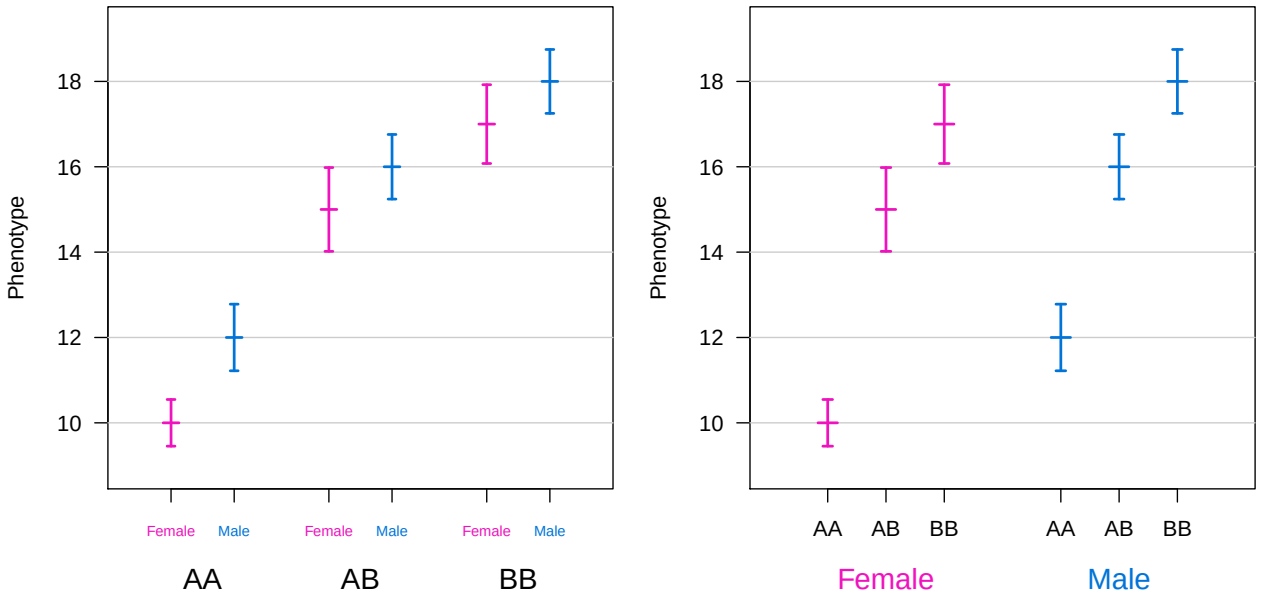
To ease comparisons, you want to put the things to be compared next to each other.

These two figures show the average phenotype value for three different genotypes, and also split by sex. In the version on the left, the two sexes are next to each other; in the version on the right, the three genotypes are next to each other.

There are always trade-offs in data visualization. If you are most interested in comparing the two sexes within each genotype, you should go with the version on the left. If you are most interested in comparing the three genotypes (controlling for sex), you should go with the right version.

Ease comparisons

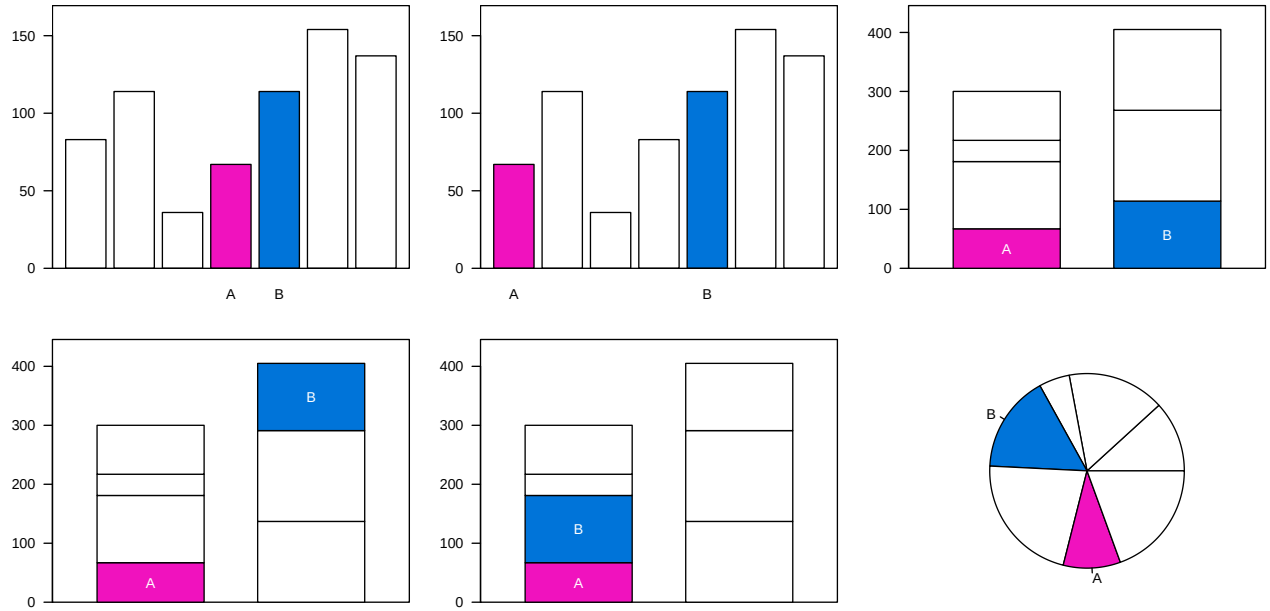
(add a bit of color)



Color can be useful to help guide comparisons.

But note that here I've used stereotypical color choices. In a way, this graph reinforces that "pink for girls; blue for boys" stereotype. It has the advantage that it might be easier to remember, but that's probably not worth the disadvantage of reinforcing an unfortunate stereotype, and so I'd avoid this choice of colors, now. We could instead go with, say, green and purple.

Which comparison is easiest?



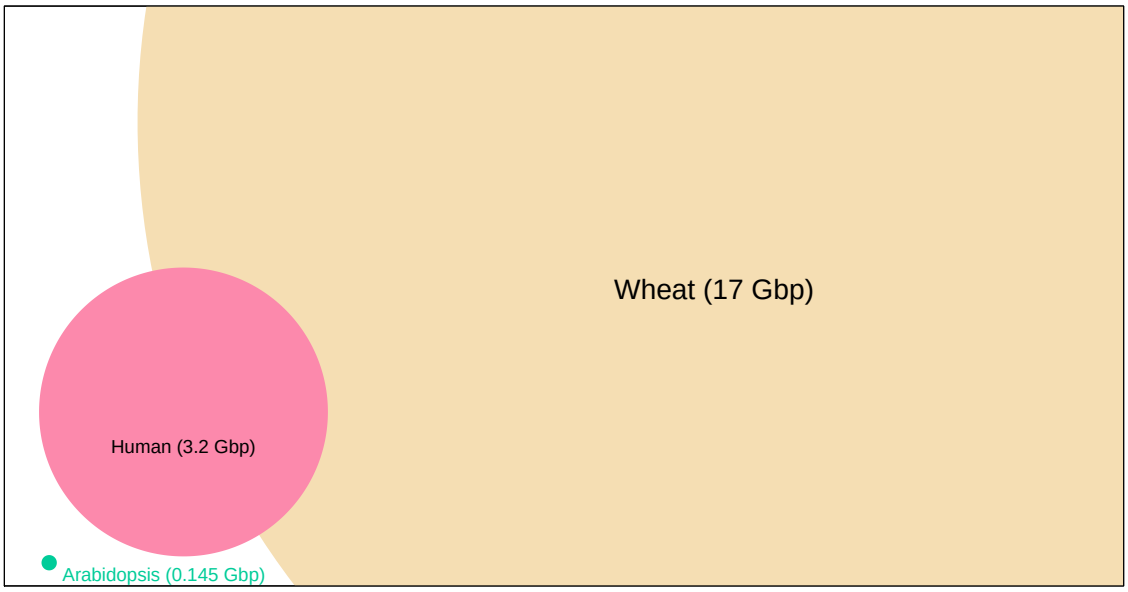
There are a variety of different plots in which you might be comparing two values, A and B. Which one is easiest?

It's easiest if you're comparing positions, and if they're right next to each other. Move them apart or compress them or offset the lengths, and they become harder. Hardest of all is the pie chart, where you're comparing areas or maybe angles.

This is important for when you are designing a visualization. There are often a variety of comparisons that you want to be making. But think about which ones are the most important ones.

Don't distort the quantities

(value $\propto$ radius)

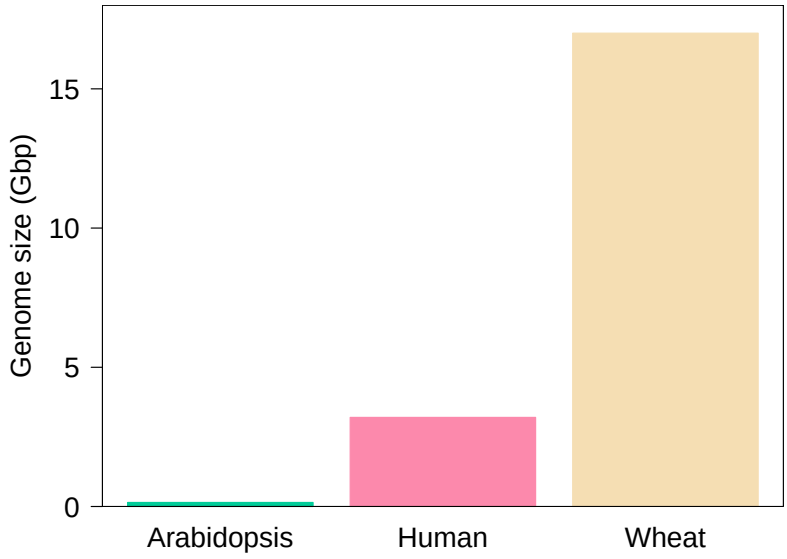


Another key principle is to not distort the quantities. Here, the sizes of the genomes are proportional to the diameters of the circles. But one naturally focuses on the areas of the circles.

If you change the figure so that the values are proportional to the areas, you get a more realistic view of the data. The wheat genome is like $5\times$ that of the human, and the human genome is like $20\times$ that of arabidopsis (a model plant). But it's still pretty hard to see that without showing the numbers, because humans just aren't so good at comparing areas.

Don’t use areas at all

(value $\propto$ height)



This plot is not nearly so interesting, but it does a much better job of conveying the values.

Encoding data

Quantities

- Position
- Length
- Angle
- Area
- Luminance (light/dark)
- Chroma (amount of color)

Categories

- Shape
- Hue (which color)
- Texture
- Width

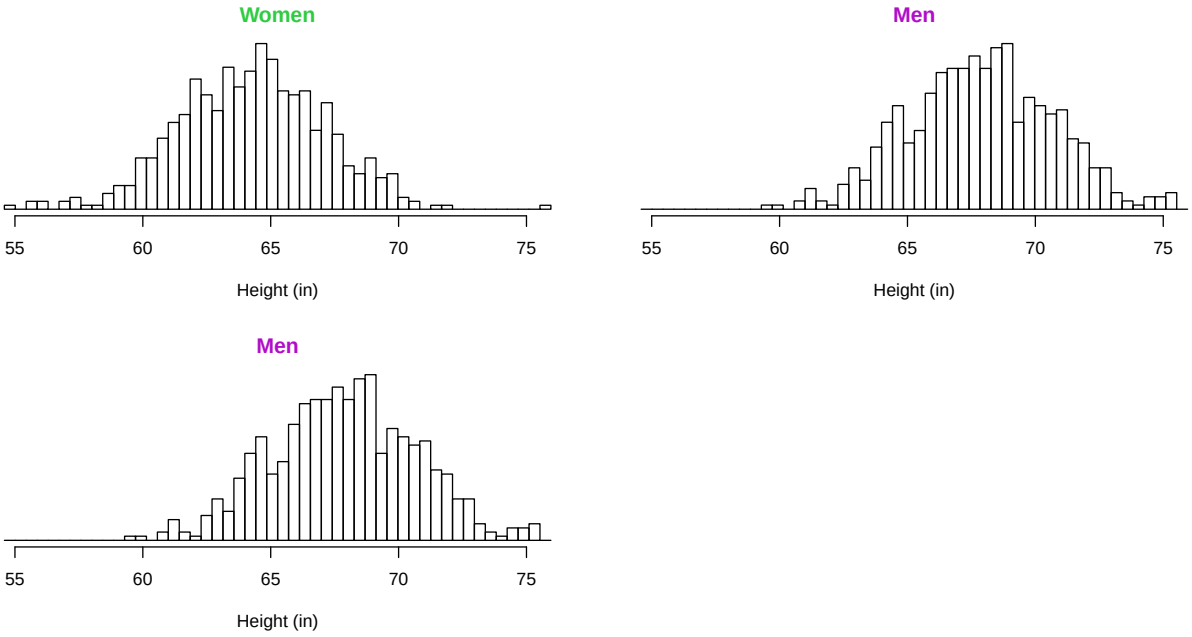
In any data visualizations, you're encoding quantities or categories using visual features. There are a variety of choices you can make, and they vary enormously in the ease of the reader of decoding the values.

Position is easier then length which is easier than angle which is maybe easier than area which is easier than shading or color.

Shape is easier than color which is easier than texture (like cross-hatching) which is easier than line width.

Ease comparisons

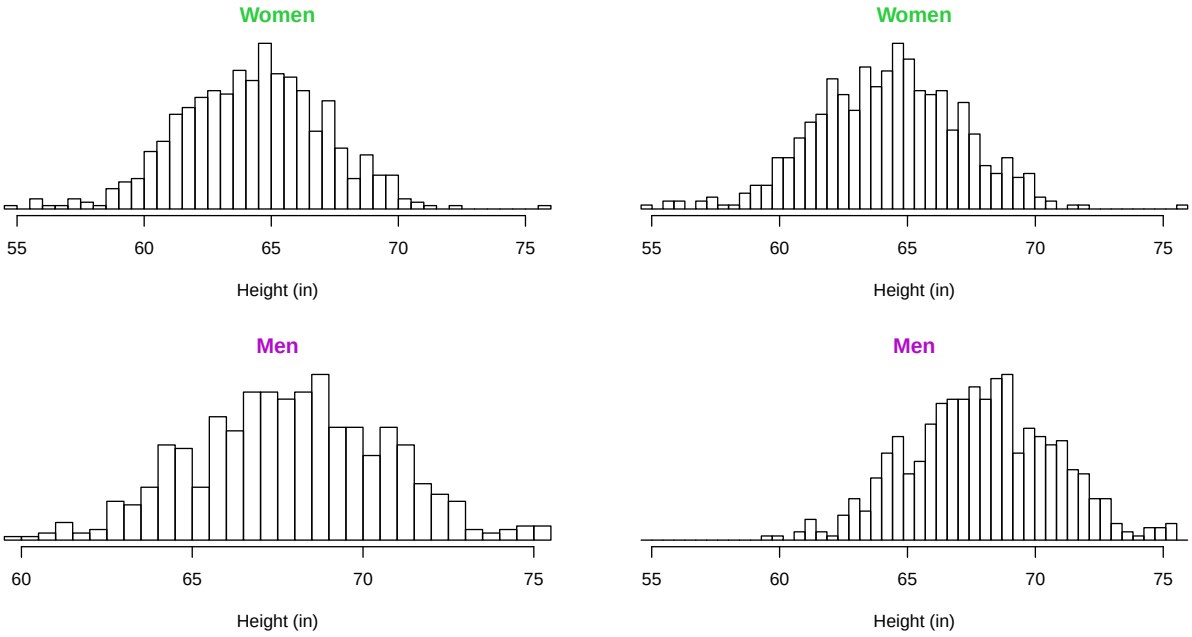
(align axes)



Another important technique is to align axes to ease comparisons. If you're comparing the heights of women and men, it is much easier to do so with them stacked on top of each other rather than side-by-side.

Ease comparisons

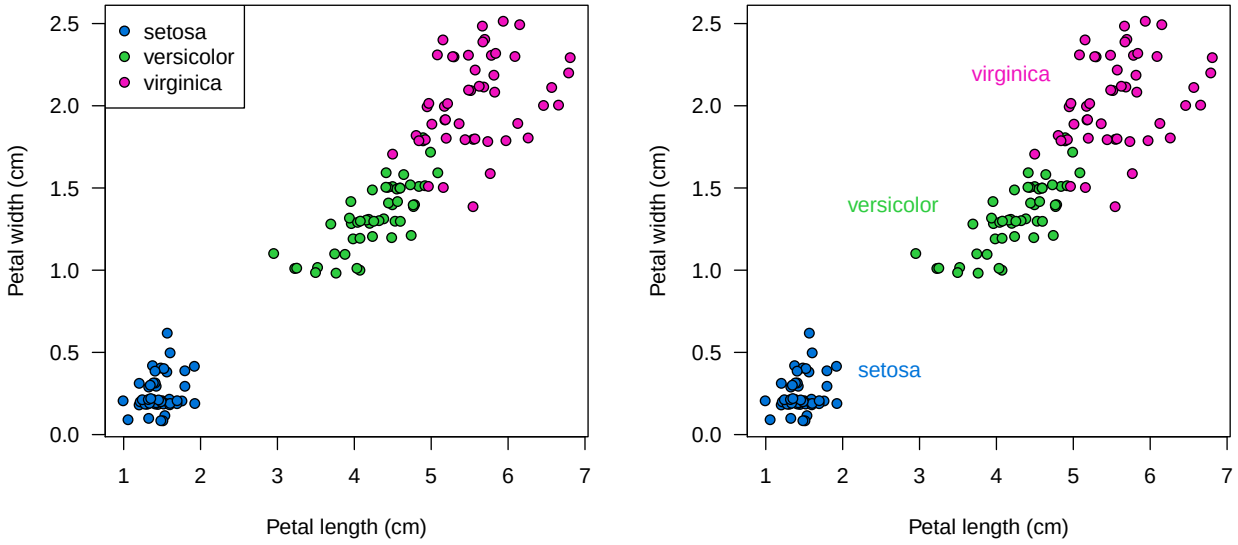
(use common axes)



You also want to use common axes as much as possible, as in the panels on the left.

If you allow the axes to differ, as in the panels on the right, the reader has a much harder time comparing the values.

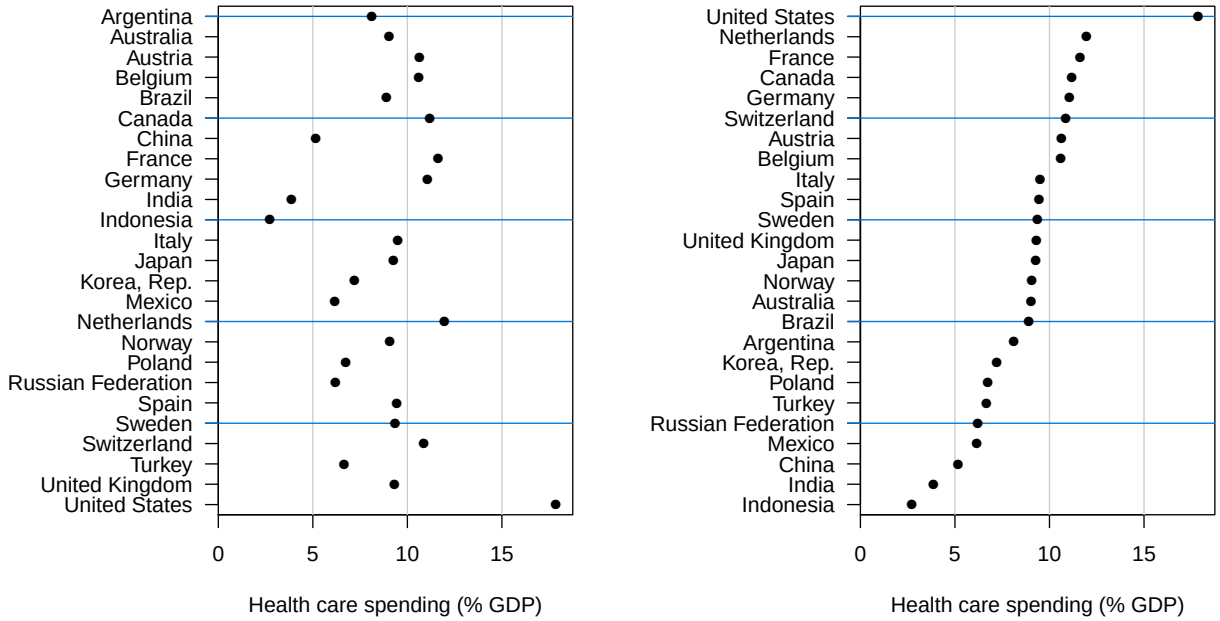
Use labels not legends



Another thing I really like to emphasize is the value of placing labels directly in the plot (as in the right panel) rather than using a separate legend (as in the left panel) or even worse explaining the colors in the figure caption.

It is arguably more difficult to place labels directly on the plot. It generally requires some extra effort to get the labels right. But for a figure in a paper or talk, it's often worth the effort.

Don’t sort alphabetically

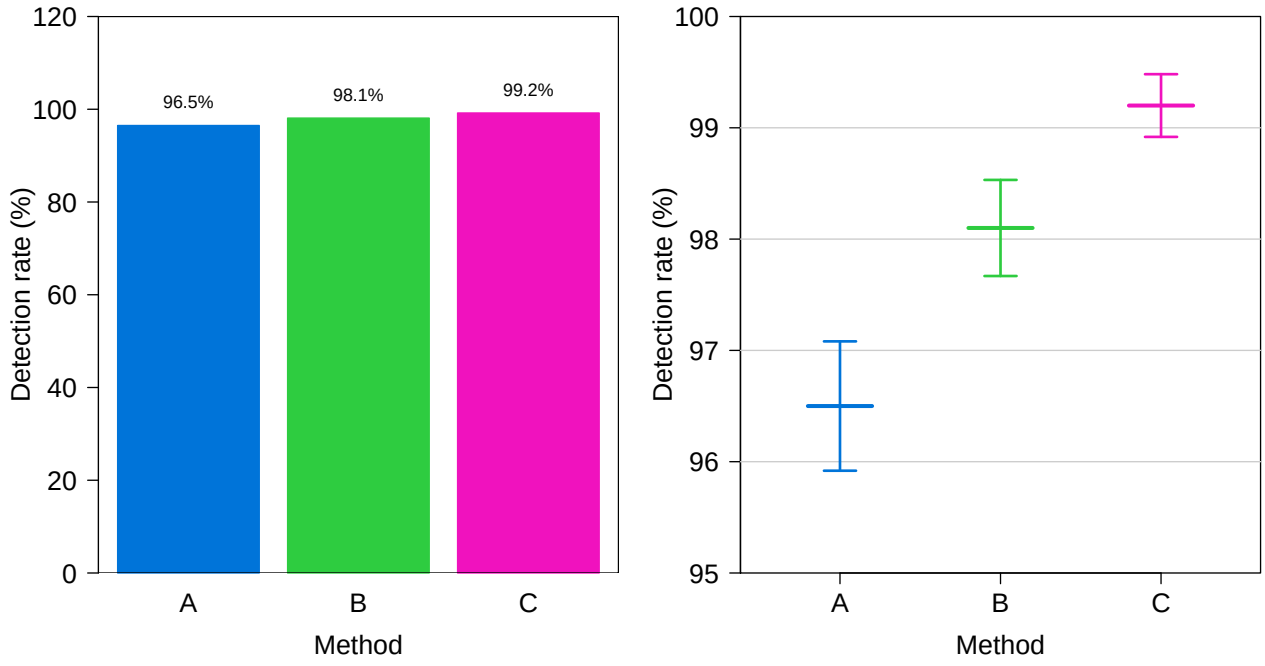


Never sort things alphabetically. It’s basically always better to sort by the value, or sort by the value of one variable and have related panels in that same order.

In the panel on the right, you can immediately see important relationships (such as by health care spending differences by continent).

The alphabetical order in the panel on the left is good only for looking up a particular country, and that’s not worth it.

Must you include 0?



A common question is whether you should include 0 in the y-axis. In bar charts, I think it is misleading to not include 0. But much of the time, it's better to not use bars and just use line segments or dots, and then in that case it can be useful to focus just on the interval with the data.

Here, we're looking at detection rates near 100%, and it's much better to focus on the high-end of the range, so we can see the differences. Thus, it's best to omit the bars.

I'd also say that if you have something that's constrained to 0–100%, it's best to have the y-axis constrained to those limits. Having the range go above 100% or below 0% is a bit silly.

The choice of whether you should include 0 or not is partly dependent on the audience. Will they be misled by using a focused interval, or not?

A bad table

	$b/c = 10.0$		$b/c = 10.0$		$b/c = 100.0$	
N	$r^\star$	G	$r^\star$	G	$r^\star$	G
3	2	0.2	2	2.225	2	22.47499
4	2	0.26333	2	2.88833	2	29.13832
5	2	0.32333	3	3.54167	3	35.79166
6	3	0.38267	3	4.23767	3	42.78764
7	3	0.446	3	4.901	3	49.45097
8	3	0.50743	4	5.5765	4	56.33005
9	3	0.56743	4	6.26025	4	63.20129
10	4	0.62948	4	6.92358	4	69.86462

This is a really bad table. It's actually taken from the Journal of the American Statistical Association, which is the flagship journal for the largest society of statisticians in the US. Statisiticians can be really terrible at data visualization. They are often lazy and revert to tables, and their tables are often terrible, like this one.

There are way too many digits shown. We almost never can measure things to more than 3 digits. And you can just tell, looking down each column, than no more than 3 digits are needed to tell the key story.

Also, leading 0's are omitted. If we need five digits in 0.38267, then we should show the same precision in 0.44600.

Related to that, the numbers are centered, rather than being aligned at the decimal point. Always align numbers at the decimal point.

Fewer digits

	$b/c = 10.0$		$b/c = 10.0$		$b/c = 100.0$	
N	$r^\star$	G	$r^\star$	G	$r^\star$	G
3	2	0.20	2	2.2	2	22
4	2	0.26	2	2.9	2	29
5	2	0.32	3	3.5	3	36
6	3	0.38	3	4.2	3	43
7	3	0.45	3	4.9	3	49
8	3	0.51	4	5.6	4	56
9	3	0.57	4	6.3	4	63
10	4	0.63	4	6.9	4	70

Here's the same table, corrected. Two digits is really sufficient; the previous table had 3× as many digits as were necessary. Also, don't drop the ending 0's and line things up the numbers at the decimal point.

Yuck!

	1990		2005		2010		p value
	n	Rate (95% CI)	n	Rate (95% CI)	n	Rate (95% CI)	
(Continued from previous page)							
Globally							
<75 years							
Incidence	6 353 868	159·22 (145·32–174·98)	9 288 048	167·45 (150·96–187·11)	10 469 624	168·75 (152·43–187·09)	0·208
Prevalence	13 234 062	324·26 (288·74–374·96)	20 187 246	358·58 (317·58–412·79)	23 052 804	366·93 (328·04–420·66)	0·086
MIR	..	0·359 (0·318–0·409)	..	0·293 (0·249–0·332)	..	0·254 (0·212–0·287)	<0·001
DALYs lost	63 991 864	1543·96 (1452·03–1728·25)	74 855 520	1326·17 (1172·08–1388·74)	73 293 552	1163·448 (1011·43–1232·19)	<0·001
Mortality	2 301 435	57·38 (54·12–64·27)	2 734 251	49·16 (43·60–51·55)	2 668 499	42·89 (37·65–45·81)	<0·001
≥75 years							
Incidence	3 725 067	3173·50 (2932·14–3422·23)	5 446 077	3082·97 (2819·52–3372·55)	6 424 911	3113·00 (2850·95–3403·57)	0·361
Prevalence	4 681 276	3974·37 (3609·66–4441·23)	8 308 337	4700·18 (4239·37–5256·84)	9 972 153	4835·38 (4382·63–5433·92)	0·005
MIR	..	0·634 (0·575–0·709)	..	0·543 (0·476–0·607)	..	0·500 (0·439–0·560)	<0·001
DALYs	22 018 520	18665·35 (17 464·55–20 408·51)	27 096 178	15 300·36 (13 987·78–16 317·62)	28 938 754	14 053·63 (12 761·98–15 088·12)	<0·001
Mortality	2 359 013	2033·21 (1888·78–2233·65)	2 950 719	1678·65 (1528·60–1807·22)	3 205 682	1545·29 (1412·76–1685·12)	<0·001
All ages							
Incidence	10 078 935	250·55 (229·70–273·25)	14 734 124	255·79 (232·10–283·88)	16 894 536	257·96 (234·40–284·11)	0·335
Prevalence	17 915 338	434·86 (389·45–496·84)	28 495 582	490·13 (436·60–557·52)	33 024 958	502·32 (451·26–572·18)	0·047
MIR	..	0·461 (0·415–0·518)	..	0·386 (0·336–0·432)	..	0·348 (0·299–0·390)	<0·001
DALYs lost	86 010 384	2062·74 (1949·53–2280·29)	101 951 696	1749·59 (1568·67–1830·82)	102 232 304	1554·02 (1373·94–1642·26)	<0·001
Mortality	4 660 449	117·25 (111·51–129·68)	5 684 970	98·53 (89·02–103·86)	5 874 182	88·41 (79·84–94·41)	<0·001

*p value for the difference in age-adjusted rates between 1990 and 2010 only.

Table 1: **Age-adjusted annual incidence and mortality rates (per 100 000 person-years), disability-adjusted life-years (DALYs) lost, prevalence (per 100 000 people), and mortality-to-incidence ratio (MIR) by age groups in high-income and low-income and middle-income countries, and globally in 1990, 2005, and 2010**

Feigen et al., Lancet 383:245-255, 2014, Table 1

Here's a table from the Lancet. Epidemiologists are expert in creating terrible tables. There is so much that is wrong with this table.

What was wrong with that?

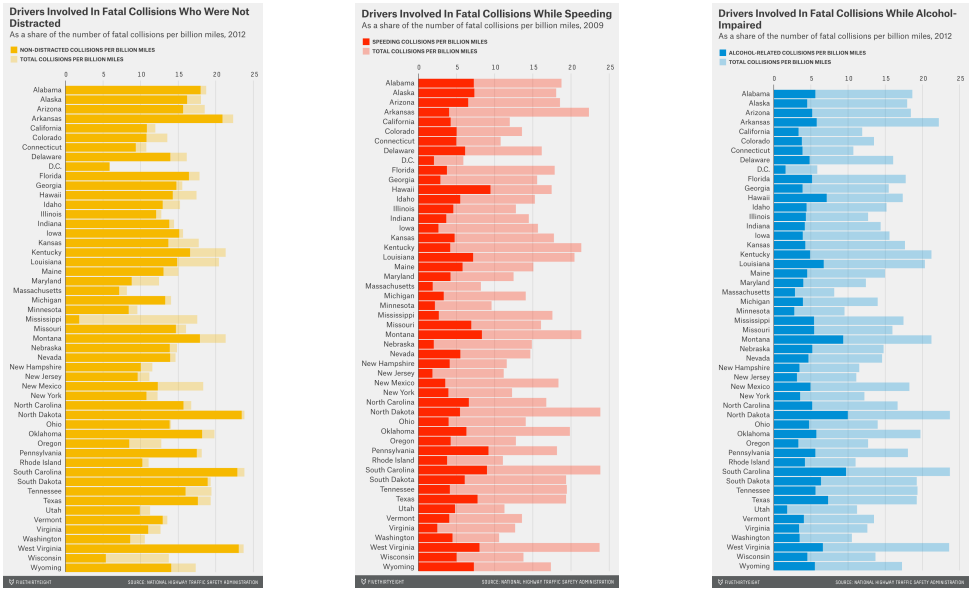
- Way too many digits.
- Numbers aren't aligned.
- Numbers to be compared aren't anywhere near each other.
- The interesting comparisons are horizontal rather than vertical.
- It would be much better as a multi-panel figure.

Here's a summary of what was wrong with that Lancet table.

It's interesting to note that it tends to be easier to compare numbers vertically rather than horizontally. So arrange the table so that the key comparisons are with numbers near each other, and ideally going up-and-down rather than side-to-side.

Even better would be to show these data as a graph. Seldom is one interested in the detailed quantitative values; rather, you'd be looking for qualitative differences which would be much easier to see in a graph.

One last example



fivethirtyeight.com/datalab/which-state-has-the-worst-drivers

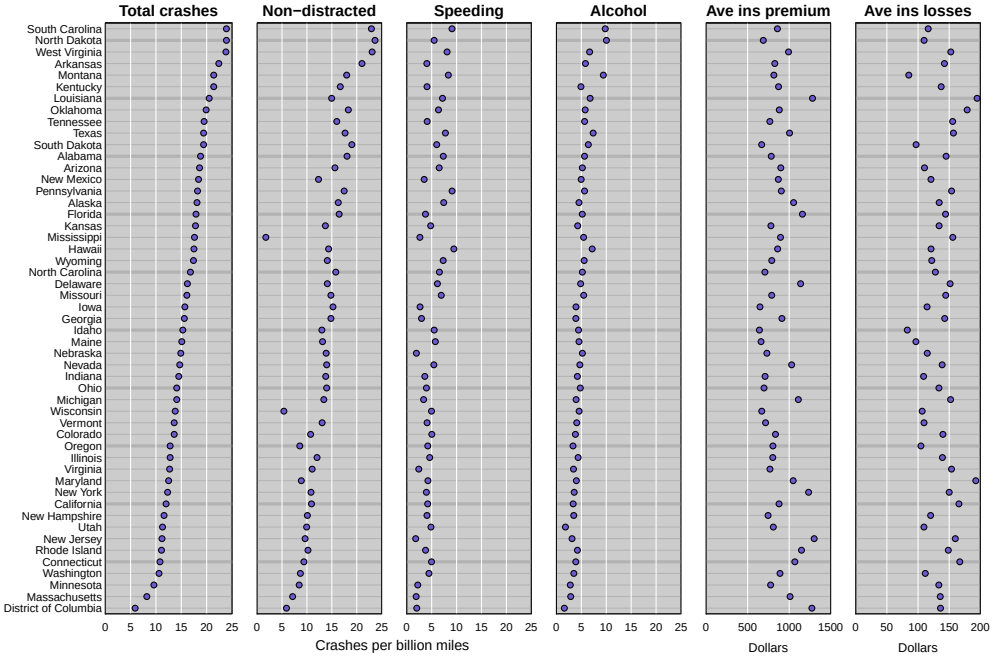
This is a set of graphs from the fivethirtyeight blog. These were shown in a single column, quite far apart from each other. There were actually five panels like this.

There is so much wrong with these graphs.

First, note the alphabetical order of the states. It would be much better to order the states by a key feature (like total crashes) and then apply that order to all of the figures.

There are a lot of interesting features in these data, but they are really hard to detect in these graphs.

An alternative

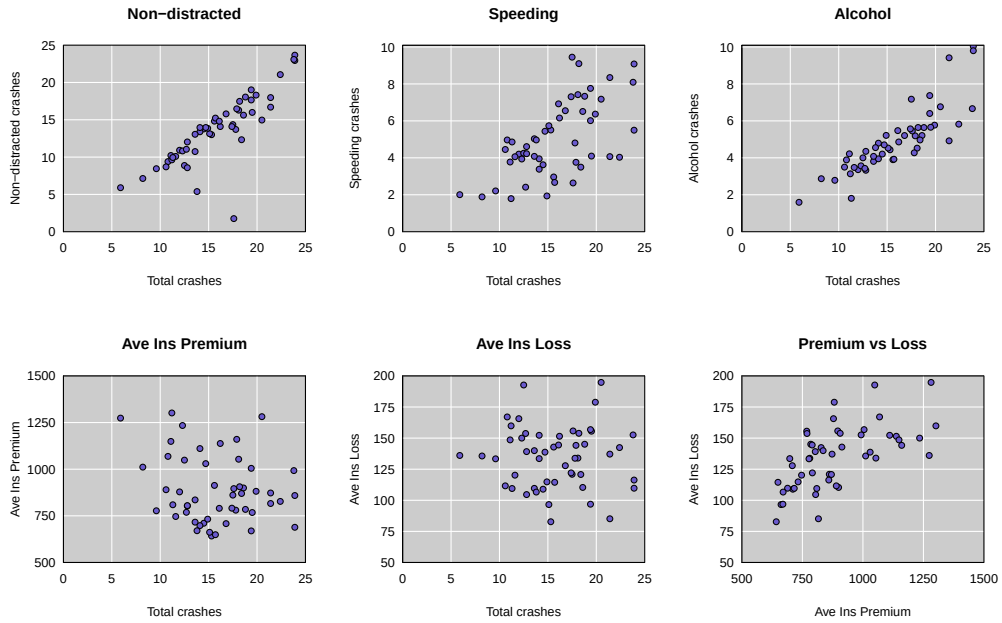


I would much prefer a set of side-by-side dot charts. These are a sort of visual table. You sort the states by the total number of crashes and then apply that to each of the other measured variable.

You can now more easily see the clear outliers of Mississippi and Wisconsin in the "non-distracted" crashes. Also for crashes due to speeding, there seem to be two groups of states.

You can clearly see that there are associations between total crashes and the number of crashes of different types, but that these are not associated with average insurance premiums or losses.

Scatterplots



And if you're interested in relationships among the outcomes, it's much better to study them directly with scatterplots. A disadvantage is that you can't see which states are which. But Mississippi and Wisconsin clearly stand out in the top-left panel, and you can more easily see the two groups of states in the top-center panel. And you can see the relationships (or lack) among variables more directly.

Summary I

- Show the data
- Avoid chart junk
- Consider taking logs and/or differences
- Put the things to be compared next to each other
- Use color to set things apart, but consider color blind folks
- Use position rather than angle or area to represent quantities

Summary II

- Align axes to ease comparisons
- Use common axis limits to ease comparisons
- Use labels rather than legends
- Sort on meaningful variables (not alphabetically)
- Must 0 be included in the axis limits?
- Use scatterplots to explore relationships

Inspirations

- Hadley Wickham
- Naomi Robbins
- Howard Wainer
- Andrew Gelman
- Dan Carr
- Edward Tufte

These are some people who have inspired my understanding of data visualization, and this lecture.

Further reading

- ER Tufte (1983) The visual display of quantitative information. Graphics Press.
- ER Tufte (1990) Envisioning information. Graphics Press.
- ER Tufte (1997) Visual explanations. Graphics Press.
- A Gelman, C Pasarica, R Dodhia (2002) Let's practice what we preach: Turning tables into graphs. The American Statistician 56:121-130
- NB Robbins (2004) Creating more effective graphs. Wiley
- Nature Methods columns: `bit.ly/points_of_view`

Here is some recommended reading.