DOI: 10.1177/0094306119880195
http://cs.sagepub.com

# REVIEW ESSAYS

## See It with Figures

John Levi Martin
*University of Chicago*
*jlmartin@uchicago.edu*

*Data Visualization: A Practical Introduction*, by **Kieran Healy.** Princeton, NJ: Princeton University Press, 2019. 272 pp. $40.00 paper. ISBN: 9780691181622.

The short story is that Kieran Healy's *Data Visualization: A Practical Introduction* is a gentle introduction to the effective display of social science data using the R package ggplot2. It is beautifully put together, achingly clear, and effective. It takes pains not to repeat classic works on principles of visualization like that of Edward Tufte (e.g., *The Visual Display of Quantitative Information*), more detailed manuals like that of Hadley Wickham (e.g., *Ggplot2*), or arguments about general approaches like that of Leland Wilkinson (*The Grammar of Graphics*)—all of which are, however, explicitly discussed. Instead, this book is for the person with general programming facility who cannot manage to pick up ggplot on his or her own, or who will make bad choices if s/he does. Since I fell precisely into this category, I had ordered the book before it was released. If you fall into this category, you probably have already done so too. If you haven't, you should. And even if you don't fall into this category, chances are you should have it by your side.

The secret to using R, my friend Adam Slez confided to me, is learning how to think like Hadley Wickham. But Wickham has a famously "tidy" way of thinking—a place for everything, and everything in its place—and many of us find that trying to think like him hurts. Healy translates the principles of Wickham's system for us less-tidy thinkers. (This includes his making the otherwise baffling hieroglyphic of the "pipe" operator clear, which fits the ggplot way of thinking, even though it cuts against the overall R feel.) Further, while the book has exercises and points one to just the right places for further development, one doesn't actually have to do the exercises to learn from the book (in contrast to texts that really are just a set of puzzles that force you to teach yourself). You can use this book to really master ggplot, or you can read it *as a book*—that is, as opposed to having to sit by the computer trying each part out as you read—and return to it for practical guidance when necessary.

What is the goal of Healy's book? Its subtitle should be taken at face value: this is a *practical introduction*. Regarding the practical aspect: this is not the place for a detailed discussion of a philosophy of visualization. The initial chapter on general principles is excellent, but brief—Healy isn't interested in replicating ideas that are found elsewhere, and so he concentrates on the issues that are likely to be relevant to those using ggplot for social science work.

Regarding the introductory aspect: when it comes to the practice of visual display, there is a difference between the use of visualizations for exploration or for the production of new results (or hypotheses), and their use for communication of known results. Healy deliberately stops where these two diverge. This is an introduction to low-dimensionality data visualization: how one examines one or two variables at a time, perhaps dividing a data set by a third as one does so. (It also has an extremely effective introduction to geographical mapping.) The book, then, sets out to be a practical introduction, and it is a resounding success in these terms. If you want to learn these principles, the book teaches you painlessly. If you want to find code in the book to

copy and adapt *without* learning all the principles, you can do that too.

It might then seem perverse to discuss what a book deliberately does *not* do, but it makes sense in this venue: *Contemporary Sociology* does not usually review text books, but it seeks to be a place for the discussion of the state of the field. In keeping with this spirit, I propose to consider not only the degree of success that Healy has had in achieving his goal, but some of the externalities of that achievement. While I think Healy's book will lead to a net improvement—more people will have a better idea of what they are doing, and they will do it better—there are downsides to the increasing ggplottification of the sociological mind, which Healy's book will no doubt accelerate.

There might be many consequences of this shift, most welcome, some pernicious.[1] Here I want to suggest that ggplot has led to a real confusion over the distinction between descriptive and explanatory statistics. Of course, this isn't ggplot itself, except insofar as it facilitates people making certain types of graphs easily that they otherwise might not make. Blaming ggplot for this is like blaming McDonald's for obesity—it just makes it easier for us to do something we want to do, at least once we've tried it. But the ease with which one can plot bivariate relations between continuous variables (or a set of such relations, if the data are broken down as groups) has led to an inundation of papers with charts like Figure 1, which demonstrates a theoretically interesting relation between $x$ and $y$—but neither shows the raw data nor tests any model. It may be, as some have argued, that the distinction between description and explanation is not always a helpful one. But if we are to have explanations that come in the form of models, we need to distinguish these from unprocessed data.

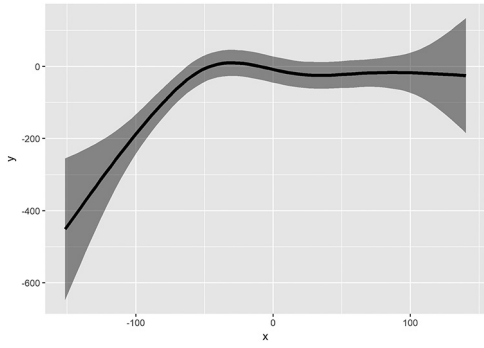


**Figure 1.** "A Very Interesting Relation."

It is not that there is anything wrong with the use of these smoothing models, which have long been used by excellent methodologists outside of sociology. However, the ease with which ggplot allows the construction of such figures has led to some perverse consequences. We have a set of data (which I have simulated[2]), and we run the command to make the same type of nice visualization that we saw in Figure 1, and we find undeniable evidence of absolutely no relation (see Figure 2). Yet, if we simply drop the last observation arbitrarily and rerun the command, we find something extremely interesting (Figure 3)! Huh? Well, the reason is that, by default, ggplot automatically stops using the more computationally intense

[1] In addition to the non-trivial environmental impact of the gray background default (toner is chock-full of deadly VOCs!), two minor irritations come from ggplot's defaults. One has been the provision of charts with teeny tiny axis labels—preserved in my figures for your edification and amusement—and the other has been the tendency to misuse valuable journal space by putting legends outside the main frame of a figure, reducing the available space for the more important presentation of results. (One can put a legend inside by specifying coordinates, but, as far as I know, there is no automated part of legend.position() that searches for appropriate blank space.) An intended and, I think, generally valuable consequence of the reliance on ggplot has been the replacement of multiple lines on the same plot with facets (smaller independent panels). At the same time, the ease of such presentation has also generated disastrous examples of the presentation of model results, where all variables are plotted on the same irrelevant scale. We will actually see an example of this in Healy's book below. We can't fault ggplot—every program has defaults—but it may unsettle the emerging ggplorthopraxy to choose to decouple the $y$ axes' scalings.

[2] All code is available at http://home.uchicago.edu/~jlmartin/Programs.htm. Around one in four or five runs of this simulation comes out this stark.

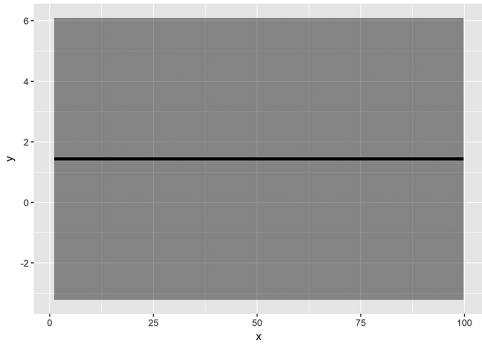


**Figure 2.** "An Uninteresting Relation."

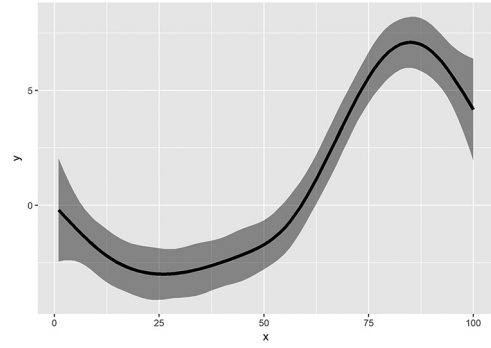


**Figure 4.** "A Stronger Relation."

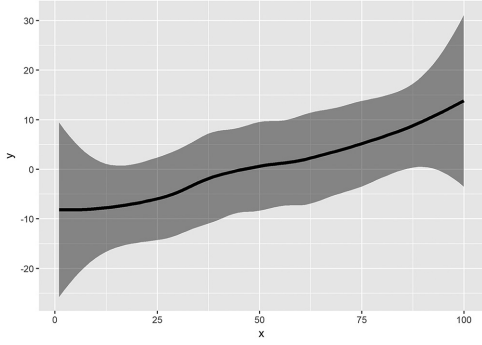


**Figure 3.** "A More Interesting Relation."

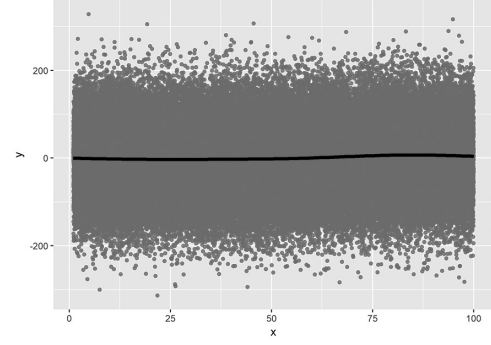


**Figure 5.** "A Strong Relation?"

loess smoothing once one has 1000 cases and instead switches to generalized additive models using a cubic spline.

Figure 4 displays a stronger and crisper relationship, one that seems to have different implications from Figure 3. Someone who took Figure 3 as a guide and wanted to hazard a guess as to what value of $x$ would maximize $y$ would propose an $x$ of 100, but someone looking at Figure 4 would conclude that one expects $y$ to be maximized at around 80. Actually, Figure 4 comes from running the same command on data coming from the same (true) generating function that was used to make the data in Figure 3, but with many more observations. And in fact, the ribbon drawn in Figure 4 fits entirely within the confidence intervals of Figure 3! Statistics isn't fooled, even if your eye was.

But now let me draw the exact same data a bit differently (Figure 5). The confidence intervals are still being drawn—but they are so close together that you can't even see them. Almost no cases fall within this interval, it should be noted (not that they should, but still . . . ). The seemingly strong relationship now appears completely trivial. What is going on? In the—very popular—representation of Figure 4, the bivariate result is presented as a *model*, even though we are looking for a purely descriptive relationship. That is, we allow the GAM to fit the relationship however it can, as opposed to testing an a priori theory. Our eye tends to read the narrow ribbon as suggesting that most of the *cases* are going to be in this ribbon, when all ggplot is trying to tell us is that—if the model is correct—the conditional means *in the population* should be within these lines. It tells us nothing about the conditional *variance*. What the second representation is telling us is that if you are wondering

whether $x$ is an important predictor of $y$, the answer is no. The conditional variance of $y$ is so high that the fact that the central tendency moves a bit up and down probably doesn't matter for most purposes. By asking ggplot to print the raw data as well as the smoothed line, it rescales to the actual limits of $y$, as opposed to the limits of the conditional mean of $y$ on $x$.

I propose, then, that this common use of ggplot (and any other program that does this) to make ribbons for predicted means has a general tendency to confuse people as to the differences between three things: first, actually visualizing the data, second, examining predictions from models, and third, examining the confidence of imputation of a prediction line to the population. Healy (p. 134) points out that these smoothers are in fact models and directs us to consider the ways in which they are fit; even more, he urges using *more than one* and comparing the implications. (And, I think it goes without saying, he did not simply mean, ''choose the one you like best''!) He gives example code to make this possible for others. But he passes quickly over the potential perversities of such visualizations.

For example, it is very common to see results graphed in the way shown in Figure 6. The idea here is that we are seeing a prediction for our best guess as to what the dependent variable should be as the predictor of interest changes. These beautiful diagrams are often taken to be an important way of presenting results, far superior to a simple model coefficient and standard error. It is true that for nonlinear models, or for linear models with interactions, we may find it difficult to interpret the implications of our model simply from looking at a table of results. But the illusion that we have visual inspection of the data-space can lead us to place more weight on claims about particular values of our predictor than is warranted.[3] We have, right now, an odd inconsistency in the discipline: we treat p-values are some sort of unreliable voodoo but rely on confidence intervals uncritically,

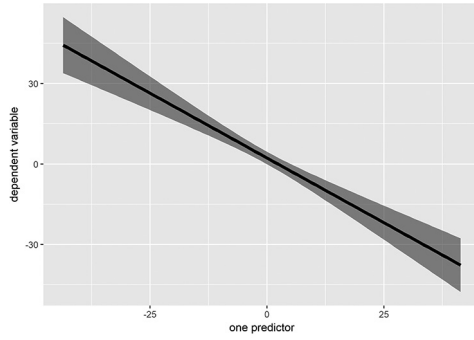


**Figure 6.** ''Confidence in Our Predictions.''

even though, in most cases, these are more or less the same thing—they're a lot about sample size and a bit about fit, and they don't, by themselves, handle our problem of getting the model right.

There is something deceptive about feeling that one is *seeing*—since what we are seeing is really only our coefficient and standard error. And there is, I posit, an indefensible security that we feel when we see tight confidence intervals around a nonhorizontal line. Indeed, Healy (p. 142) even makes the slip of seeming to equate plotting a confidence interval (on the one hand) and showing one's subjective ''degree of confidence'' in the result (on the other hand). For the case graphed in Figure 6, we would feel quite confident asserting not only that this predictor was strongly associated with a decrease in the dependent variable, but saying things like ''those at -40 on the predictor have high values on the dependent variable.''

But, by construction, for these simulated data, the predictor really has a strongly *positive* direct relation with the dependent variable! Not only that, but the *bivariate* (descriptive) relation is *also* positive! The dependent variable is a simple linear function of the predictor of interest as well as two other variables. Figure 6 is the slope from a model that omits one of the three variables (and it is a rare day when we can be confident that we have *not* omitted a relevant predictor!). Figure 7 includes that slope and confidence interval (or ''misplaced confidence interval'') from Figure 6, adds the data, and also draws a dashed line indicating the *true*

[3] I have pointed to the ease with which excellent methodologists can make confident statements about non-existent areas of their data space in *Thinking Through Statistics*.

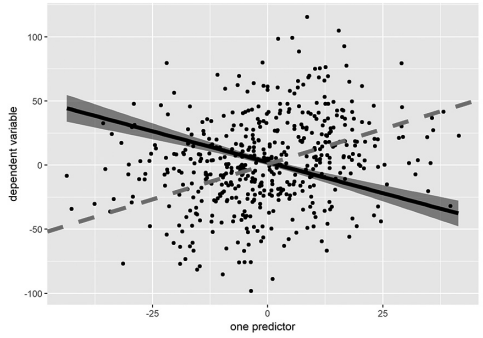


**Figure 7.** "Misplaced Confidence."

relation. The true slope is positive ($b$ = .88) and the bivariate slope is as well ($b$ =.57), but the slope from the misspecified model is negative ($b$ = -1.12). All relations are highly statistically significant, even though there are only 500 cases; this isn't about everything being noise. But I don't think many people looking just at the first version would ever imagine such a reversal of the actual relations as a possibility.

Healy, understanding these sorts of problems, gives the reader a number of important and clear principles. He warns us about the complexity of model interpretation in general (p. 141) and emphasizes that visualization is not a substitute for really understanding the underlying model.[4] Most important, he also notes that we should show our data when possible, comparing the predicted values to the original points (pp. 142–43). Indeed, it should be emphasized that the new techniques of shifting from tables to plots (and the easy layer-based approach of R and ggplot) have made it much easier to actually show data, and Healy *always* shows the original data (and *not* just predictions) where it is practicable.

Yet Healy rarely walks through the problems that can come from reliance on visual displays. An important exception is his treatment of geographic data. Without bemoaning the proliferation of pointless shadings of the continental 48 that fill journal submissions, Healy (p. 194) insists that we ask ourselves before making such maps whether our data really are spatial. Even when they are, Healy (p. 197) demonstrates that organized facet plots can be much more enlightening at showing change than a set of maps (the book is worth it for this one wonderful example alone!). Even more, Healy points to the problems that can arise when making comparisons across geographic units that contain very different numbers of persons (p. 188). But Healy is less interested in examining the limits of the bivariate relation as line-plus-confidence-interval, and he does not discuss ways of displaying data with a limited number of response categories, especially dichotomies, when the inability to do a straightforward scatterplot is often used as a justification for replacing description with model results.[5]

[4] That said, there is one place (pp. 147–48) where Healy himself demonstrates the problems that can come from simply feeding model results into a pipeline. The question is how to take results from an R model (lm, in this case) and display them meaningfully. The problem here is that the independent variables have very different scalings, which means that the coefficient sizes do not accord with their statistical significance. But the grinder he feeds them to doesn't know that. The first step rounds all numbers in the table the same way, so that all p values are "0," and the second step produces a dot plot of estimates from coefficients that makes it appear that the highly significant variables that happen to have low absolute values of their coefficients are in fact zero. Still, not as bad as me reversing the summation and product operators in my equation for latent class analysis in *Thinking Through Statistics*! We all make mistakes.

[5] He does, however, spend some time analyzing weaknesses of less sociological figures, the sort that might be in a newspaper article or a business powerpoint. He (pp. 216–17) excoriates two-vertical-axis plots with a passion that I cannot fathom, seemingly implying that because one *could* try to minimize the variation in one of the two by a dishonest scaling, there can be no honest scaling; and he proposes fixing the beginning of both series at 100 and comparing proportional changes. This sort of solution might be appropriate for ratio data (those with a true zero), but not even always then. I would imagine that climate-change deniers would be delighted to examine trend graphs that used this standardization to compare the simultaneous changes in the $CO_2$ load in the atmosphere to the rise in mean average temperature (degrees Kelvin)!

This practical introduction, then, is one that is oriented to graphing low-dimensional distributions and model results. But this notion of graphing bivariate relations as predictions-and-data-points becomes more and more difficult to follow as the number of dimensions increases. There are two ways of responding to this problem. The first, one that many statisticians will favor and that, I think, Healy implicitly leans toward, is to be extremely suspicious of multivariate models and to shake one's head sadly when results from highly leveraged models are interpreted. There is much to be said for this response. It is not only statistically conservative, but it fits a way of doing social science—one that is oriented to crisper questions, more convenient research designs, and proof via visual inspection—that is widely appreciated *outside* of sociology. (If you want to publish in *Science*, say, bivariate relations are often preferred if possible!) However, sociology has a way of thinking, going back to Durkheim's *Suicide*, in which contenders battle it out by adjudicating between possible interpretations of *one* relation between variables by *adding* another variable. We could split the sample, but when the number of contenders gets large, or each brings many variables, then this becomes impracticable. Moving sociology away from this paradigm will not be easy. If Healy thinks this is the way to go, he will need to bang a drum more vigorously than he has done here.

Pushing the field in that direction probably cannot hurt; even if we don't really go there, we need repeated caution about multivariate models. But it seems that it might be a lost opportunity if we were to consider only data sets possessing such elegant simplicity. There is another approach, which is to use visualization to try to better understand the nature of our high-dimensional data *before* attempting to fit models. There is valuable work coming from the computer science realm, broadly understood, on ways of constructing such visualizations, including ones that end up printable in black and white on regular paper. These have their own difficulties and obscurities (for example, the ''perplexity'' parameter for the t-SNE algorithm), but this is probably not the time for sociology to do an about-face and flee multivariate data structures. What we now need is a practical introduction to high-dimensionality data visualization, one that builds on the work here on low-dimensionality data visualization. Healy seems like just the person to do it.

## What *Should* Historical Sociologists Do All Day? *Starving the Beast*, the Reagan Tax Cuts, and Modes of Historical Explanation

Elizabeth Popp Berman[1]
*University of Michigan*
*epberman@umich.edu*

Monica Prasad, along with collaborators like Isaac Martin and Ajay Mehrotra (e.g., Martin, Mehrotra, and Prasad 2009), has made fiscal sociology—the sociology of taxation—a thriving part of the discipline. Her first book showed how different national patterns of taxation help explain the variable strength of neoliberalism across nations (Prasad 2006). Her second identified progressive taxation as key to producing both democratized credit and a weak welfare state in the United States (Prasad 2012). More generally, her work has been critical for helping us understand how political

*Starving the Beast: Ronald Reagan and the Tax Cut Revolution*, by **Monica Prasad.** New York: Russell Sage Foundation, 2018. 338 pp. $35.00 paper. ISBN: 9780871546920.

[1] Thanks to Dan Hirschman, Jeremy Levine, and Isaac Martin for comments on previous versions of this essay, and to Damon Mayrl and Nick Wilson both for comments and for sharing multiple drafts of their paper, ''What Do Historical Sociologists Do All Day? Methodological Architectures in Historical Sociology.''