Harvard Data Science Review • Issue 3.3, Summer 2021

'The Art of Graphics Isn't Hard to Master' (pace Elizabeth Bishop)

Antony Unwin¹

¹Institute of Mathematics, University of Augsburg, Augsburg, Germany

Published on: Jul 30, 2021

DOI: <https://doi.org/10.1162/99608f92.01df2b32>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Graphics is an important part of statistical analysis and complements modeling. It is encouraging to see researchers attempting to formalize part of this relationship. The paper proposes ideas for improving graphical tools for users of business graphics systems, people with expertise in other fields but with less statistical knowledge. Jessica Hullman investigates the display of uncertainty in graphics and recommends more to be done in that direction. Andrew Gelman is, amongst his many interests, an active Bayesian statistician, and he emphasizes the role a Bayesian approach could play. Both have much to contribute and it is positive to have experts with different backgrounds cooperating.

The title of the paper refers to "interactive exploratory data analysis." The term 'interactive' means different things to different people, and the authors take a relatively limited view. They are aiming to assist less statistically-experienced analysts, so this is reasonable, but they would have benefited from having worked with more advanced interactive tools and from the book *Interactive Graphics for Data Analysis* (Theus et al, 2008).

Data analysis has always needed initial exploratory work to understand the data and to sort out whatever unexpected features may arise. This has become more important as the data sets that can be handled have become bigger, especially in terms of numbers of variables. Hullman and Gelman play down this initial stage, although it is where graphics can be particularly valuable. They could also emphasize more the value of interacting with experts who have background knowledge of the data. Interactive EDA means interacting with people as well as with graphics, to learn about what unexpected features may imply and what can be done about them. There is an illustration in the article's first example, a real data set of property sales in Ames, Iowa. There are three obvious outliers in the scatter plot and the supporting article for the data set explains that these are partial sales that should be removed. Indeed, the 18% of sales that are not classified as 'Normal' in the variable 'Sale Condition' might be removed. Interaction with experts would also be useful to decide which variables to include in a deeper analysis. The dataset is a moderately large one with just under 3000 cases and around 80 variables. There is no mention of the number of variables or what would be recommended to help select the ones to include in an analysis.

Hullman and Gelman concentrate on graphical inference, assuming that initial EDA has already been carried out. They point out the dangers of over-interpreting graphical results and suggest two improvements in software: systematic structuring of information using a Bayesian approach and representation of uncertainty. Both are promising proposals and worth pursuing. Their value will depend on their being part of an established practice of sound data analysis. That means cooperating with experts with knowledge of the subject matter of the data, drawing careful and informative graphics, and checking results in all manner of ways, not just with formal statistical models, but using other data and other variables.

There is not space in Hullman and Gelman's article to discuss the four data sets used in the examples in detail. Amongst other things, this means it is not always clear whether data are in some sense real or have been

simulated. It would have been better if the authors had concentrated on one example and explained it in more depth.

Adding inference tools to graphics requires that the graphics are good. The paper's Figure 1 shows displays of the Ames Housing data set using three different softwares. (Experts in using those systems might have drawn different displays.) Figure 1c has several weaknesses that probably do not do the software justice. The authors (2021, this issue) say of Figure 2a that the "Trellis plot of housing sale prices by neighborhood might invoke comparisons to a normal or log-normal distribution." That may be true, but their second point that it provides "a visual check for a main effect of neighborhood" is more important. The graphics are too small to read directly, but the quality is good and it is possible to zoom in. There are 20 separated bars for each neighborhood, and they are labelled from 18K to 414K in 16 steps of 18K and three steps of 36K (one at the beginning, two near the end). Would it not have been easier and more sensible to draw a standard histogram with equal bin widths of 20K or 25K and no gaps between the bars? How the authors recommend comparing bins of equal drawn widths for unequal actual widths with normal or log-normal distributions would be interesting to know. Figure 2c is a residual plot from the same data set and is dominated by the three outlying points. The increasing variability with increasing price is downplayed (and not referred to). Of Figure 2d the authors write, "Trellis plot of sale price by lot configuration and neighborhood enables, among other effects, a visual check for an interaction between lot configuration and neighborhood." This is puzzling, if not misleading, as the plot is actually one of the total price of all sales by lot configuration and neighborhood. The bar for an Inside configuration in College Creek represents 188 sales with a total value of \$38.6M, while the bar beside it represents 1 sale of \$220k for an FR3 configuration. Perhaps this is just an example of the kind of unrecognized aggregation the authors warn against.

Several of the graphics use small multiples to make comparisons. This is an excellent idea and can work well. The graphics to be compared should have the same scales and sizes and be properly aligned. With reliable software, these conditions are commonly met by default. So it is unsettling that in Figure 4, the plots are not always vertically aligned and the scales are mostly different. In Figure 5 the vertical alignment is fine, but the scales are different. In Figure 8e, the plots are not precisely horizontally aligned and the spaces between the plots are unequal. Some of these points may seem minor, but they are unnerving, just like the lower limit of the vertical scales in Figures 8b, c, d: why is -50 drawn so big in each of them? Producing good graphics nowadays is easier than it used to be, but you still have to do the work, check the defaults, make sure the software is doing what is required.

Figure 7 returns to the Ames Housing data and is an enlarged, but cropped, version of Figure 2d with "standard uncertainty intervals" added. How to interpret the medians of total sale prices by lot configuration and what use they might be is unclear. There may be good reason, but the authors do not explain. That the intervals are huge is hardly surprising, given the small numbers of bars, although readers ought to be told what the term "standard uncertainty interval" means here. The graphic does show the technical possibility the authors want to display,

but it would be more convincing if we knew why it made sense to plot this statistic and those intervals, and knew what the intervals were.

When adding a display of uncertainty to a graphic you have to define uncertainty and explain what is shown. Sometimes there are many alternatives. The following figure shows two displays for the four Ames neighborhoods mentioned in the article. The plot on the left shows boxplots for sales prices, where the width of each box is proportional to the square root of the number of observations. Like Figure 2a in the article it suggests that prices are generally higher in two of the neighborhoods and lower in the other two. The plot on the right is an ordered spineplot showing the proportion of one story houses, a possible explanatory variable. The neighborhoods have been ordered by those proportions and there is little relationship between the variations in the two plots. Different comparisons could be made in each plot and different intervals would be appropriate. What would the authors recommend?

Figure 1. Sale prices and proportions of single story houses in four neighborhoods of Ames, Iowa.

Looking at two or more plots simultaneously increases cognitive load. Adding uncertainty displays to the individual plots would increase it more. Interactive software packages like Data Desk and JMP that include linking between windows lessen the load and support exploration across graphics. In the Ames housing example you could add displays of the variables recording the overall quality and condition of the properties and, possibly, several others. Using many graphics at once is what truly interactive EDA is all about and considering several variables at once is related to what the article describes as the first step of Bayesian Data Analysis: "Setting up a full probability model—a joint probability distribution for all observable and unobservable quantities in a problem." How easy would that be here?

Uncertainty displays can definitely be of value for a single graphic display and might encourage users to study a display in more depth. As Battle and Heer (2019) point out in their review article: "The observed cadence of

analyses is surprisingly slow compared to popular assumptions from the database community." Others who do not use graphics much may have similar misconceptions. Lower time spent on a task is not always an ideal criterion, as Hullman and Gelman remind us. Graphics need time and effort from both their designers and their readers. Designers of a graphic may think it can be understood instantly, but that may not be what users experience. Even graphics that have a "signal so large that it 'hits you between the eyes'" (to quote the authors) may offer additional information that could be identified with a thorough study. Some educational authorities encourage the teaching of Close Reading in schools (Close reading, 2021), Close Viewing should be encouraged as well.

Supporting non-expert users in understanding graphics better and making better use of graphics is a worthy aim. More sophisticated software tools can play a part, but ensuring that domain knowledge is considered, that good graphics are drawn, and that users know how to interpret those graphics come first. There is a great deal of good advice on how to draw graphics and considerably less on how to interpret graphics. Mary Eleanor Spear (1969) put it well over 50 years ago: "there is quite a difference between simply looking at a chart and seeing it."

I applaud the authors' efforts and many of their ideas, but they should build on a sounder basis.

Disclosure Statement

Antony Unwin has no financial or non-financial disclosures to share for this article.

References

Battle, L. & Heer, J. (2019). Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. *Computer Graphics Forum*, 38(3), 145–159. <https://doi.org/10.1111/cgf.13678>

Close reading. (2021). In *Wikipedia*. (2021). https://en.wikipedia.org/wiki/Close_reading

Spear, M. (1969). *Practical charting techniques*. McGraw Hill

Theus, M. & Urbanek, S. (2008). *Interactive graphics for data analysis*. London: CRC Press

©2021 Antony Unwin. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.