Data visualization: practices

Rob Davies

2023-11-27

## 1 PSYC401: Data visualization – practices

* My name is Dr Rob Davies, I am an expert in communication, individual differences, and methods

|  |
| --- |
| Tip |
| **Ask me anything**:   * questions during class in person or anonymously through slido; * all other questions on discussion forum |

## 2 Introduction: Data visualization – perspectives

* In this *PSYC401* class, we look at *practices* or how we get visualization done
* In the linked *PSYC403* class, we focus on *perspectives* or how to think about visualization

## 3 Our lesson plan

1. Identify your goals
2. Think about your audience
3. Develop reflectively
4. Implement good practice

## 4 Our learning objectives: — what are we learning about?

We are working together to help you:

1. *Goals* — Formulate questions you can ask yourself to help you to work effectively
2. *Audience* — Understand the psychological factors that affect your impact
3. *Development* — Work reflectively through a development process
4. *Implement* — Produce visualizations in line with best practice

## 5 Our assessment targets: — how do you know if you have learned?

We are working together so you can:

1. *Goals* — Identify a set of targets for a development process in your professional teams
2. *Audience* — Explain what you need to do to make a visualization effective
3. *Development* — Locate yourself within the stages of the development process
4. *Implement* — Produce visualizations that look good and are useful

## 6 What are our goals – Questions to help you to work effectively

|  |
| --- |
| Tip |
| * We begin by thinking about the questions you will ask yourself when you need to decide *what you will do* * We build, here, on the insights developed by A. Gelman & Unwin (2013). |

## 7 What are our goals?

* Why don’t we just use the *good enough* easy to produce plots in Excel? *Why bother?*
* Why don’t we just produce a summary table? *Why bother?*
* Are we engaged in making beautiful graphics or informative displays or both? *What are we doing?*
* In *PSYC403*, we look at *Perspectives: the context and history of thinking about visualization*

## 8 Visualize to enable comparison

|  |
| --- |
| Figure 1: Scatterplot of the relation between reaction time and days in the sleepstudy data |

|  |
| --- |
| Figure 2: The relation between reaction time and days: here, we plot the data for each participant separately |

## 9 What are our goals? — What are our jobs?

* *Data visualization* workers: we may aim to get and keep the attention of our audience, to tell a story, to persuade our viewers
* *Data analysis* workers: we may aim to enable our audience to understand our data, our findings, and to discover more for themselves

## 10 What are our goals? — Where or when are we in our process?

* Sometimes in a *workflow*, we are quickly sketching draft visualizations: exploring, for ourselves, or with others, what we can see in our data
* Sometimes, we are ready to present our visualization to a wider audience: we aim to share a polished visual object

## 11 What are our goals? — Discovery

*Discovery goals*

1. Do we need an overview? – To get a sense of what is in the data, and to check our assumptions
2. Are we looking for the unexpected? – Comparing groups to check for variability, exploring data open to surprises

## 12 What are our goals? — Communication

*Communication goals*

1. What do we need our audience to understand?
2. What story are we telling?
3. Do we need to attract attention or stimulate interest?

## 13 Think about your audience – An evidence based account of what works

|  |
| --- |
| Tip |
| * We will produce more effective visualizations if we think about how our audience *sees*, and what they *expect* (Franconeri et al., 2021) * Check out the *PSYC403 Perspectives* lecture for more in-depth explanation; here, I present a selective summary |

## 14 The human visual system is highly developed

* Your audience can look at your visualization
* And quickly and easily extract statistical information from what you show
* You look at a scatterplot and see the minimum, maximum and mean heights of the points

|  |
| --- |
| Franconeri et al. (2021) Fig. 2 |

## 15 Communicating uncertainty is critical

* As scientists, we think about uncertainty all the time
* We quantify and typically show uncertainty over estimates e.g. average differences
* We should also show *and think about* outcome variability

|  |
| --- |
| Zhang et al. (2023): The difference between uncertainty over estimates and uncertainty over the predictability of outcomes |

## 16 Consider accessibility from the start

* The first row shows a scatterplot encoded with two colors, green and orange
* People with typical vision can see that the green dots have a steep positive correlation and the orange dots make a flat line
* We use colour blindness friendly colour palettes

|  |
| --- |
| Franconeri et al. (2021) Fig.5 |

## 17 Development – Work reflectively through a development process

|  |
| --- |
| Tip |
| * Your first question is always going to be: (why) do we need to make a plot? * Your answer will evolve through a development process that will gradually reveal the characteristics of your data |

## 18 The benefits of investing in the development process

* Identifying your goals enables you to understand what you are doing and why
* Through the development process, you may create different versions — *iterations* — of a plot
* This iterative work benefits both you and your audience (A. Gelman et al., 2002; Kastellec & Leoni, 2007)

|  |
| --- |
| Tip |
| * As you iterate, **reflect** on what your *goals* are, what your *audience* needs and expects, and *how* each plot version moves you closer to effective discovery or communication * This reflection uncovers what is interesting, useful and beautiful about your data |

## 19 Scientific thinking and data visualisation

We can use text and tables to communicate specific values but visualizations help us to:

* stimulate thinking
* discover what is unexpected
* communicate scale and complexity
* make comparisons to show how results *vary*
* display uncertainty about estimates

## 20 Anscombe (1973): visualizations show data features quickly and vividly

### 20.1 data columns

x1 x2 x3 x4 y1 y2 y3 y4  
1 10 10 10 8 8.04 9.14 7.46 6.58  
2 8 8 8 8 6.95 8.14 6.77 5.76  
3 13 13 13 8 7.58 8.74 12.74 7.71  
4 9 9 9 8 8.81 8.77 7.11 8.84  
5 11 11 11 8 8.33 9.26 7.81 8.47  
6 14 14 14 8 9.96 8.10 8.84 7.04

### 20.2 x-variables

x1 x2 x3 x4   
 Min. : 4.0 Min. : 4.0 Min. : 4.0 Min. : 8   
 1st Qu.: 6.5 1st Qu.: 6.5 1st Qu.: 6.5 1st Qu.: 8   
 Median : 9.0 Median : 9.0 Median : 9.0 Median : 8   
 Mean : 9.0 Mean : 9.0 Mean : 9.0 Mean : 9   
 3rd Qu.:11.5 3rd Qu.:11.5 3rd Qu.:11.5 3rd Qu.: 8   
 Max. :14.0 Max. :14.0 Max. :14.0 Max. :19

### 20.3 y-variables

## 21 Anscombe (1973): visualizations show data features quickly and vividly

|  |
| --- |
| Figure 3: All 4 of the Anscombe (1973) x,y datasets are identical when examined using summary statistics but we see how they vary when we use scatterplots to visualize them |

## 22 Matejka & Fitzmaurice (2017) give us the Datasaurus dozen

|  |
| --- |
| Figure 4: All 12 Matejka & Fitzmaurice (2017) x,y datasets (via jumpingrivers (n.d.)) have the same mean and standard deviation summary statistics but we only understand how the data are structured when we plot them and can look at the structure |

## 23 Develop visualizations to discover and communicate variability in outcomes

|  |
| --- |
| Figure 5: In this plot we show data on the impact of sleep deprivation on reaction time, from Belenky et al. (2003; via Bates et al., 2015). We can see how reaction time slows with increasing deprivation on average (grey line) but that the rate of slowing varies between individuals |

## 24 Reflect on kinds of uncertainty

* Scientists are often faced with the challenge of conveying uncertainty to their audiences (Hofman et al., 2020):

1. *Inferential uncertainty* — the degree to which a particular summary statistic (e.g., a population mean) is known to the scientist
2. *Outcome uncertainty* — how much individual outcomes vary (e.g., around the mean, regardless of how well it has been estimated)

* Inferential uncertainty can be reduced by collecting and analyzing more data, whereas *outcome uncertainty cannot*

## 25 As we work, reflect on the challenges of visualizing uncertainty

* The process through which we understand the world is characterized by assumptions, limitations, extrapolations, and generalizations, and this brings uncertainty (Van Der Bles et al., 2019)
* We often face the challenge of communicating this

|  |
| --- |
| Zhang et al. (2023): The difference between uncertainty over estimates and uncertainty over the predictability of outcomes |

## 26 The challenges of uncertainty

* Non-expert people will tend to overstate the impact of interventions and understate the variability of outcomes
* when they see visualizations like error bars that show
* mean and standard error values, that focus on inferential uncertainty (Hofman et al., 2020)
* Expert scientists also overestimate the impact of interventions when they see standard visualizations that focus on inferential uncertainty: *the illusion of predictability*
* We can stimulate more accurate understanding if we show outcome variability (Zhang et al., 2023)

## 27 Variation and uncertainty — the importance, the challenges

Vasishth & Gelman (2021):

The most difficult idea to digest in data analysis is that conclusions based on data are almost always uncertain, regardless of whether the outcome of the statistical test is statistically significant or not

## 28 Variation and uncertainty — the importance, the challenges

a. Gelman (2015):

We must move beyond the idea that effects are ‘there’ or not and the idea that the goal of a study is to reject a null hypothesis. As many observers have noted, these attitudes lead to trouble because they deny the variation inherent in real social phenomena, and they deny the uncertainty inherent in statistical inference

## 29 We use visualizations to help us to see and understand the variation and the uncertainty in our data

* Results will vary: we should expect changes over time, or differences between individuals or between groups
* Knowledge is uncertain: outcomes will vary even when the average effect is precisely estimated
* We have the responsibility to accept and to express this uncertainty

## 30 Implement – Produce visualizations in line with best practice

|  |
| --- |
| Tip |
| * We combine our creative thinking with the flexibility of the Grammar of Graphics to produce effective plots |

## 31 {ggplot2} means: the *Grammar of Graphics* Plot 2

* When we use the {ggplot2} to draw plots, we are using tools developed with a philosophy of visualization in mind (Wickham, 2010; Wilkinson, 2013): *The Grammar of Graphics*
* A grammar is a system of rules that allows people to collaborate and individuals to create
* We do not need to think about the grammar when we produce visualizations
* But it will help you to know that when we puzzle over *how* we do things, there are always reasons *why* we do things

## 32 A simple plot has many elements

* data and aesthetic mappings
* statistical transformations
* geometric objects
* scales

|  |
| --- |
| Figure 6: A scatterplot showing the potential association between health literacy and vocabulary |

## 33 We begin with data: here, from the Health Comprehension project

# A tibble: 6 × 12  
 participant\_ID mean.acc mean.self study AGE SHIPLEY HLVA FACTOR3 QRITOTAL  
 <fct> <dbl> <dbl> <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 studyone.1 0.49 7.96 studyo… 34 33 7 53 11  
2 studyone.10 0.85 7.28 studyo… 25 33 7 60 11  
3 studyone.100 0.82 7.36 studyo… 43 40 8 46 12  
4 studyone.101 0.94 7.88 studyo… 46 33 11 51 15  
5 studyone.102 0.58 6.96 studyo… 18 32 3 51 12  
6 studyone.103 0.84 7.88 studyo… 19 37 13 45 19  
# ℹ 3 more variables: GENDER <fct>, EDUCATION <fct>, ETHNICITY <fct>

## 34 A simple plot has many elements

### 34.1 Plot with no objects

* When we code a plot, we tell R we want:
* to use ggplot() to create a plot
* using the data-set clearly.one.subjects
* and the variables SHIPLEY, HLVA

|  |
| --- |
| Figure 7: Scatterplot showing association between health literacy and vocabulary |

### 34.2 Code for the plot

ggplot(data = clearly.one.subjects, aes(x = SHIPLEY, y = HLVA))

* We bring the data-set and the variables
* We declare the *aesthetic mappings*:

1. SHIPLEY score x-axis (horizontal: left-to-right position)
2. HLVA score y-axis (vertical: bottom-to-top position)

## 35 A simple plot has many elements

### 35.1 Plot with only objects

* When we code a plot, we tell R we want:
* to use a geometric object, like geom\_point
* to display the data aesthetic mappings

|  |
| --- |
| Figure 8: Scatterplot showing association between health literacy and vocabulary |

### 35.2 Code for the plot

ggplot(data = clearly.one.subjects, aes(x = SHIPLEY, y = HLVA)) +  
 geom\_point()

* We add the geom\_point() to tell R to draw the information about the SHIPLEY and HLVA scores as points
* Each point represents information about one participant in the clearly.one.subjects data-set

1. SHIPLEY score x-axis (horizontal: left-to-right position)
2. HLVA score y-axis (vertical: bottom-to-top position)

## 36 When we use {ggplot2} we work in *layers*

* The grammar of graphics define the components of a plot: the data, the mappings, and the geometric object
* Together, the data, mappings, and geometric object form a *layer*
* A plot may have multiple layers

## 37 When we use {ggplot2} we are in control and we can be creative

|  |
| --- |
| Tip |
| * Having a system of graphics: with components, layers and rules * Releases us to be *creative*: changing a single feature at a time |

## 38 Plot with layers: add a smoother

### 38.1 Plot with smoother

* Build a plot *layer by layer*
* We can begin by using points to display the vocabulary and health literacy scores for each person
* We add a layer using a *smoother* to show the average association between vocabulary and literacy

|  |
| --- |
| Figure 9: Scatterplot showing association between health literacy and vocabulary |

### 38.2 Code for the plot

ggplot(data = clearly.one.subjects, aes(x = SHIPLEY, y = HLVA)) +  
 geom\_point() +  
 geom\_smooth()

* We add the geom\_smooth() to tell R to represent the average trend for the association between SHIPLEY and HLVA scores
* The line is drawn by {ggplot2} which calculates a statistical transformation
* Here, the transformation summarizes the association for different ranges of SHIPLEY vocabulary scores

## 39 Defaults and arguments

clearly.one.subjects %>%  
 ggplot(aes(SHIPLEY, HLVA)) +  
 geom\_smooth() +  
 geom\_point()

* The {ggplot2} library supplies default values
* So we do not need to tell R how to do every thing
* We do not need to tell R that the points in a scatterplot:
* should represent the data aesthetic mappings in *Cartesian* (x-horizontal, y-vertical) 2-dimensional space
* and should be black in colour

clearly.one.subjects %>%  
 ggplot(aes(SHIPLEY, HLVA)) +  
 geom\_smooth() +  
 geom\_point(colour = "darkgrey", size = 3)

* We can over-ride the defaults by supplying arguments, entering values *inside the brackets* in the function calls
* geom\_point(colour = "darkgrey", size = 3) tells R we want:

1. dark grey points when the default is black
2. points that are 3x larger than the default size

## 40 When we use {ggplot2} we are in control and we can be creative

|  |
| --- |
| Tip |
| * We can add layers, control the appearance of each component * To construct more *effective* plots * The plots can be more effective because we develop them in an iterative process * in which we reflect on our *goals* and the *needs of our audience* |

## 41 We can use colour

### 41.1 Using colour

* When we code a plot, we tell R we want:
* to display data about people with different education levels
* distinguishing education level by colour

|  |
| --- |
| Figure 10: Using colour |

### 41.2 Code for the plot

clearly.one.subjects %>%  
 ggplot(aes(x = SHIPLEY, y = HLVA,  
 group = EDUCATION, colour = EDUCATION)) +  
 geom\_smooth(method = "lm", se = FALSE,  
 linewidth = 2, alpha = .75) +  
 geom\_point(size = 3)

* group = EDUCATION, colour = EDUCATION tells R to:

1. group the data by EDUCATION level
2. colour the points for people with different levels of education in different colours

## 42 Method, size, transparency

clearly.one.subjects %>%  
 ggplot(aes(x = SHIPLEY, y = HLVA,  
 group = EDUCATION, colour = EDUCATION)) +  
 geom\_smooth(method = "lm", se = FALSE,  
 linewidth = 2, alpha = .75) +  
 geom\_point(size = 3)

* method = "lm", se = FALSE tells R what method to use to draw the smoother line
* linewidth = 2 makes the width of the smoother line 2 x larger than the default
* alpha = .75 makes the line .75 x the opacity of the default (i.e. a. bit more transparent)
* Learn to edit: shape, size, transparency and colour

## 43 We *facet* plots to enable comparisons

### 43.1 Using facets

* It is often easier to *compare* trends
* By presenting a separate plot for each condition or group
* Showing the separate plots in a grid side-by-side

|  |
| --- |
| Figure 11: The association between health literacy and vocabulary varies by education level |

### 43.2 Code for the plot

clearly.one.subjects %>%  
 ggplot(aes(x = SHIPLEY, y = HLVA,  
 group = EDUCATION, colour = EDUCATION)) +  
 geom\_smooth(method = "lm", se = FALSE,  
 linewidth = 2, alpha = .75) +  
 geom\_point(size = 3) +  
 facet\_wrap(~ EDUCATION)

* facet\_wrap(~ EDUCATION) tells R to split the data by EDUCATION level
* And show a separate plot for each EDUCATION level group side-by-side for easy comparison

## 44 We can guide our audience

### 44.1 Labelled plot

* We do not present visualizations in isolation
* We present plots embedded in the context of labels and titles
* We use the text to guide the viewer

|  |
| --- |
| Figure 12: A labelled plot |

### 44.2 Code for the plot

clearly.one.subjects %>%  
 ggplot(aes(x = SHIPLEY, y = HLVA)) +  
 geom\_smooth(method = "lm", se = FALSE,  
 colour = "darkgreen", linewidth = 2, alpha = .75) +  
 geom\_point(size = 3, colour = "lightgreen") +  
 labs(x = "Vocabulary (Shipley)", y = "Health literacy (HLVA)",  
 title = "Scatterplot showing how higher vocabulary\npredicts higher health literacy on average")

* We use the labs() function to add: the plot title and the labels for the x-axis and y-axis
* We edit the title so that the viewer can see what we want them to see
* We use \n to make the title fit on two lines

## 45 We annotate plots to direct attention

### 45.1 Annotated plot

* We can direct the attention of our audience to key features of our data
* By adding annotations like text and lines

|  |
| --- |
| Figure 13: An annotated plot |

### 45.2 Code for the plot

clearly.one.subjects %>%  
 ggplot(aes(x = SHIPLEY, y = HLVA)) +  
 geom\_smooth(method = "lm", se = FALSE,  
 colour = "darkgreen", linewidth = 2, alpha = .75) +  
 geom\_point(size = 3, colour = "lightgreen") +  
 labs(x = "Vocabulary (Shipley)", y = "Health literacy (HLVA)") +  
 geom\_hline(yintercept = mean(clearly.one.subjects$HLVA),  
 linetype = "dashed",  
 linewidth = 2,  
 colour = "grey",  
 alpha = .85) +  
 annotate("text", x = 27, y = 9.3, label = "Mean HLVA", colour = "grey") +  
 theme\_bw()

* geom\_hline() adds a line to show mean health literacy
* annotate("text" ...) adds a text label :::

## 46 Extensions free our creativity

### 46.1 Complex plot

* The power of the *Grammar of Graphics* lies in the rules
* Developers can use the rules to expand our capacity to visualize data
* We add marginal histograms to our scatterplot to visualize associations *and* distributions

|  |
| --- |
| Figure 14: A scatterplot showing the potential association between health literacy and vocabulary |

### 46.2 Code for the plot

plot <- clearly.one.subjects %>%  
 ggplot(aes(x = SHIPLEY, y = HLVA)) +  
 geom\_smooth(method = "lm", se = FALSE,  
 colour = "darkgreen", linewidth = 2, alpha = .75) +  
 geom\_point(size = 3, colour = "lightgreen") +  
 labs(x = "Vocabulary (Shipley)", y = "Health literacy (HLVA)")  
  
ggMarginal(plot, type = "histogram", fill = "lightgreen",  
 xparams = list(binwidth=2), yparams = list(binwidth=1))

* ggMarginal(plot, type = "histogram") enables us to show the distribution of scores on each variable
* This helps our viewer to process the association *and* information about each variable (Franconeri et al., 2021)

## 47 Choose your plot theme

* We can choose a theme to adapt the look of the whole plot to suit our needs or the needs of our audience

|  |
| --- |
| Figure 15: theme\_dark() |

|  |
| --- |
| Figure 16: theme\_bw() |

|  |
| --- |
| Figure 17: theme\_classic() |

## 48 Summary

You start your work with these questions:

1. What are our goals?
2. What does our audience need or expect?

You develop your visualization in a reflective process:

1. Begin with a quick draft to show the distributions or make the comparisons you think about first
2. Then reflect, and edit: does this enable me to discover sources of variability in my data?
3. Then reflect, and edit: does this enable me to effectively communicate what I want to communicate?
4. Then reflect, and edit: does this look good? – do my viewers tell me this works well?

|  |
| --- |
| Tip |
| I can only show you the potential for creative and effective visualization   * experiment and find what looks good and is useful to you * seek out information – good places to start are:   <https://ggplot2.tidyverse.org/index.html>  <https://r-graph-gallery.com> |

## 49 References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21. <https://doi.org/10.2307/2682899>

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., & Balkin, T. J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, *12*(1), 1–12. <https://doi.org/10.1046/j.1365-2869.2003.00337.x>

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, *22*(3), 110–161. <https://doi.org/10.1177/15291006211051956>

Gelman, a. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective. *Journal of Management*, *41*(2), 632–643. <https://doi.org/10.1177/0149206314525208>

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let’s practice what we preach. *The American Statistician*, *56*(2), 121–130. <https://doi.org/10.1198/000313002317572790>

Gelman, A., & Unwin, A. (2013). Infovis and statistical graphics: Different goals, different looks. *Journal of Computational and Graphical Statistics*, *22*(1), 2–28. <https://doi.org/10.1080/10618600.2012.761137>

Hofman, J. M., Goldstein, D. G., & Hullman, J. (2020). *How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results*. 112. <https://doi.org/10.1145/3313831.3376454>

jumpingrivers. (n.d.). *Datasets from the Datasaurus Dozen*. <https://jumpingrivers.github.io/datasauRus/>

Kastellec, J. P., & Leoni, E. L. (2007). Using Graphs Instead of Tables in Political Science. *Perspectives on Politics*, *5*(4), 755–771. <https://doi.org/10.1017/S1537592707072209>

Matejka, J., & Fitzmaurice, G. (2017). *CHI ’17: CHI Conference on Human Factors in Computing Systems*. 1290–1294. <https://doi.org/10.1145/3025453.3025912>

Van Der Bles, A. M., Van Der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). *Communicating uncertainty about facts, numbers and science* (Vol. 6).

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342. <https://doi.org/10.1515/ling-2019-0051>

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, *19*(1), 3–28. <https://doi.org/10.1198/jcgs.2009.07098>

Wilkinson, L. (2013). *The Grammar of Graphics*. Springer Science & Business Media.

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, *120*(33), e2302491120. <https://doi.org/10.1073/pnas.2302491120>