Publication ready graphics

2941 - Data Challenge

Sam Clifford

2022-01-27

Introduction

About this practical session

In the week 1 lecture session on exploratory data analysis we introduced visualisation with the histogram, x-y plots and other scatter plot techniques, and touched on Tufte's principles of graphical excellence.

This practical will investigate what makes a good and a bad graph, in order to help you generate publication quality graphics. You will first critique a plot produced through the provided code. Activities 2a and 2b are about producing a better and a worse version of the given plot.

- Assumed skills
  - Writing R code into a script file
  - Reading and writing ggplot2 code
  - Identifying things that are visually pleasing
- Learning objectives
  - Identifying the link between code and the graph it produces
  - Being able to critique a graph
  - Understanding why and how data is encoded and decoded visually
  - Understanding the subjectivity of what is aesthetically pleasing
- Professional skills
  - Creating high quality graphics

Group formation

You will be allocated to groups of 3-5 in Zoom breakout rooms.

A reminder of expectations in the prac:

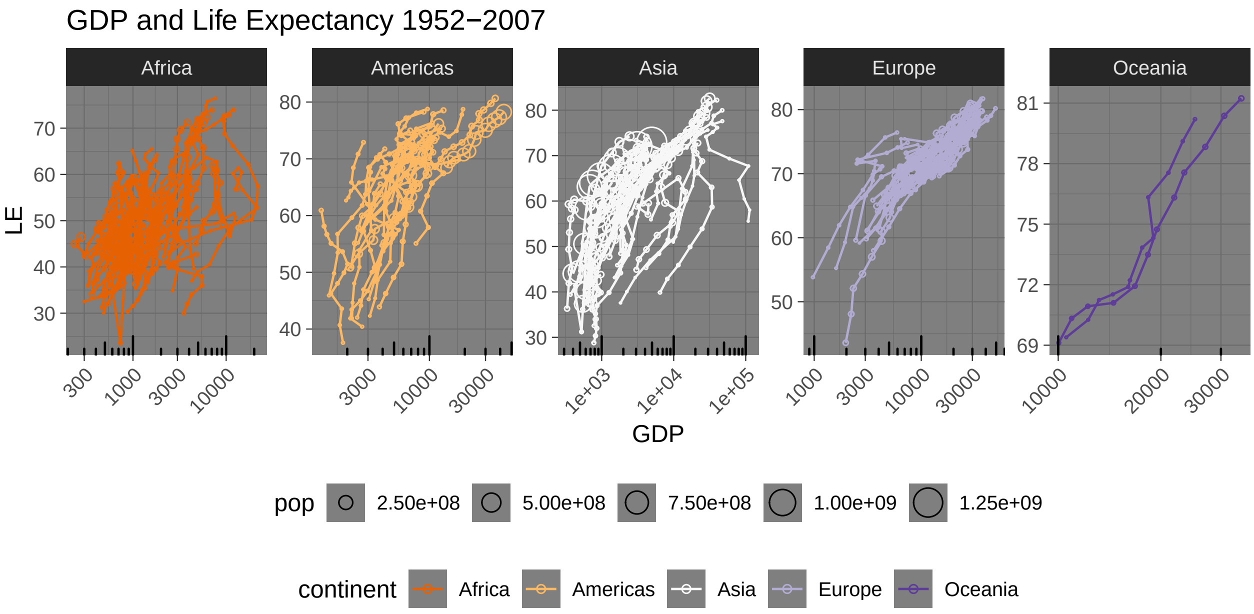
- Keep a record of the work being completed by including meaningful comments as you modify the `R/vis_gapminder.R` script and filling your answers in the `doc/solutions.md` file.
- Allow everyone a chance to participate in the learning activities, keeping disruption of other students to a minimum while still allowing for fruitful discussion
- All opinions are valued provided they do not harm others
- Everyone is expected to do the work, learning seldom occurs solely by watching someone else do work

Activity 1 - Building an attempt at a plot

We will be looking at the gapminder data set as found in the gapminder package (Bryan 2017). This data has been collected from countries around the world and contains data on life expectancy, population and GDP per capita for 142 countries from 1952 to 2007.

Exercise: Have one person in the group fork the github repository at https://github.com/samclifford/2491_prg and add the team members as collaborators.

Exercise: Run the provided code to produce a plot showing how the relationship between GDP, life expectancy and population vary over time and continent. Save the plot as both a PDF and PNG and zoom in to see the pixelation in the PNG image.



Exercise: Discuss, within your group, what you think is good and bad about this plot. Does it conform to Tufte's principles of graphical excellence? Is it easy to interpret? Does it show the relationship we are interested in? List *three* important improvements that are needed for this graph to be useful. This should take no longer than ten minutes.

Exercise: As a group, discuss what you think each line of code does. You may wish to answer as comments in your code (everything after a # is a comment) or in a separate document.

Activity 2

Split your group in two and have each half tackle one of the following two activities. Ensure that you regularly save, commit and push your changes.

You may choose to either modify the code given in the R script or create your own graph from scratch. Make sure your code is written in your script file with appropriate comments.

Some things you may wish to consider:

- what makes annotations meaningful?
- how can the overall theme be modified to improve legibility?
- what is the most appropriate geometry to show your relationship of interest?
- what aesthetics (Figure 5, Kunz and Hurni 2011) draw the eye in helping highlight key context?

You may wish to sketch the graph by hand before attempting to write the R code to generate it. This may help you and your group come to an agreement about the plot you want to make and will help the tutors understand what you're aiming for when you ask them for help.

If you get stuck, look at the following resources for help

- ggplot2 documentation
- R Graphics Cookbook (Chang 2018)
- Chapter 3 and 4 of R for Data Science (Wickham and Grolemund 2020)
- RStudio cheatsheets (RStudio 2012)

Activity 2a – Making a better graph

Based on the ideas discussed, build a graph which your group believes better shows the relationship between life expectancy and GDP. Think first about what story you want your plot to tell; are you interested in trends over space and/or time? Are you interested in a particular continent or even just one country? Do you want to show a snapshot in a particular year, or compare across two or more years?

Exercise: What is the story you wish to tell with the graph? Are you showing only a subset? What are the additional variables that provide context?

Exercise: Make a plot, save it to your computer and write comments in your code or standalone document that outline what the changes you made were and why.

Activity 2b – Making a worse graph

Make a new graph as in the previous activity but make it as bad as possible while still attempting to honestly show the information (i.e. don't add things to the plot which can't be derived from the variables in the plot). Your plot should be an honest attempt to show the data poorly, rather than a deliberately unreadable mess. Think of something you'd expect to see in a newspaper staffed with well-intentioned but unskilled staff.

Exercise: Consider the principles of graphical excellence and how can we go against them to make a terrible plot. Think about what was bad about the plot provided earlier. Consider abusing the ability to map graphical options (e.g. color, fill, line type, point size) to our variables of interest.

Exercise: Make a plot, save it to your computer and write comments in your code or standalone document that outline what the changes you made were and why.

Activity 3 – Group discussion

Exercise: Have each half of the group present their best/worst graph from the last activities to the other half. What did you identify as good and bad and how have you attempted to represent the relationship?

Activity 4 – Room discussion

Exercise: Have each group present their best and/or worst graph from the last activities. What did they identify as good and bad and how has each group attempted to present the relationship?

Tidy up

Make sure you save your R script, and anything else you have produced and ensure everyone in your group has a copy. Email your best and worst graphs to Dr Sam Clifford.

Further reading

Key ideas introduced by Tufte (1983) are summarised by Pantoliano (2012). Some of the history of data visualisation is summarised well by Friendly (2005) and Friendly (2006). Tufte's website is well worth exploring, particularly the discussion on how the visual presentation of information could have helped avert the *Challenger* disaster (Tufte 1997).

References

Bryan, Jennifer. 2017. *Gapminder: Data from Gapminder*. https://CRAN.R-project.org/package=gapminder.

Chang, Winston. 2018. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. 2nd ed. O'Reilly Media. https://r-graphics.org/.

Friendly, M. 2005. "Milestones in the History of Data Visualization: A Case Study in Statistical Historiography." In *Classification: The Ubiquitous Challenge*, edited by C. Weihs and W. Gaul, 34–52. New York: Springer. http://www.math.yorku.ca/SCS/Papers/gfkl.pdf.

———. 2006. "A Brief History of Data Visualization." In *Handbook of Computational Statistics: Data Visualization*, edited by C. Chen, W. Härdle, and A Unwin. Vol. III. Heidelberg: Springer-Verlag. http://www.datavis.ca/papers/hbook.pdf.

Kunz, Melanie, and Lorenz Hurni. 2011. "How to Enhance Cartographic Visualisations of Natural Hazards Assessment Results." *The Cartographic Journal* 48 (1): 60–71. https://doi.org/10.1179/1743277411y.0000000001.

Pantoliano, Mike. 2012. "Data Visualization Principles: Lessons from Tufte." 2012. https://moz.com/blog/data-visualization-principles-lessons-from-tufte.

RStudio. 2012. "RStudio Cheat Sheets." 2012. https://www.rstudio.com/resources/cheatsheets/.

Tufte, Edward R. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

———. 1997. "Visual and Statistical Thinking." In *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press. https://www.edwardtufte.com/tufte/books_textb.

Wickham, Hadley, and Garrett Grolemund. 2020. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. http://r4ds.had.co.nz.