Nicola Rennie

Programming in public

How sharing code on social media led to new opportunities

About Me

Academic background in statistics

Experience in data science consultancy

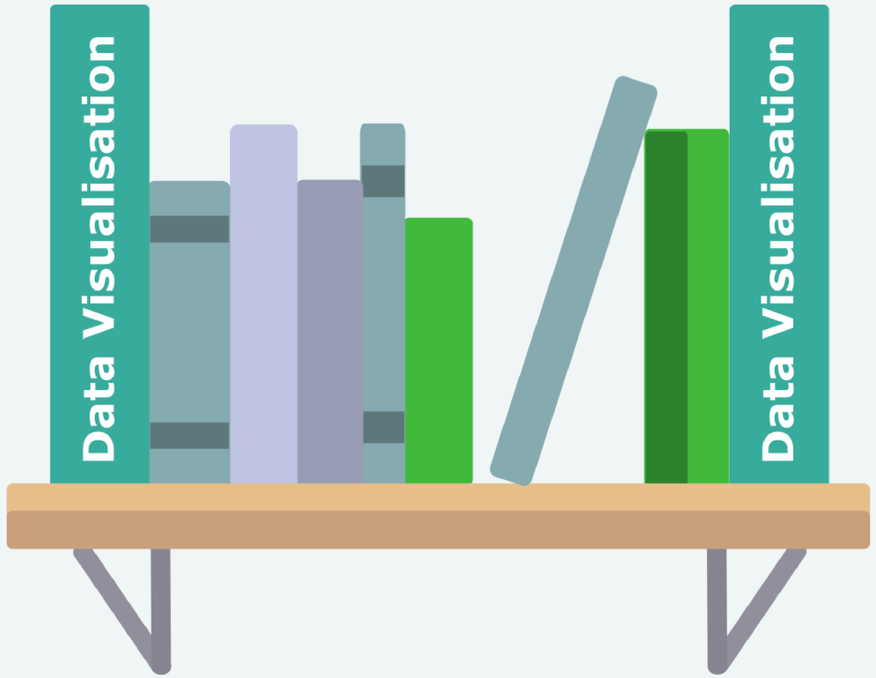
Lecturer in Health Data Science in Lancaster Medical School

Interested in ML, reproducible research, communicating data effectively,...

Learning in public

Sharing what you learn as you learn it

Data visualisation



Goals

How do I visualise (big) data?

Plotting averages disguises patterns, a lot of data over space and time

---

How do I make better charts with R?

Previously a Python user, but existing codebase was in R

How do I work with a wide range of data?

Exposure to a wider range of data sources than just those I had been working with

#TidyTuesday

Explore the data

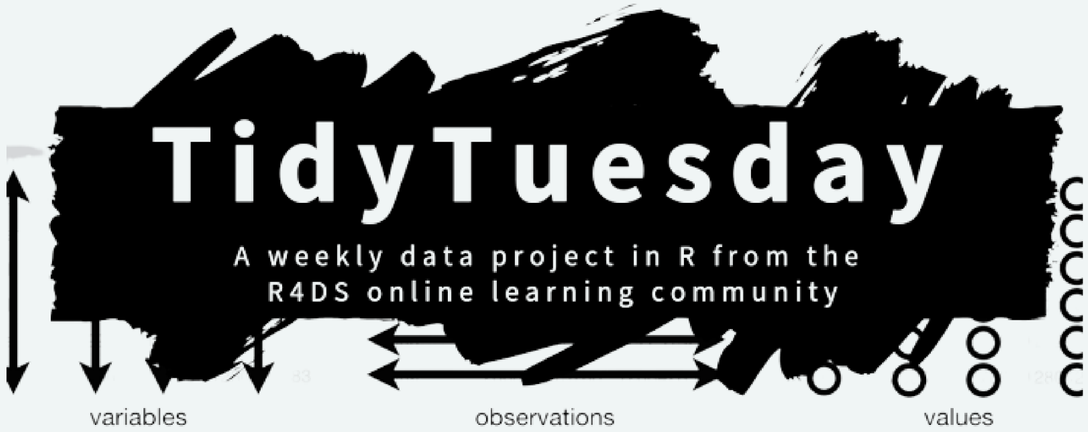
A new open source data is published each week, from various sources.

Create a visualisation

or a model, or an app, or something else. Build it in R or in another programming language.

Share with others

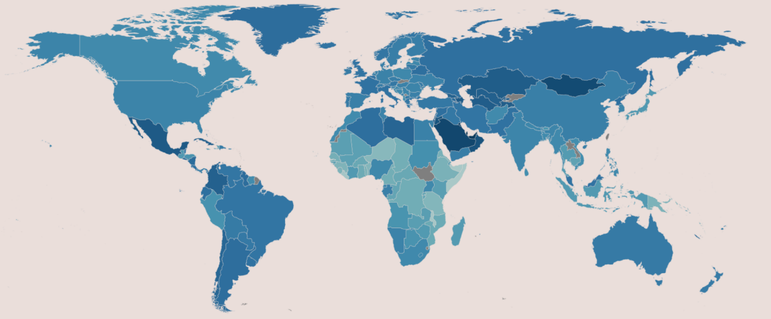
Share your chart, and your code, with others on social media or via Slack (~15,000 people).



Doctors in an ageing population

This map show the number of doctors per thousand people, rescaled by the percentage of the population aged over 70, revealing which countries* may be more likely to struggle in providing care for an elderly population.

* using most recent available data for each country.

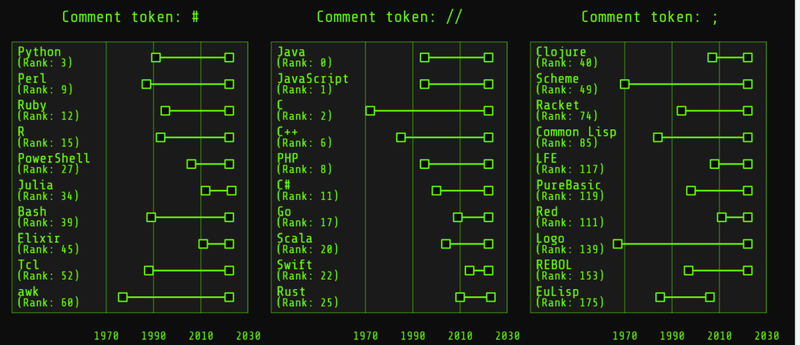


N. Rennie | Data: Gapminder

Fewer doctors — More doctors

Programming Languages

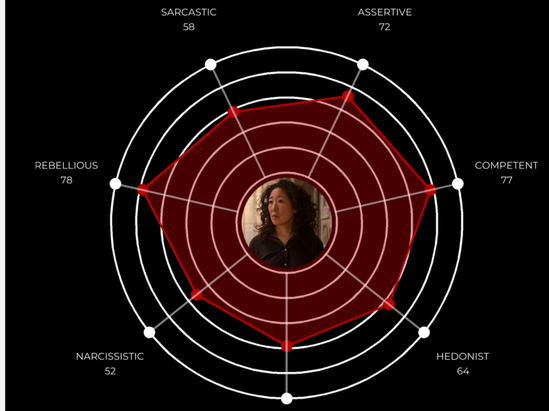
Of the 4,303 programming languages listed in the Programming Language DataBase, 205 use //, 101 use #, and 64 use ; to define which lines are comments. 3,831 languages do not have a comment token listed.



@nrennie35 fosstodon.org/@nrennie nrennie

KILLING EVE

EVE POLASTRI

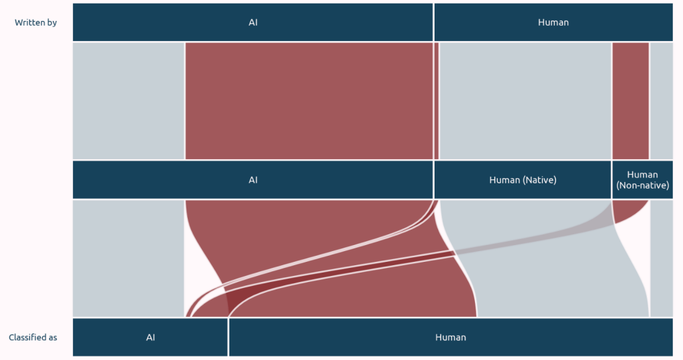


N. Rennie | Data: Open-Source Psychometrics Project

Can AI detect AI?

Generative AI makes it incredibly easy to generate large amounts of content in a short space of time. But can AI also be used to detect when content was written by AI? Liang et al. tested almost 1,000 documents across seven different GPT detectors (Crossplag, GPTZero, HFOpenAI, OriginalityAI, Quil, Sapling, and ZeroGPT).

When text was written by AI, it was incorrectly classified as human in 69% of cases. When text was written by a human, it was only incorrectly classified as AI 18% of the time. However, that 18% isn't split evenly. Native English speakers were only classified as AI in 3% of cases. Non-native English speakers were classified as AI in 61% of cases.

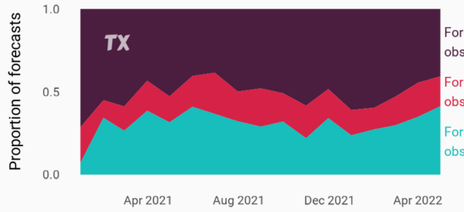


Data: GPT Detectors Are Biased Against Non-Native English Writers. arXiv: 2304.02819
Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, James Zou.
Chart: @nrennie35 fosstodon.org/@nrennie nrennie

Higher or lower?

Forecasts for higher temperatures tend to overestimate how hot it's going to be. Across 81,496 forecasts for high temperatures estimated 12 hours previously, 34,629 over-estimated the temperature. This compares to 24,995 cases of under-estimating the temperature.

N. Rennie | Data: USA National Weather Service



Outcomes

Found new ways of doing things

Seeing how other people approached the same data set.

Write better and tidier code

Sharing code by default forces you to write it better the first time around.

Met a community of people

Leading to speaker invitations, conference presentations, as well as new jobs and projects.

Developing in public

Creating a data visualisation style guide
(and the R and Python packages to implement it...)

The Project

Royal Statistical Society

Organisation for all statisticians and data analysts, who advocate for the use of statistics and data in society.

Survey responses

Graphics should be more distinctive and be more accessible to a wider audience

---

Development team

Co-authored with Andreas Krause (Idorsia) and Brian Tarran (RSS).

The Project

Write a style guide

Define style and visualisation guidelines for publications

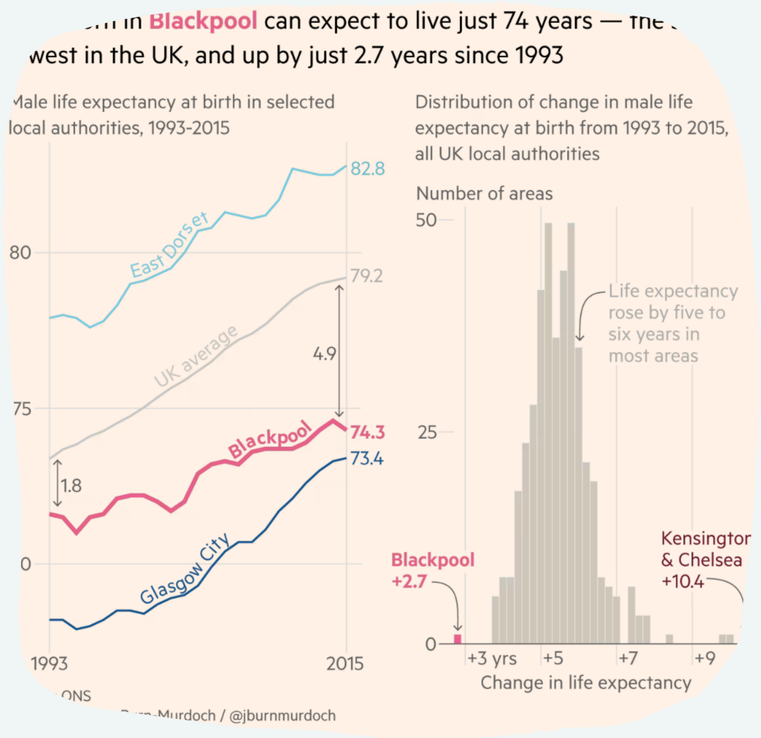
Publish the guide

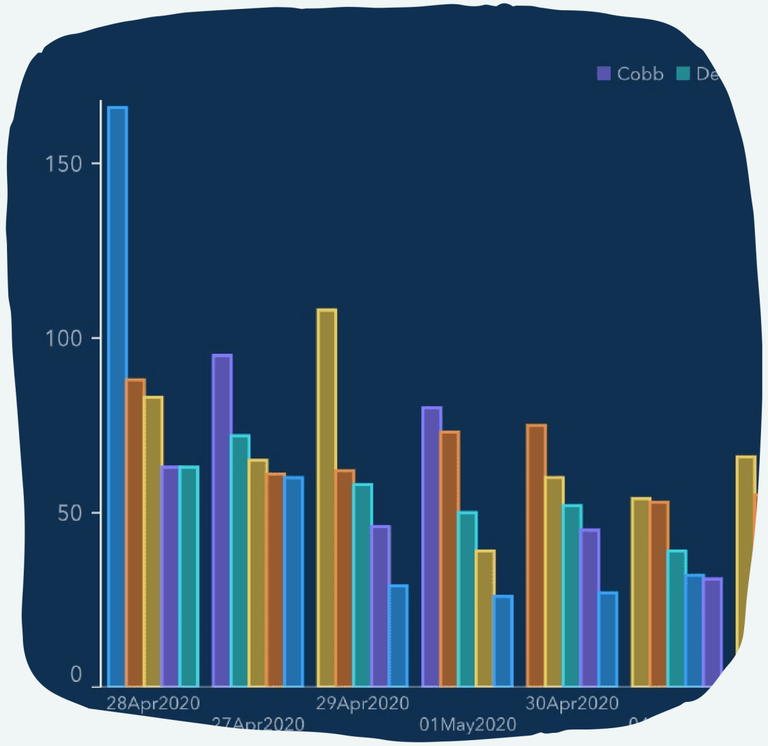
Develop a website for users to explore the guide (which they can contribute to)

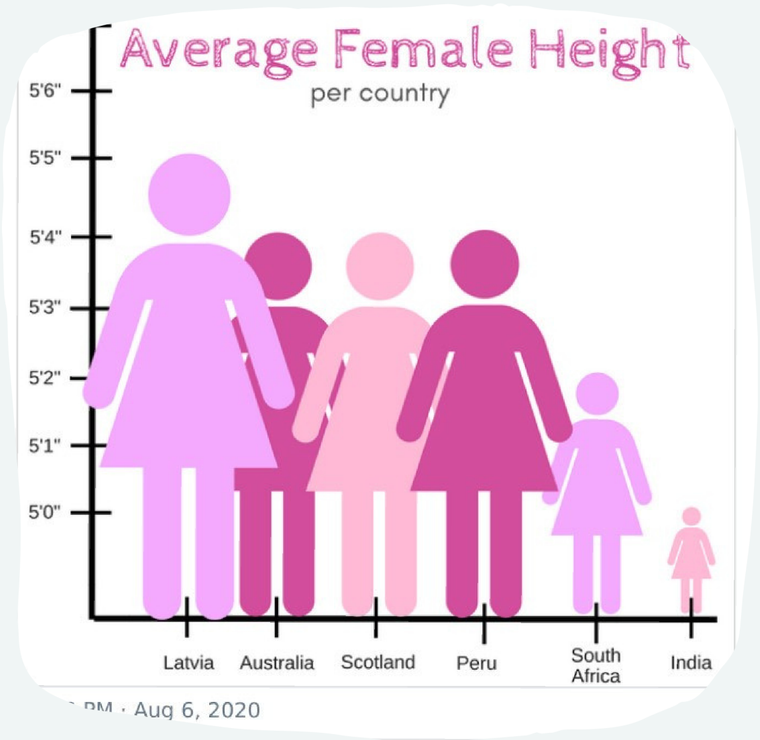
Build software

Build R and Python packages to help authors implement the styling

Defining style guidelines







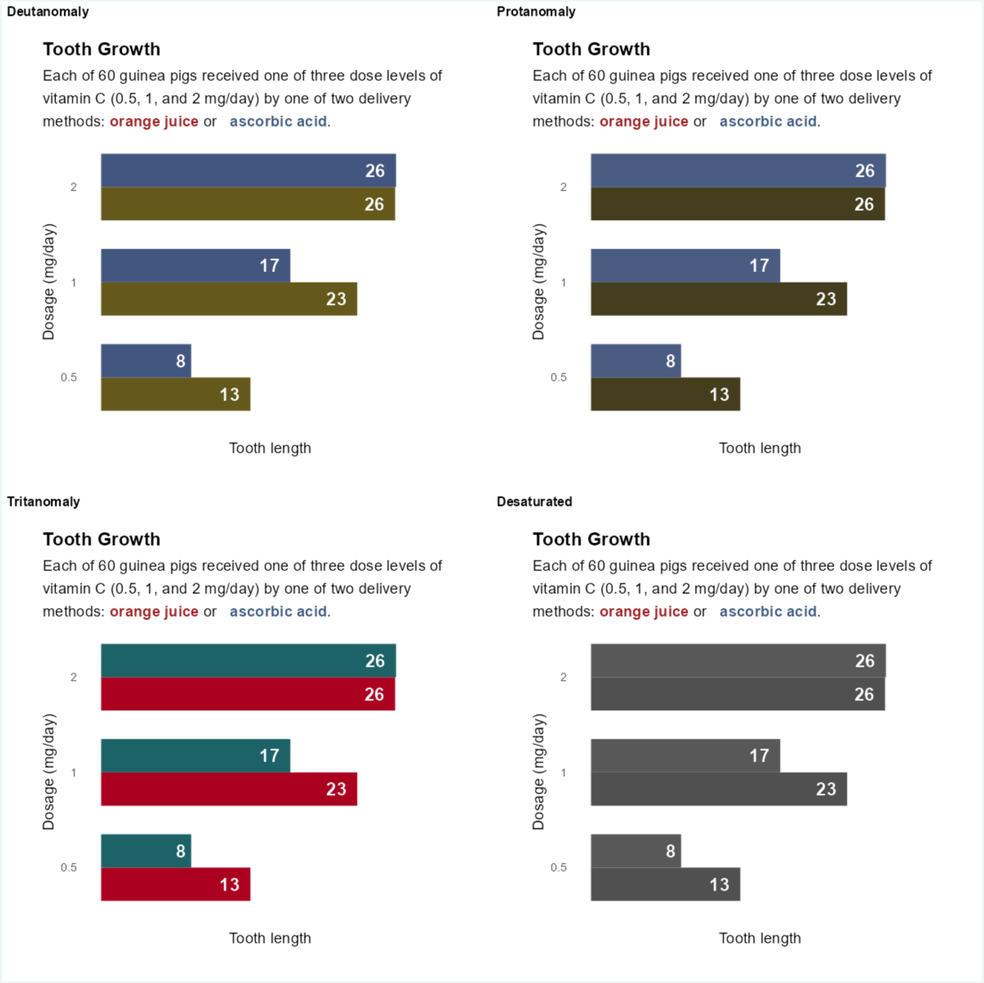
Defining style guidelines

Which type of chart should I make?

How do I structure the chart?

How do I make it accessible?

- Appropriate use of colours that are colourblind-friendly
- Good use of annotations
- Choosing readable font families and sizes
- Adding alt text



Developing a website

Open source software

Both the guide itself and the website source code are publicly available

Community contributions

The guide will be edited and added to in the future by the community

100% reproducible

Code to produce example charts is embedded in the website

Building an R package

R is one of the more common tools

Encouraging the use of open source tools, to improve reproducibility.

Build a style package

Minimise the amount of work that authors have to do to implement the recommended styling

Linked to website

Website installs the R package to demonstrate the examples

Outcomes

Developed in public

Feedback incorporated earlier and easy for others to contribute

---

Challenges

Addressing issues and feature requests from non-GitHub users in a transparent way

---

Longer term development

New tools and methods for creating charts, and overseeing how authors respond to the guide

Work with me!

PhD Opportuntities

Biases and inequalities in machine learning for healthcare

Dr N. Rennie, Prof J. Knight
Link: findaphd.com/phds/project/biases-and-inequalities-in-machine-learning-for-healthcare/?p159858

Late-onset and vascular epilepsy: a data science approach to improve understanding and care

Prof H. Emsley, Dr N. Rennie
Link: findaphd.com/phds/project/phd-fellowship-late-onset-and-vascular-epilepsy-a-data-science-approach-to-improve-understanding-and-care/?p159859

nicola-rennie

nrennie.rbind.io

chicas.lancaster-university.uk