Introduction

Data visualization 2023/2024

Matteo Ceccarello

2025-01-12

> Yes, I see.
>
> – *most English speakers when they understand*

Sight is the most prominent of the five senses for most people. Evolution made the human visual system extremely well suited at discovering visual patterns in what we see.

In this course we will study how to leverage these characteristics to present our data in informative, compelling, and convincing ways.

Before delving into more technical matters, we will consider a few more motivanting examples.

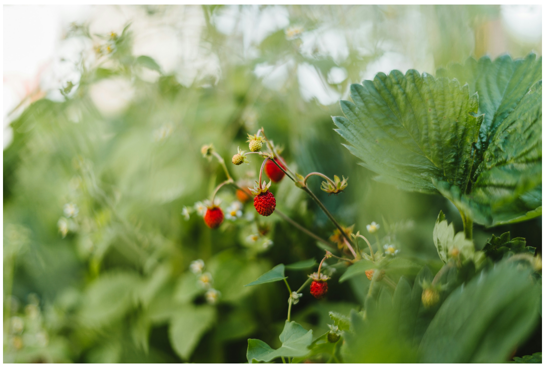


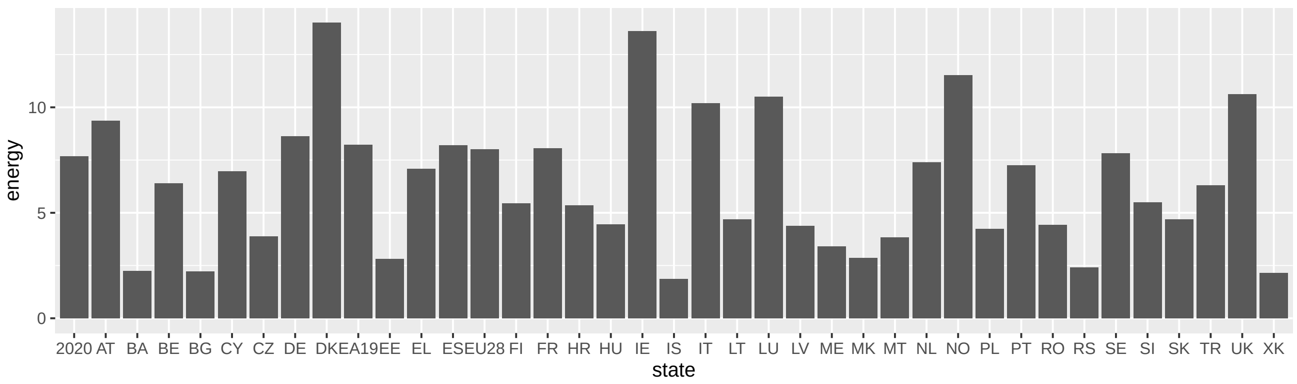
Figure 1: How long does it take you to spot the berries in the picture?

Consider the following table, which reports the energy productivity (i.e. the amount of money "produced" for each Euro spent) of the European Union countries in 2014. Which is the country with the highest energy productivity? And the second highest?

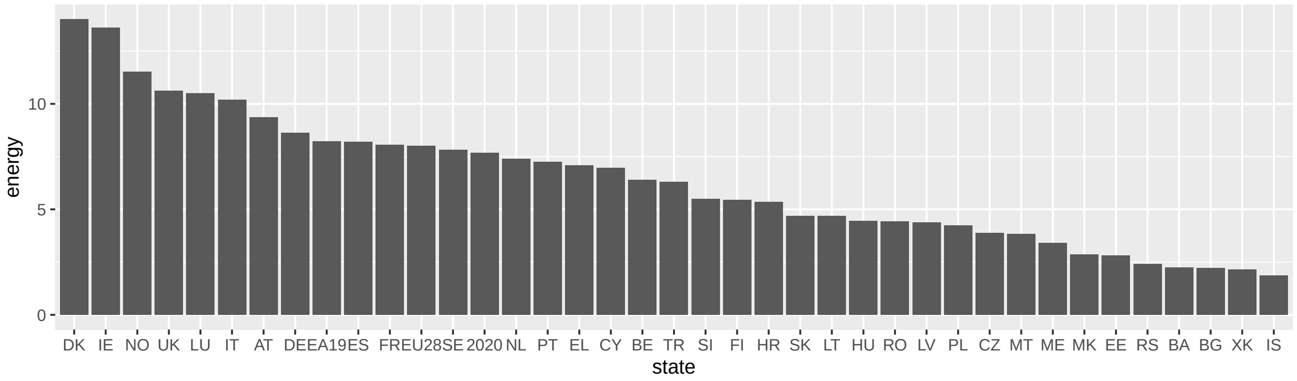
state	energy	state	energy	state	energy	state	energy
AT	9.374	EL	7.087	IT	10.182	PT	7.254
BA	2.240	ES	8.191	LT	4.680	RO	4.430
BE	6.388	2020	7.672	LU	10.506	RS	2.420
BG	2.226	EU28	8.012	LV	4.379	SE	7.821
CY	6.968	FI	5.453	ME	3.396	SI	5.494
CZ	3.879	FR	8.059	MK	2.853	SK	4.686
DE	8.616	HR	5.355	MT	3.826	TR	6.291
DK	14.007	HU	4.461	NL	7.400	UK	10.620
EA19	8.230	IE	13.620	NO	11.534	XK	2.140
EE	2.806	IS	1.866	PL	4.232		NA

How long did it take you to answer these questions? The main issue here is that you have to scan the entire table, keeping in your memory the largest value seen so far.

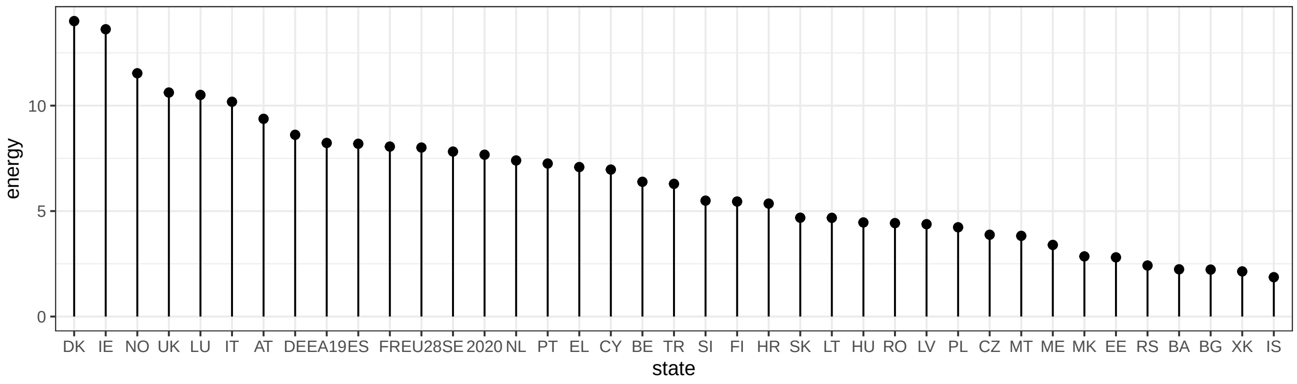
The same information can instead be presented in the following way.



Now it is easier to find the most productive country. But what if we wanted to ask other ranking-related questions, for instance about the fifth-country in the ranking? In order to make the figure more effective in *supporting finding the answer* to these ranking questions we can sort the columns:

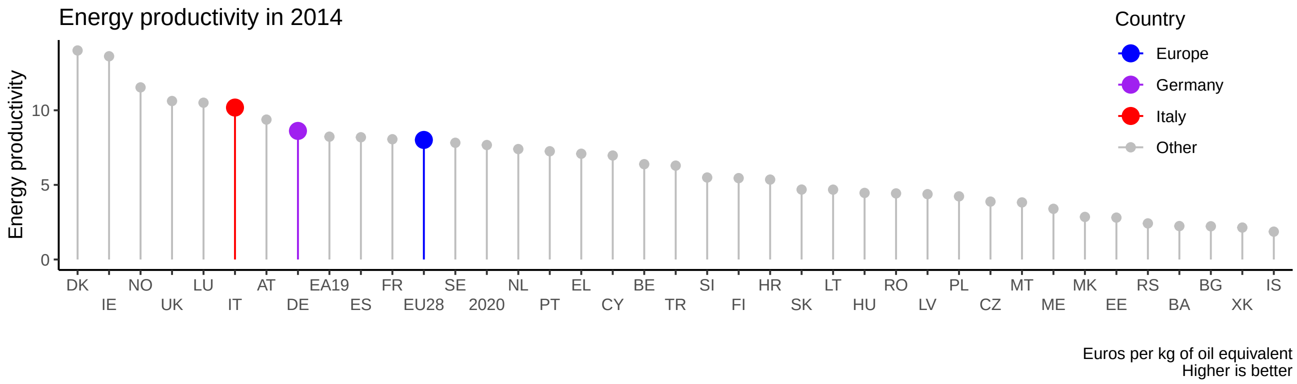


Now things are much clearer. Still, the plot is quite crowded. How about replacing the classic bars with some *lollipops*?



Still, the picture is a bit dull. We can try to add some color, but first we should ask ourselves the question: to which end are we going to use color? Maybe our goal is to compare the energy productivity of Italy and Germany with the average of all the European countries.

To do so we can make all the lollipops gray so that *they provide context*, and highlight with different colors the information we want our readers to focus on.



We also fixed the labels, added a title and a legend.

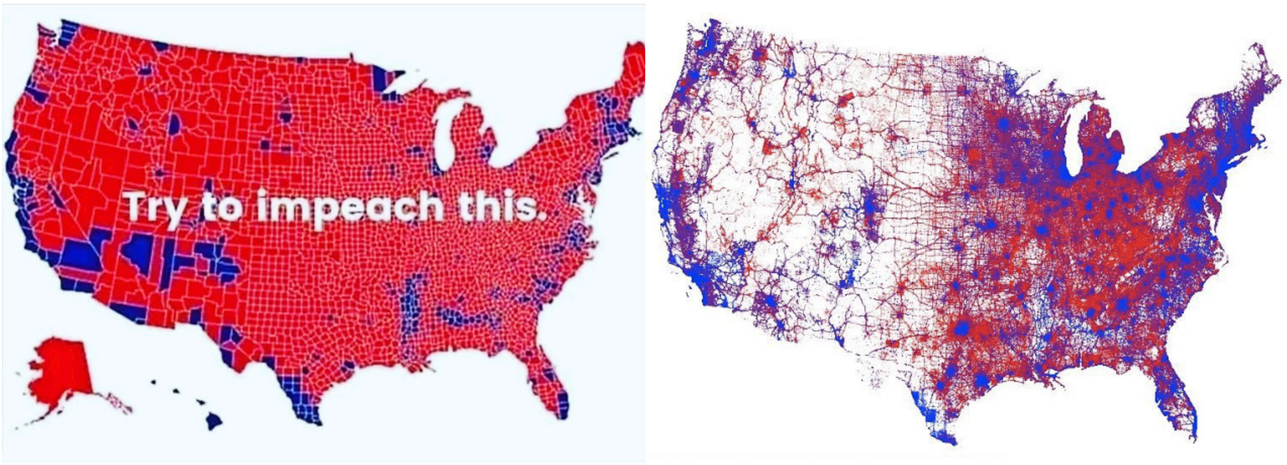
There are a couple of things that are particularily notable in this example:

- We have to choose a story we want to tell
- Raw data in and by itself is not very meaningful
- We have to display just the information we need to support our story

Of course, in a scientific setting the last point **must not** imply that we cherry pick our results! Rather it means that while we consider all the data we have we should make it clear which data is supporting our conclusions and which data is instead providing context.

Indeed, being clear without being misleading is one of the things to watch out for when designing visualizations.

The following figure reports a choropleth map of the United States on the left It was used in 2019 by Donald Trump, who at the time was facing the prospect of an impeachment, to try to make the point that he had overwhelming popular support. The map on the left shows all the counties of the United States, colored red where the Republican party won the majority.



The thing is that landmass does not vote, and most of the counties in the Midwest of the US are very sparsely populated. The map on the right tries to take into account this fact: it is made of many tiny dots, positioned according to the population density distributed according to the votes. Clearly the situation is much more balanced.

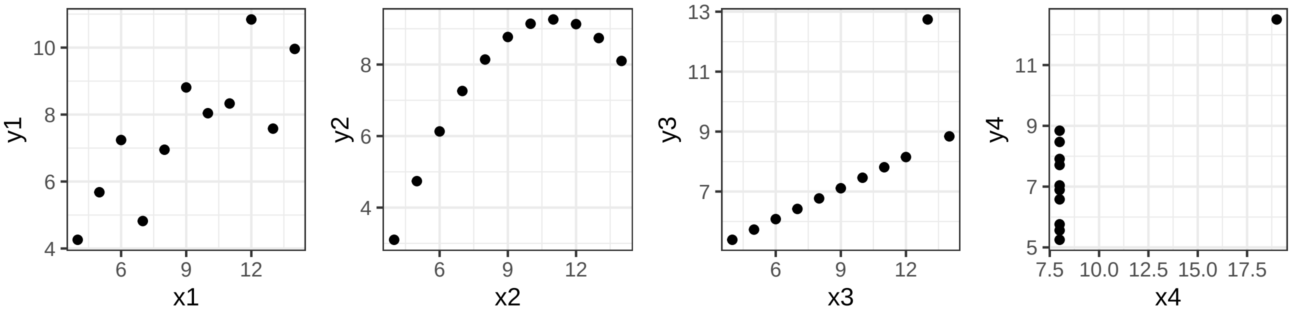
The above is an example of how *the same* data can be used to produce graphics that convey opposite messages. But then, why even bother to use visualization if it is so easy to misuse it? First, knowing the grammar and language of graphics allows us to be proficient and attentive *readers* of graphs: this is an extremely useful skill when reading and reviewing paper. Second, data is much better at *summarizing data* than summary statistics alone.

Consider the following summaries of a 2 dimensional dataset

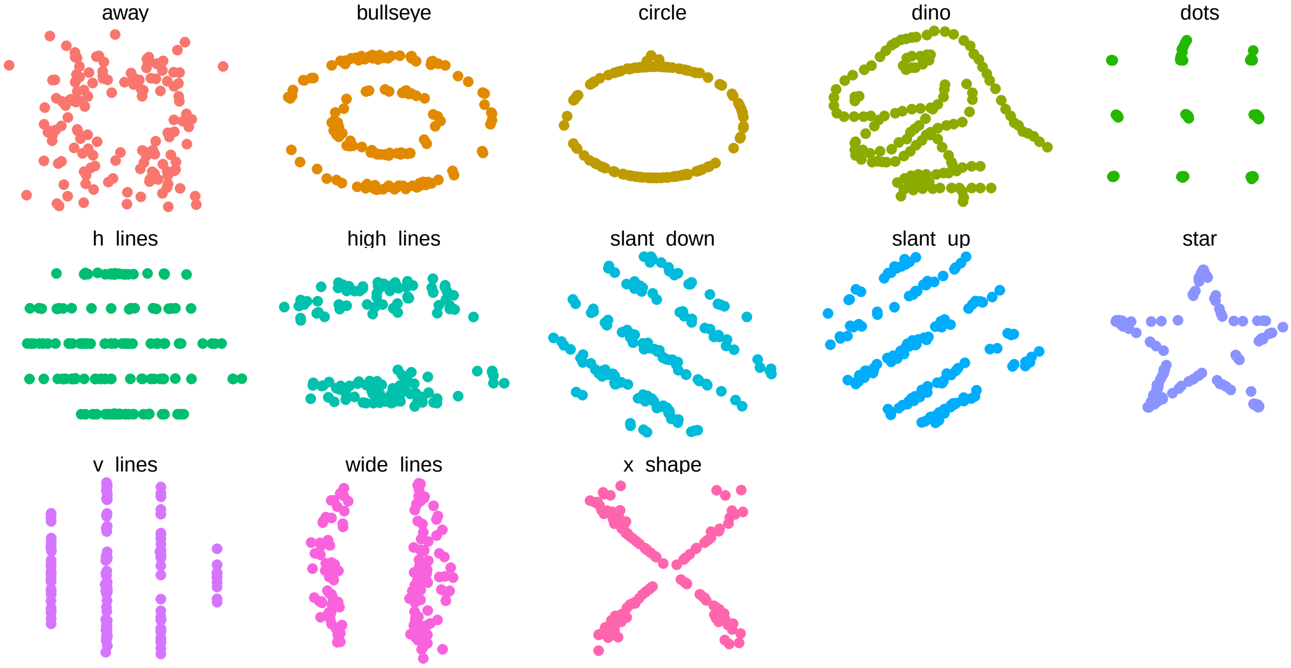
- Mean 9, 7.5009091
- Standard deviation 3.3166248, 2.0315681
- Correlation coefficient 0.8164205

How many datasets are uniquely identified by those statistics?

As it turns out: quite a lot. The four datasets in the margin are called Anscombe's quartet.



Those datasets are not alone: you can concoct datasets with the same summary statistics in all shapes and forms, including a dinosaur!



About this course

In this course we will study techniques to *explore* and *visualize* data in order to *communicate* with people.

In particular we will adopt a *programming based* workflow that aims to be flexible and reproducible.

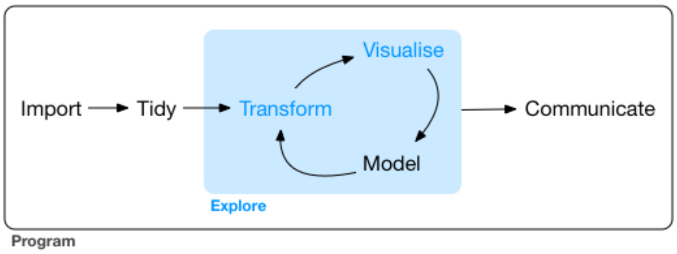


Figure 2: A data exploration workflow. Source: https://r4ds.had.co.nz/explore-intro.html

About me

- Matteo Ceccarello
- email: matteo.ceccarello@unipd.it
- https://www.dei.unipd.it/~ceccarello/
- Course material: https://stem.elearning.unipd.it/course/view.php?id=11140

Reference material

- *Data Visualization. A practical introduction.* Haley.
- *Fundamentals of Data Visualization.* Wilke.
- *R for Data Science.* Wickham.
- *A layered grammar of graphics.* Wickham.