Quantitative Tutorials - Session 2

Data visualization

23/11/22

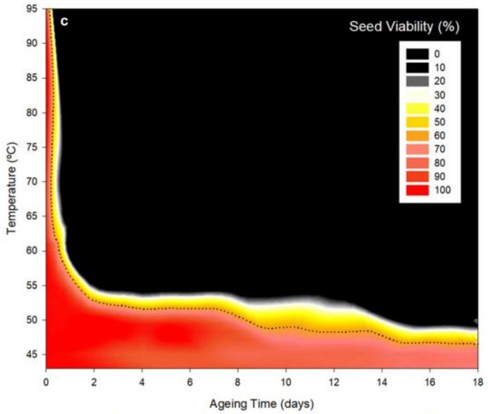
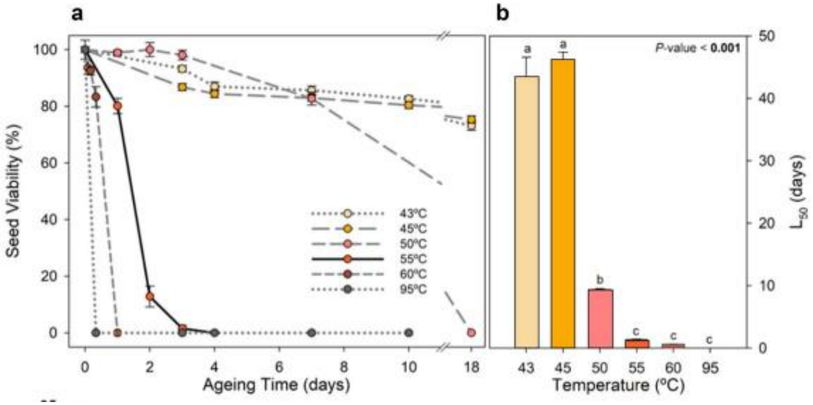


Fig. 2 Influence of temperature and aging time in artificially aged seeds of *C. edulis*. **a** Viability loss of *C. edulis* seeds at the different tested temperatures and 87% RH. Data is represented as Mean ± SE (n = 6). **b** L_{50} (loss of 50% viability) at the different tested temperatures. Different letters indicate statistically significant differences (*P*-value < 0.05). **c** Contour plot of seed viability considering temperature and aging days. Dotted line represents the L_{50}, 50% viability loss



Dr. Erola Fenollosa

BALLIOL COLLEGE UNIVERSITY OF OXFORD

SaiGo TEAM

UNIVERSITY OF OXFORD

Department of BIOLOGY

IRBio Institut de Recerca de la Biodiversitat

@Erola_Fenollosa

Data visualization objectives

Non-exclusive data visualization aims

- Communicate
- Explore
- Data in its context
- Find patterns and outliers

Visual analysis purposes

- Detect trends, patterns and outliers
- Compare
- Establish relationships

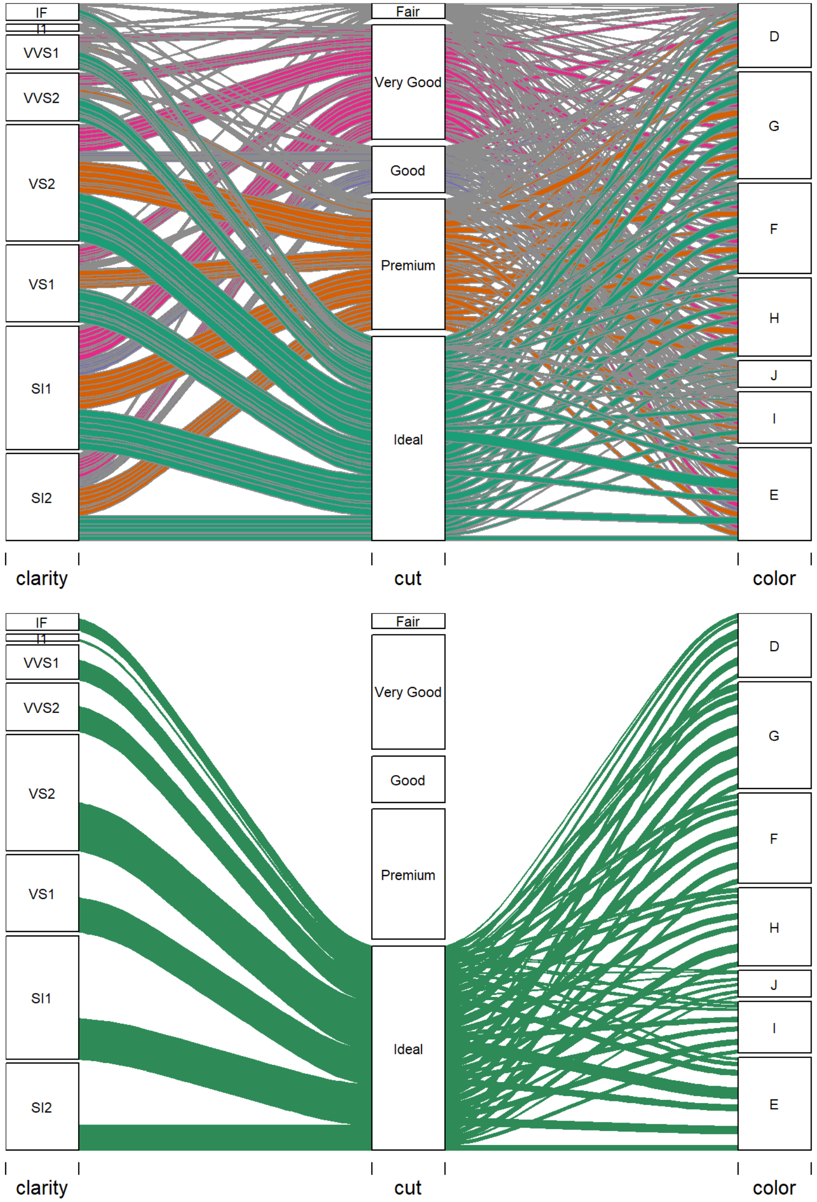
Visual codes:
position, shape, colour, movement

But, the human eye has limitations!

More detailed information and pre-session tasks: https://github.com/erolafr/Quantitative-Tutorials-2022/blob/main/Session2/Session2.md

Visualizations for data analysis

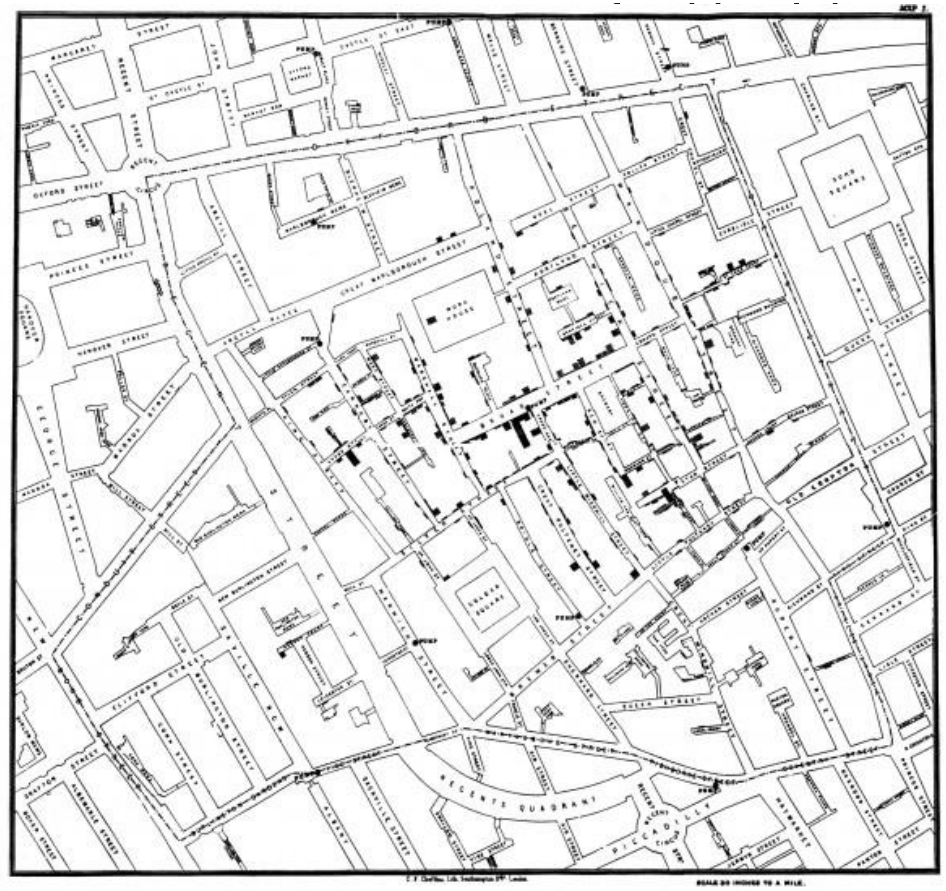
Alluvial Plot



https://www.bryanschafroth.com/post/2018/05/13/data-preprocessing-diamonds-dataset/

Spatial analysis

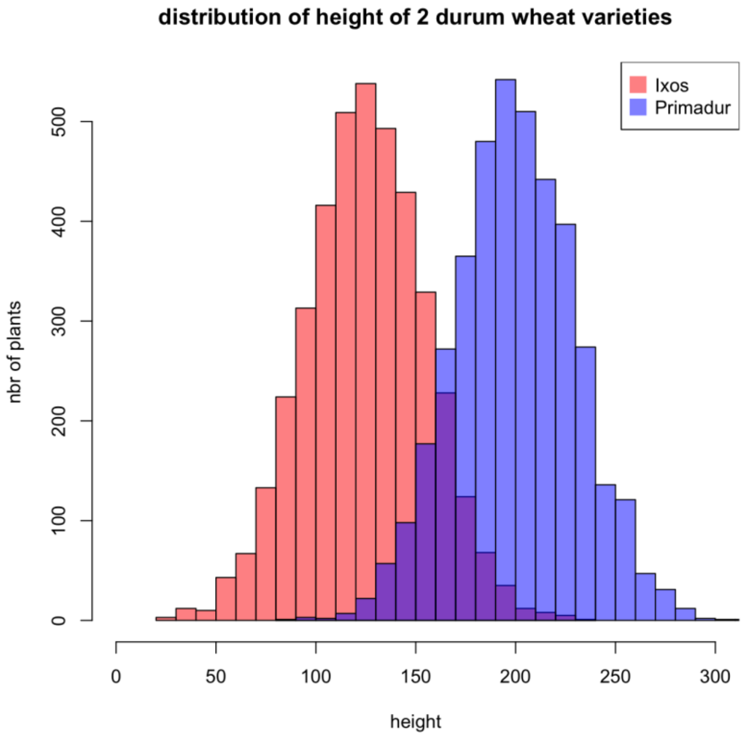
Today we know that cholera is spread through water, but in the early 1800s people weren't sure. John Snow's cholera map helped to show that contaminated wells were at the center of outbreaks. His research helped save countless lives and set the foundation for the field

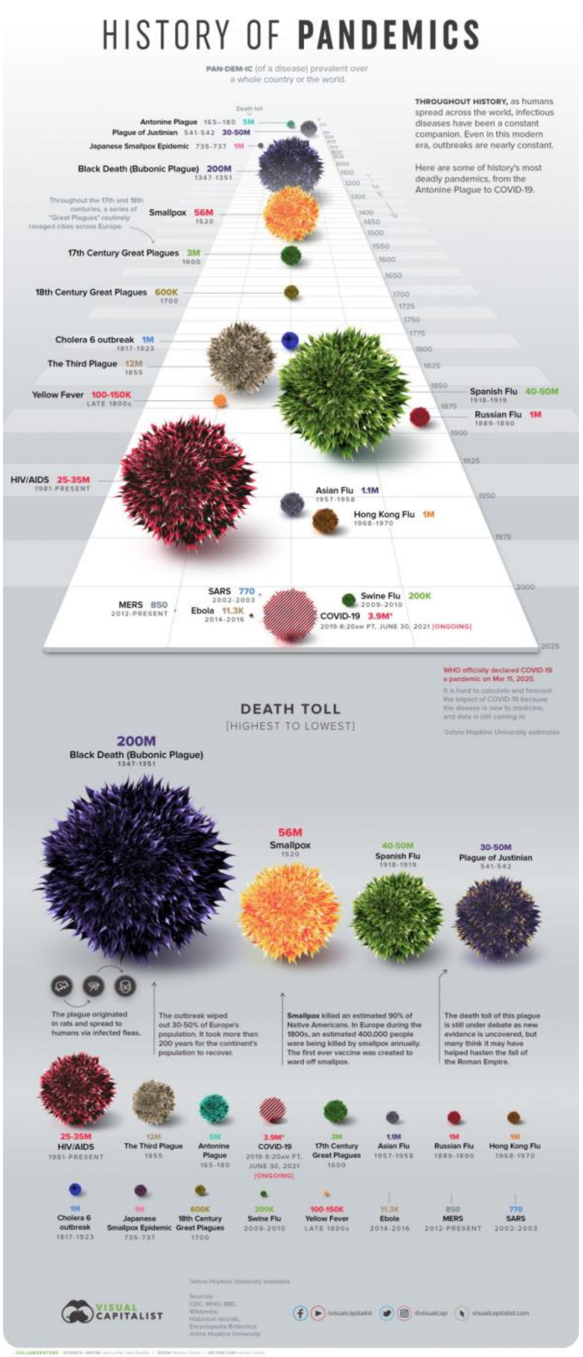


https://rockcontent.com/blog/12-great-visualizations-that-made-history/

Good visualizations

- Simple and effective
- High information with low ink
- Intuitive (colours, axis)
- Honest





THROUGHOUT HISTORY, as humans spread across the world, infectious diseases have been a constant companion. Even in this modern era, outbreaks are nearly constant.

Here are some of history's most deadly pandemics, from the Antonine Plague to COVID-19.

Throughout the 17th and 18th centuries, a series of "Great Plagues" routinely ravaged cities across Europe.

WHO officially declared COVID-19 a pandemic on Mar 11, 2020.

It is hard to calculate and forecast the impact of COVID-19 because the disease is new to medicine, and data is still coming in.

*Johns Hopkins University estimates

The plague originated in rats and spread to humans via infected fleas.

The outbreak wiped out 30-50% of Europe's population. It took more than 200 years for the continent's population to recover.

Smallpox killed an estimated 90% of Native Americans. In Europe during the 1800s, an estimated 400,000 people were being killed by smallpox annually. The first ever vaccine was created to ward off smallpox.

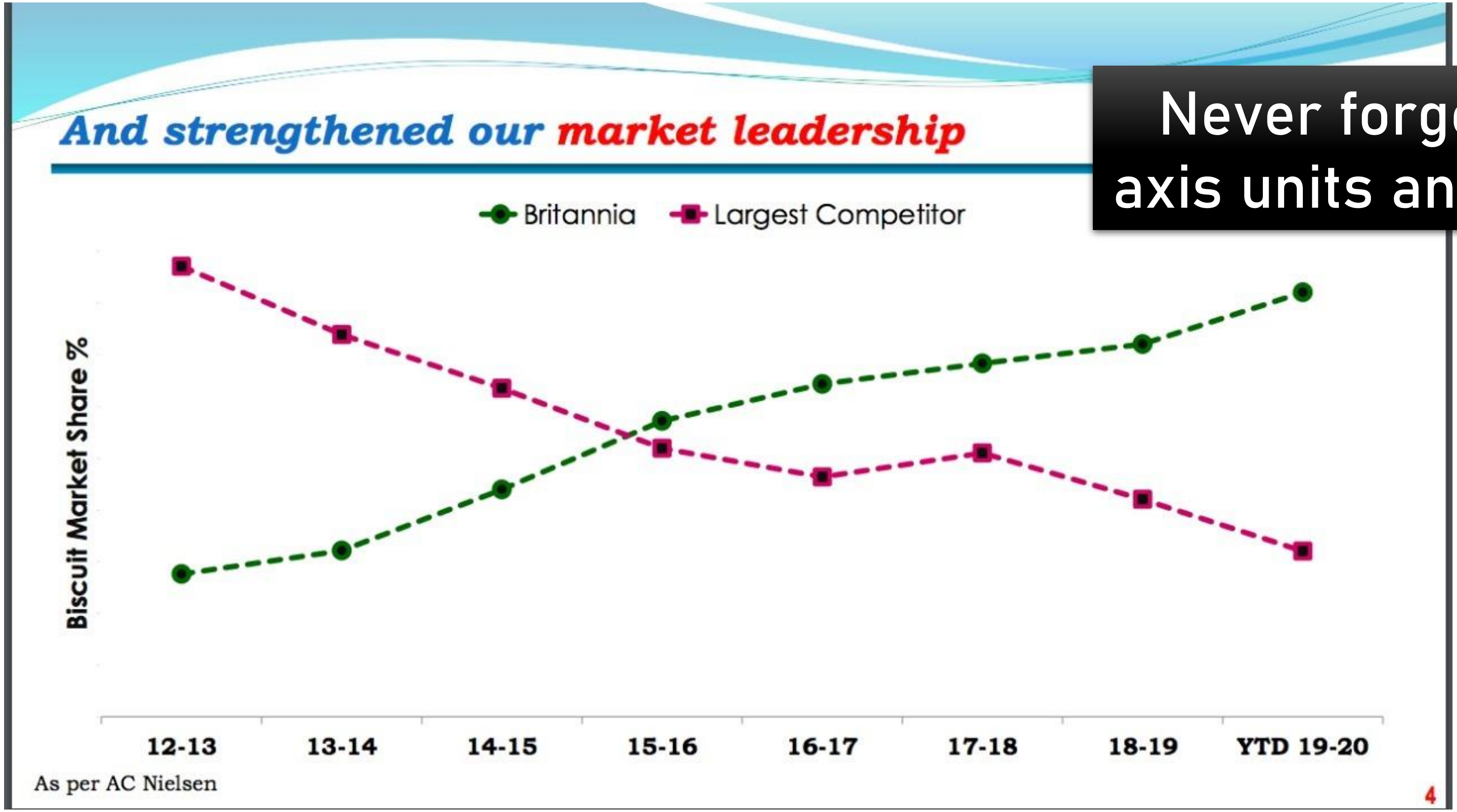
The death toll of this plague is still under debate as new evidence is uncovered, but many think it may have helped hasten the fall of the Roman Empire.

*Johns Hopkins University estimates

Sources:
CDC, WHO, BBC,
Wikipedia,
Historical records,
Encyclopedia Britannica
Johns Hopkins University

https://www.visualcapitalist.com/history-of-pandemics-deadliest/

CASE 1: What do you think about this visualization?



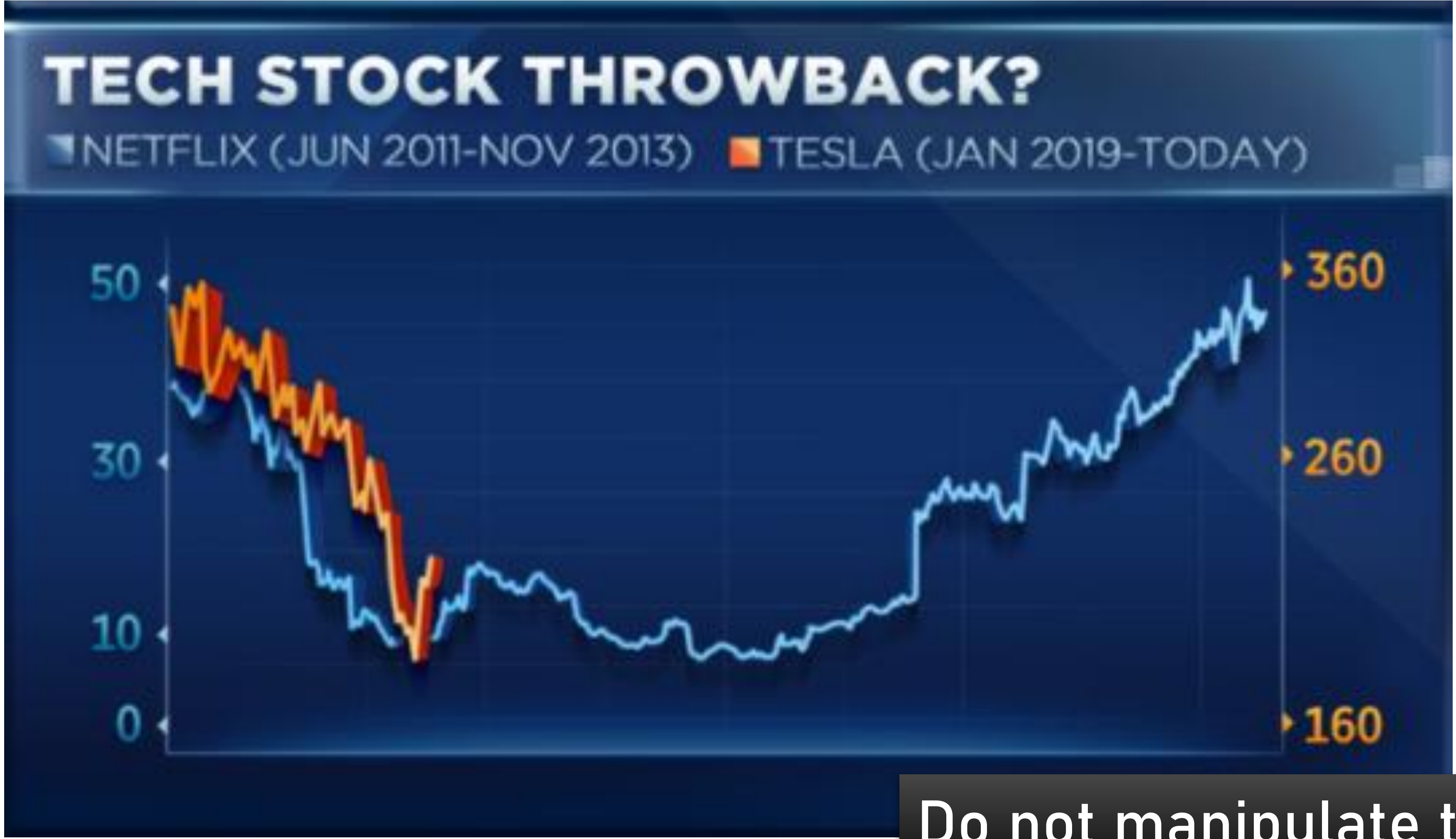
And strengthened our market leadership

As per AC Nielsen

Never forget the axis units and scale

https://at.tumblr.com/badvisualisations/this-is-from-britannias-investor-presentation/o7hi1vwlyu96

CASE 2: What do you think about this visualization?

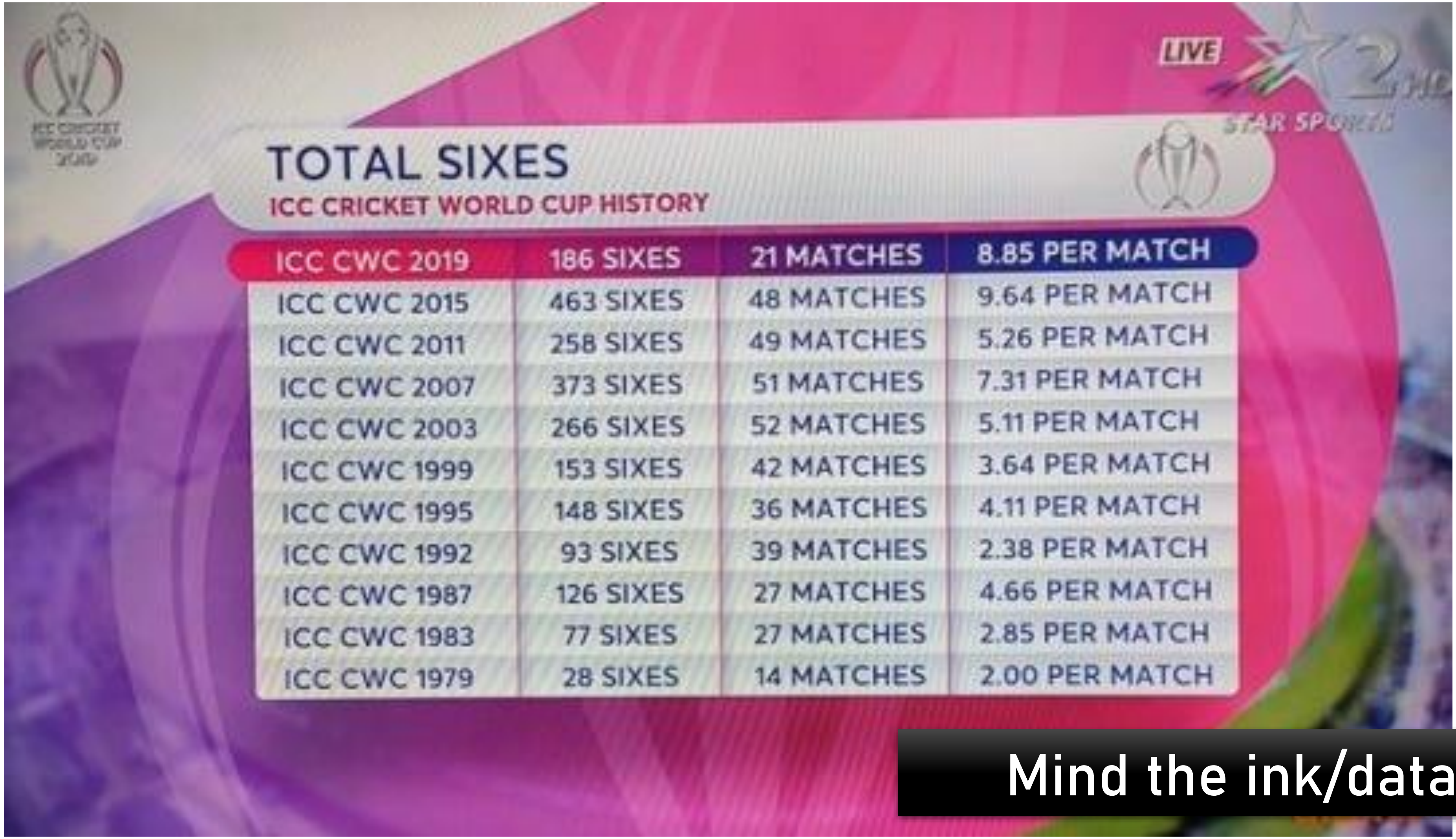


X axis missing

Do not manipulate the data to fit an idea. Use from-0 and proportional axis

https://www.cnbc.com/2019/06/12/tesla-looks-like-netflix-did-in-2011and-it-may-see-a-similar-r

CASE 3: What do you think about this visualization?



TOTAL SIXES

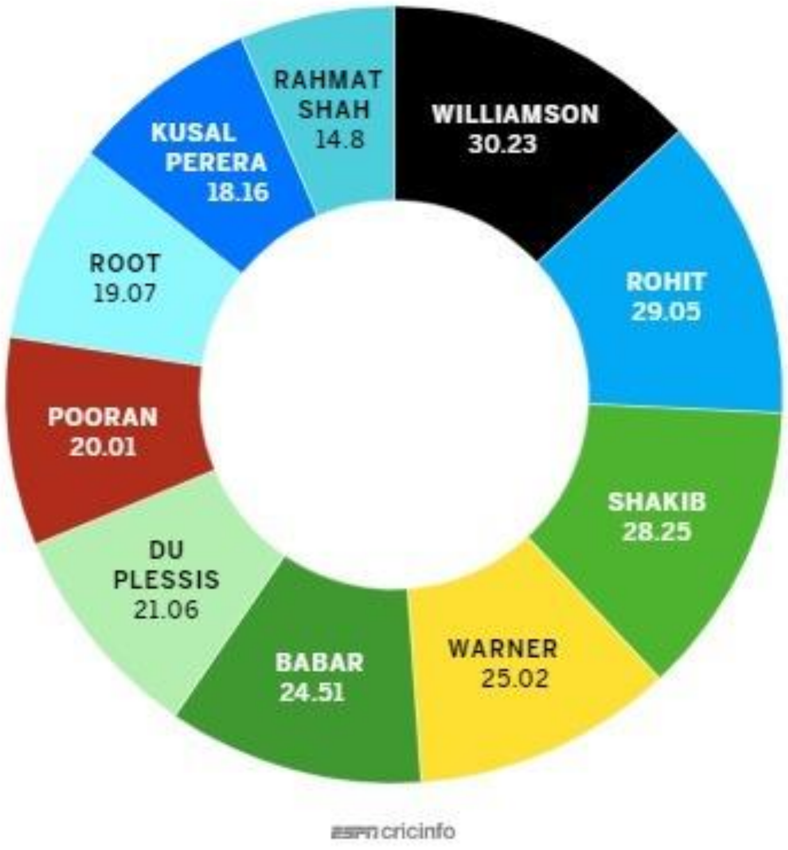
ICC CRICKET WORLD CUP HISTORY

ICC CWC 2019	186 SIXES	21 MATCHES	8.85 PER MATCH
ICC CWC 2015	463 SIXES	48 MATCHES	9.64 PER MATCH
ICC CWC 2011	258 SIXES	49 MATCHES	5.26 PER MATCH
ICC CWC 2007	373 SIXES	51 MATCHES	7.31 PER MATCH
ICC CWC 2003	266 SIXES	52 MATCHES	5.11 PER MATCH
ICC CWC 1999	153 SIXES	42 MATCHES	3.64 PER MATCH
ICC CWC 1995	148 SIXES	36 MATCHES	4.11 PER MATCH
ICC CWC 1992	93 SIXES	39 MATCHES	2.38 PER MATCH
ICC CWC 1987	126 SIXES	27 MATCHES	4.66 PER MATCH
ICC CWC 1983	77 SIXES	27 MATCHES	2.85 PER MATCH
ICC CWC 1979	28 SIXES	14 MATCHES	2.00 PER MATCH

Mind the ink/data ratio

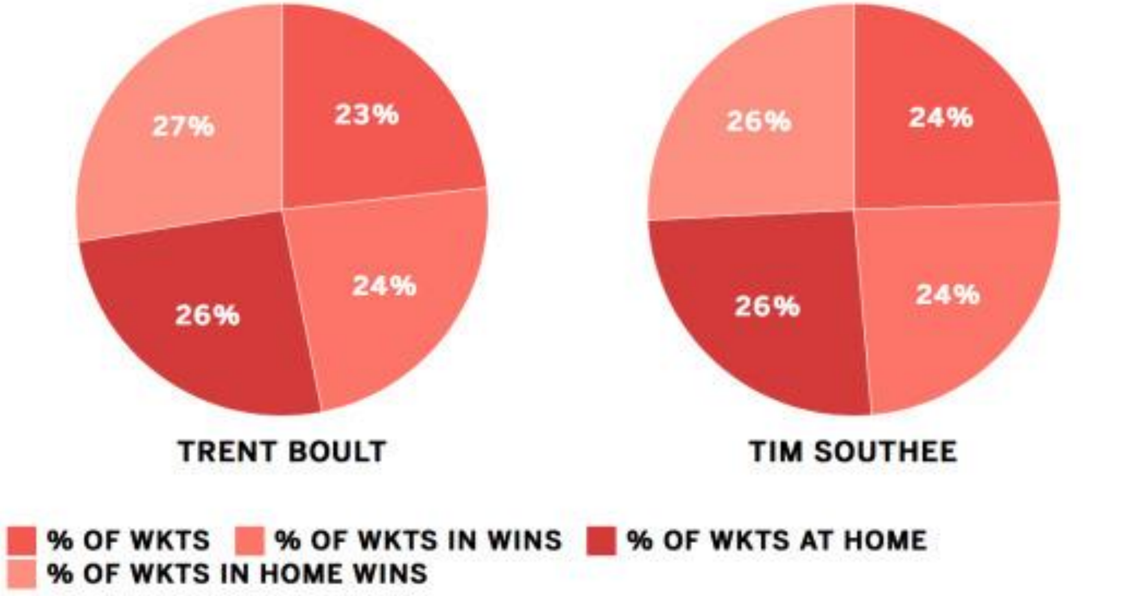
https://at.tumblr.com/badvisualisations/this-is-a-screengrab-from-star-sports/1p97g4l7r6hj

CASE 4: What do you think about these visualizations?



THE WORLD CUP'S BIG GUNS

% OF TEAM'S RUNS SCORED BY TOP SCORER



BOULT AND SOUTHEE'S SHARE OF TEST WICKETS AMONG NEW ZEALAND BOWLERS

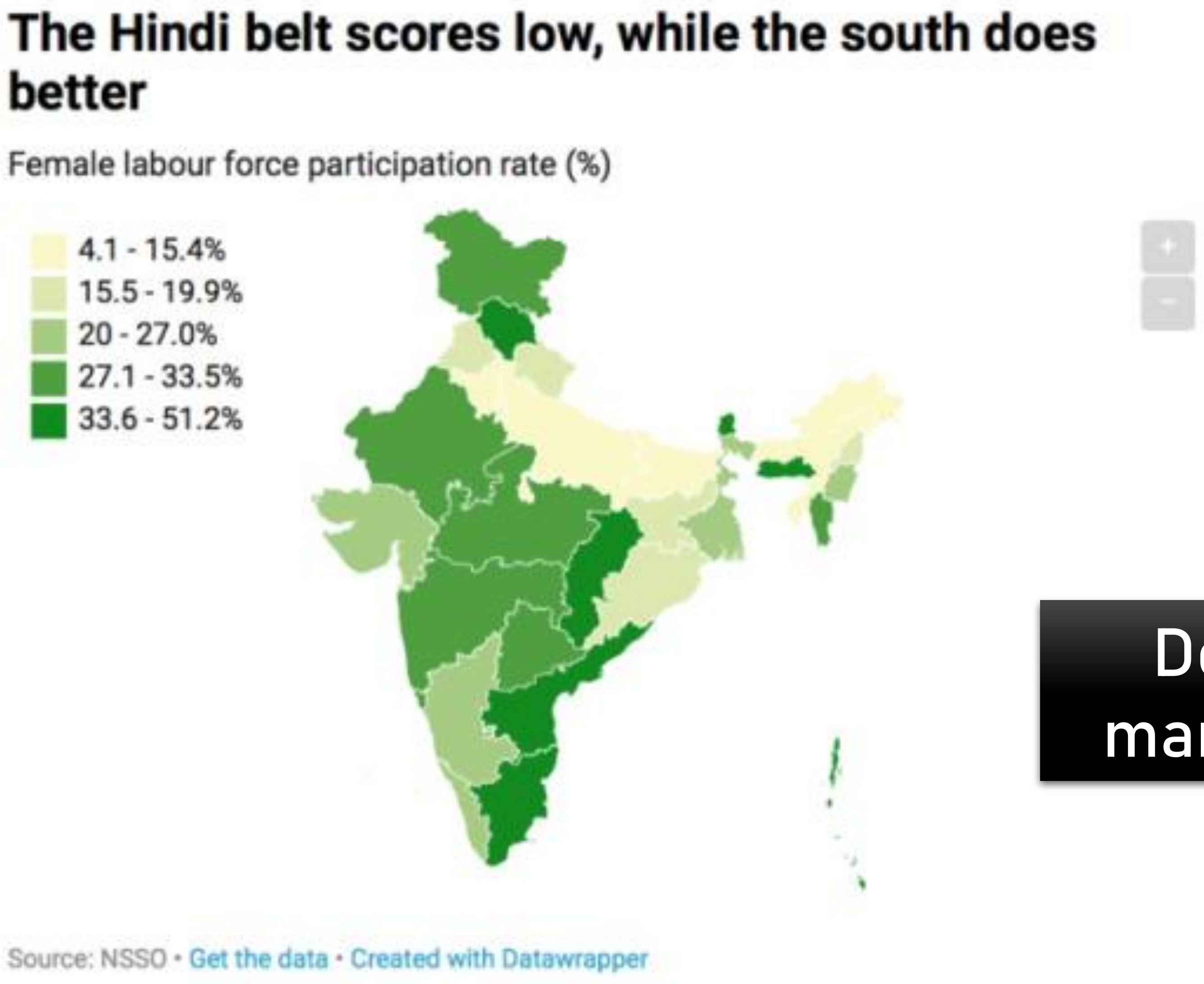
Pie charts are a bad choice

Mind the color

https://www.espncricinfo.com/story/_/id/27143430/kane-williamson-hand-steadies-new-zealand-ship

http://www.thecricketmonthly.com/story/1181193

CASE 5: What do you think about this visualization?



The Hindi belt scores low, while the south does better

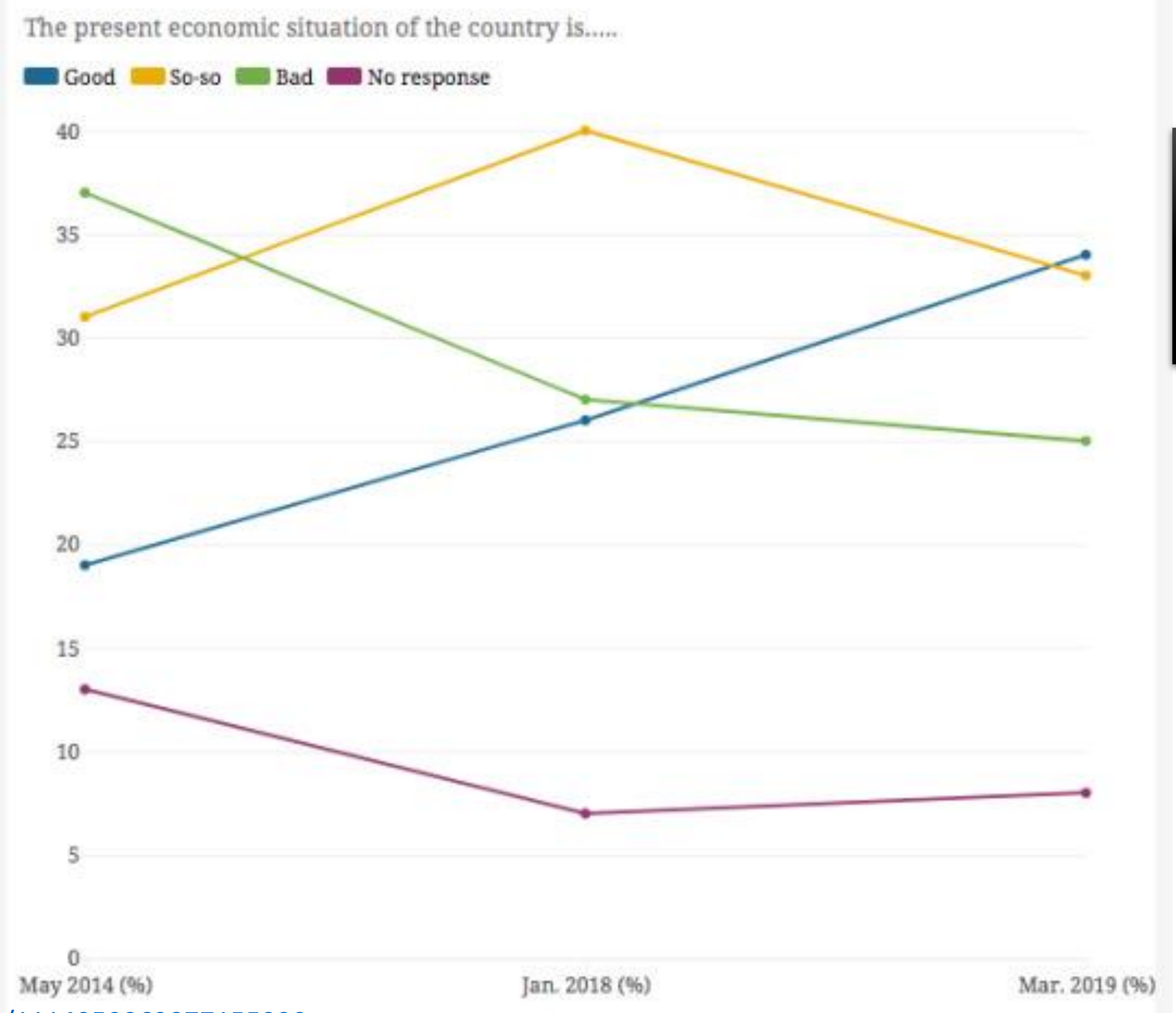
Female labour force participation rate (%)

Source: NSSO • Get the data • Created with Datawrapper

Do not be manipulative

https://www.livemint.com/news/india/india-s-workforce-is-masculinising-rapidly-1560150389726.html

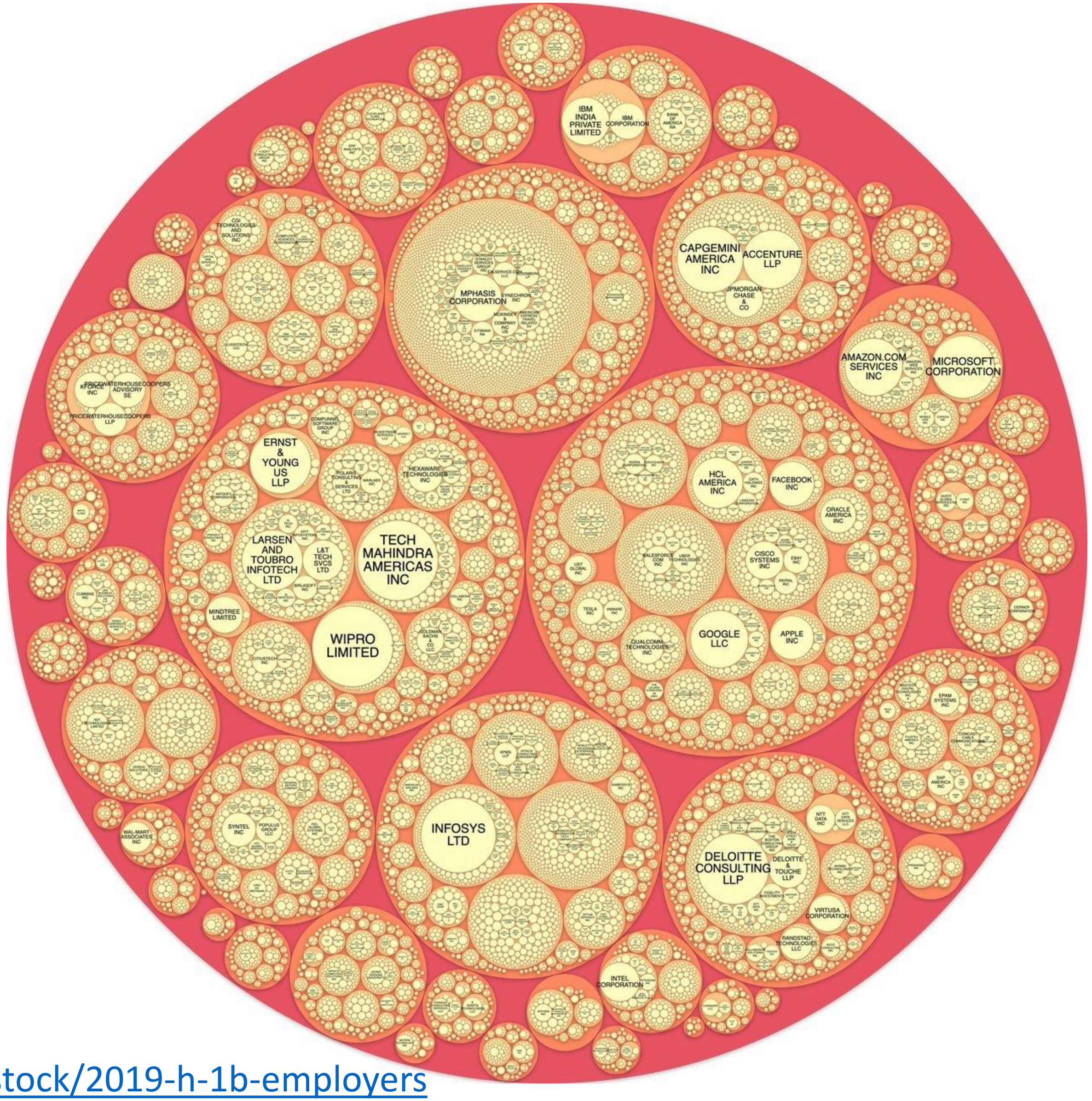
CASE 6: What do you think about this visualization?



Do not treat numerical as categorical

https://twitter.com/MilanV/status/1114950062877155329

CASE 7: What do you think about this “pizza” visualization?

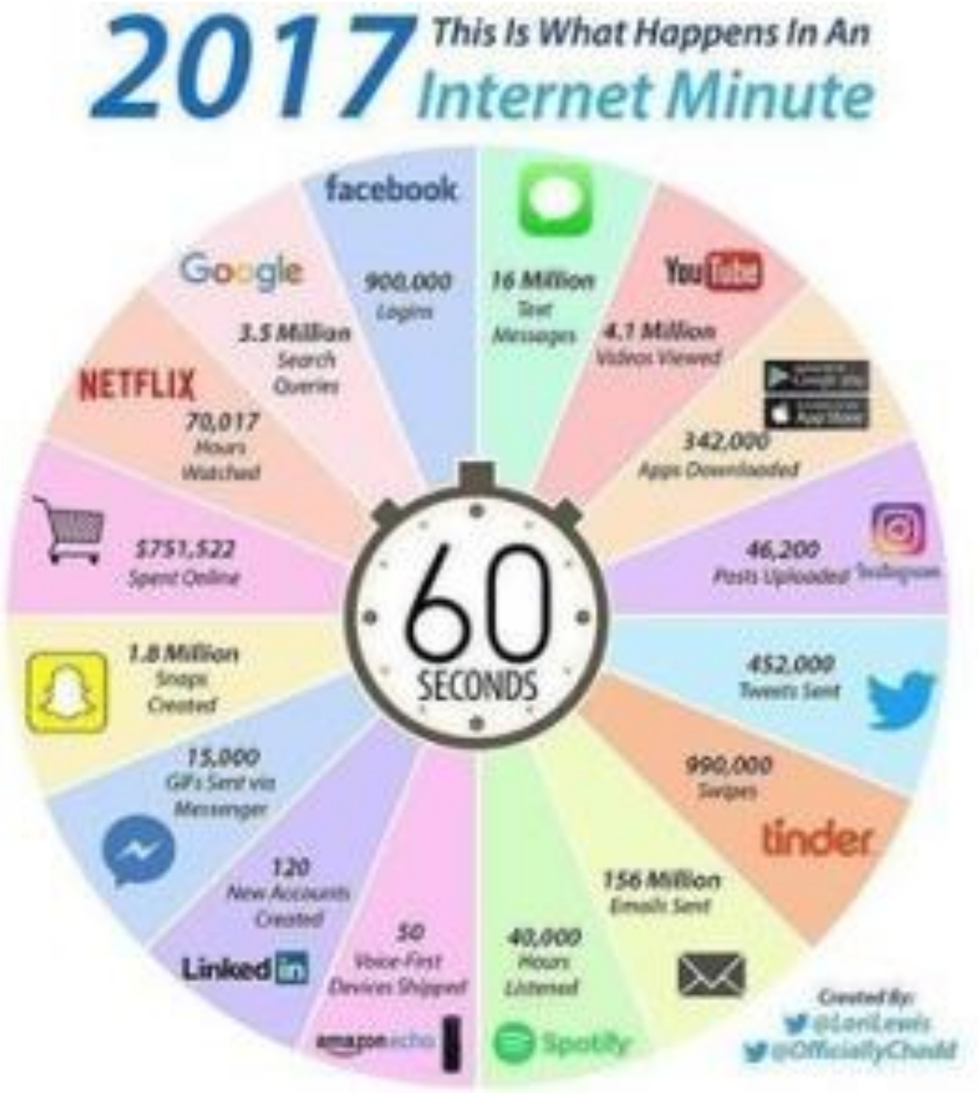


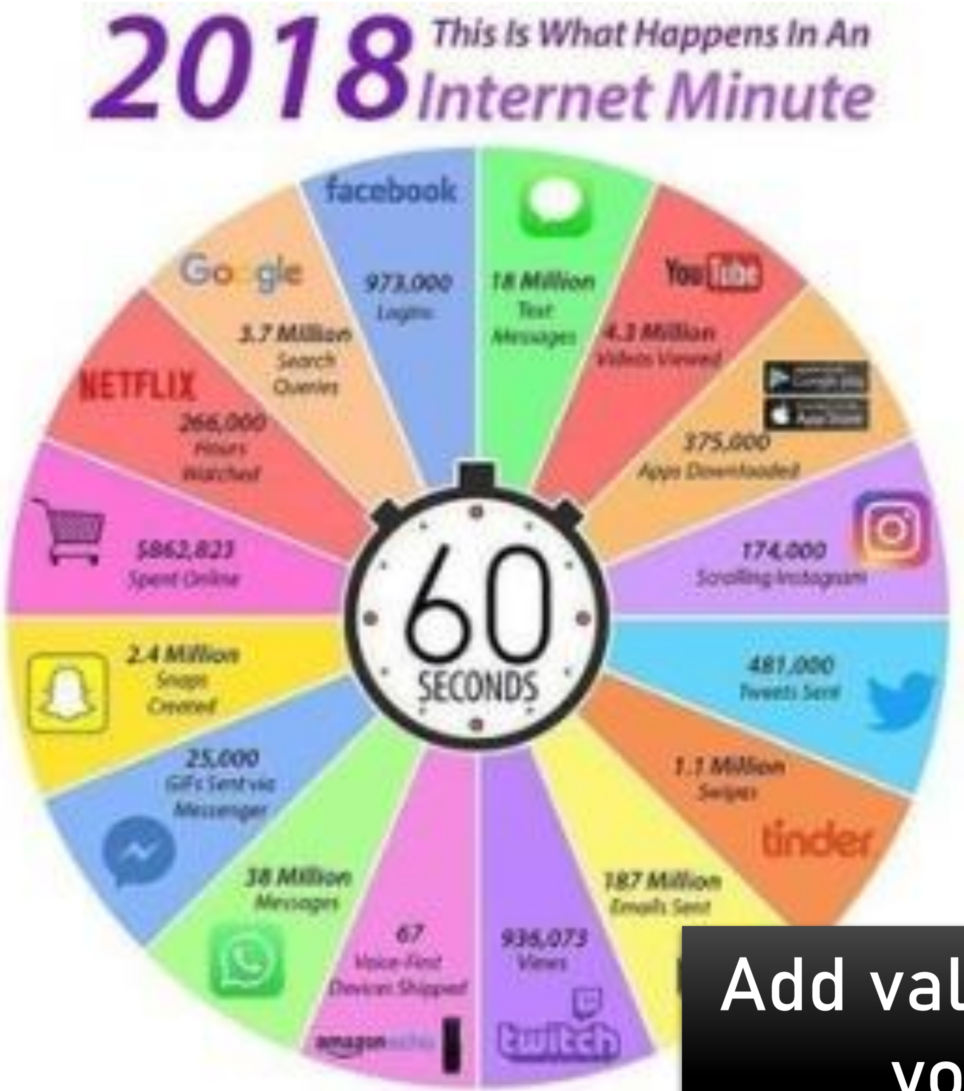
Add value with your visualization

Add legend if necessary

https://observablehq.com/@mbostock/2019-h-1b-employers

CASE 8: What do you think about this visualization?

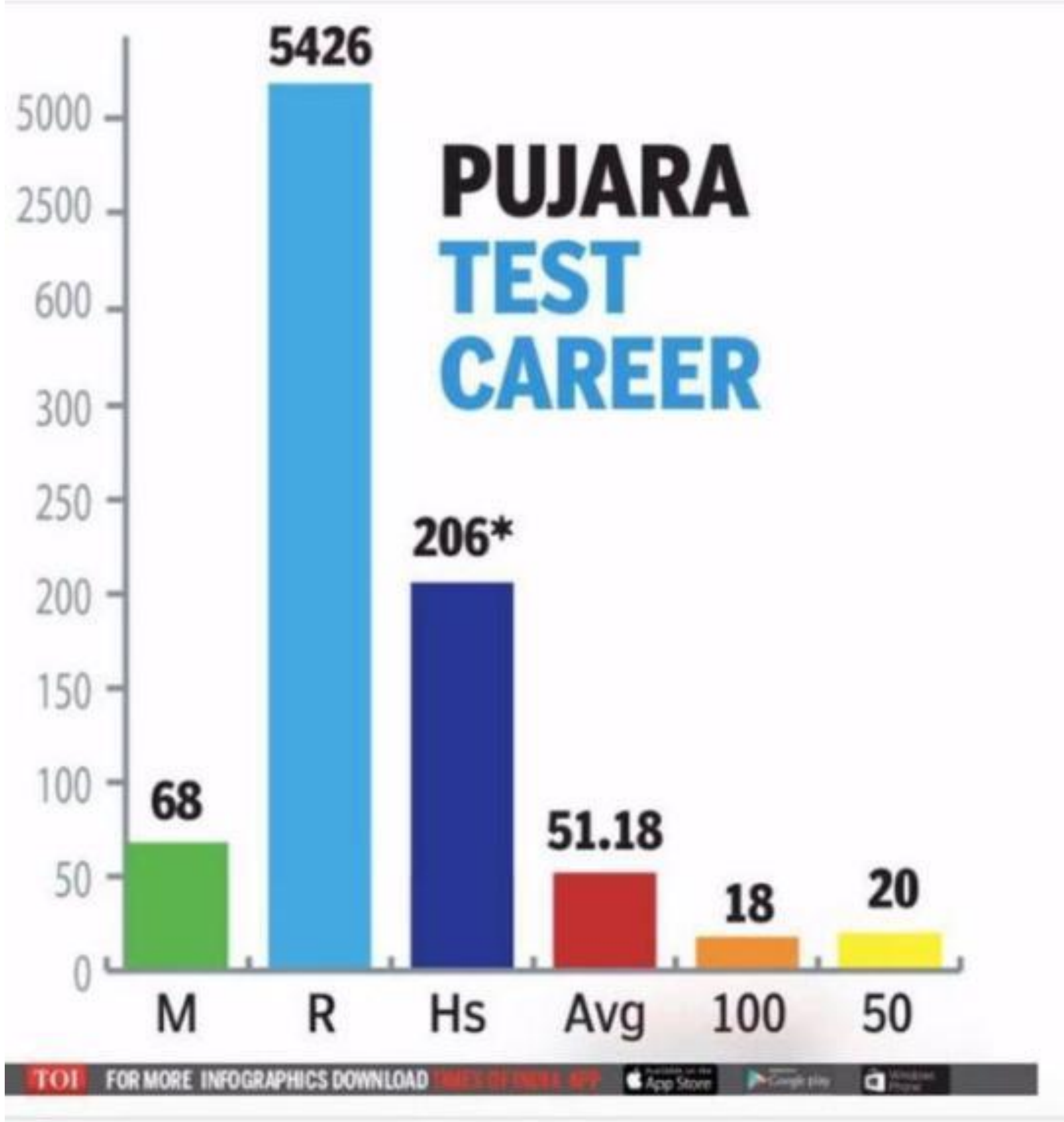




Add value with your visualization

https://at.tumblr.com/badvisualisations/regular-readers-of-this-tumblr-may-not-require/x3opxh7vdrt0

CASE 9: What do you think about this visualization?



Do not mix data types

Via Bhaskar Dasgupta on LinkedIn

CASE 10: What do you think about these visualizations?

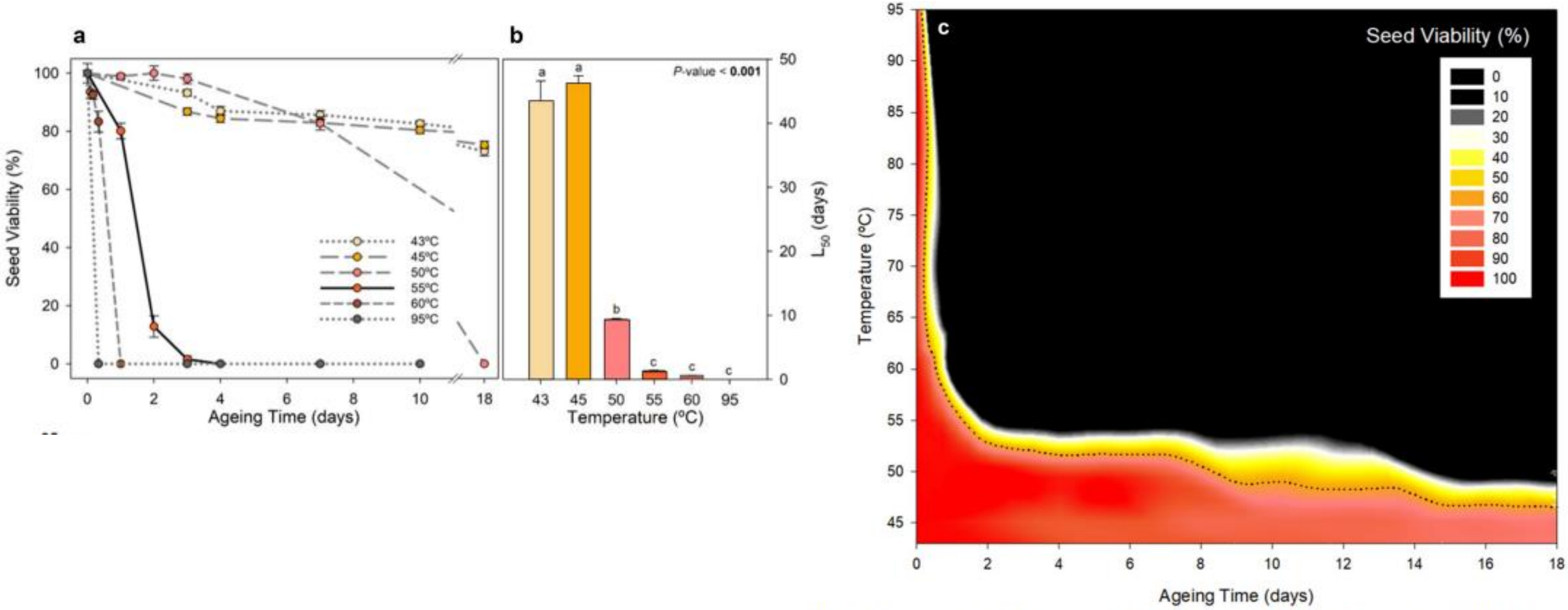
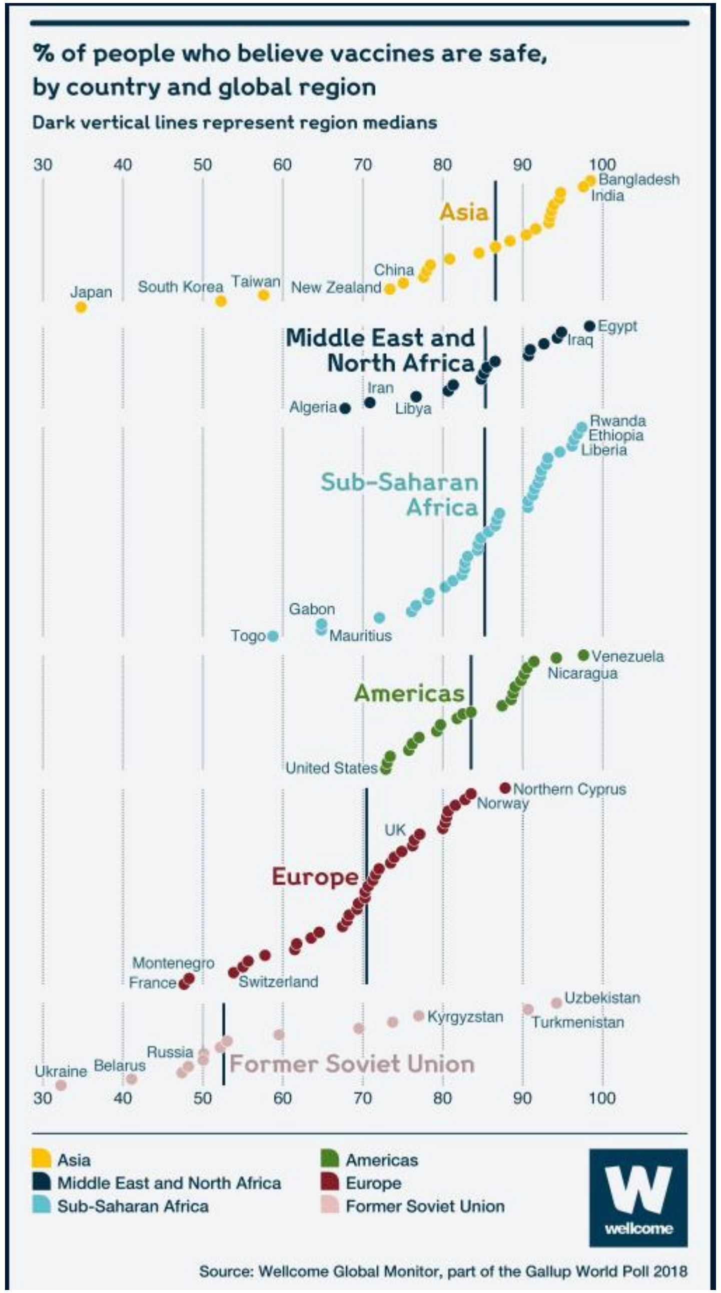


Fig. 2 Influence of temperature and aging time in artificially aged seeds of *C. edulis*. **a** Viability loss of *C. edulis* seeds at the different tested temperatures and 87% RH. Data is represented as Mean ± SE ($n = 6$). **b** L_{50} (loss of 50% viability) at the different tested temperatures. Different letters indicate statistically significant differences (*P*-value < 0.05). **c** Contour plot of seed viability considering temperature and aging days. Dotted line represents the L_{50}, 50% viability loss

Fenollosa et al., 2020 Plant Methods

Transforming a visualization

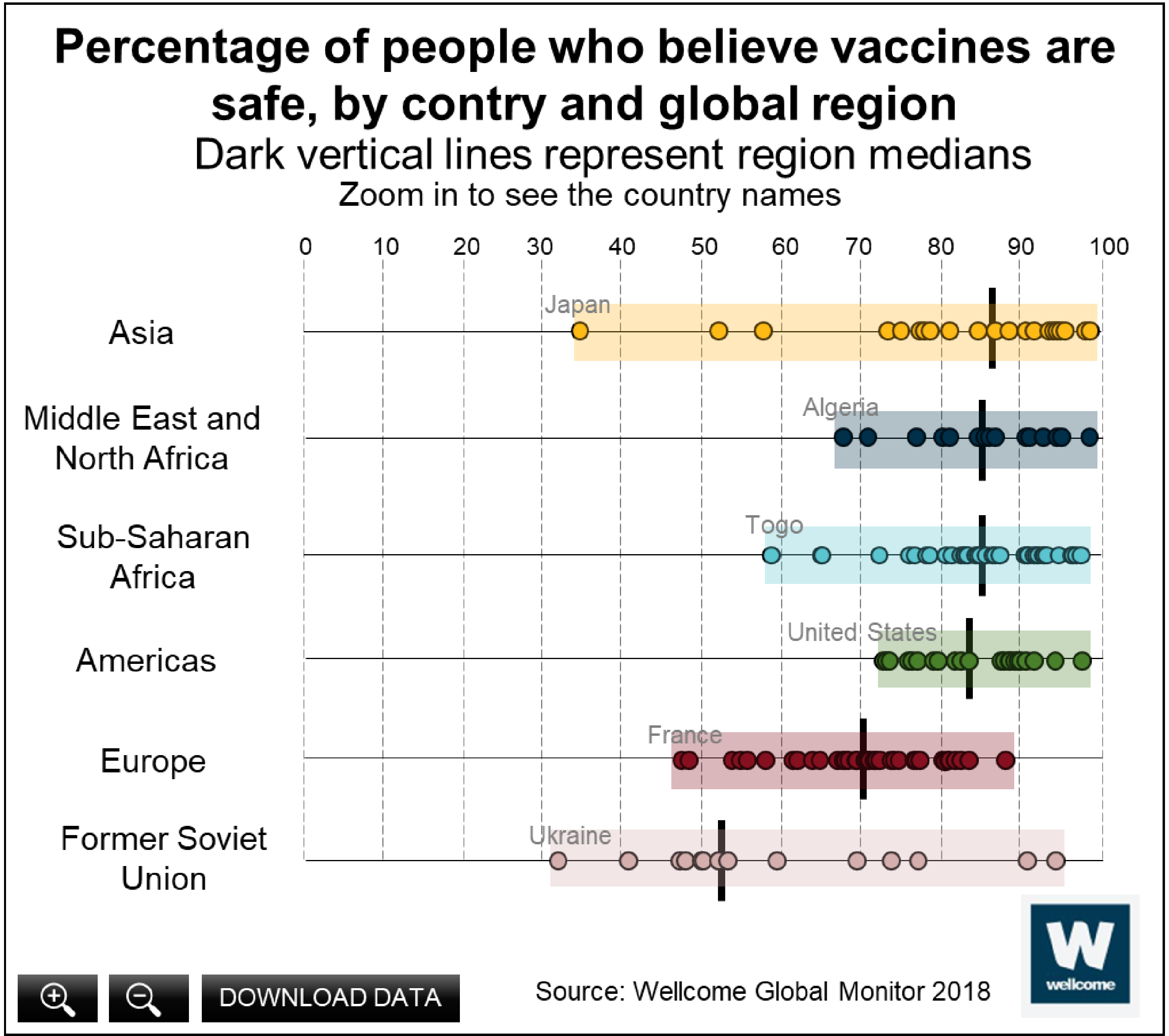


% of people who believe vaccines are safe, by country and global region

Dark vertical lines represent region medians

Source: Wellcome Global Monitor, part of the Gallup World Poll 2018

My proposal



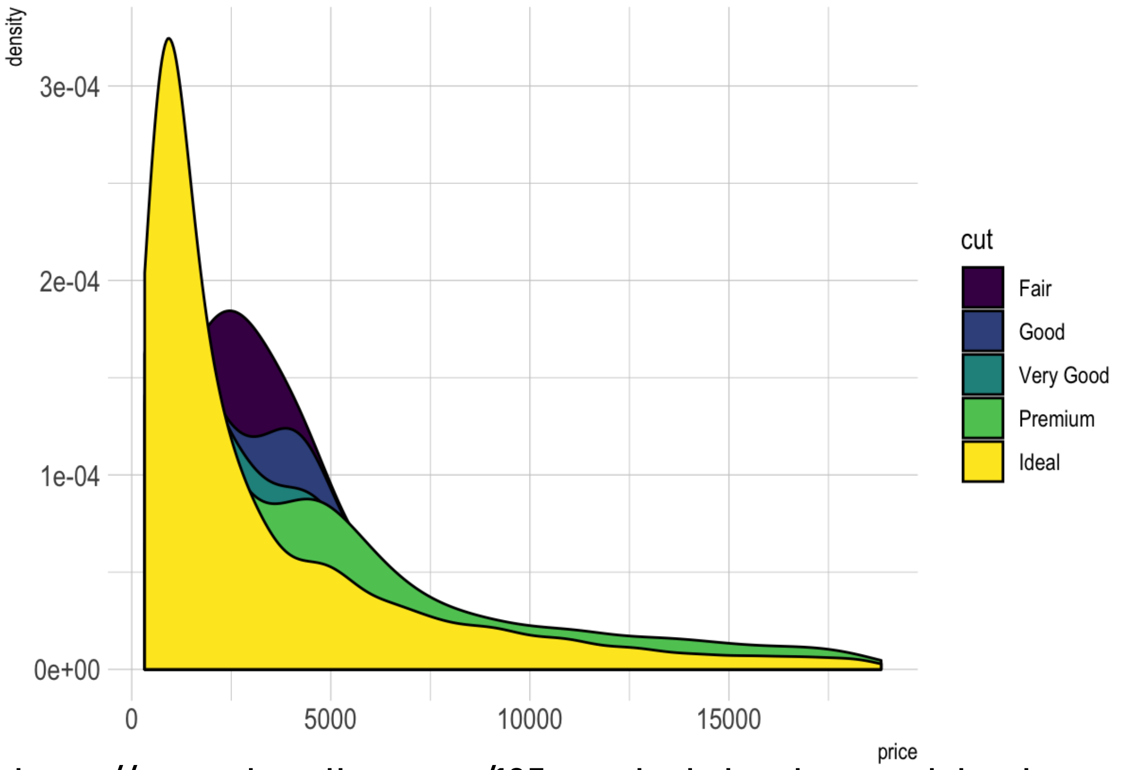
Percentage of people who believe vaccines are safe, by contry and global region

Dark vertical lines represent region medians

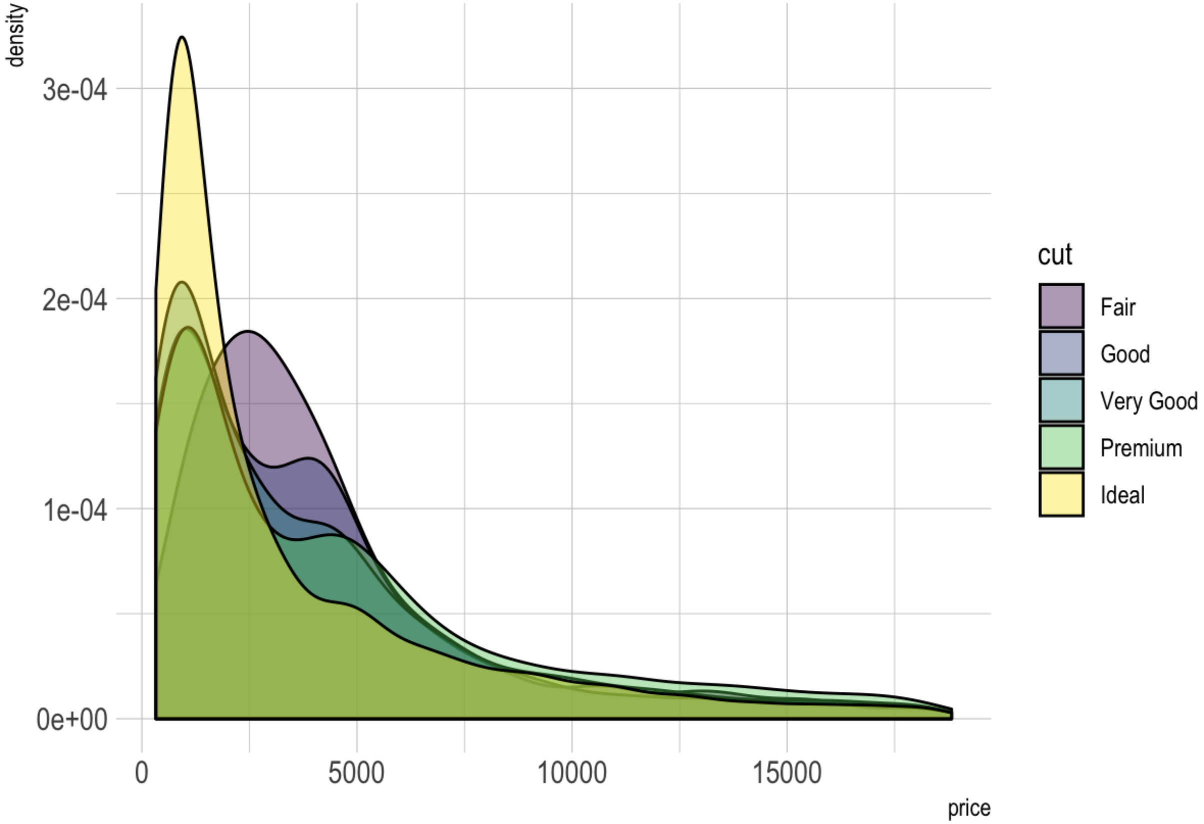
Zoom in to see the country names

Source: Wellcome Global Monitor 2018

What’s the problem with this plot?



https://r-graph-gallery.com/135-stacked-density-graph.html



Which is the best visualization for the diamont dataset?

How would you improve now your own creations?

Statistics supporting ideas

By now we are working with just two variables of interest (categorical and numerical or both numerical):

- Both numerical: Correlation significance and adjustment to test relationships
- Numerical and categorical: ANOVA to compare groups

Next sesion

We will cover one of the main methods for contrasting groups: ANOVA (linear models)

1. Choose two numerical and one categorical variable from your dataset

2. Think about a research question and writte it down

3. Test with a a linear model (lm, aov, glm, etc.)

    - Perform an ANOVA to compare categorical-numerical

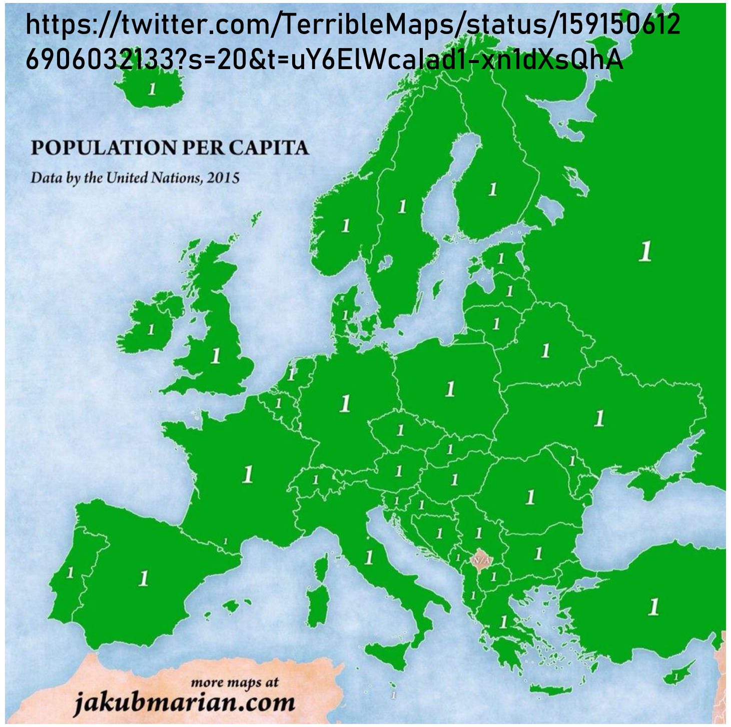
    - Perform a Pearson correlation for the two numericals.

Additional: Create a GitHub profile and start your portfolio

See more:

- https://twitter.com/terriblemaps
- https://twitter.com/amazingmap
- https://www.connectedpapers.com/

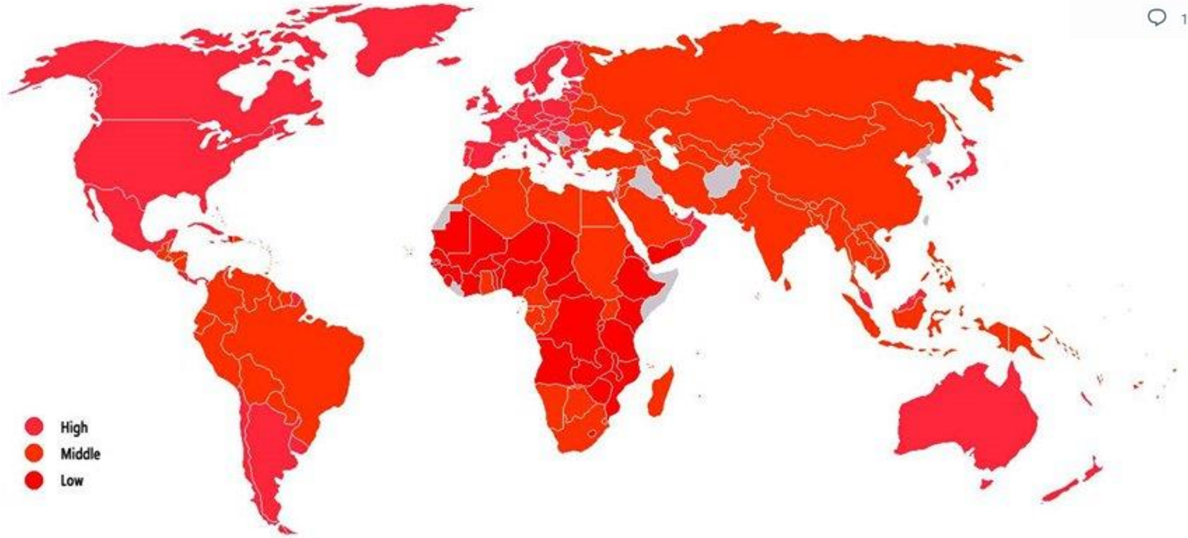
https://twitter.com/TerribleMaps/status/1591506126906032133?s=20&t=uY6ElWcalad1-xn1dXsQhA



Terrible Maps @TerribleMaps · 24 oct.
Locations Johnny Cash claims to have been in "I've been everywhere"

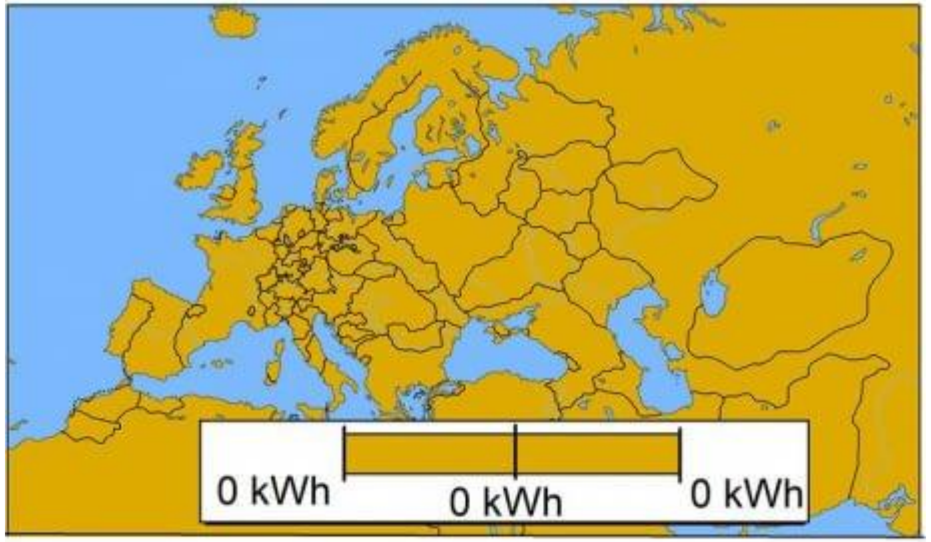
194 2.369 25,3 mil

World Incidence of Color Blindness



https://twitter.com/TerribleMaps/status/1592618803523117057?s=20&t=uY6ElWcalad1-xn1dXsQhA

Electricity consumption in Europe in 1507



https://twitter.com/TerribleMaps/status/1583067072858181635?s=20&t=uY6ElWcalad1-xn1dXsQhA