CMPU 100 · Programming with Data

Introduction to Visualization

Class 9

What's the point of visualization?

Exploratory

Explanatory

What's the point of visualization?

Exploratory

To further your own understanding of your results

Explanatory

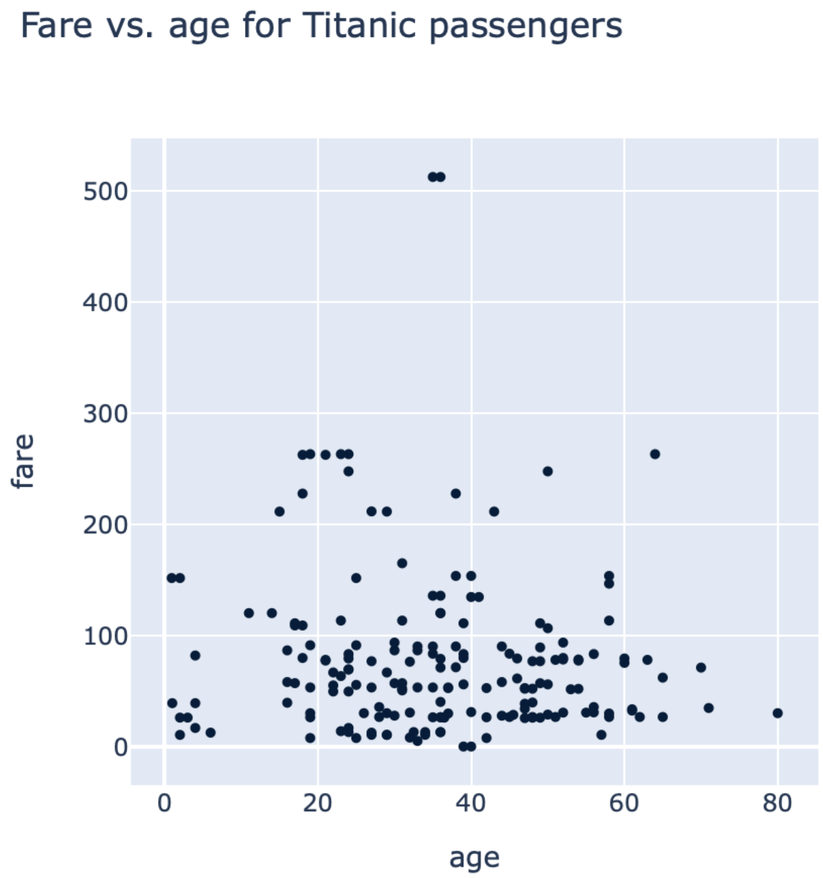
What's the point of visualization?

Exploratory — *To further your own understanding of your results*

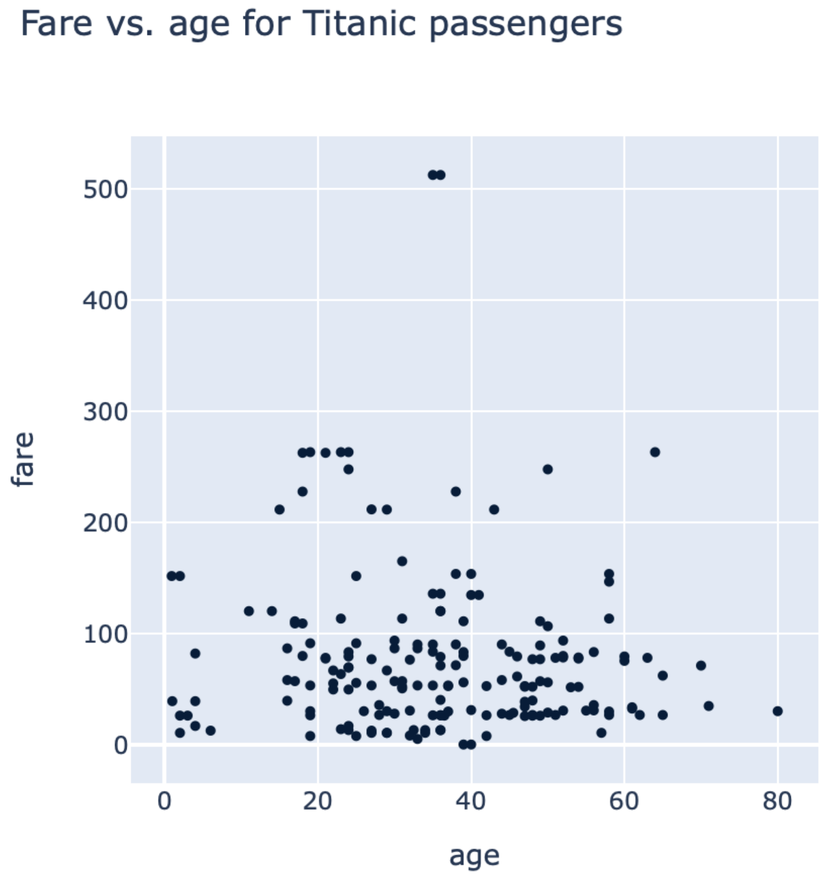
Explanatory — *To communicate your results with others*

Age	*Fare*
38	71.2833
35	53.1
54	51.8625
4	16.7
58	26.55
58	26.55
34	13
⋮	⋮

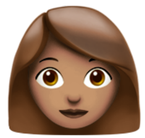
Age	*Fare*
38	71.2833
35	53.1
54	51.8625
4	16.7
58	26.55
58	26.55
34	13
⋮	⋮



Age	Fare
38	71.2833
35	53.1
54	51.8625
4	16.7
58	26.55
58	26.55
34	13
⋮	⋮



Looks like older people didn't spend more than younger people.



These datasets all have several statistical properties in common – means, standard deviations, and correlations

These datasets all have several statistical properties in common – means, standard deviations, and correlations

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

These datasets all have several statistical properties in common – means, standard deviations, and correlations

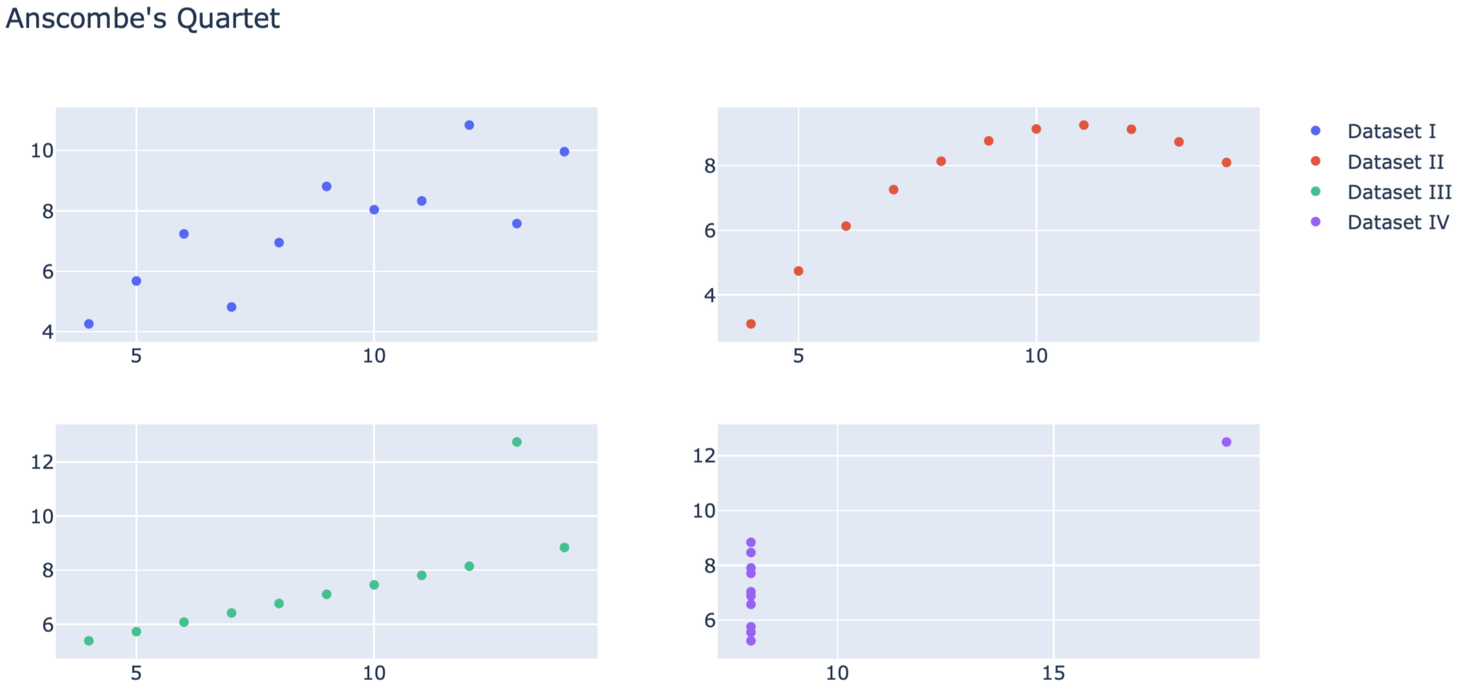
But they look very different when plotted!

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

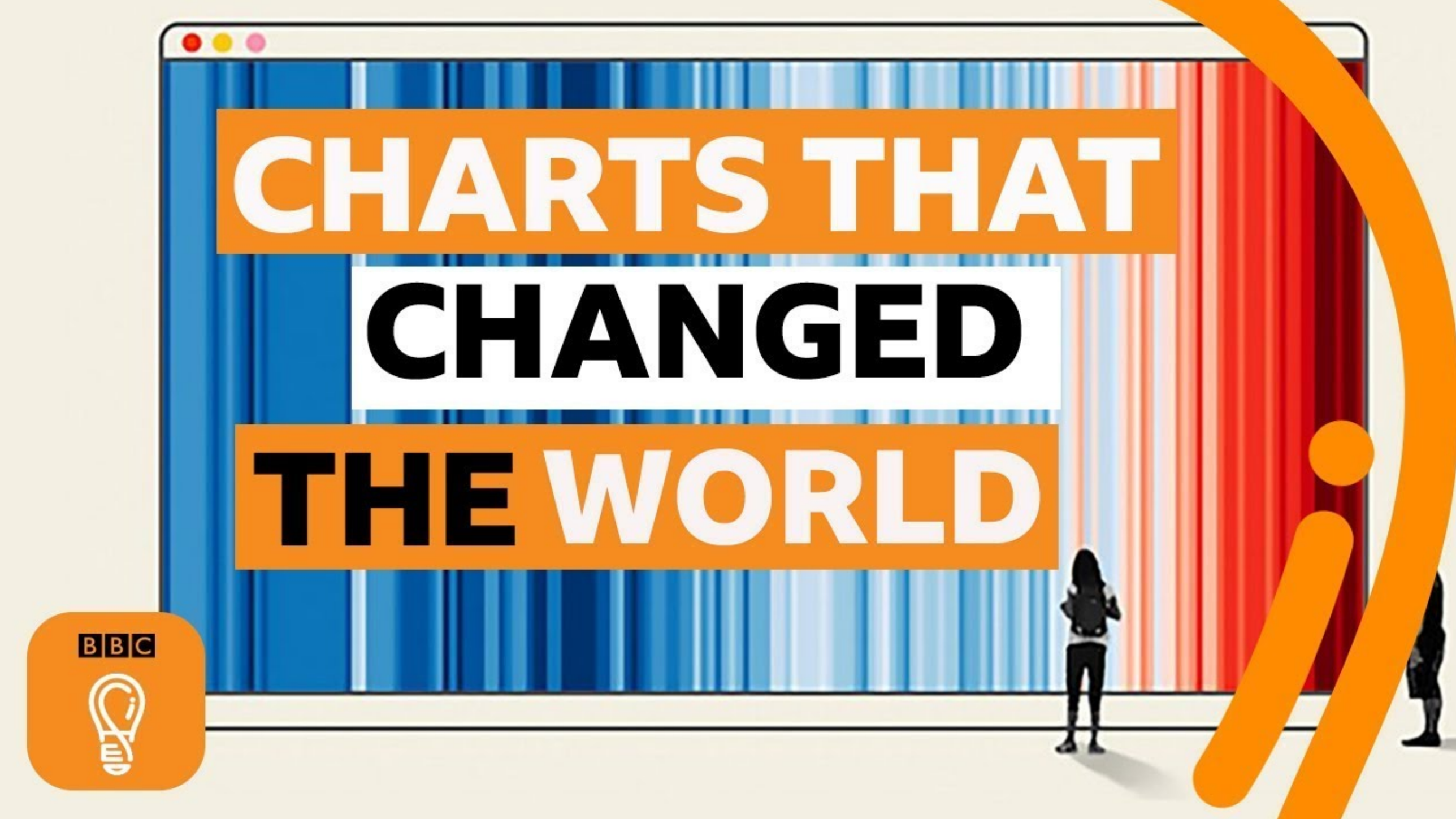
x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

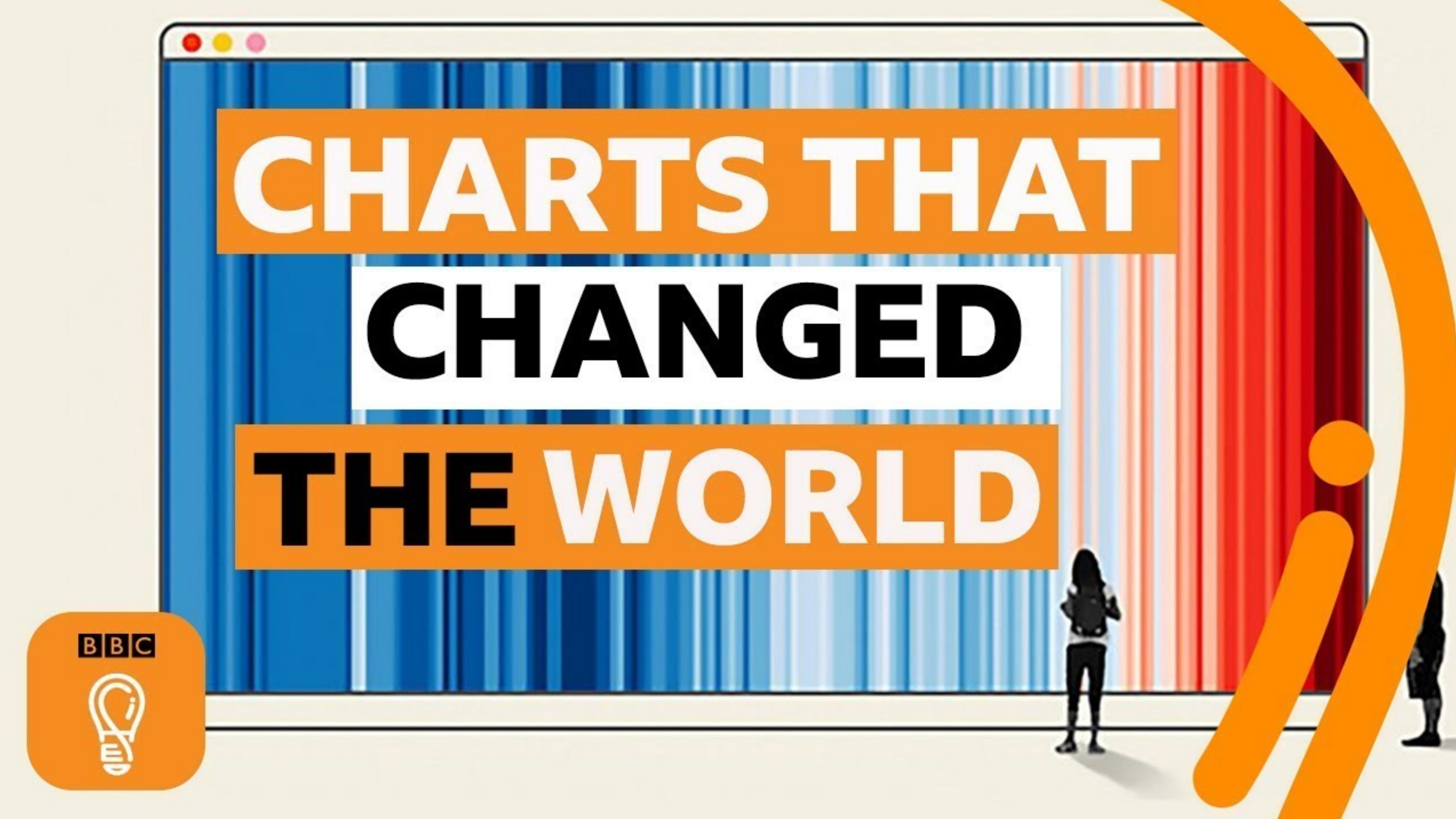
x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89



Visual explanations

Our ideas of what makes an effective – and honest – visualization build on centuries of work to communicate the information contained in data.


CHARTS THAT
CHANGED
THE WORLD
BBC


CHARTS THAT
CHANGED
THE WORLD
BBC

ArcGIS Blog | Mapping | ArcGIS Pro | Dec 03, 2020

Something in the water: the mythology of Snow's map of cholera

By Kenneth Field

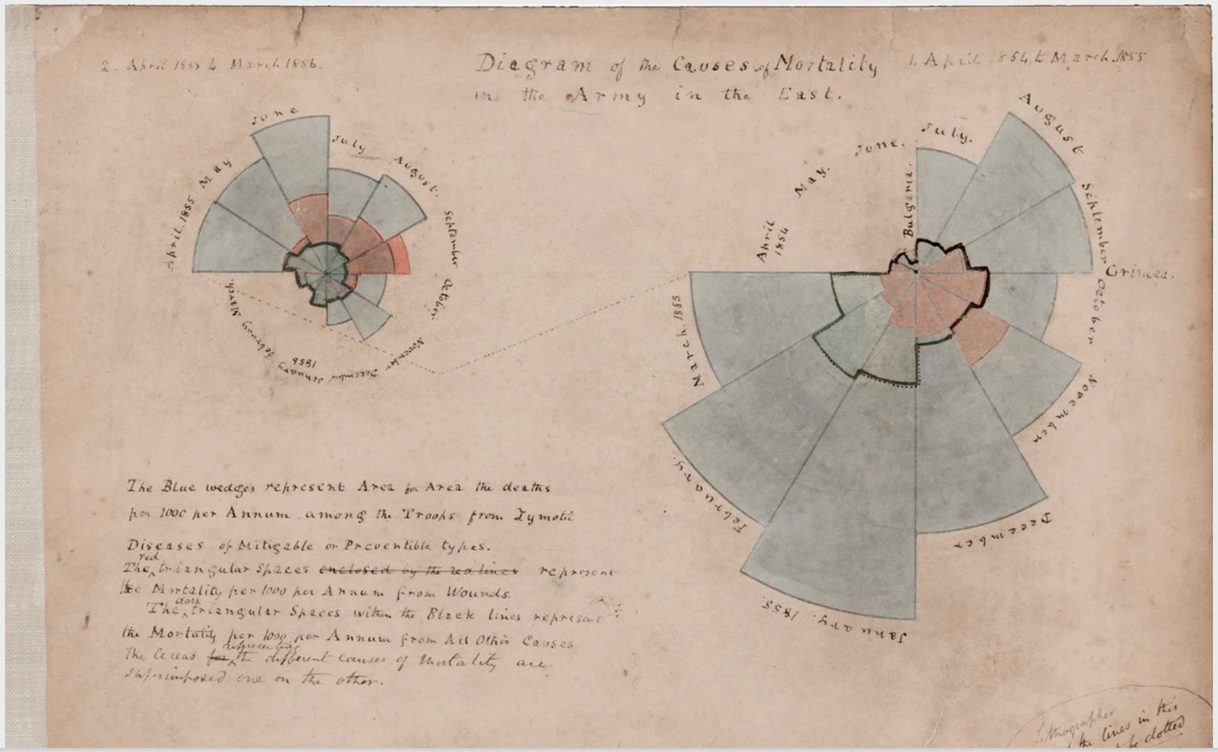
If there are positives to take from 2020 (a big ask I know but bear with me), the spotlight has shone brightly on the use of maps, charts, and data that help us understand COVID-19. It's quite literally been a viral moment for geography and cartography. Not that maps weren't vital visual and analytical tools beforehand, but they have taken centre stage this year. And of course, examining the role of geography, and cartography in contemporary infectious disease epidemiology has also meant looking at the past for examples. There's definitely been something in the water in 2020 as inevitably the story of Dr John Snow and his map of cholera showing the outbreak in Soho, London in 1854 has frequently been recounted. You may know the story. Or you may think you know the story. Let's see.

AUGUST 1, 2022 | 4 MIN READ

How Florence Nightingale Changed Data Visualization Forever

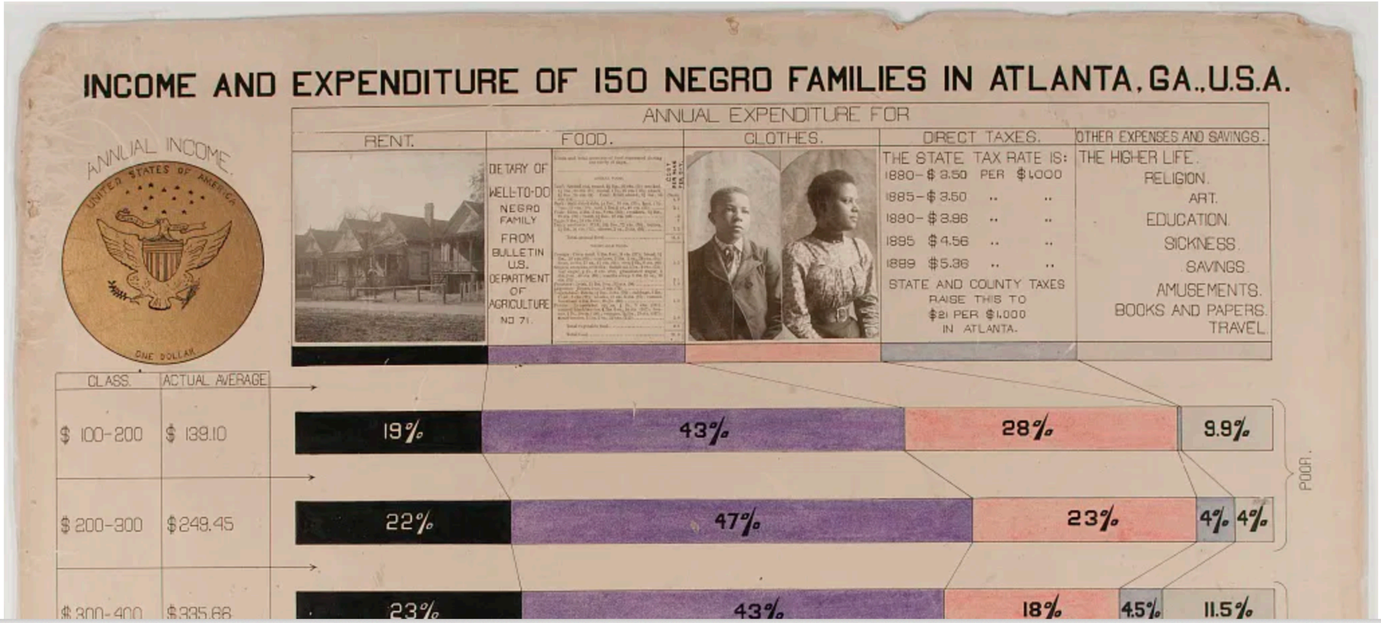
The celebrated nurse improved public health through her groundbreaking use of graphic storytelling

BY RJ ANDREWS



COLLECTIONS / IMAGES

W. E. B. Du Bois' Hand-Drawn Infographics of African-American Life (1900)



INTELLECTUAL AND DEVELOPMENTAL DISABILITIES
2012, Vol. 50, No. 2, 169–178

DOI: 10.1352/1934-9556-50.2.169

Who Was Deborah Kallikak?

J. David Smith and Michael L. Wehmeyer

Abstract

The Kallikak Family was, along with *The Jukes: A Study in Crime, Pauperism, Disease, and Heredity*, one of the most visible eugenic family narratives published in the early 20th century. Published in 1912 and authored by psychologist Henry Herbert Goddard, director of the psychological laboratory at the Vineland Training School for Feebleminded Children in Vineland, New Jersey, *The Kallikak Family* told the tale of a supposedly "degenerate" family from rural New Jersey, beginning with Deborah, one of the inmates at the Training School. Like most publications in the genre, this pseudoscientific treatise described generations of illiterate, poor, and purportedly immoral Kallikak family members who were chronically unemployed, supposedly feebleminded, criminal, and, in general, perceived as threats to "racial hygiene." Presented as a "natural experiment" in human heredity, this text served to support eugenic activities through much of the first half of the 20th century. This article reviews the story of Deborah Kallikak, including her true identity, and provides evidence that Goddard's treatise was incorrect.

Key Words: *history of intellectual disability; Vineland Training School; Kallikak family; Deborah Kallikak; Henry Herbert Goddard*

University of Reading

PARTNERING FOR THE PLANET

What the science says | What we are doing | What you can do | Climate Education | Our campus | Our research | Our programmes | Climate resources | Events

Partnering for the planet › Climate resources › Climate stripes

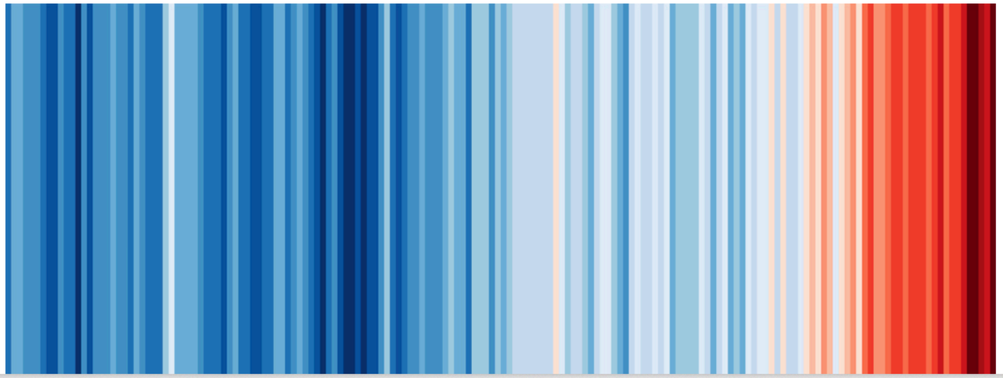
Climate stripes

Representing global temperature rise over two centuries

What are the stripes?

No words. No numbers. No graphs. Just a series of vertical coloured bars, showing the progressive heating of our planet in a single, striking image.

The climate stripes were created by Professor Ed Hawkins at the University of Reading in 2018



A few more

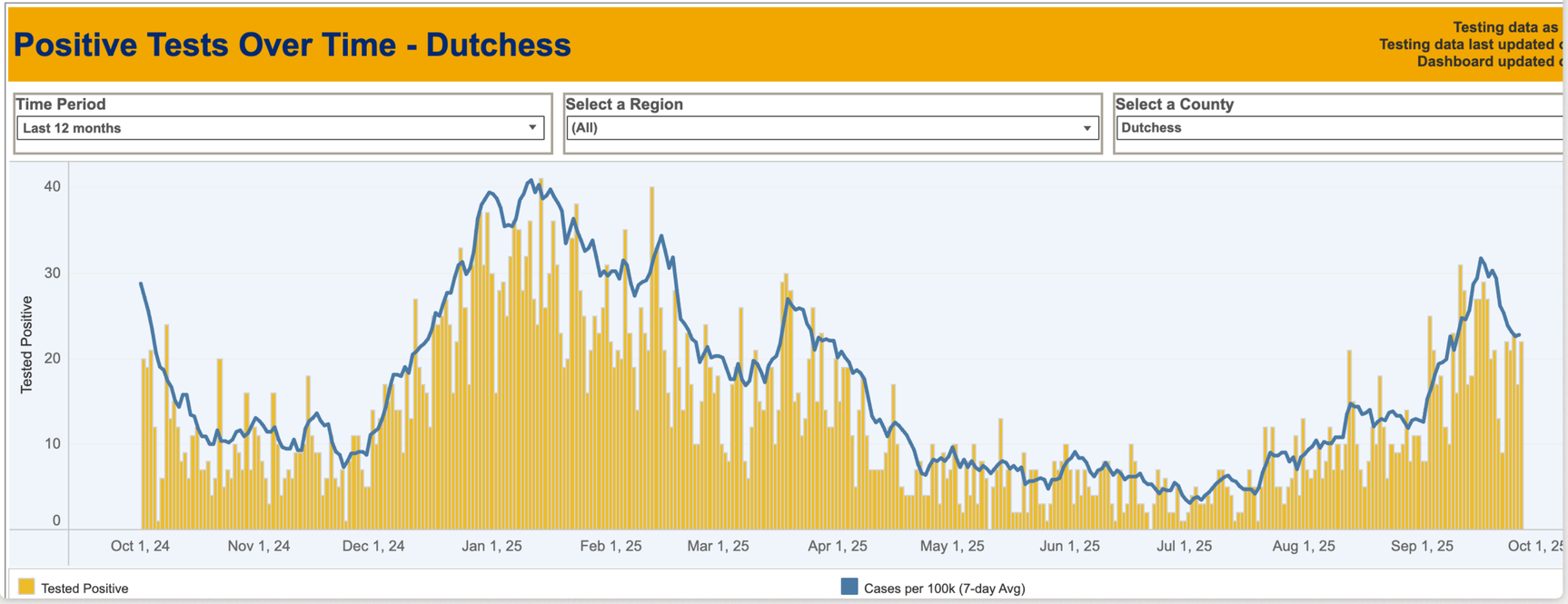
Maya Lin's Vietnam War Memorial, Washington, DC, 1982

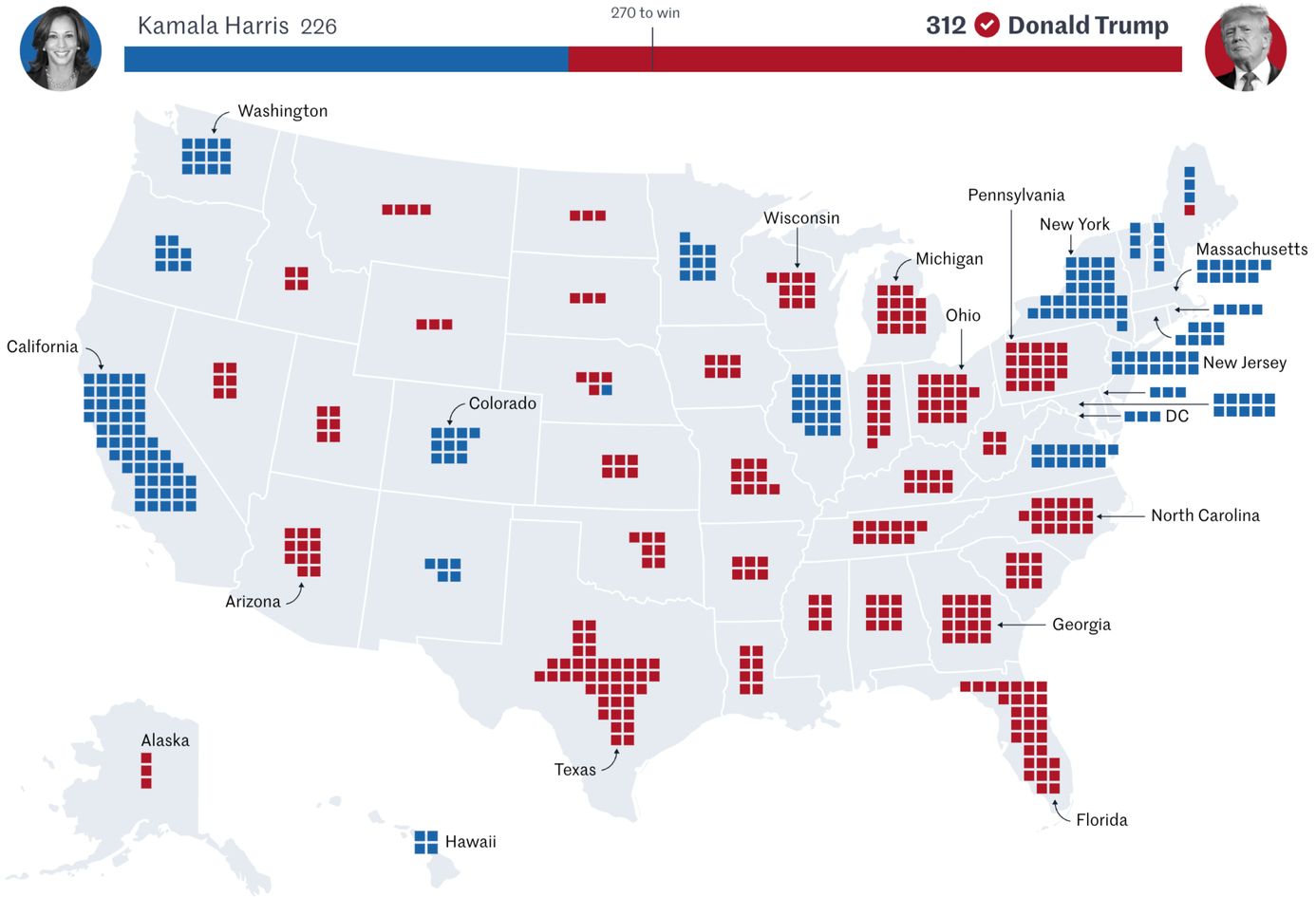
“From a distance the entire collection of names of 58,000 dead soldiers arrayed on the black granite yields a visual measure of what 58,000 means, as the letters of each name blur into a gray shape, cumulating to the final toll. When a viewer approaches, these shapes resolve into individual names. … We focus on the tragic information; absent are the big porticoes, steps and stairs, and other marble paraphernalia usually attached to grand official monuments.”

Edward Tufte, *Envisioning Information*

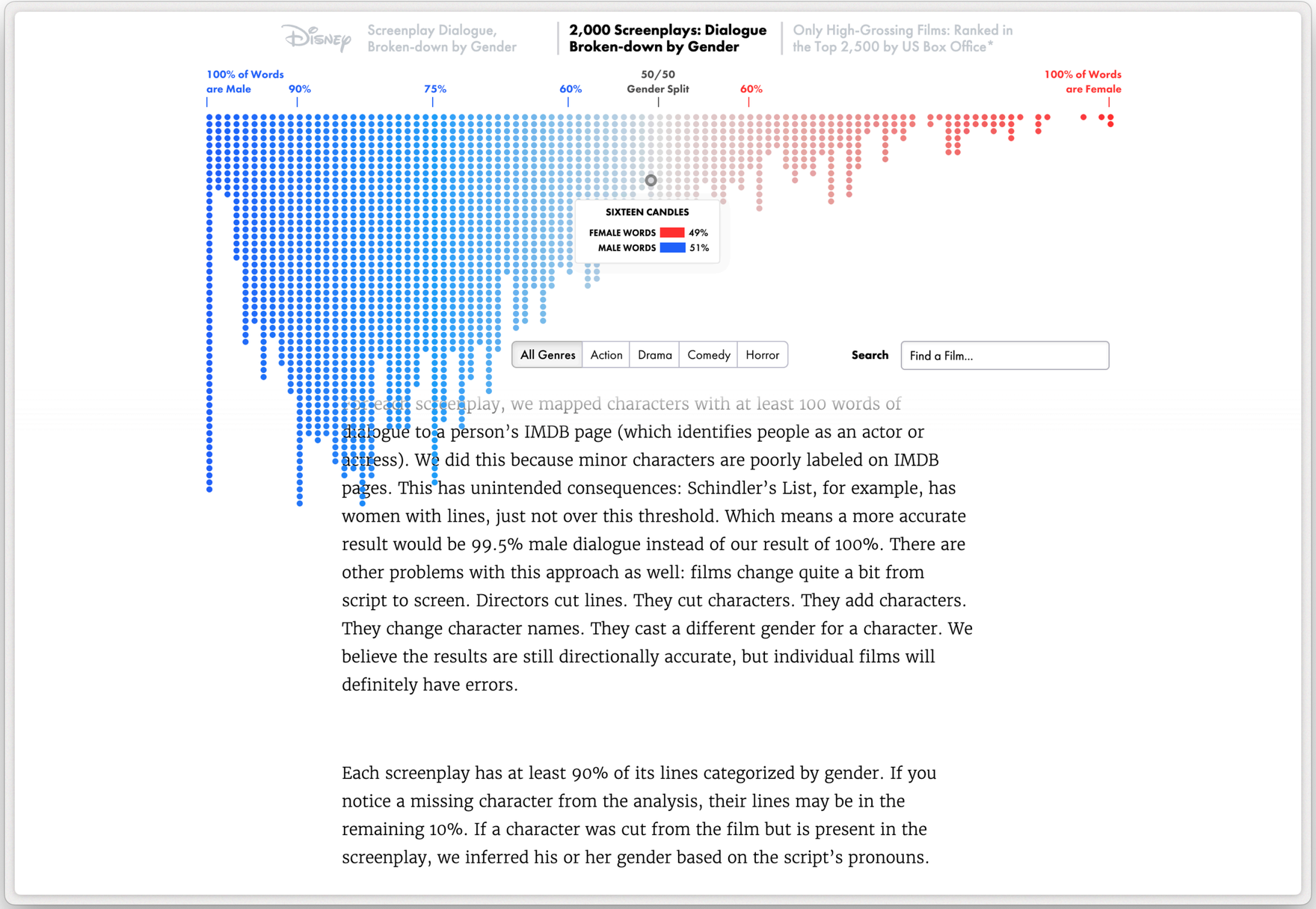
Positive Tests Over Time, by Region and County

POSITIVE TESTS OVER TIME, BY REGION AND COUNTY





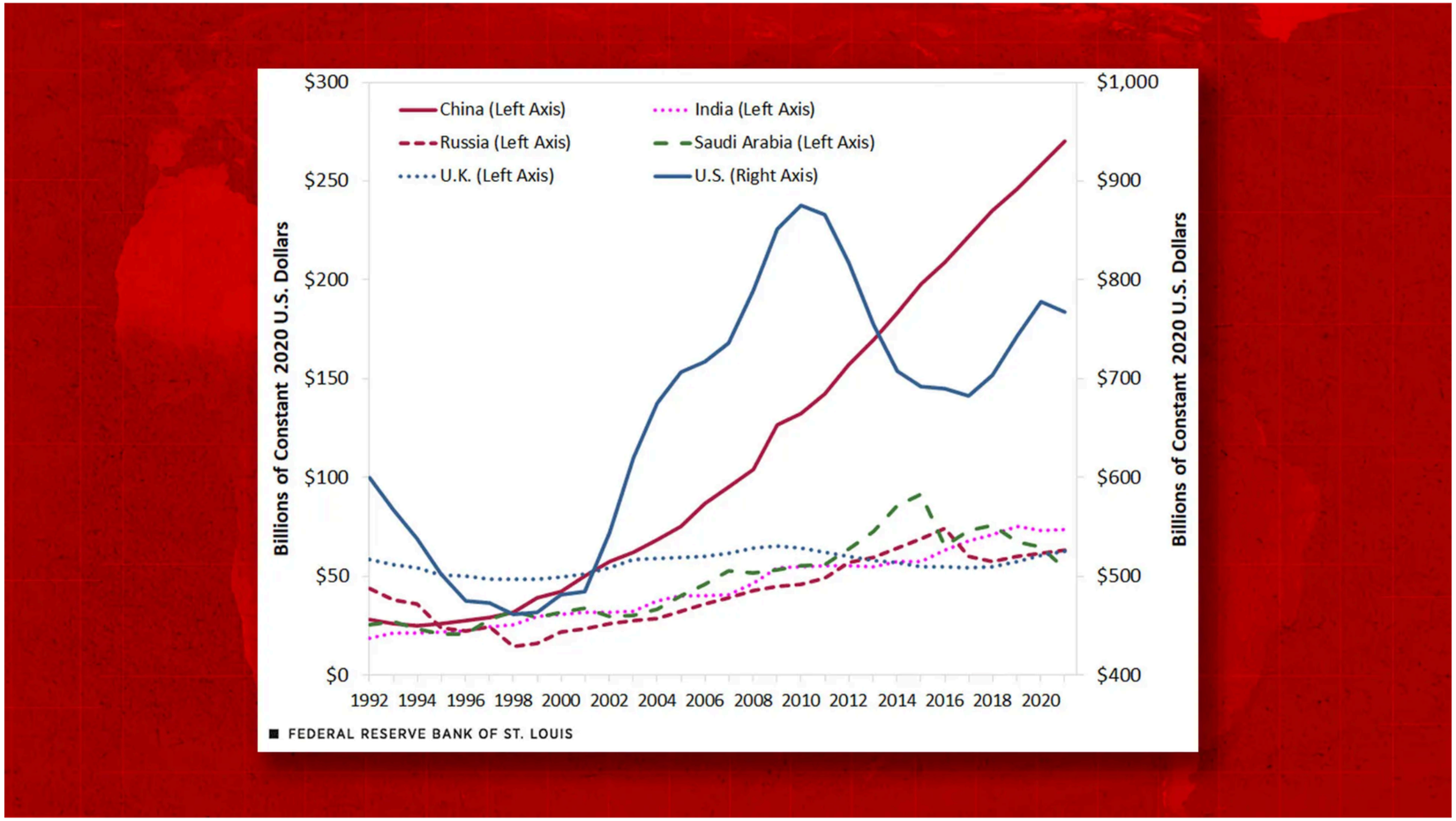
States are ranked according to the number of electors they have, and classified according to their tendency to vote more for one of the two



For each screenplay, we mapped characters with at least 100 words of dialogue to a person's IMDB page (which identifies people as an actor or actress). We did this because minor characters are poorly labeled on IMDB pages. This has unintended consequences: Schindler's List, for example, has women with lines, just not over this threshold. Which means a more accurate result would be 99.5% male dialogue instead of our result of 100%. There are other problems with this approach as well: films change quite a bit from script to screen. Directors cut lines. They cut characters. They add characters. They change character names. They cast a different gender for a character. We believe the results are still directionally accurate, but individual films will definitely have errors.

Each screenplay has at least 90% of its lines categorized by gender. If you notice a missing character from the analysis, their lines may be in the remaining 10%. If a character was cut from the film but is present in the screenplay, we inferred his or her gender based on the script's pronouns.

By Ben Norton Published 2023-01-23



Post

Richard McElreath
@rmcelreath.bsky.social

Follow

Several ppl sent me this extraordinary chart crime. Thank you I love it

Search

Trending

Tyreek Hill
Jets
Mike Johnson
Chuck Schumer
Bob Melvin
Jimmy Kimmel

Privacy • Terms • Help

English

Bluesky

Create account
Sign in

Defense Against Dishonest Charts

Charts are a window into the world. When done right, we gain an understanding of who we are, where we are, and how we can become better versions of ourselves. However, when done wrong, in the absence of truth, charts can be harmful.

This is a guide to protect ourselves and to preserve what is good about turning data into visual things.

We start with chart anatomy; then we look at how small changes can shift a point of view; this takes us to misleading chart varieties; and we finish with reading data and next steps.

Chart Anatomy

To defend against dishonest charts, you must understand them. You must take them apart and put them back together. Data forms the foundation and the following visual elements build on that foundation.

Visualization and data types

Suppose we have some data and want to create a visualization – how do we know what kind of visualization to make?

Suppose we have some data and want to create a visualization – how do we know what kind of visualization to make?

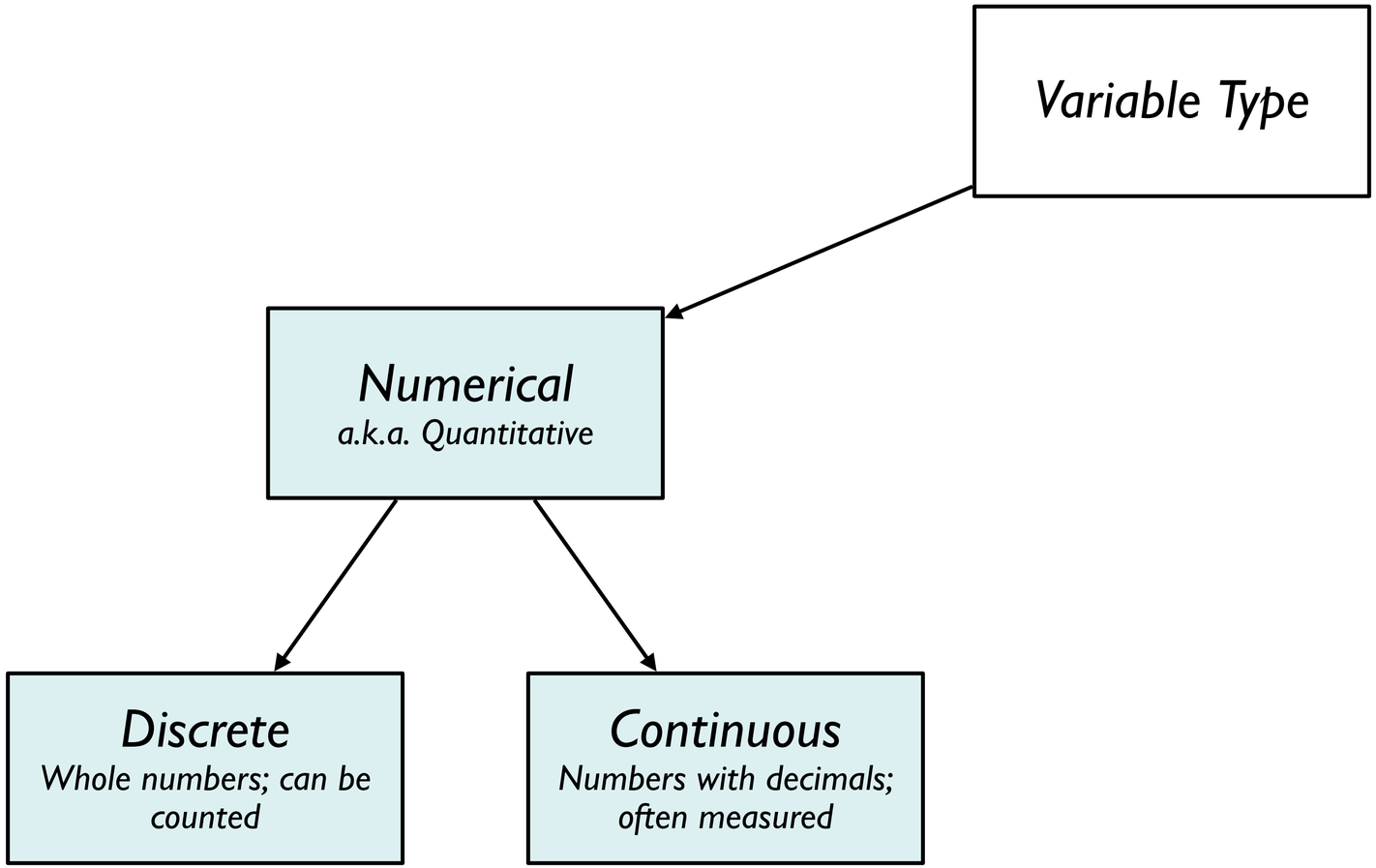
It depends on the type(s) of information we want to visualize.

For example, bar charts work for some kinds of data, but not all!

Where we are visualizing different variables (features) of a dataset, a lot depends on the type(s) of variables we're working with.

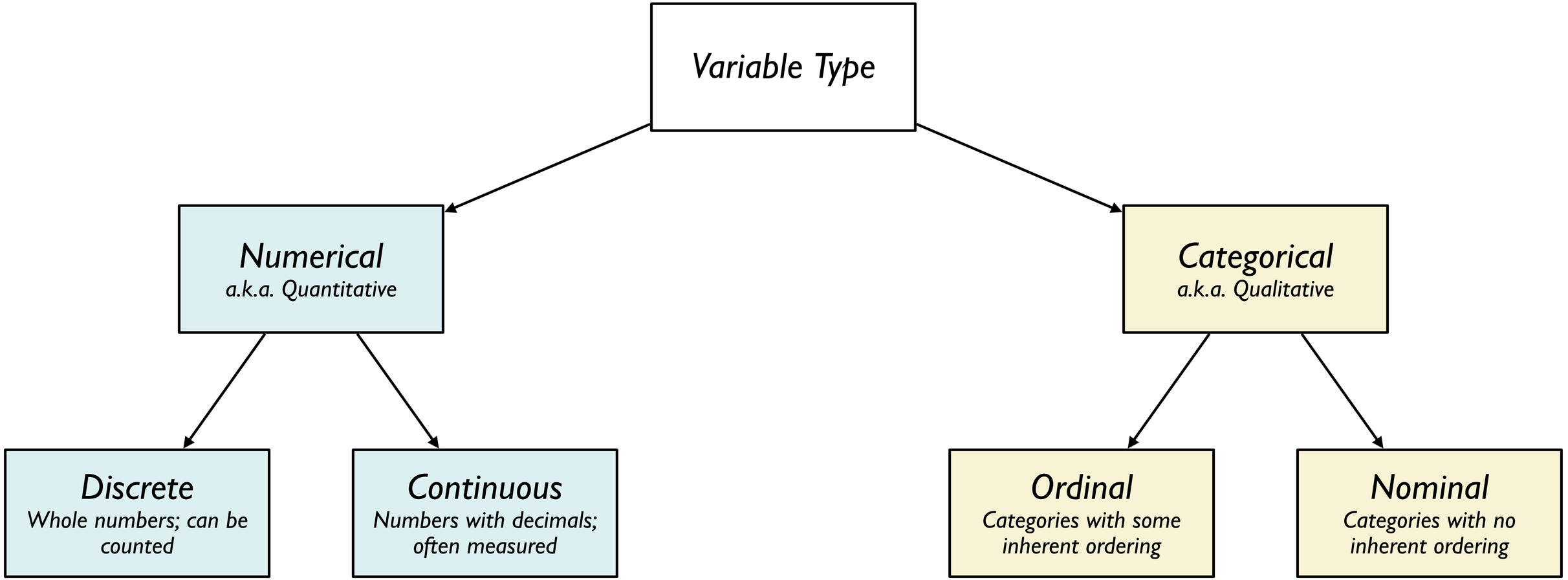
These *variable types* are different from *data types* like **`str`**, **`int`**, **`Table`**, etc.

Variable Type



e.g., number of cars, number of VC students

e.g., price, temperature, GPA, weight

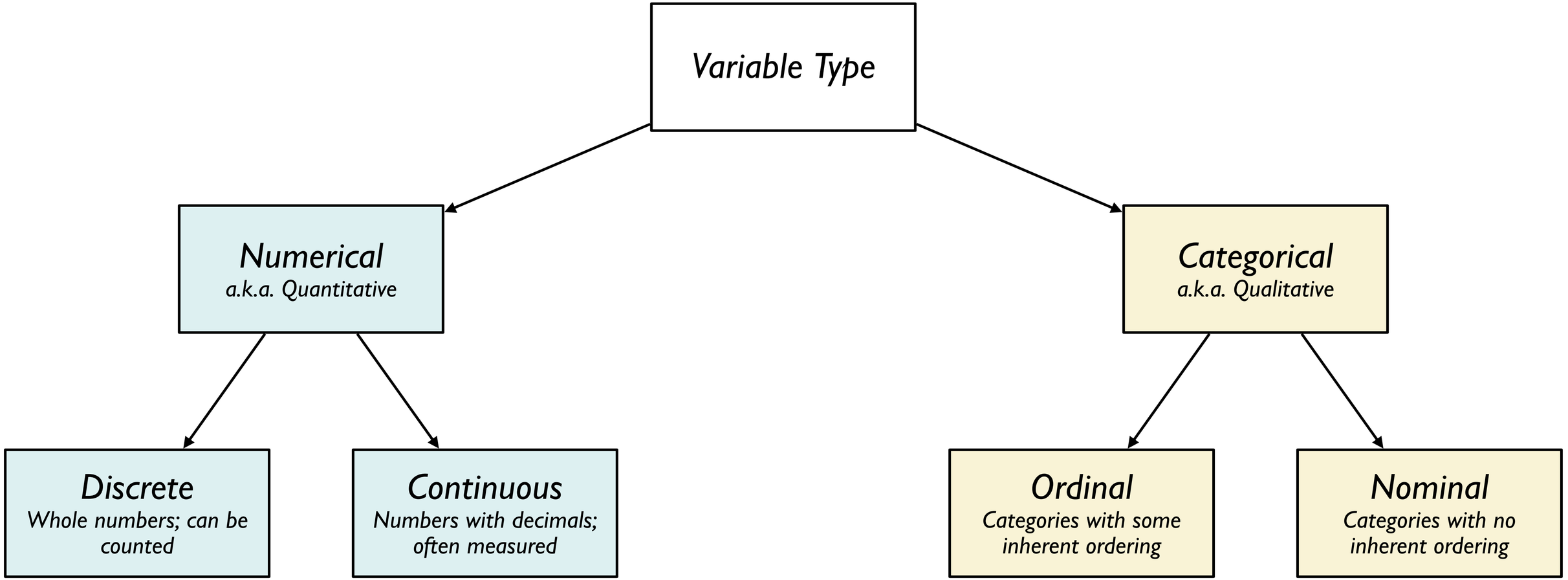


e.g., number of cars, number of VC students

e.g., price, temperature, GPA, weight

e.g., highest degree attained, Yelp stars

e.g., colors, political affiliation



e.g., number of cars, number of VC students

e.g., price, temperature, GPA, weight

e.g., highest degree attained, Yelp stars

e.g., colors, political affiliation

Can do arithmetic with

Cannot do arithmetic with

Example: Area codes

Just because a variable has numbers for values doesn't mean it's numerical.

City	*Area Code*
New York	917
Buffalo	716
Rochester	585
Albany	518
Poughkeepsie	845

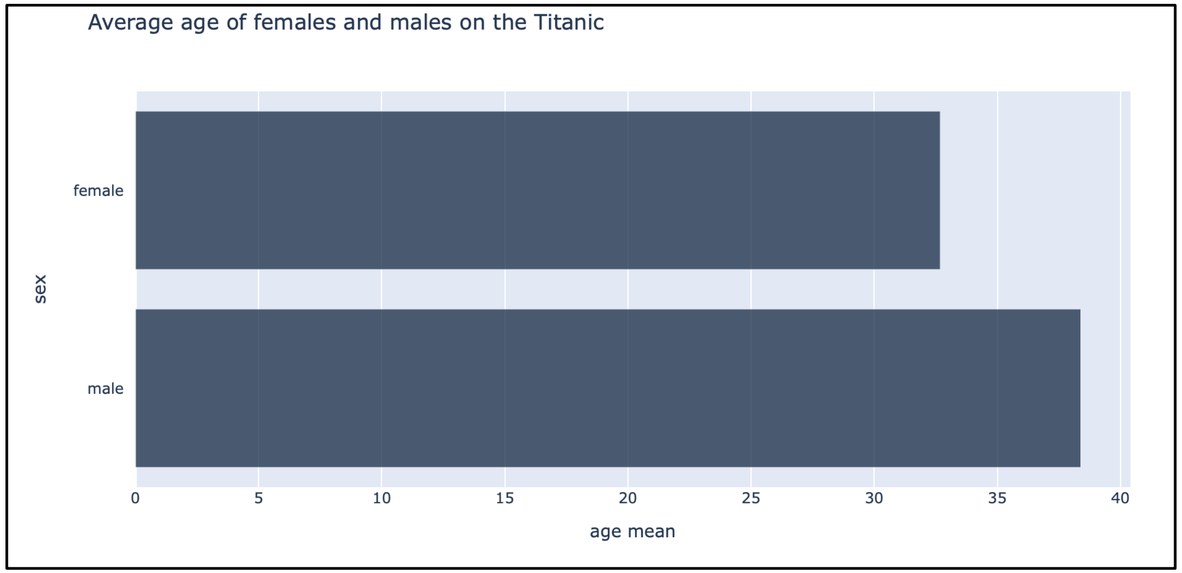
Area code is a categorical nominal variable.

While area codes are numbers, it doesn't make sense to do arithmetic with them, and there's also no (universal) order.

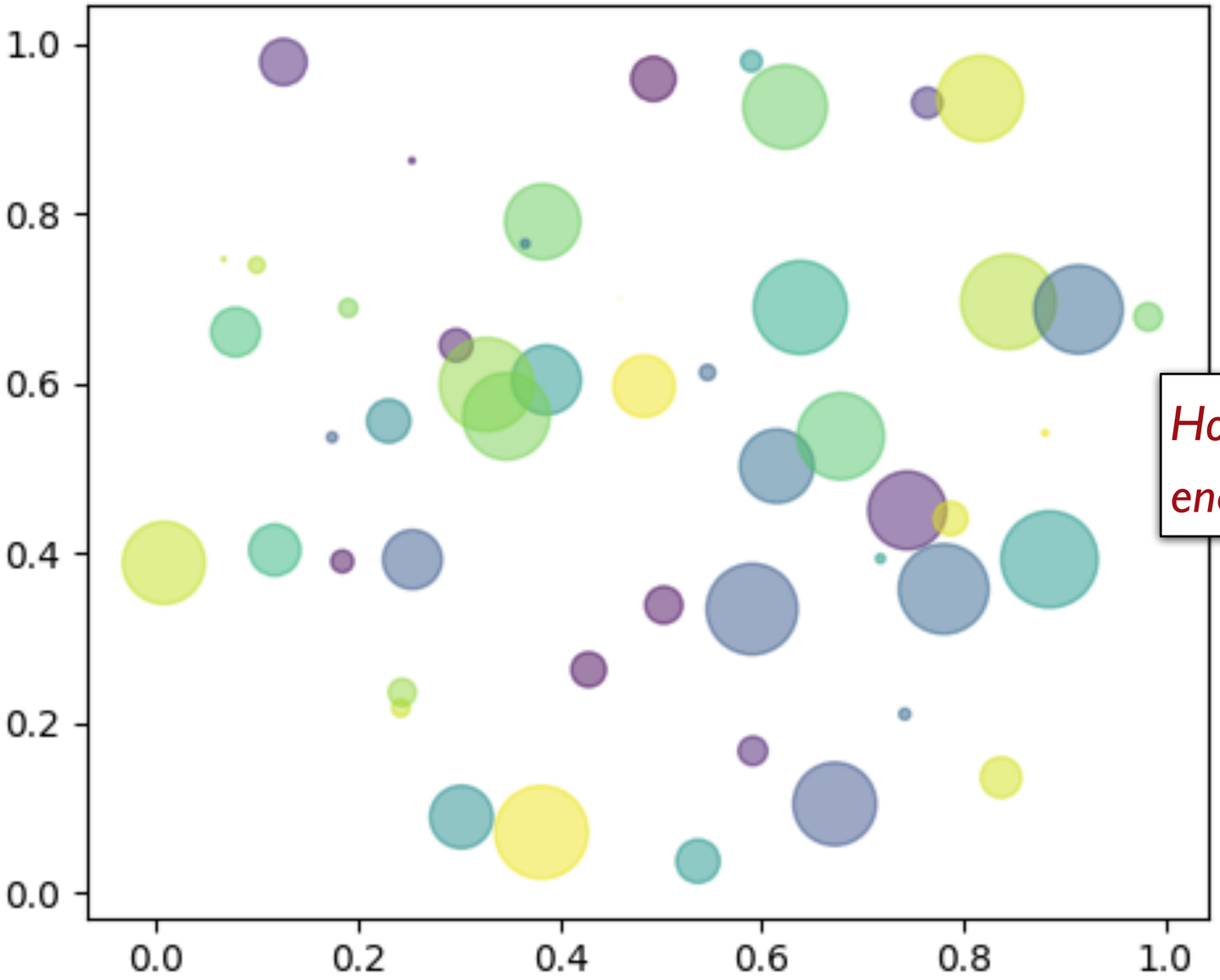
Similarly, a numerical variable might be stored as a string – in that case you'll need to convert the corresponding values to **`int`**s or **`float`**s, depending on what's appropriate.

An *encoding* is a mapping from a variable to a visual element.

For instance, in bar charts, length can visually encode a numerical variable:



Longer bar ⇒ higher average age



How many variables are encoded in this plot?

What's wrong?

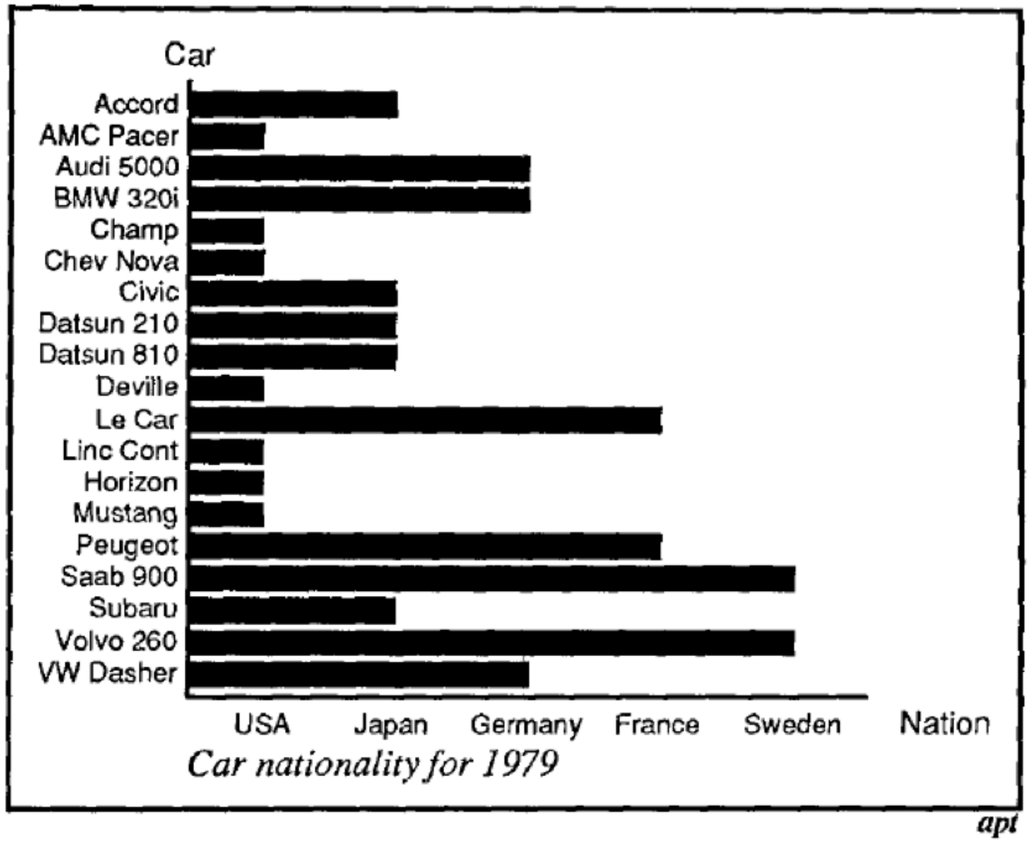


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

This graph implies that Swedish cars are, in some sense, "greater" than cars from other countries when that's not part of the data being visualized.

This is the wrong type of plot (and hence encoding) for this variable type.

Visualization in Python

As we saw in lab, the **`datascience`** module provides functions to generate plots from the data in tables.

These functions are built on top of the popular **`matplotlib`** module.

Let’s try it out – see the notebook!

Acknowledgments

This class incorporates material from:

Data 6, University of California, Berkeley (CC BY-NC-SA)