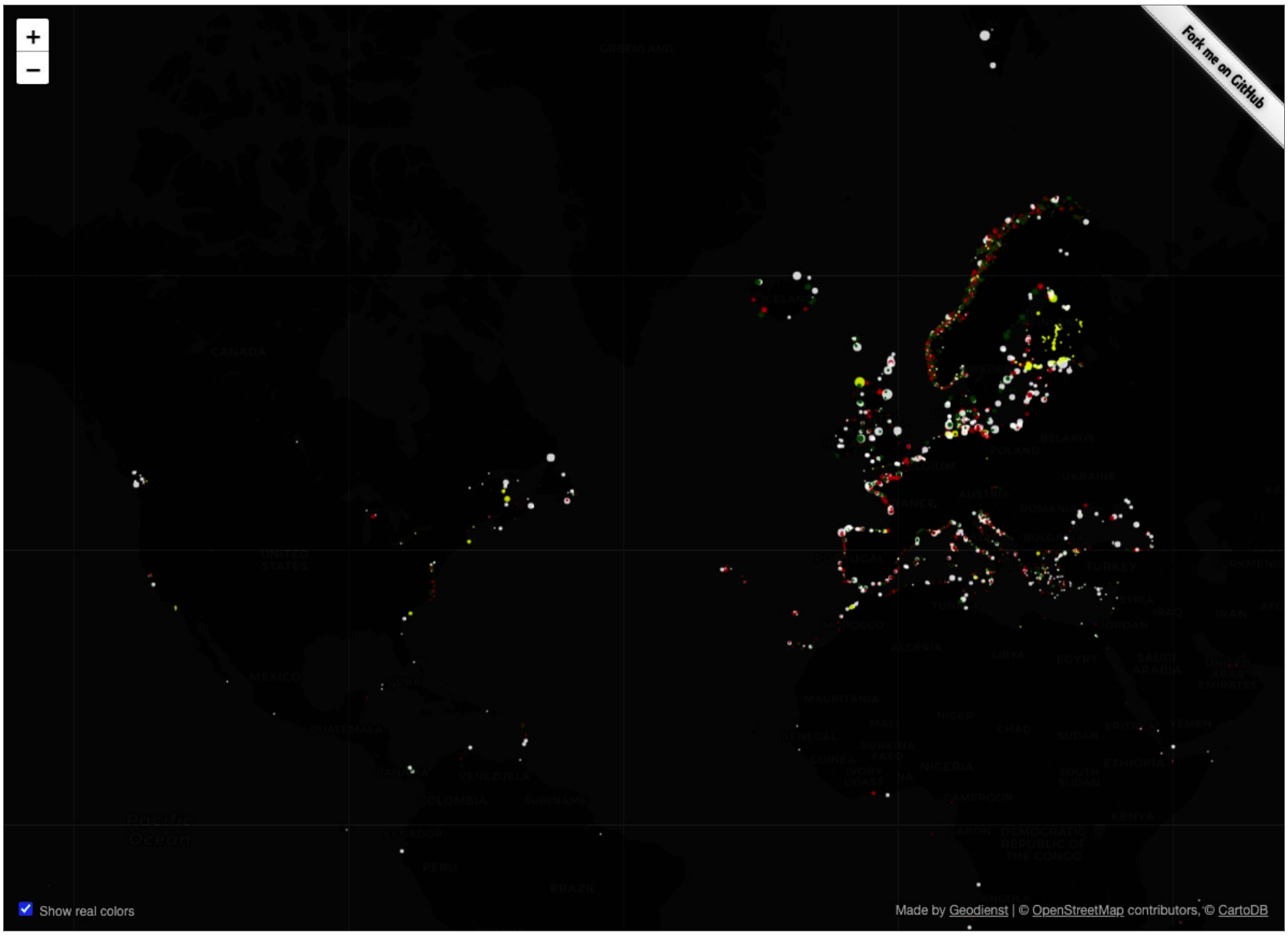
Lecture 1: Data Visualization

Today's Visualization



Today's lecture slides

Lighthouses

Full animated world map at https://geodienst.github.io/lighthousemap/

- Color true to real lighthouse
- Timing of blinks true to real lighthouse
- Size of dot corresponds to visibility range
- Data drawn from OpenStreetMap – completion and correction through crowdsourcing

  Related, and also excellent: https://twitter.com/i/status/1462095711508516865

Syllabus and Structure of the Course

After this course, you will…

- Be able to describe the key design guidelines and techniques used for the visual display of information.
- Understand how to best use the capabilities of visual perception in a graphic display.
- Understand the principles of interactive visualizations.
- Understand how Machine Learning techniques can determine data structure and pattern.
- Explore and critically evaluate a wide range of visualization techniques and applications.

We will primarily work with:

- Tamara Munzner: Visualization Analysis & Design (ISBN 978-1466-50891-0)
- Leland Wilkinson: The Grammar of Graphics, 2nd ed (ISBN 978-0387-24544-7) Wilkinson can be acquired free through the library's ebook access to SpringerLink. That way you can also get a cheap print-on-demand copy.

Syllabus and Structure of the Course

We meet in-person (or over Zoom if we are mandated back online) Wednesdays 14.00 - 16.00.
In addition, all course information can be found on Blackboard.
I can be reached on mvejdemojohansson@gc.cuny.edu, and will happily schedule meetings if you need them.
Between meetings, you will read assigned chapters from the textbook, and work on smaller assignments and semester projects.

Syllabus and Structure of the Course

Semester-long assignments

- Data Visualization Project – you take some dataset and develop a data visualization of that dataset with communicative intent.
- Blog Post – you take a recent paper from the IEEE Vis conference, and summarize it for a popular audience.

Weekly assignments

- Read assigned textbook chapters
- Find a recent published data visualization that embodies some concept in the assigned readings. Prepare to present that graphic to the class, and to demonstrate to the class what it does particularly well, and what could be improved.

Occasional assignments

- Recreate a published visualization using some data visualization tool.

Syllabus and Structure of the Course

Semester Schedule (may be changed as we go)

Date	Lecture Content	Preparation
2023-01-25	Defining Data Visualization	Munzner ch. 1, Wilkinson ch. 1-2
2022-02-01	Data Abstraction and Representation	Munzner ch. 2, Wilkinson ch. 3-4. Homework: improve the graph from the lecture slides.
2022-02-08	Task Abstraction	Munzner ch. 3, Wilkinson ch. 5
2022-02-15	Analysis and Validation	Munzner ch. 4, Wilkinson ch. 6-7. Watch: https://www.youtube.com/watch?v=Z8t4k0Q8e8Y
2022-02-22	Geometric Representation	Munzner ch. 5, Wilkinson ch. 8-9. Homework: Reproduce Minard's March on Moscow
2022-03-01	Aesthetic Mappings; Rules of Thumb	Munzner ch. 6, Wilkinson ch. 10
2022-03-8	Tabular Data, Network Data	Munzner ch. 7, 9, Wilkinson ch. 11-12
2022-03-15	Structure of a graphing library (Hannah Aizenman guest lecture)	
2022-03-22	Spatial Data, Geography, Maps	Munzner ch. 8, Wilkinson ch. 13
2022-03-29	Color	Munzner ch. 10
2022-04-19	Interactivity	Munzner ch. 11-12
2022-04-26	Summaries; Time, Time-series	Munzner ch. 13-14, Wilkinson Ch. 14
2022-05-03	Spaces, Graph Layout, Manifold Learning / Dimension Reduction	ISOMAP, MDS, UMAP
2022-05-10	Presentations	

What this course will *not* do

- …teach one specific data visualization platform.
  However, you should take this time to learn at least one platform well. I'm happy to help you in the process, but you pick a platform and work through tutorials to get going yourself.
- …get you ready to submit a paper to a major data vis conference yourself.
  However, by the time you finish the course, you should know a direction to go if you have this ambition.
- …deliver content to a passive student audience.
  Your participation is essential. Prepare for each lecture, collect questions and thoughts, and discuss eagerly in class.

What *is* data visualization anyway?!?

Defining Data Visualization

> The visual representation and presentation of data to facilitate understanding.

Andy Kirk (Data Visualization)

> Visual representation of datasets designed to help people carry out tasks more effectively.

Tamara Munzner

- Task-oriented / understanding
- Data-oriented
- Assistive
- Human visual system as co-processor

Why humans?

> Visual representation of datasets designed to help people carry out tasks more effectively.

Tamara Munzner

Why humans?

> Visual representation of datasets designed to help **people** carry out tasks more effectively.

Tamara Munzner

- We don’t need data vis when tasks can be fully automated.
- We might not know what questions we have in advance.

Why external representation?

> Visual **representation** of datasets designed to help people carry out tasks more effectively.

Tamara Munzner

- Replaces cognition with perception.
  We don’t need to know how the brain does pattern recognition in order to use it.

Why visual?

> **Visual** representation of datasets designed to help people carry out tasks more effectively.

Tamara Munzner

- Vision is high-bandwidth interface with our brain.
  Overview is possible due to background processing – we experience seeing everything simultaneously, and the visual system processes in parallel and pre-attentively.
- Sound: lower bandwidth, different semantics, no overview, subjective experience of sequentiality.
- Touch / haptics: low bandwidth, low record/replay capacity
- Taste, smell: no viable record/replay devices

Why all the data?

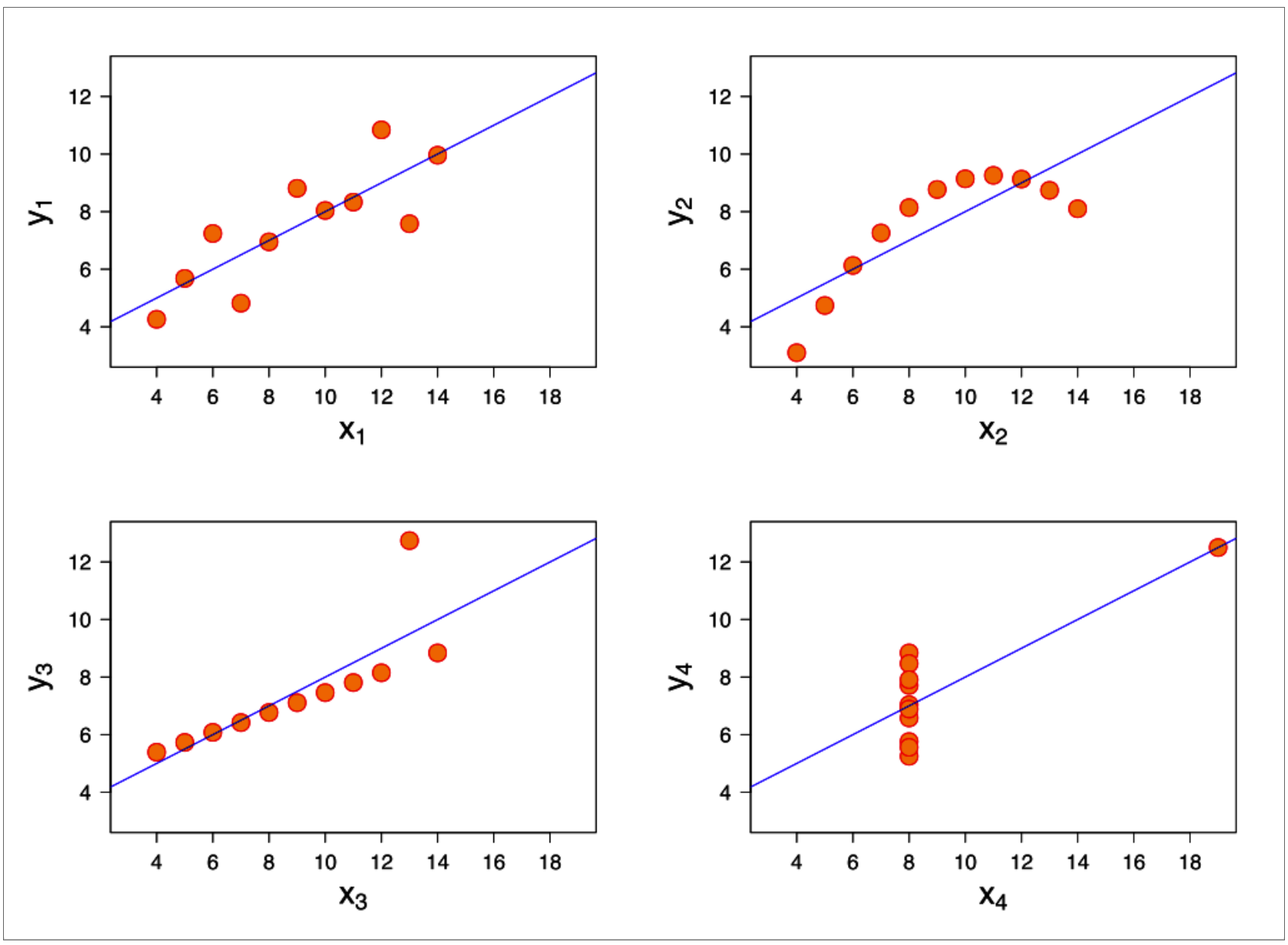
> Visual **representation of datasets** designed to help people carry out tasks more effectively.

Tamara Munzner

- Summaries inherently lose information
- 4 datasets, identical statistics

Property	Value
Mean of x	9
Sample variance of x: s_x^2	11
Mean of y	7.50
Sample variance of y: s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression: R^2	0.67

Each value exact up to at least 2 decimal places.



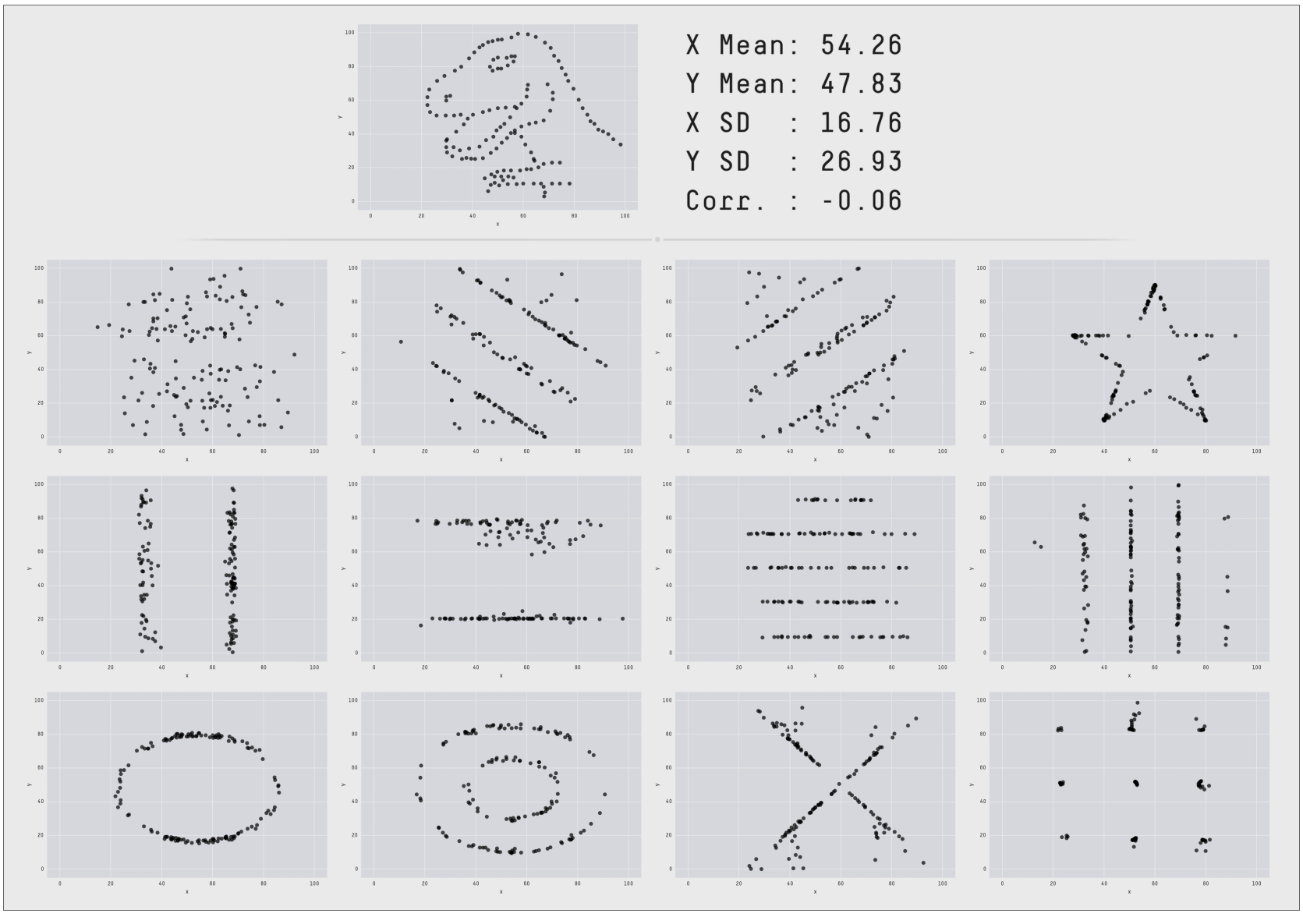
Anscombe's Quartet

Why all the data?

> Visual **representation of datasets** designed to help people carry out tasks more effectively.

Tamara Munzner

- Summaries inherently lose information
- 12 datasets, identical statistics



Datasaurus Dozen

Design: balancing constraints

Design in a context

ALL DESIGN WORK HAS A CONTEXT, WHICH PROVIDES LIMITATIONS

- Computational limits
  - computation time, system memory
- Display limits
  - pixels
  - information density
  - “ink ratio”
- Human limits
  - human time, memory, attention, capacities of vision, understanding

Analytic Scaffolds

Three different sets of questions and considerations to guide your design work.

Tamara Munzner

3 design questions:

- What?
- Why?
- How?

Andy Kirk

3 phases of understanding:

- perceiving
- interpreting
- comprehending

4 stage design process:

- formulating your brief
- working with data
- establishing your editorial thinking
- developing the design solution

3 design principles:

- trustworthy
- accessible
- elegant

Edward Tufte

- 6 design principles
- coined terminology: ink ratios, chart junk

Tufte: Fundamental Principles of Analytical Design

1. Show comparisons, contrasts, differences.
2. Show causality, mechanism, explanation, systematic structure.
3. Show multivariate data; that is, show more than 1 or 2 variables.
4. Completely integrate words, numbers, images, diagrams.
5. Thoroughly describe the evidence. Provide a detailed title, indicate the authors and sponsors, document the data sources, show complete measurement scales, point out relevant issues.
6. Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.

Edward Tufte, Beautiful Evidence, pp 122 - 139

Tufte: Data ink and chart junk

Early Tufte design guidance:

> **Tip**
>
> Measure and maximize the *data-ink ratio*. Data-ink is the non-erasable core of a graphic, and the ratio to maximize is (data-ink / total ink)

> **Tip**
>
> Eschew and eliminate chart junk – graphical decorations, textures, patterns, all of which just increase total ink without increasing data ink.

Kirk: Phases of Understanding

Visualizer control

Viewer control

Perceiving

What do I see?

- What data is shown?
- How is it represented?
- What features are observable?

Interpreting

What does it mean, given the subject?

What features are…

- interesting?
- unexpected?
- important?

Comprehending

What does it mean, to me?

- What have I learnt?
- What do I feel?
- What do I do now?

Kirk: Design Process

Four fundamental steps

1. Formulating your brief
   planning, defining and initiating your project
2. Working with data
   gathering, handling and preparing your data
3. Establishing your editorial thinking
   defining what you will show your audience
4. Developing your design solution
   making design choices about how you represent and present what it is you want to show your audience

Kirk: Design principles

Based on Dieter Rams' 10 principles of good design:

1. Good design is **innovative**
2. Good design makes a product **useful**
3. Good design is **aesthetic**
4. Good design makes a product **understandable**
5. Good design is **unobtrusive**
6. Good design is **honest**
7. Good design is **long-lasting**
8. Good design is **thorough** down to the last detail
9. Good design is **environmentally friendly**
10. Good design is as **little design** as possible

Kirk: Design principles

Principle 1. Good visualization design is **trustworthy**.

1. Good design is **innovative**
2. Good design makes a product **useful**
3. Good design is **aesthetic**
4. Good design makes a product **understandable**
5. Good design is **unobtrusive**
6. Good design is **honest**
7. Good design is **long-lasting**
8. Good design is **thorough** down to the last detail
9. Good design is **environmentally friendly**
10. Good design is as **little design** as possible

Kirk: Design principles

Principle 2. Good visualization design is **accessible**.

1. Good design is **innovative**
2. Good design makes a product **useful**
3. Good design is **aesthetic**
4. Good design makes a product **understandable**
5. Good design is **unobtrusive**
6. Good design is **honest**
7. Good design is **long-lasting**
8. Good design is **thorough** down to the last detail
9. Good design is **environmentally friendly**
10. Good design is as **little design** as possible

Kirk: Design principles

Principle 3. Good visualization design is **elegant**.

1. Good design is **innovative**
2. Good design makes a product **useful**
3. Good design is **aesthetic**
4. Good design makes a product **understandable**
5. Good design is **unobtrusive**
6. Good design is **honest**
7. Good design is **long-lasting**
8. Good design is **thorough** down to the last detail
9. Good design is **environmentally friendly**
10. Good design is as **little design** as possible

Munzner: Design questions and levels

Each level (domain/abstraction/idiom/algorithm) contained in the previous.

- **domain** situation
  - who are the target users?
- **abstraction**
  - translate from specifics of domain to vocabulary of visualization
  - **what** is shows? **data abstraction**
    - don't just draw what you're given: transform to new form
  - **why** is the user looking at it? **task abstraction**
- **idiom**
  - **how** is it shown?
    - **visual encoding idiom**: how to draw
    - **interaction idiom**: how to manipulate
- **algorithm**
  - efficient computation

Munzner: Design questions and levels

Different levels have different failure modes

- **Domain**:
  You misunderstand their needs
  - **Abstraction**:
    You're showing them the wrong thing
    - **Visual encoding / Interaction idiom**
      The way you show it doesn't work
      - **Algorithm**
        Your code is too slow

Munzner: Design questions and levels

Solution: use methods from different fields at each level.

- **Domain**:
  Observe target users using existing tools. (anthropology/ethnography)
  - **Abstraction**:
    - **Visual encoding / Interaction idiom**
      Justify design with respect to alternatives. (design)
      - **Algorithm**
        Measure system time/memory.
        Analyze computational complexity. (computer science)
    - Analyze results qualitatively.
      Measure human time with lab experiment (lab study). (cognitive psychology)
  - Observe target users after deployment (field study). (anthropology/ethnography)
- Measure adoption.

Munzner: Design questions and levels

Design questions impose a structure on an otherwise vast design space.

What?

Dataset Types

Tables (tidy data)

- Attributes (columns)
- Items (rows)

Networks

- Items (nodes)
- Links

Fields

- Value cells distributed on a shape

Data cubes / tensors

- Value cells distributed in a hypercube

Trees

- Relationships

Geometry (spatial)

- Positions

Attributes

Attribute Type

- Categorical
- Ordinal
- Quantitative (interval)

Ordering Direction

- Sequential
- Diverging
- Cyclic

Data Availability

Static
Dynamic

Munzner: Design questions and levels

Design questions impose a structure on an otherwise vast design space.

WHY?

ACTIONS

Analyze

- Consume
  - Discover
  - Present
  - Enjoy
- Produce
  - Annotate
  - Record
  - Derive

Query

- Identify
- Compare
- Summarize

Search

- Target known/unknown
- Location known/unknown

TARGETS

All Data

- Trends
- Outliers
- Features

Attributes

- One
  - Distribution
  - Extremes
- Many
  - Dependency
  - Correlation
  - Similarity

Network Data

- Topology
- Paths

Spatial Data

- Shape

Munzner: Design questions and levels

Design questions impose a structure on an otherwise vast design space.
How?

Encode

Arrange
Map from
categorical and
ordered attributes

- Color
  - **Hue**
  - **Saturation**
  - **Lightness**
- **Size**, **Angle**, **Curvature**
- **Shape**
- **Motion**
  - **Direction**
  - **Rate**
  - **Frequency**

Manipulate

Change
Select
Navigate

Facet

Juxtapose
Partition
Superimpose

Reduce

Filter
Aggregate
Embed

No One Good Answer

Data Visualization is an aesthetic field

You **will** see disagreements on what is and is not a good design.
And on what is and is not a good design principle.

Data Visualization is a communication field

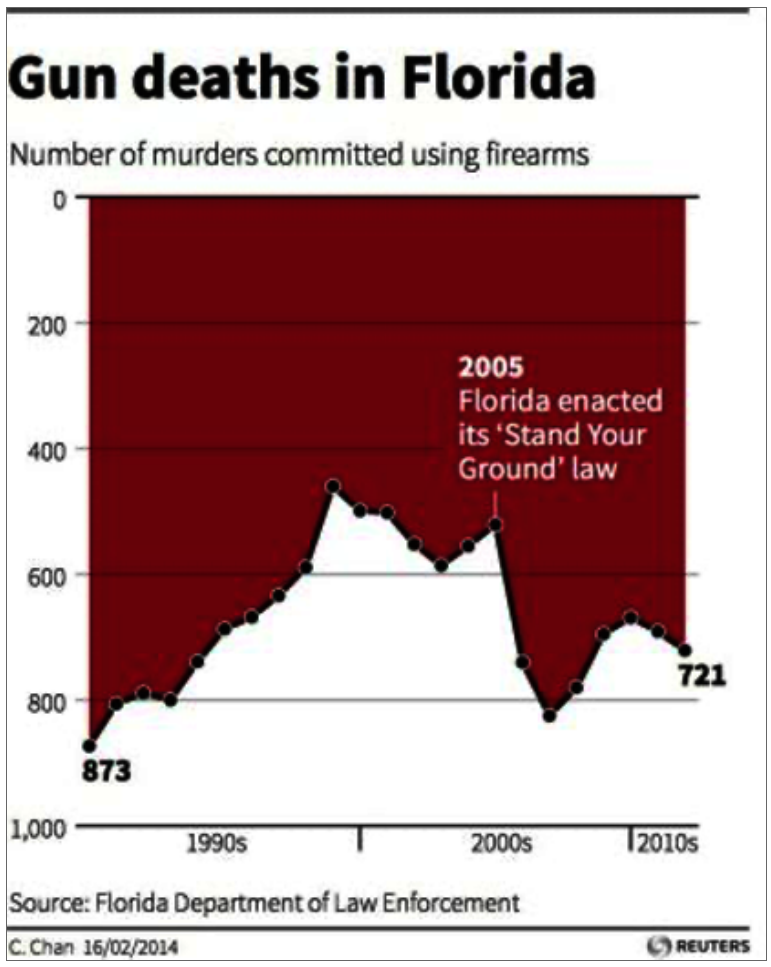
Many applications of data visualization communicate a message, either intentionally or unintentionally.
Notice how Kirk emphasize the communication, Munzner acknowledges it, and Tufte all but ignores that aspect.

Is this a good graphic?

Tufte: Look at all that chart junk! So much decorations that do not directly encode data!
Kirk: cites Jen Christiansen, Graphics Editor at Scientific American. “I found that when I developed magazine graphics according to [Tufte’s] philosophy, they were most often met with a yawn. The reality is that Scientific American isn’t required reading. We need to engage readers, as well as inform them.”
Decorations provide context for the information – it is immediately apparent what the data is about (something something razors) without impacting the trustworthiness of the data display itself.

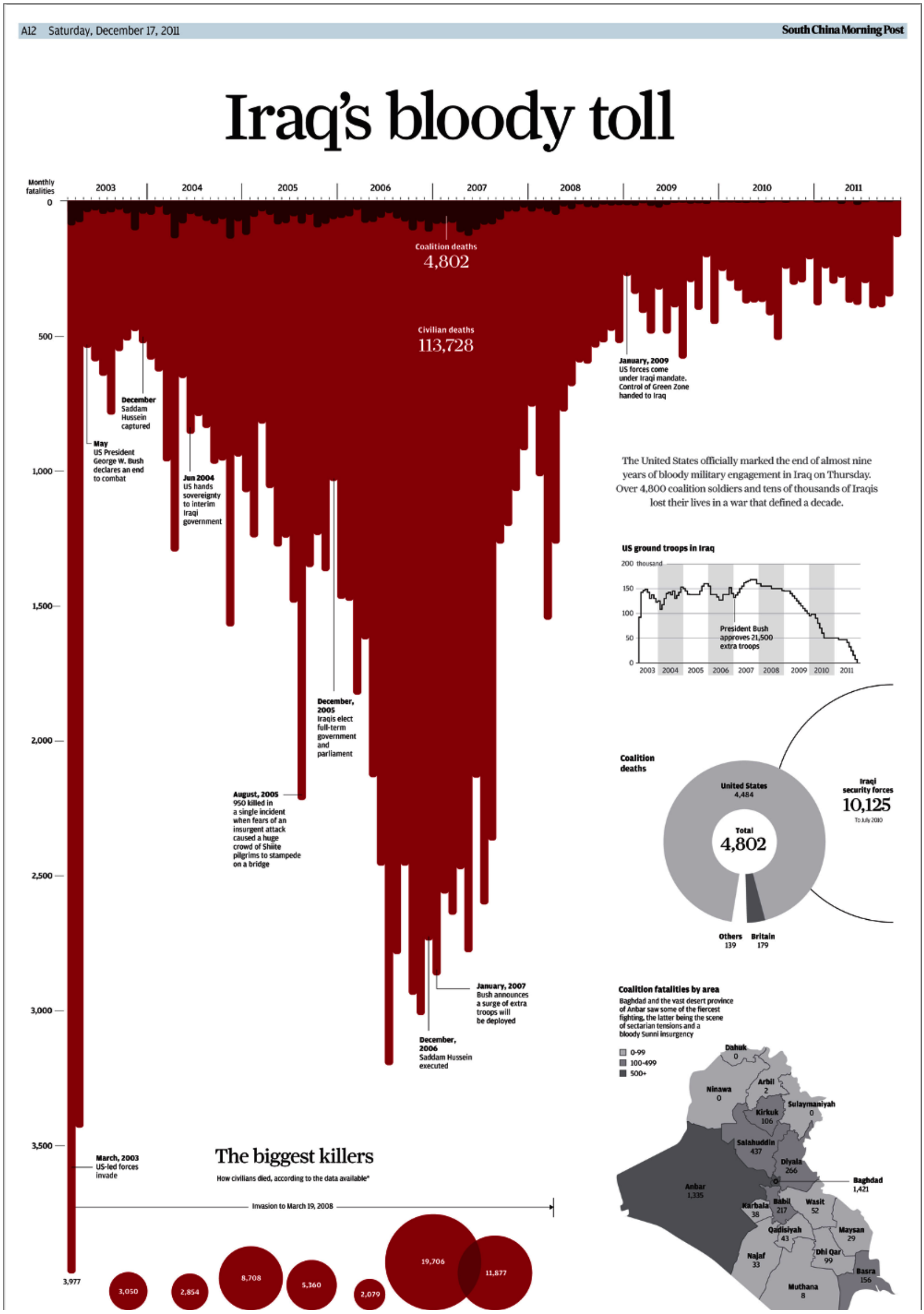
Is this a good graphic?

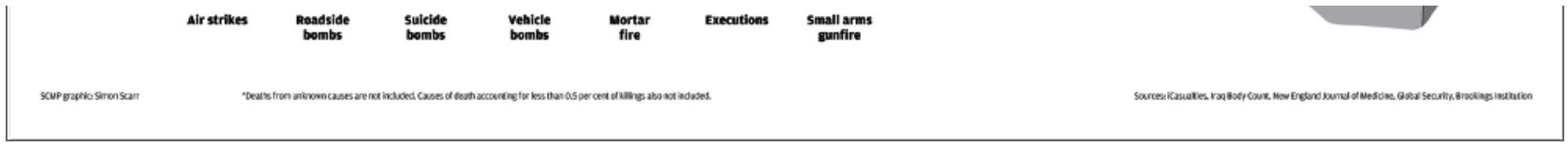
Very popular target as an example of a bad graph. The inverted y-axis is very often invoked as a condemning feature.



Is this a good graphic?

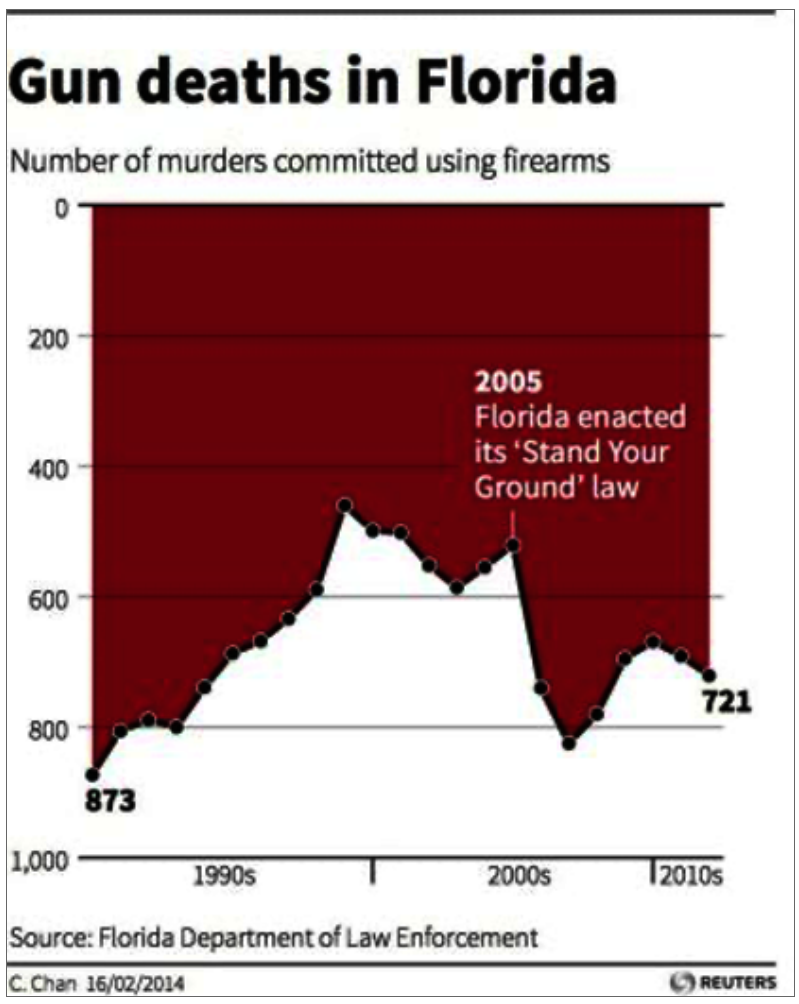
Very popular target as an example of a bad graph. The inverted y-axis is very often invoked as a condemning feature.
Kirk points out that it was designed to emulate another chart published earlier: “Iraq’s bloody toll”.
The red coloring and the inverted y-axis in combination are attempting to evoke a metaphor of blood dribbling down a wall.

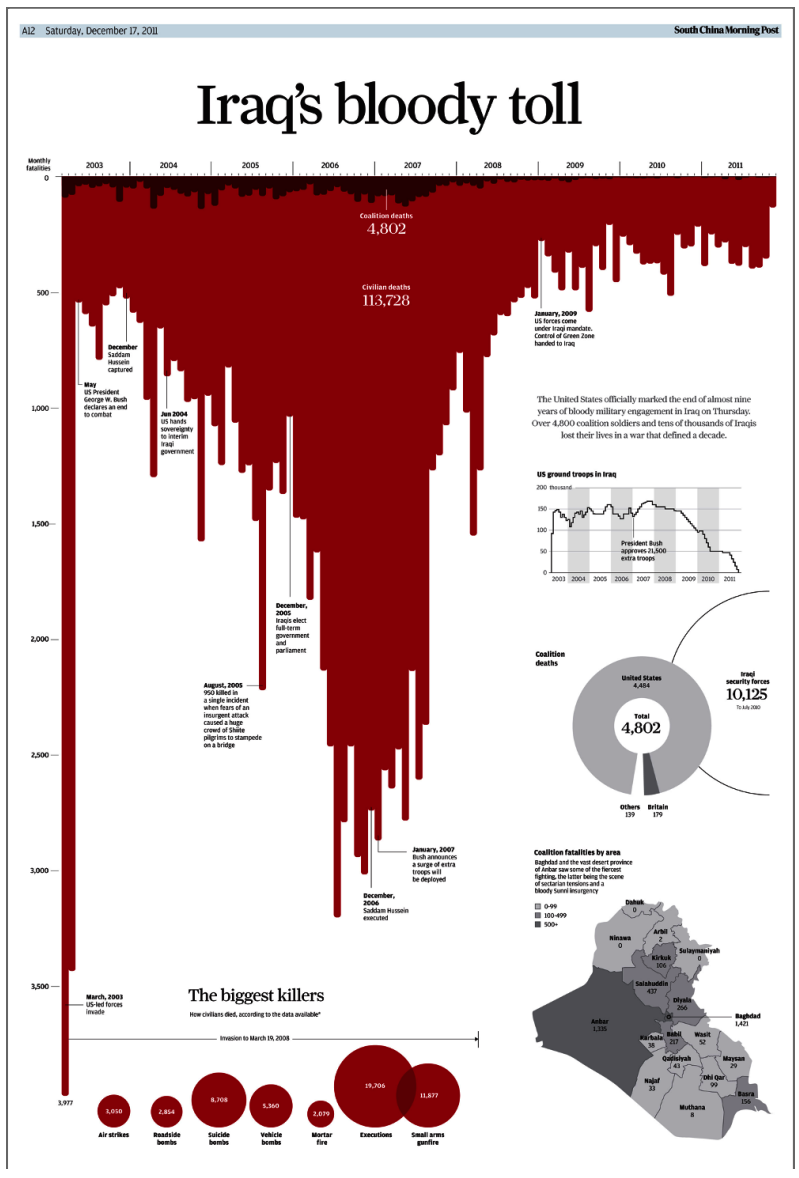



Air strikes
Roadside bombs
Suicide bombs
Vehicle bombs
Mortar fire
Executions
Small arms gunfire
SCMP graphic: Simon Scarr
*Deaths from unknown causes are not included. Causes of death accounting for less than 0.5 per cent of killings also not included.
Sources: iCasualties, Iraq Body Count, New England Journal of Medicine, Global Security, Brookings Institution

Is this a good graphic?

Very popular target as an example of a bad graph. The inverted y-axis is very often invoked as a condemning feature.
Kirk points out that it was designed to emulate another chart published earlier: “Iraq’s bloody toll”.
The red coloring and the inverted y-axis in combination are attempting to evoke a metaphor of blood dribbling down a wall.
Question: Was the intended metaphor successful? In “Gun deaths in Florida”? In “Iraq’s bloody toll”? What could have been done differently to make the message more efficiently conveyed?






SCAPP graphic: Simon Scarr
*Deaths from unknown causes are not included. Causes of death accounting for less than 0.5 per cent of killings also not included.
Sources: iCasualties, Iraq Body Count, New England Journal of Medicine, Global Security, Brookings Institution

Data Visualization Toolkits

Pick a platform - and stick with it

In this course, you will pick one platform and do all your exercises in this platform. Good options include:

- `matplotlib` (and `seaborn`)
  Python, not Grammar of Graphics
- `ggplot2`
  R, Grammar of Graphics
- `plotnine`
  Python, Grammar of Graphics
- `altair`
  Python, Grammar of Graphics
- `d3.js / ObservableJS` JavaScript, not Grammar of Graphics

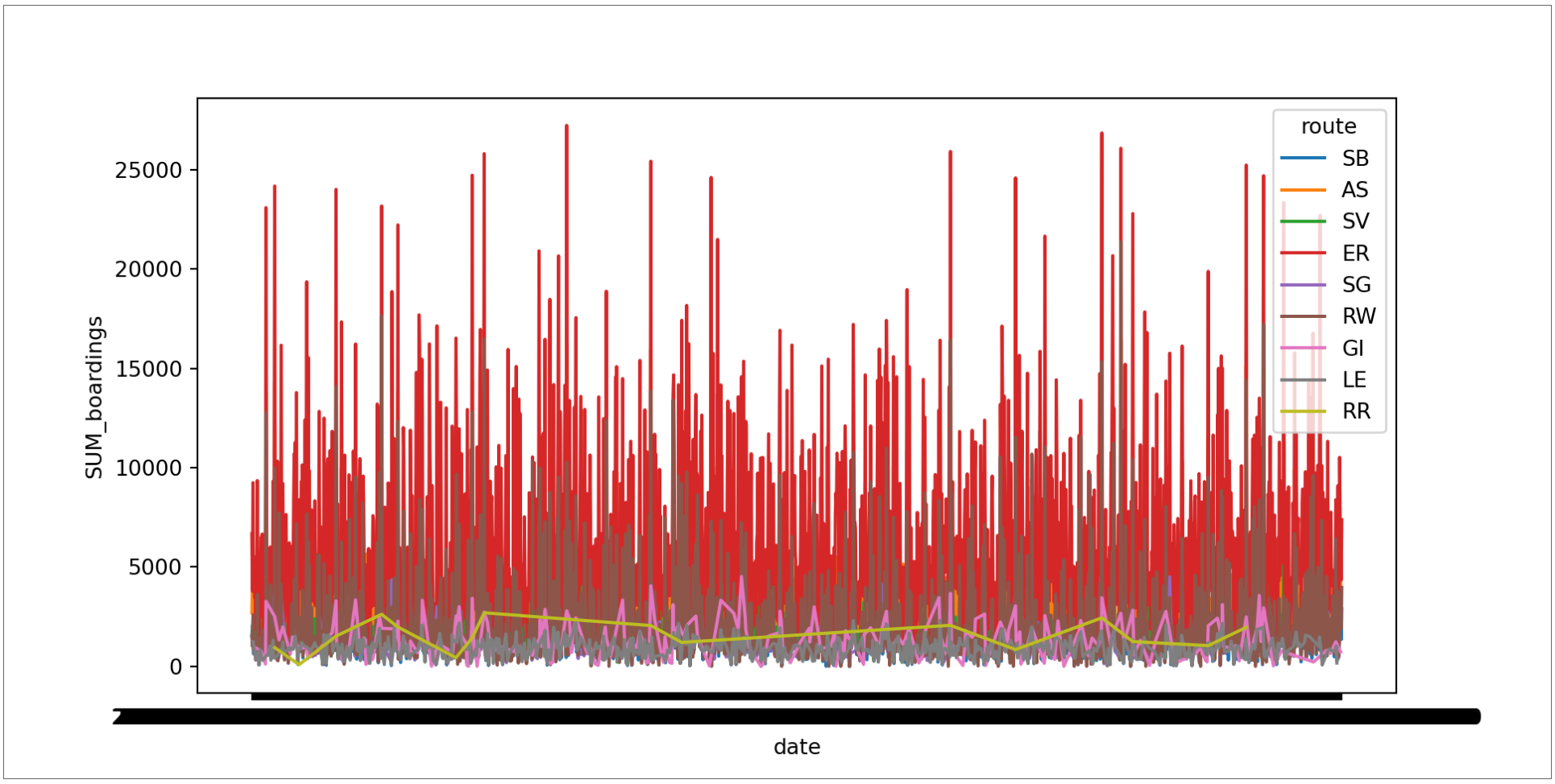
It's better to build 80% proficiency in one tool than 20% each in 3 different tools. Your next job may well use something different - and for each tool you learn, the next one is easier to learn.

Out of the box - demo visualizations

We draw on the NYC OpenData portal and collect data on traffic on the NYC Ferry network.
The data we want is available at https://data.cityofnewyork.us/Transportation/NYC-Ferry-Ridership/t5n6-gx8c and we can compose a query (to offload some computation onto the NYC OpenData servers) to extract the dily rider count:
https://data.cityofnewyork.us/resource/t5n6-gx8c.csv?select=date,route,SUM(boardings)&group=date,route&$limit=1000000
We want a linegraph of the daily ridership, by ferry route

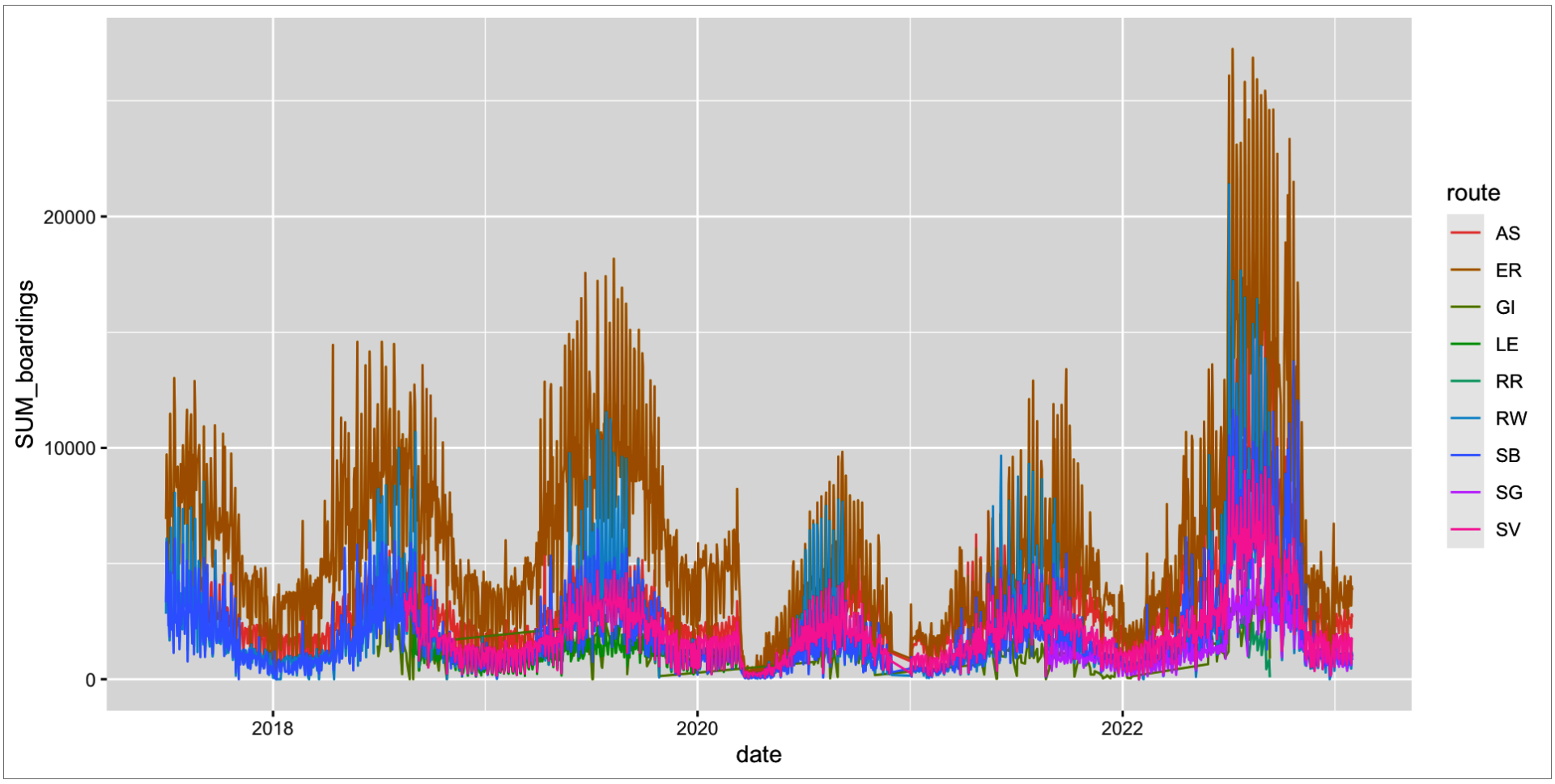
Python / matplotlib + seaborn

▶ Code



R / ggplot2

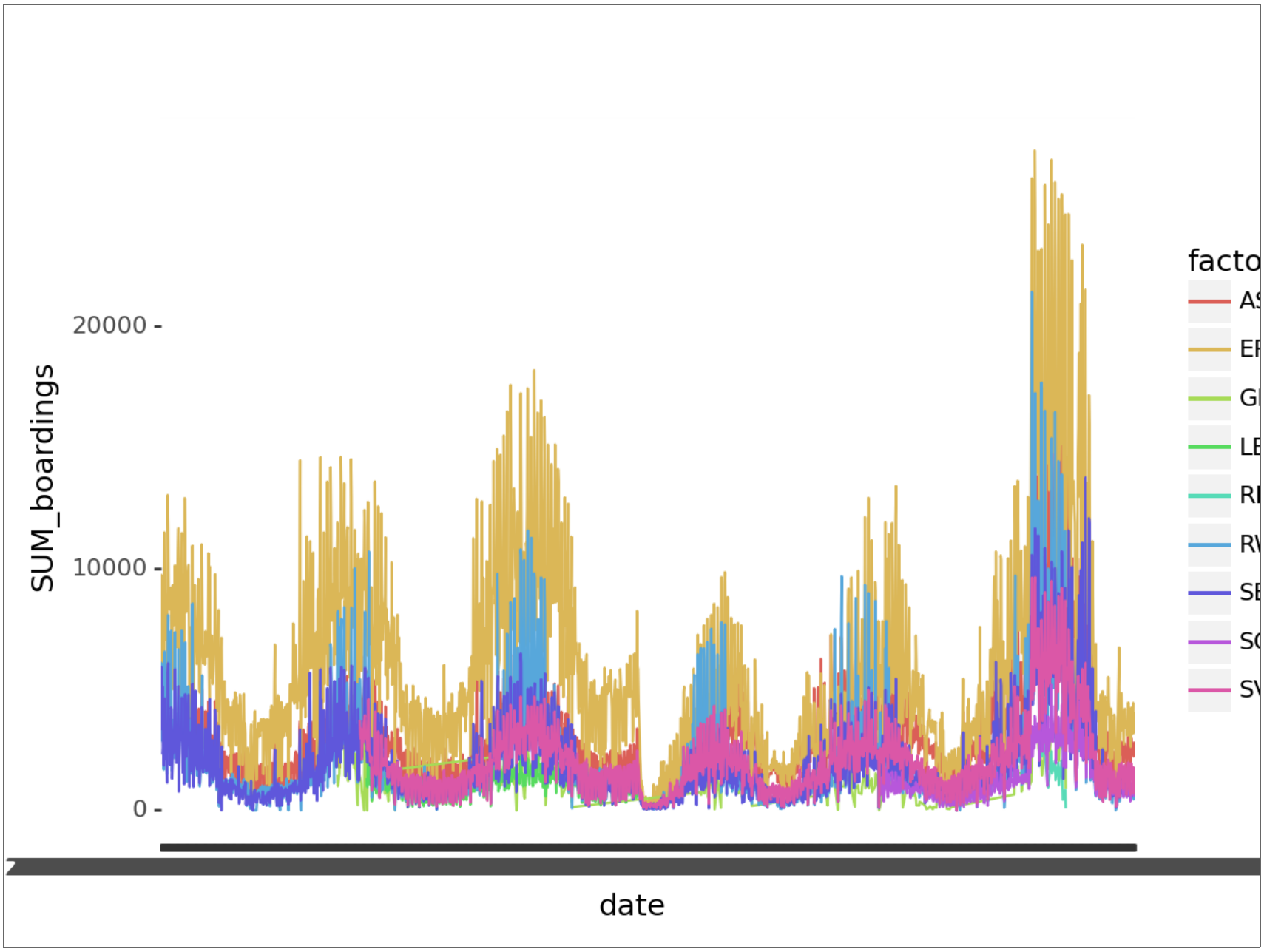
▶ Code



Python / plotnine

▶ Code

```
<ggplot: (687517027)>
```

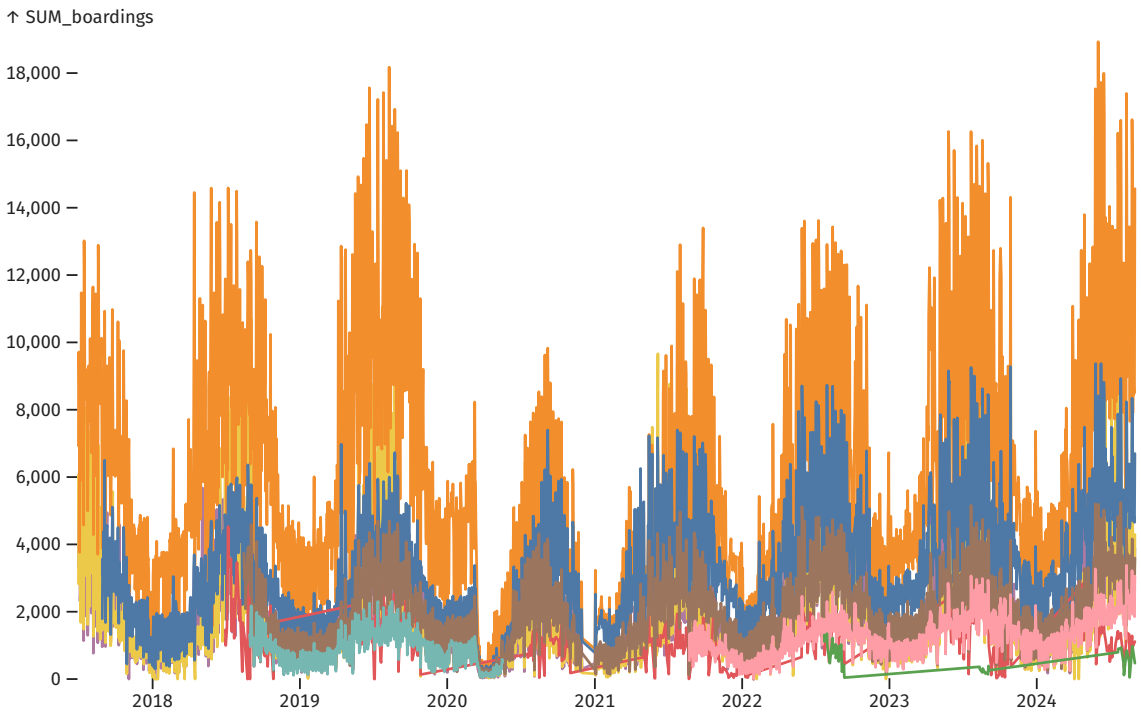


Python / Altair

- Code

JavaScript / ObservableJS+d3.js

▶ Code



Your first aesthetic critique

What differences and similarities do you see between the different “Out of the box” plots here?
What would you like to change?
What would you like to check / verify?
Would you like more (or less) binning and aggregation?
What, if any, interactive features would you like?
What, if any, labels, titles, annotations would you like to use?
What would an interesting use case for this plot be?
If you were to pull properties and features freely from all platforms (or add yourself) – how could you specify the most appropriate plot for this use case?
Homework: Reconstruct this plot in your chosen platform, and improve the things you discussed in your critique.