

MassMutual DSDP 2021:

VISUALIZATION TECHNIQUES

June 29, 2021

R. Jordan Crouser

Associate Professor of Computer Science

Smith College

Recall: deconstructing graphics

1. Find a data visualization you think is interesting
 - Some ideas: NYTimes, VisualisingData.com, Visual.ly
 - Remember to cite your source!

2. Identify the following:
 - What is the **data** that's being visualized? Where did it come from?
 - Which **data dimensions** are mapped to which **visual dimensions**?
 - How does this **shape your understanding** of the data?
 - If you **liked** the visualization: what is it doing **well**?
 - If you **disliked** the visualization: what would you **change**?

Discussion

What makes a **good** encoding?



Principle 1: expressiveness

- Encodes **all** the facts
- Example:

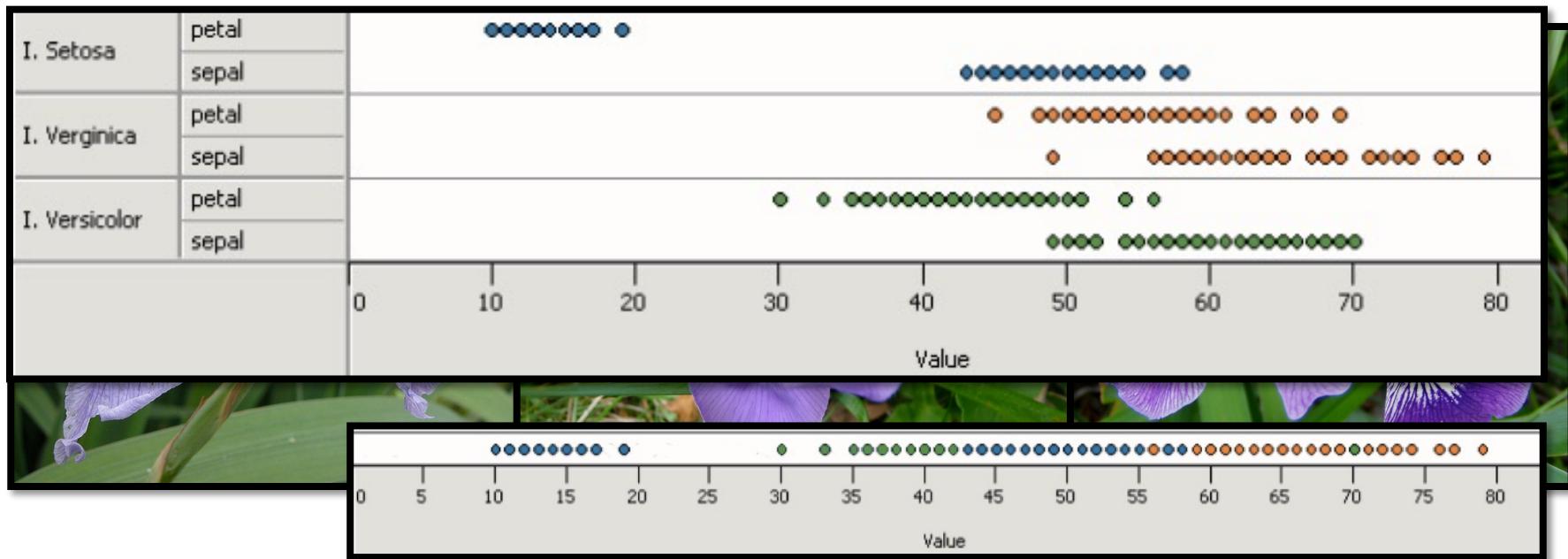
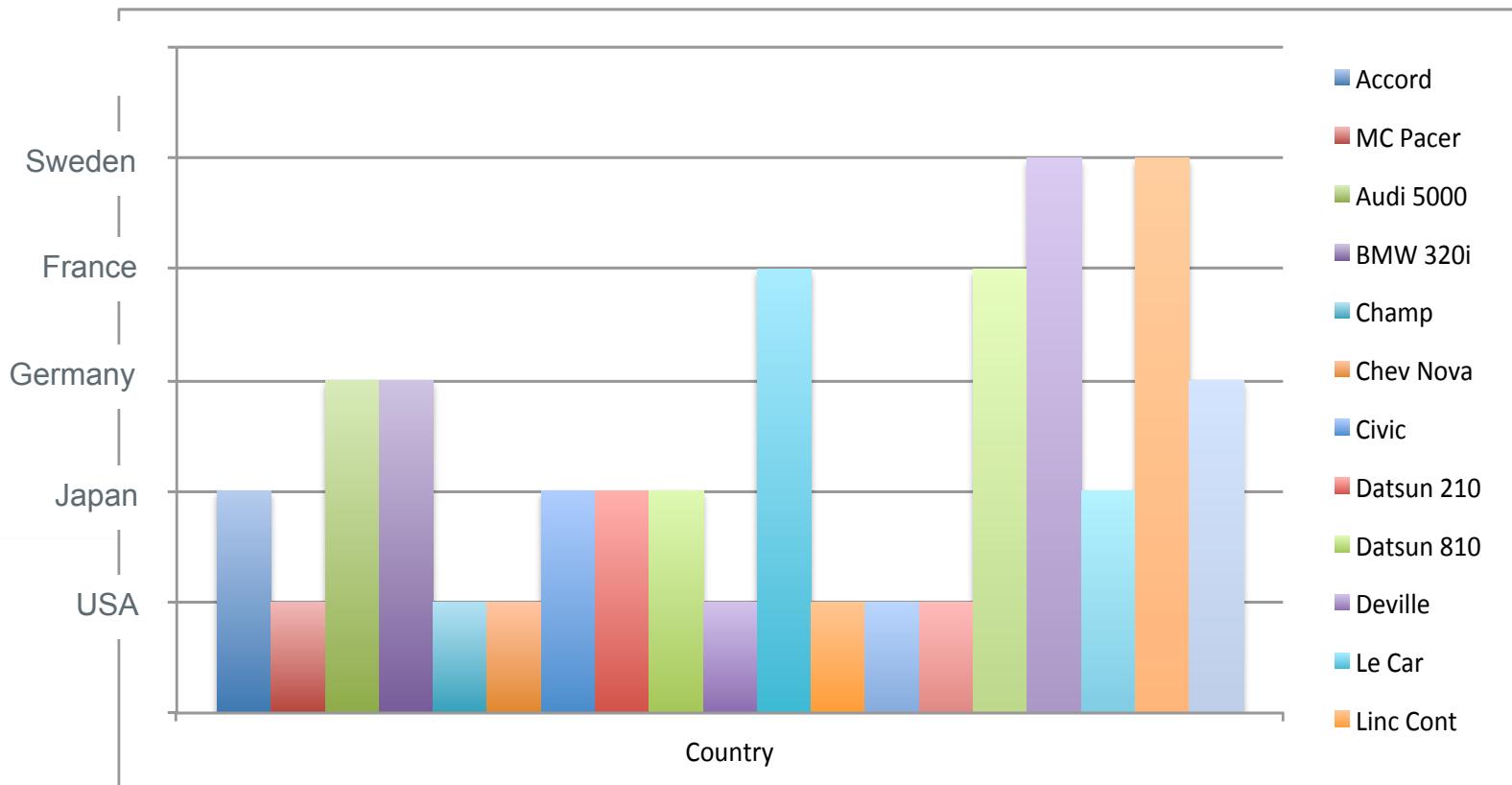


Fig. Courtesy of M Krzywinski

Principle 1: expressiveness

- Encodes **only** the facts
- Example:

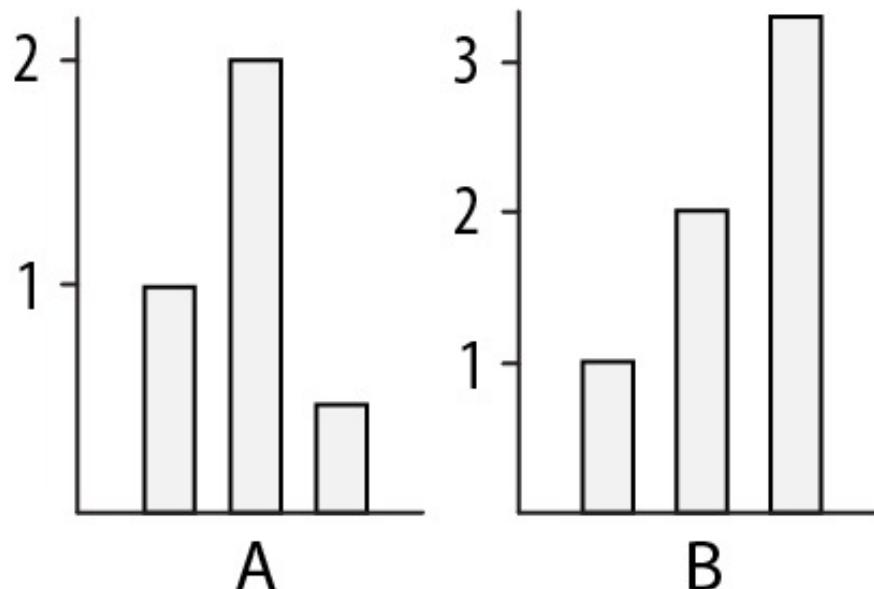


Adapted from Mackinlay J (1986) Automating the design of graphical presentations of relational information.

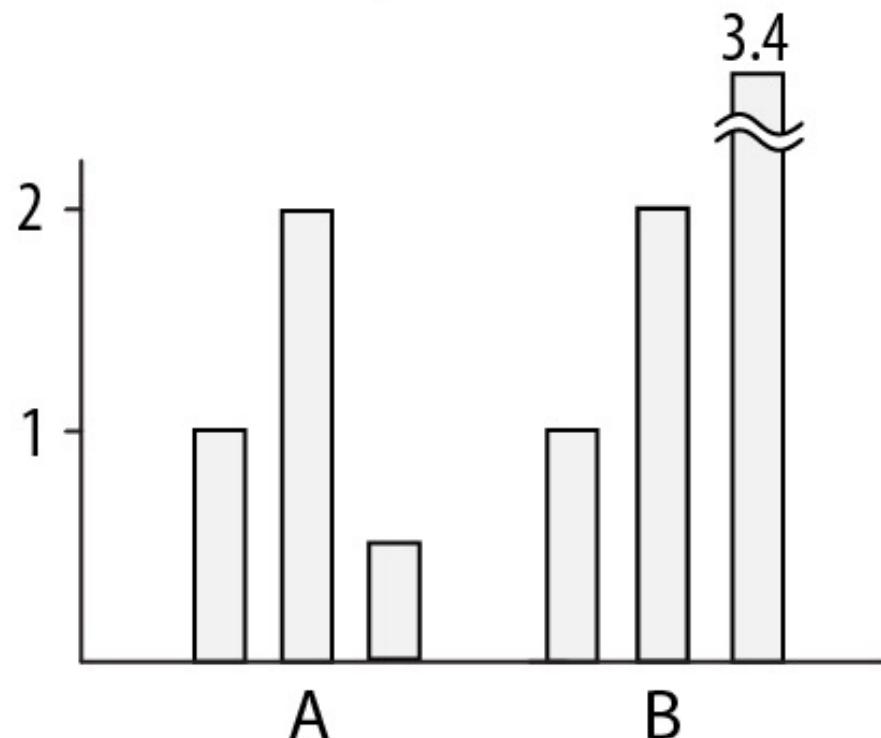
Principle 2: consistency

- Use **consistent axes** when comparing charts

misleading

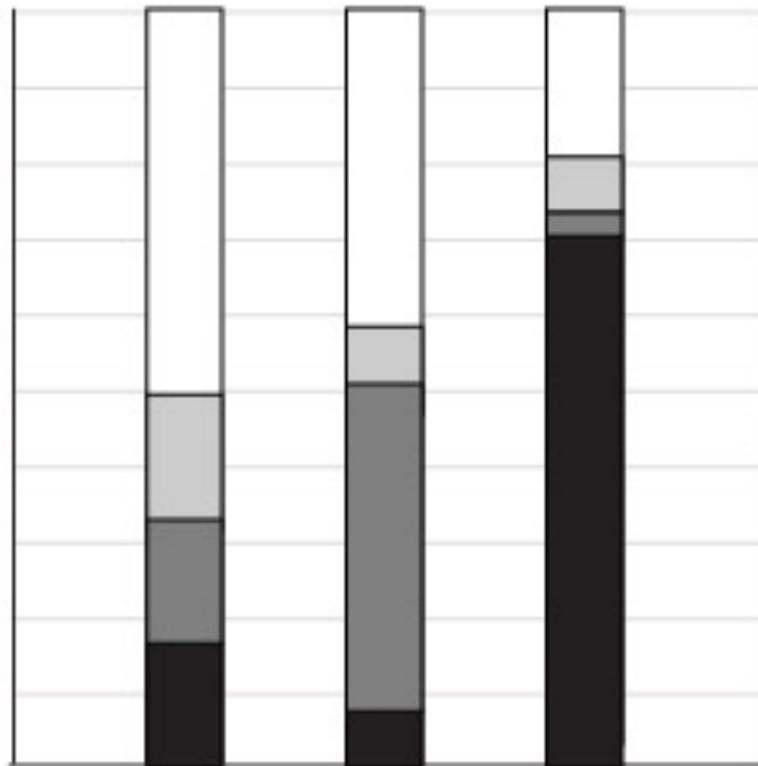


improved



Principle 2: consistency

- A note on **legends**: order items according to appearance

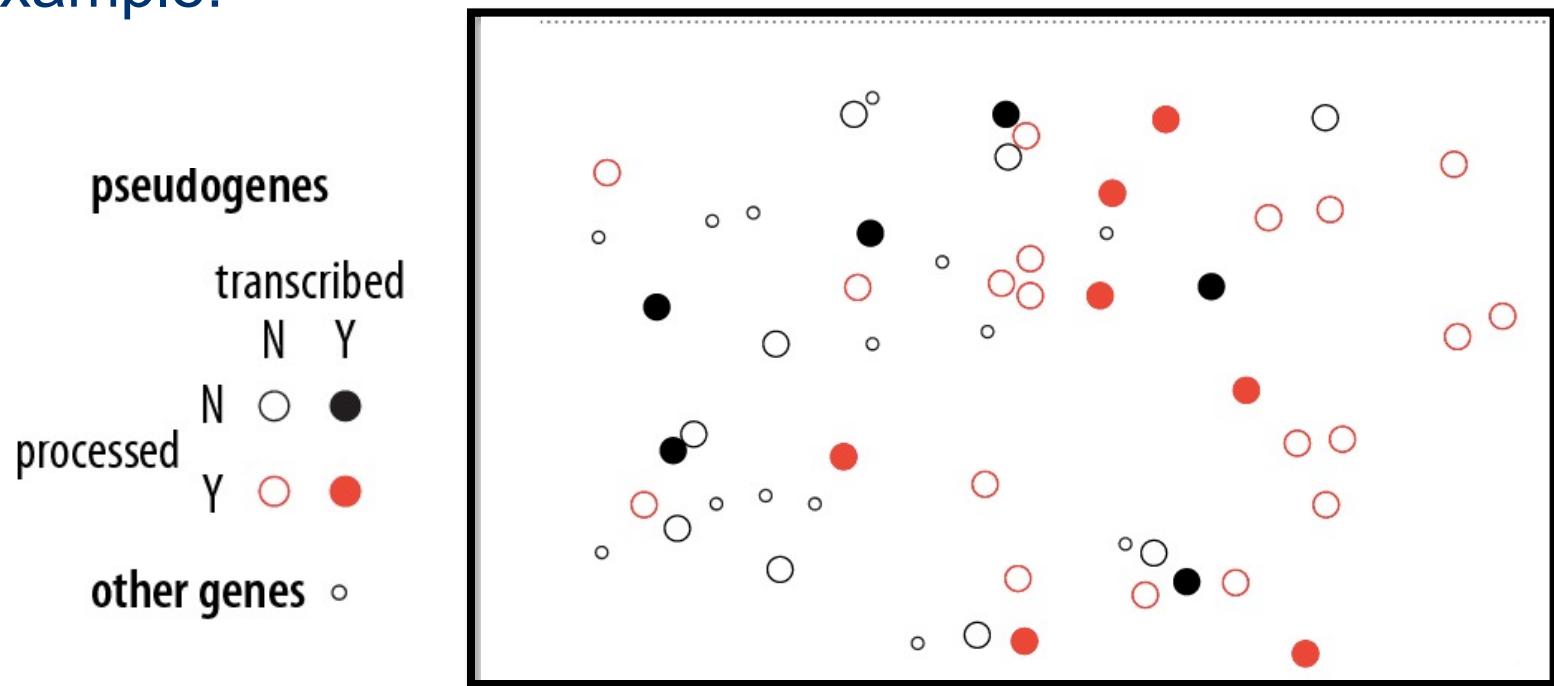


consistent inconsistent

□	A	■	A
■	B	□	B
□	C	■	C
■	D	□	D

Principle 2: consistency

- Visual variation should **reflect and enhance** the underlying variation in the data
- Avoid **visually similar** encodings for independent variables
- Example:



Principle 2: consistency

- Uniform size and alignment reduces visual complexity and aids interpretation
- Example:

variation refactored

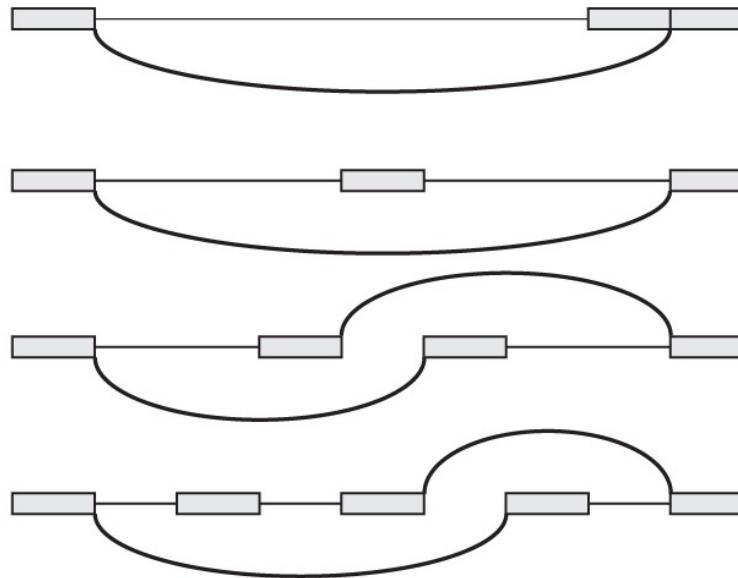
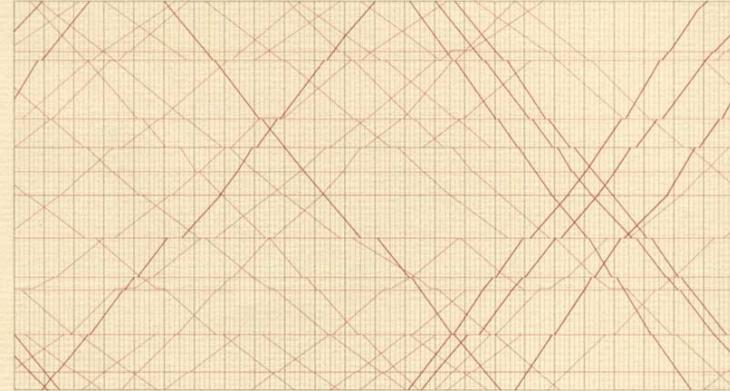


Fig. 1: Sharov AA et al. (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. Genome Res 15: 748-754.

Fig. 2: M. Krzwincki, behind every great visualization is a design principle, 2012

Tufte, 1983

“Above all else,
show the data.”



The Visual Display of Quantitative Information

EDWARD R. TUFTE

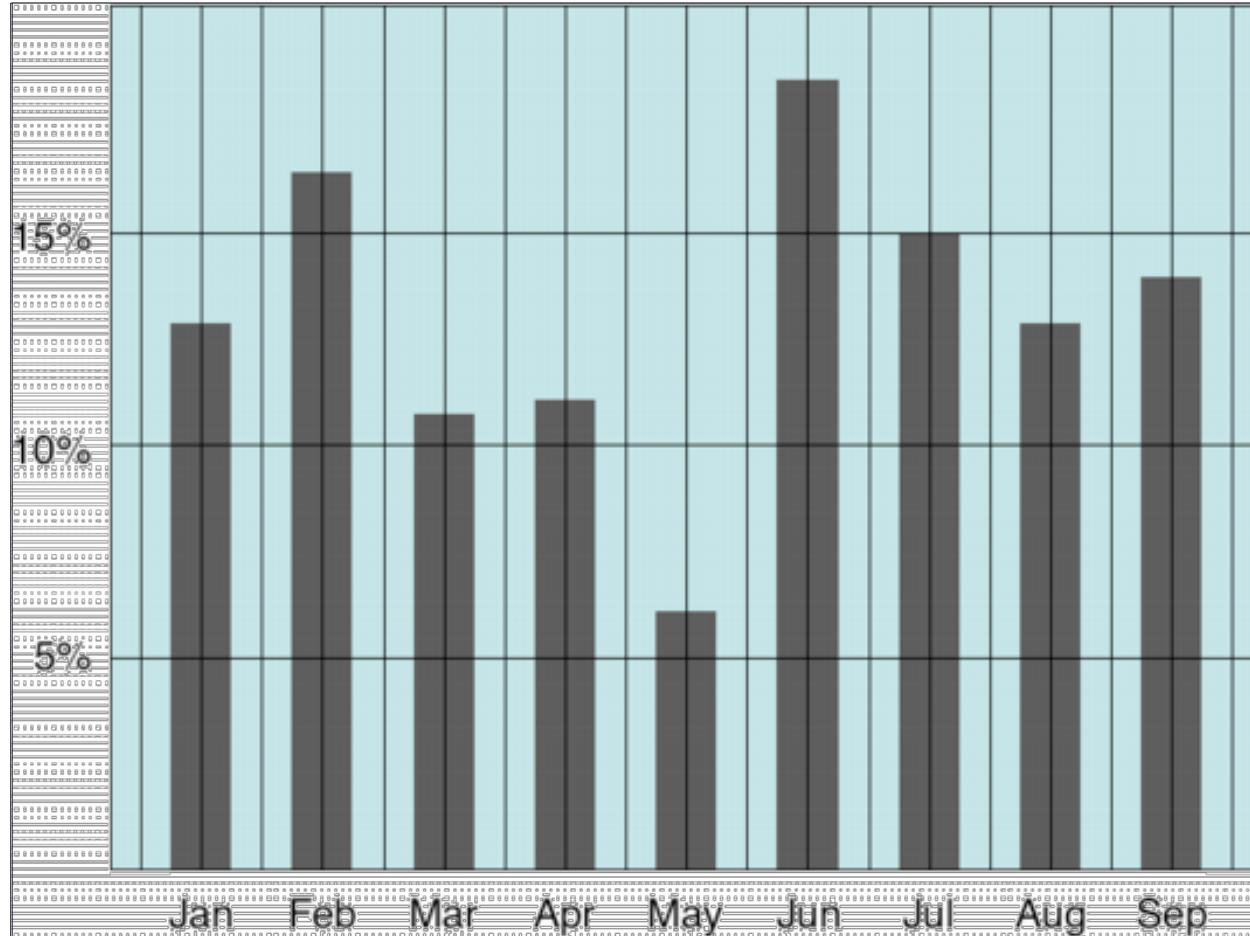
Tufte, 1983

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

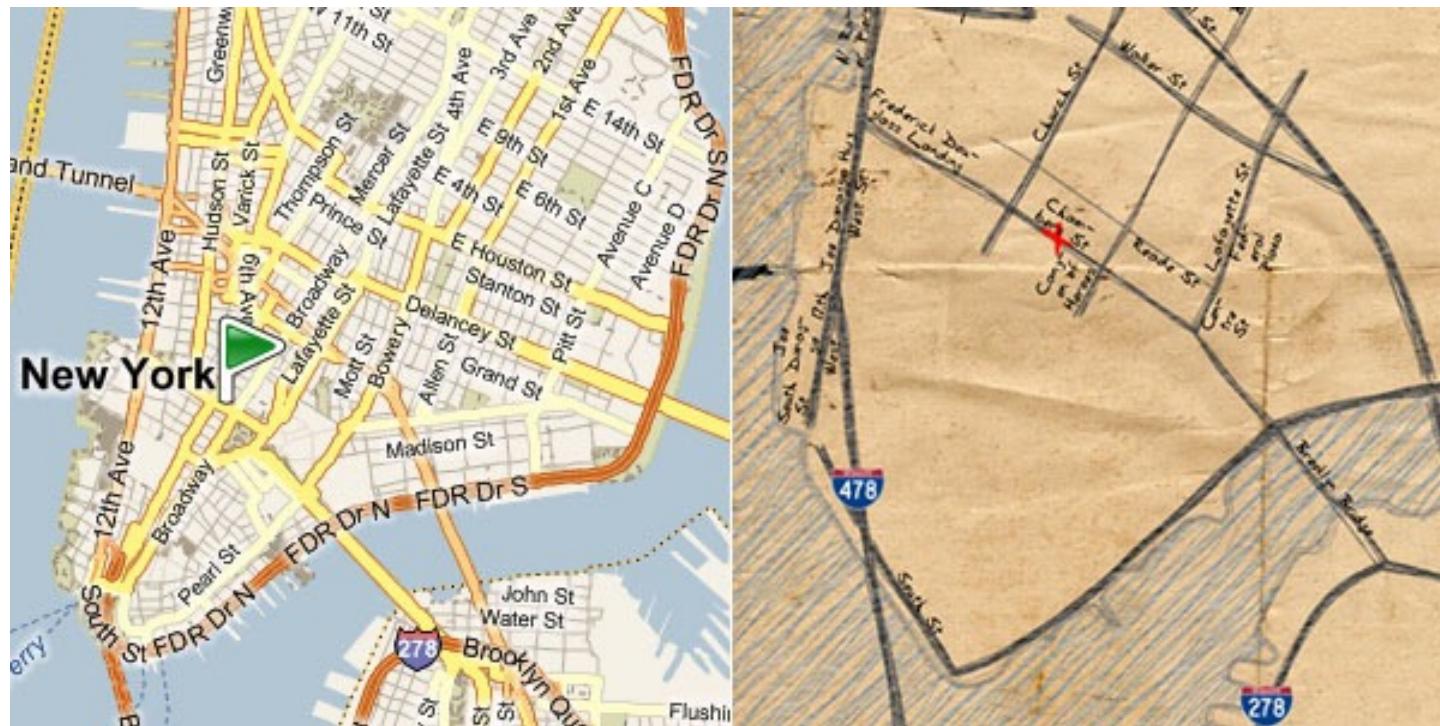
= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1 - proportion of a graphic that can be erased

Tufte: maximize the data-ink ratio



Familiar example



Discussion

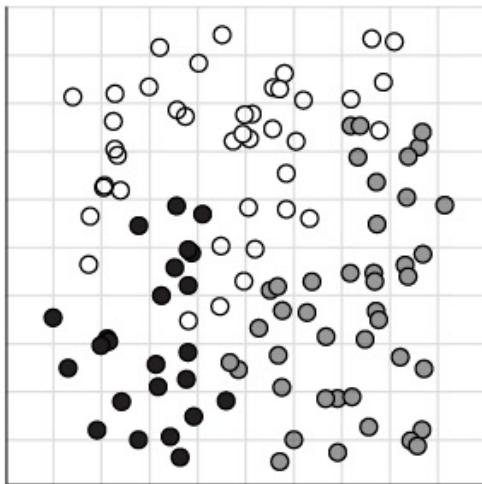
- What do you think of the data-ink ratio?
- Consider ways to **maximize** it...



Principle 3: importance ordering

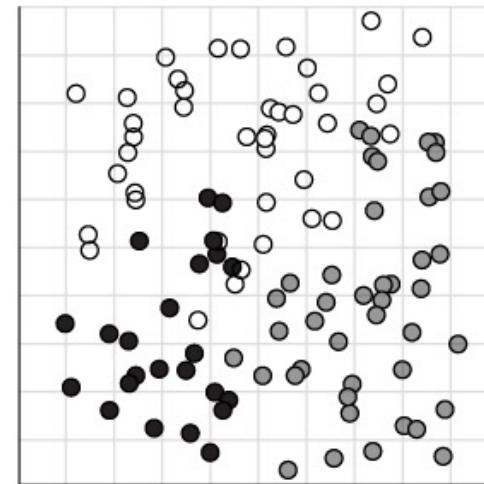
- Avoid unnecessary containment and repetition
- Example

A



Lorem ipsum dolor sit amet, consectetur adipiscing elit. In ut mauris quis tellus

B

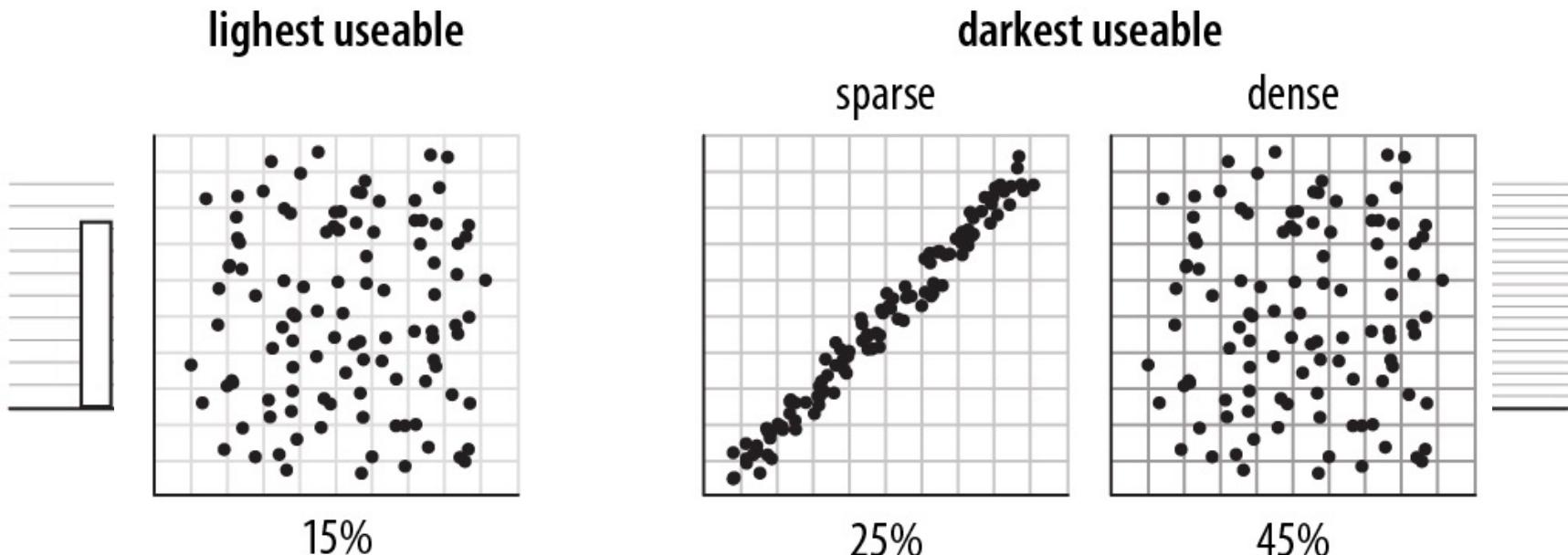


Lorem ipsum dolor sit amet, consectetur adipiscing elit. In ut mauris quis tellus

- A
- B
- C

Principle 3: importance ordering

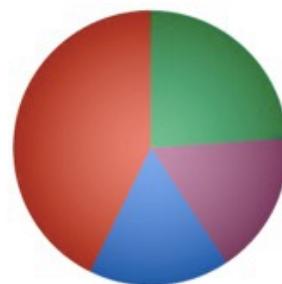
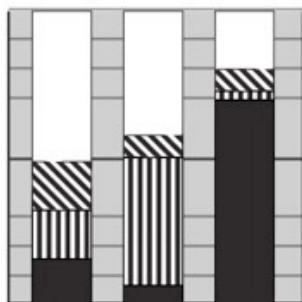
- Navigational aids shouldn't compete with data
- Avoid: **heavy axes, error bars and glyphs**



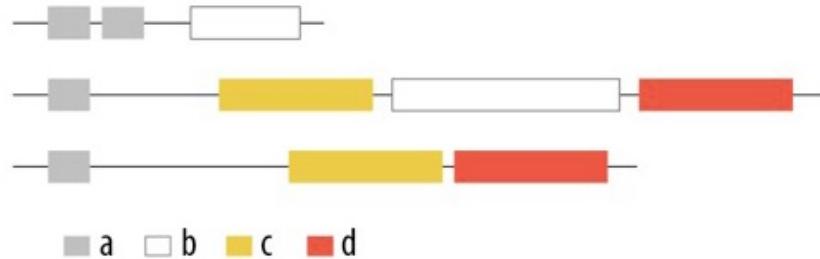
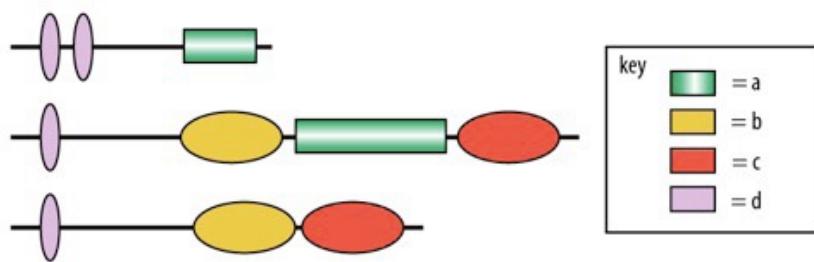
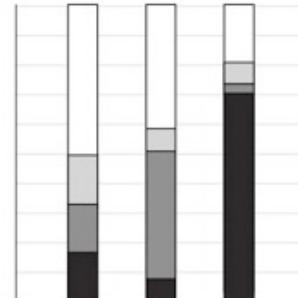
Principle 3: importance ordering

- Simplify, simplify, simplify...

chartjunk



visually concise



Sharov AA, et al (2006) Genome Res 16: 505-509.
Peterson J, et al. (2009) Genome Res 19: 2317-2323.
Thomson NR, et al. (2005) Genome Res 15: 629-640.
DB, Ko MS (2005) Genome Res 15: 748-754.

M. Krzwincki, behind every great visualization is a design principle, 2012

A caveat: “chart junk” and recall

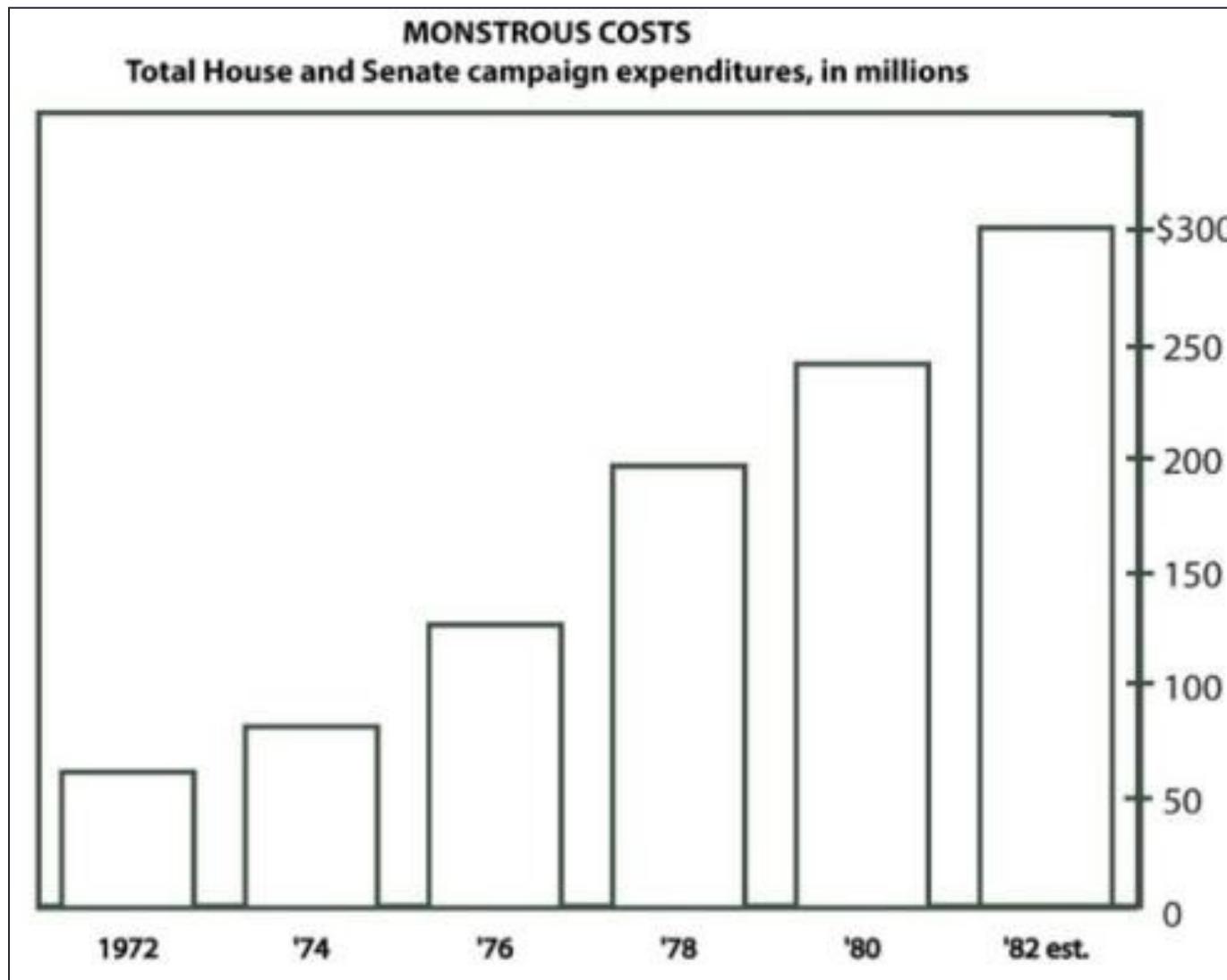


Chart junk and eye gaze

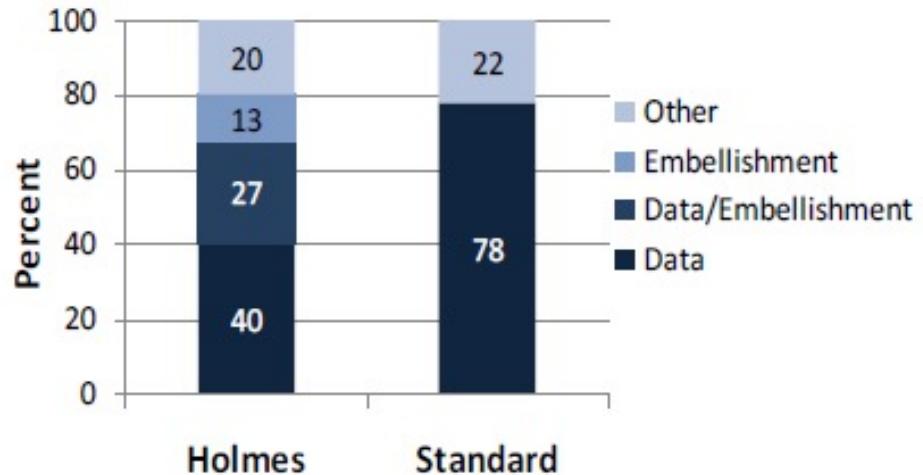
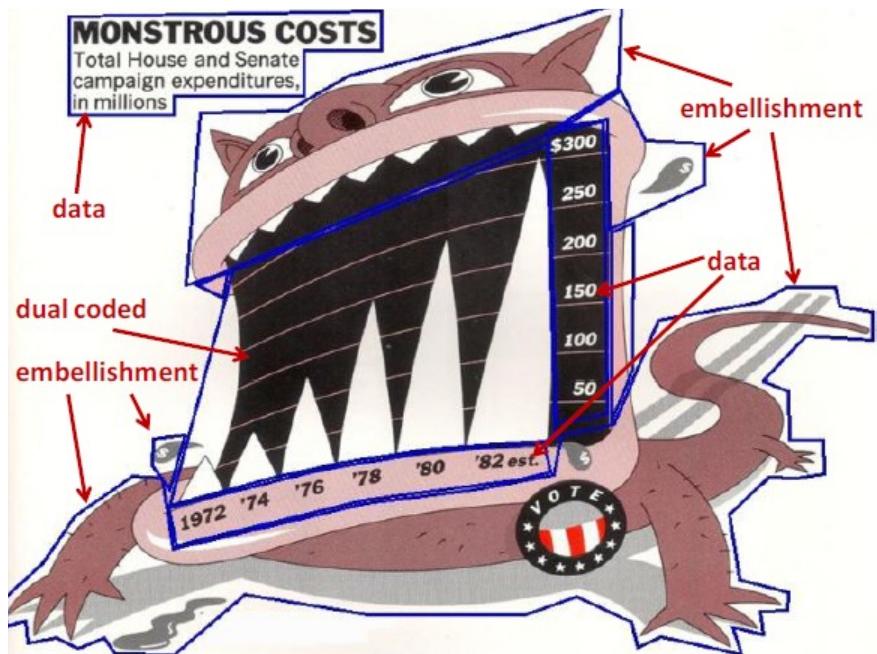


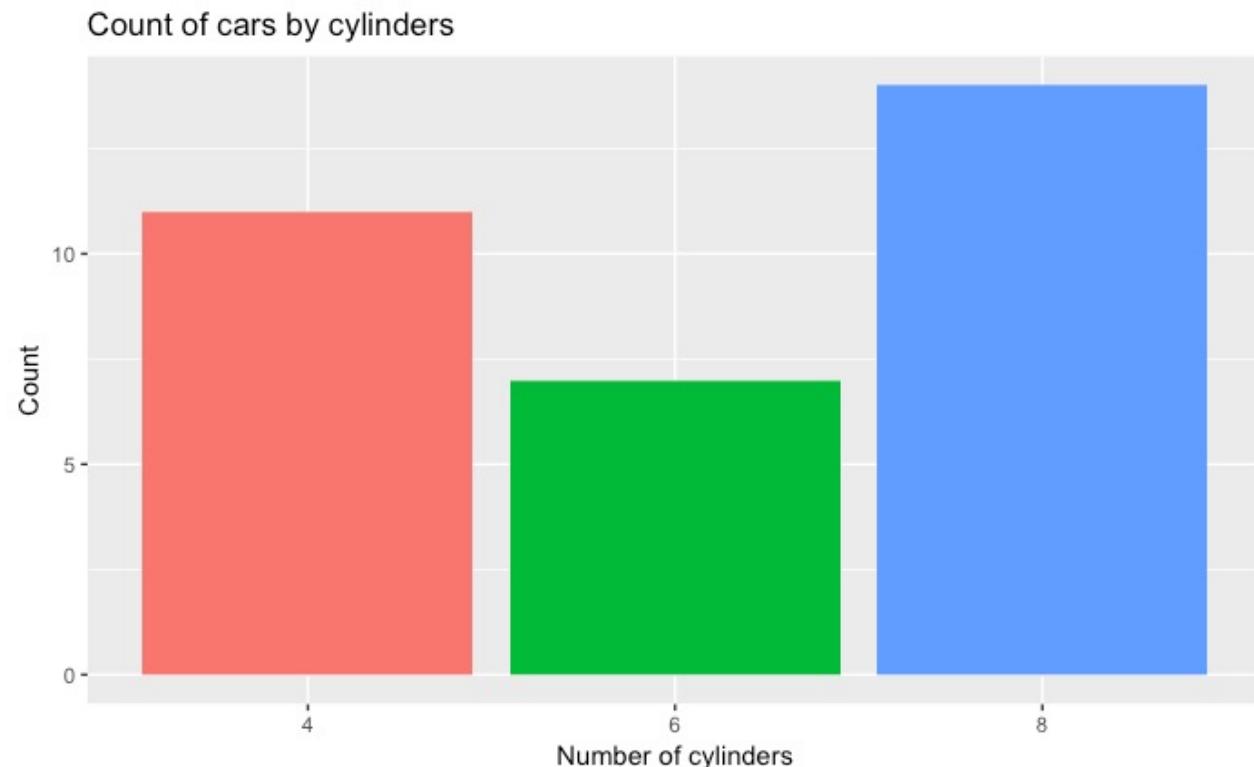
Figure 9. Percentage of on-screen time spent looking at different chart elements for Holmes and Plain charts.

What visualization techniques do you know?



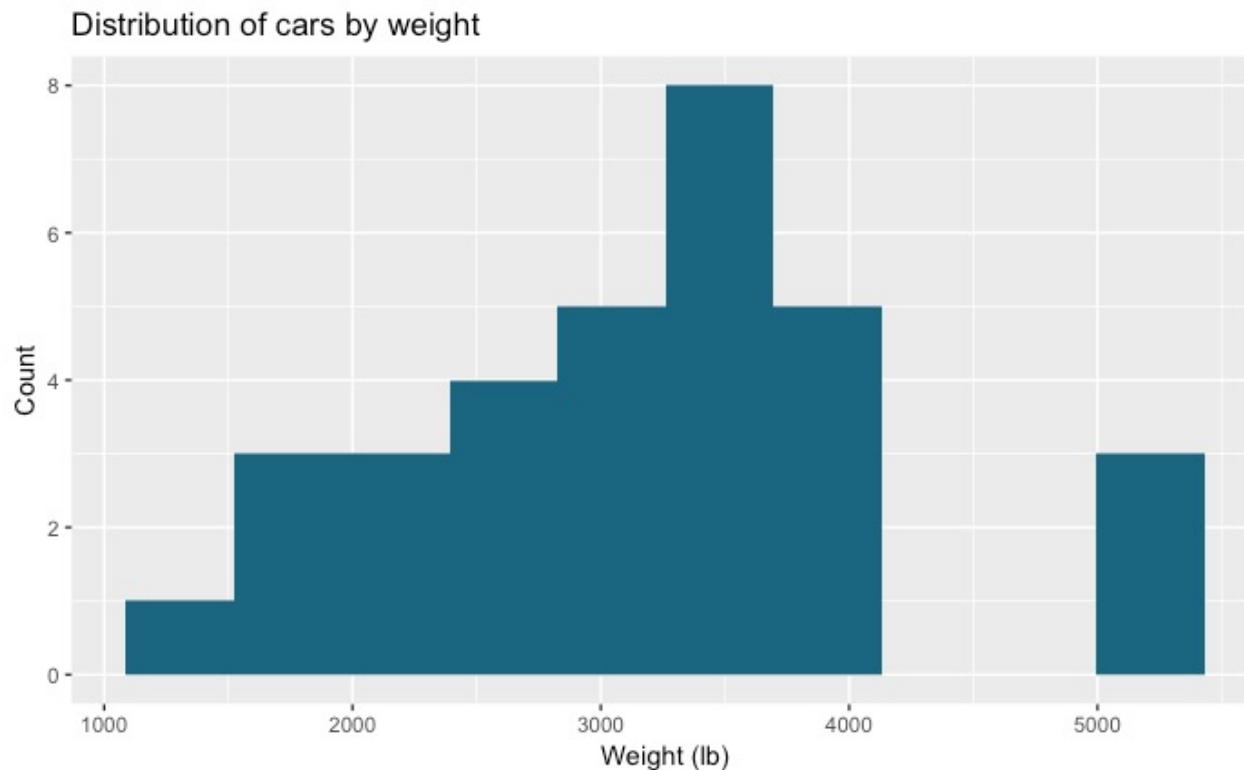
Bar chart

- Used for **comparable variables**
- Compares **quantitative** values for different categories
- Highlights **relative amounts**
- Grouped/stacked bars can break each **category** into sub-groups



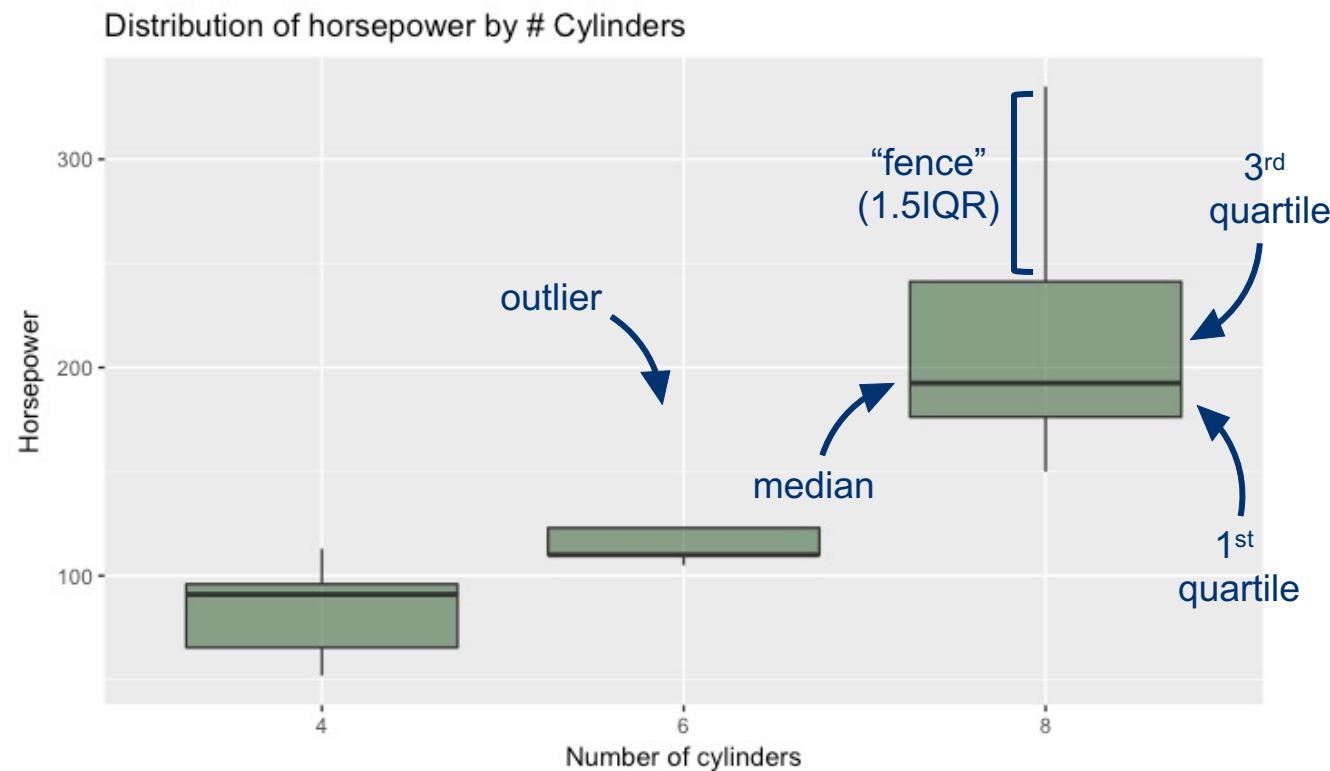
Histogram

- Looks like a bar chart... but the x-axis is **continuous**
- Y-axis shows count or relative frequency
- Highlights **distribution**
- Note: bin size makes a big difference!



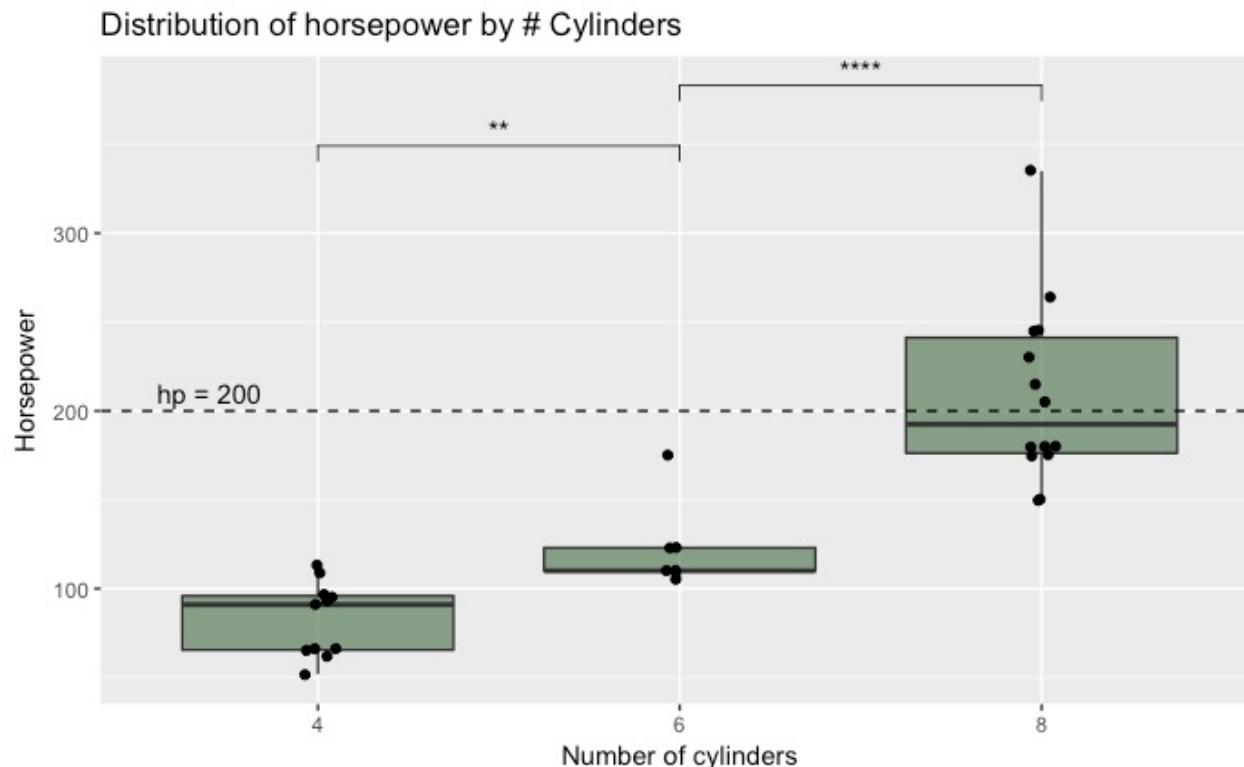
Boxplot

- Also useful for highlighting **distribution**
- Calls out key values:
 - median
 - 1st & 3rd quartiles
 - “fences”
 - outliers



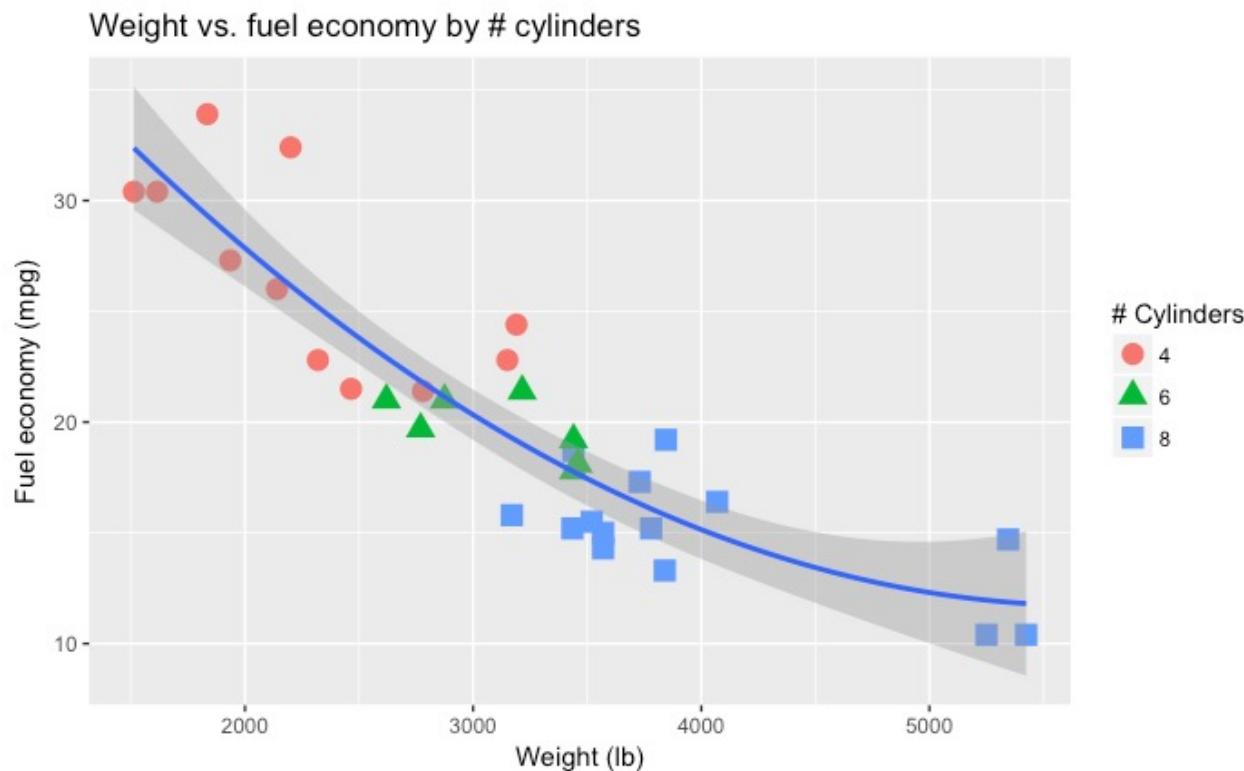
Boxplot add-ons

- Use “jitter” to show actual values
- Reference lines can help provide context
- Can use annotations to show statistical significance



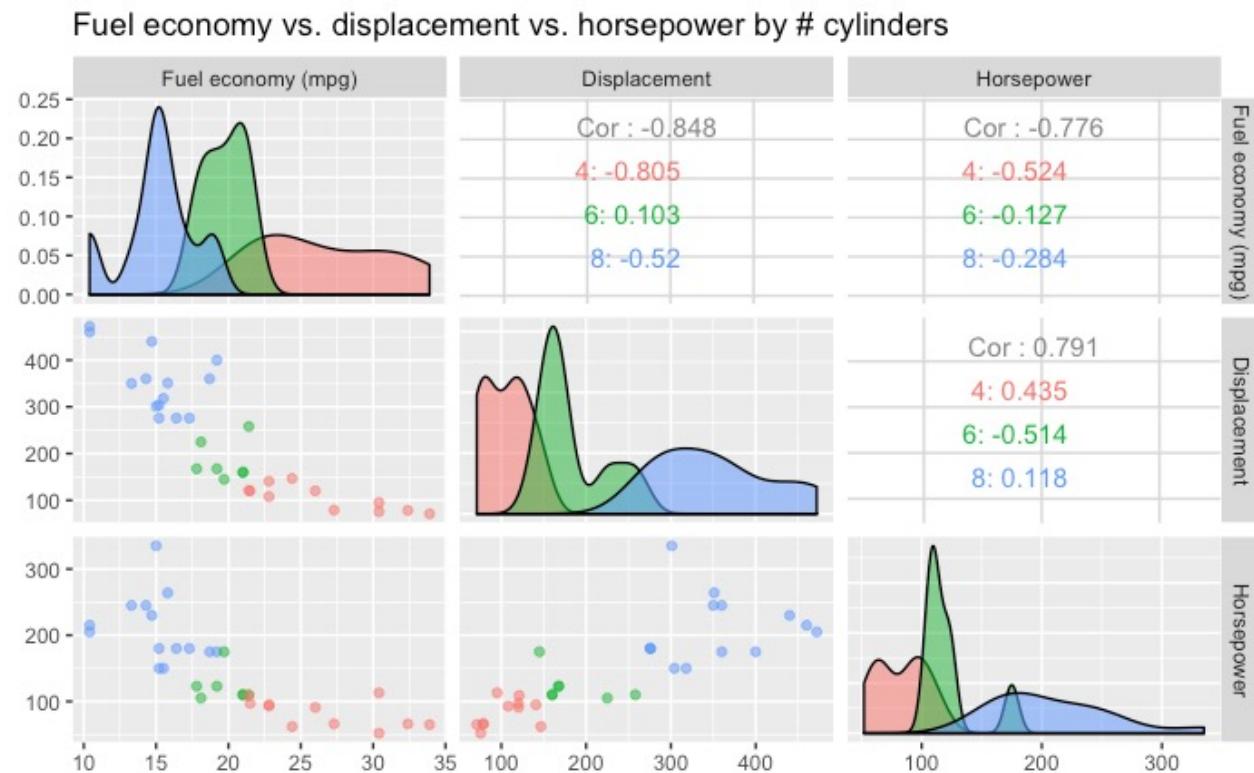
Scatterplot

- Shows the relationship between two **continuous variables**
- Each point in the plot represents an observation
- You can change color or symbol to **highlight groups**
- Sometimes useful to show a trend line (regression)



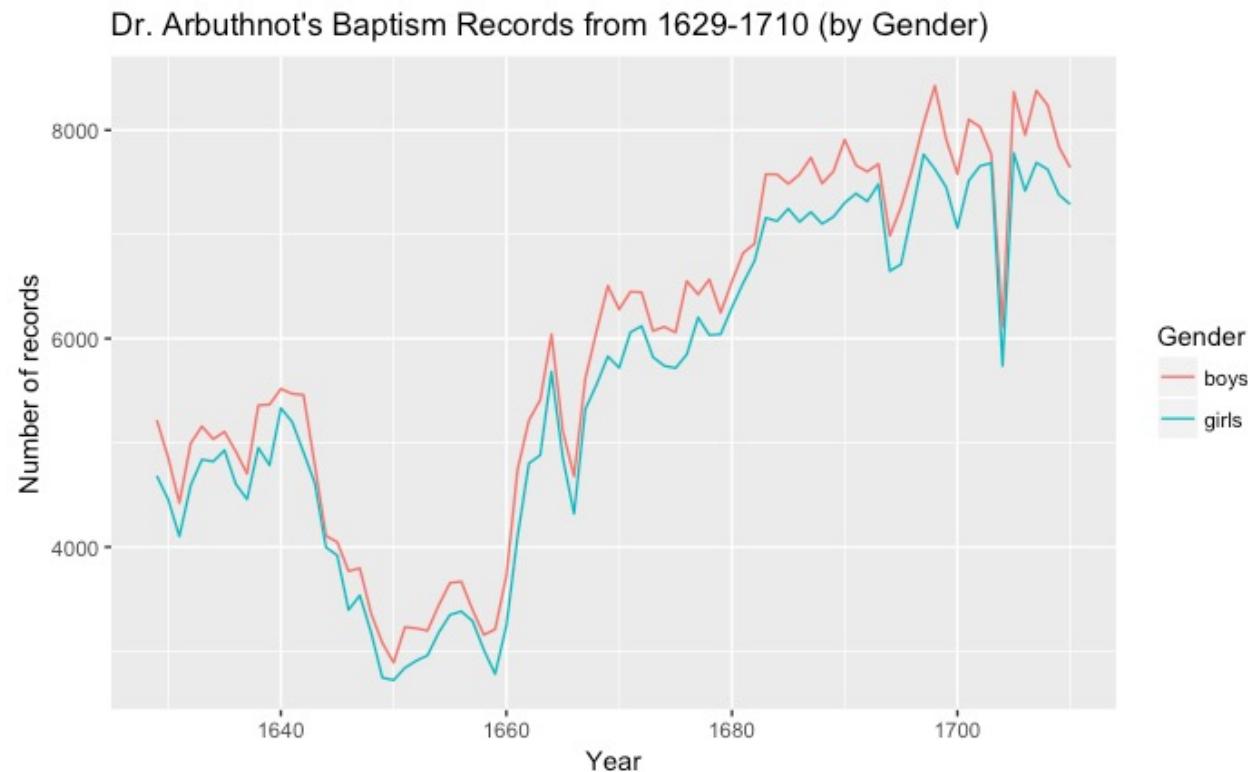
Scatterplot matrix (SPLOM)

- Scatterplots show the relationship between just two **continuous variables at a time**
- We can combine multiple scatterplots into a matrix to show additional relationships



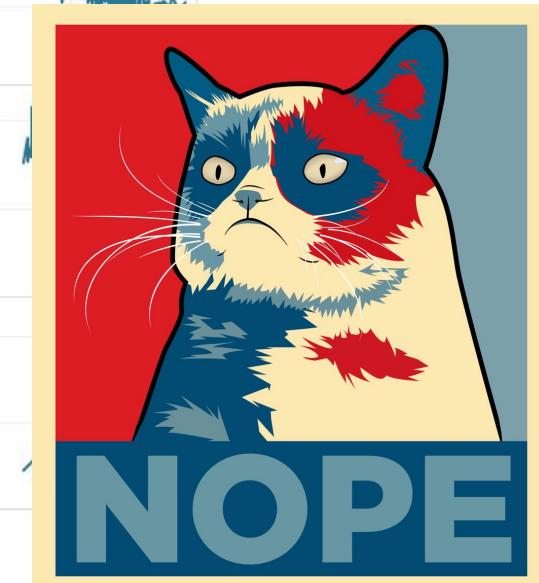
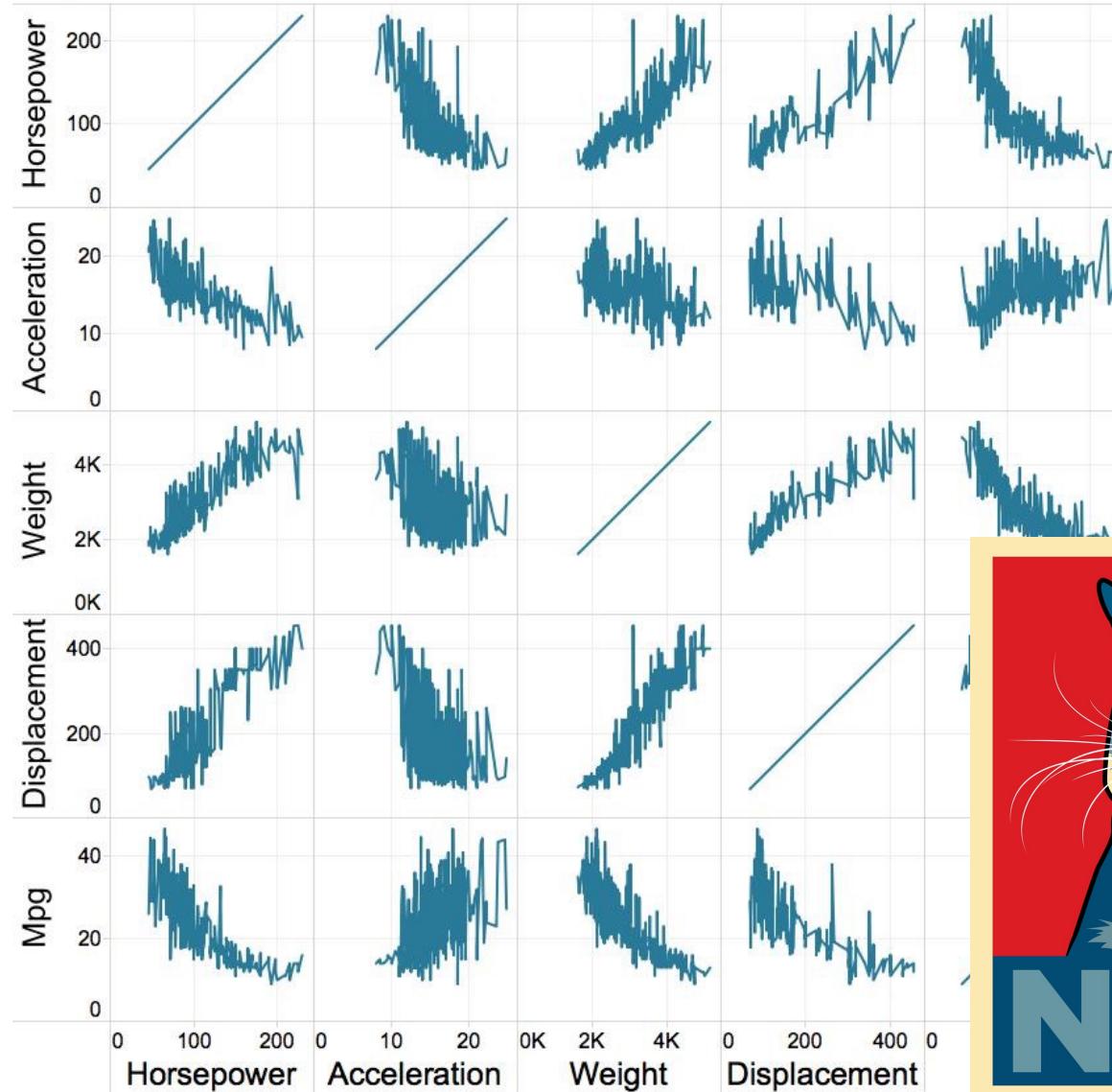
Line chart

- Shows the trend in one variable, often **over time**
- Multiple lines can show multiple variables, or the same variable for multiple observations (must have the **same scale!**)
- Highlights “position switches”

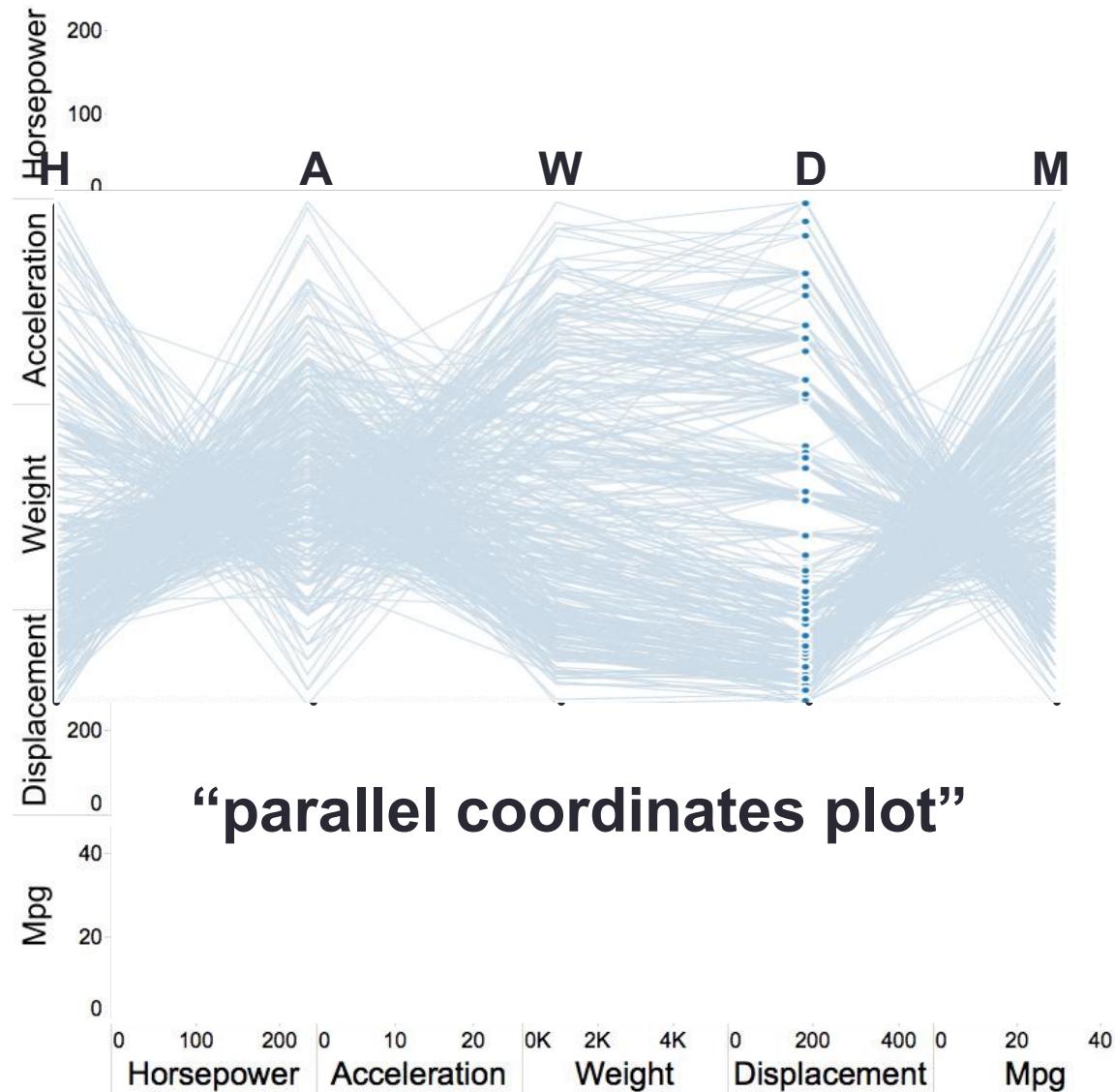


Multiple variables: line chart matrix?

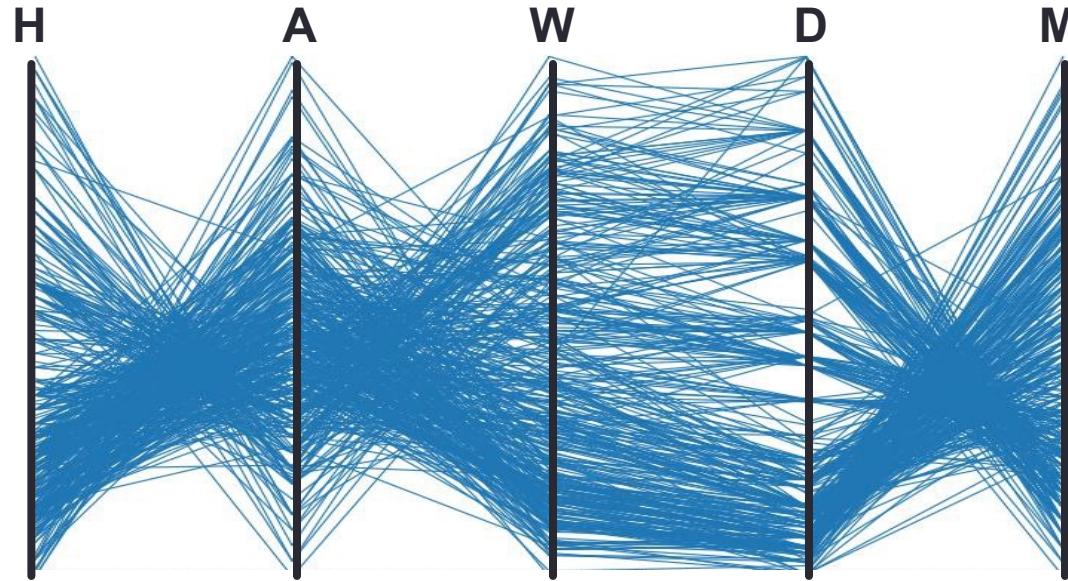
Sheet 3



Weirder idea



Morning challenge



“parallel coordinates plot”

How would you build
this using `ggplot2`?

