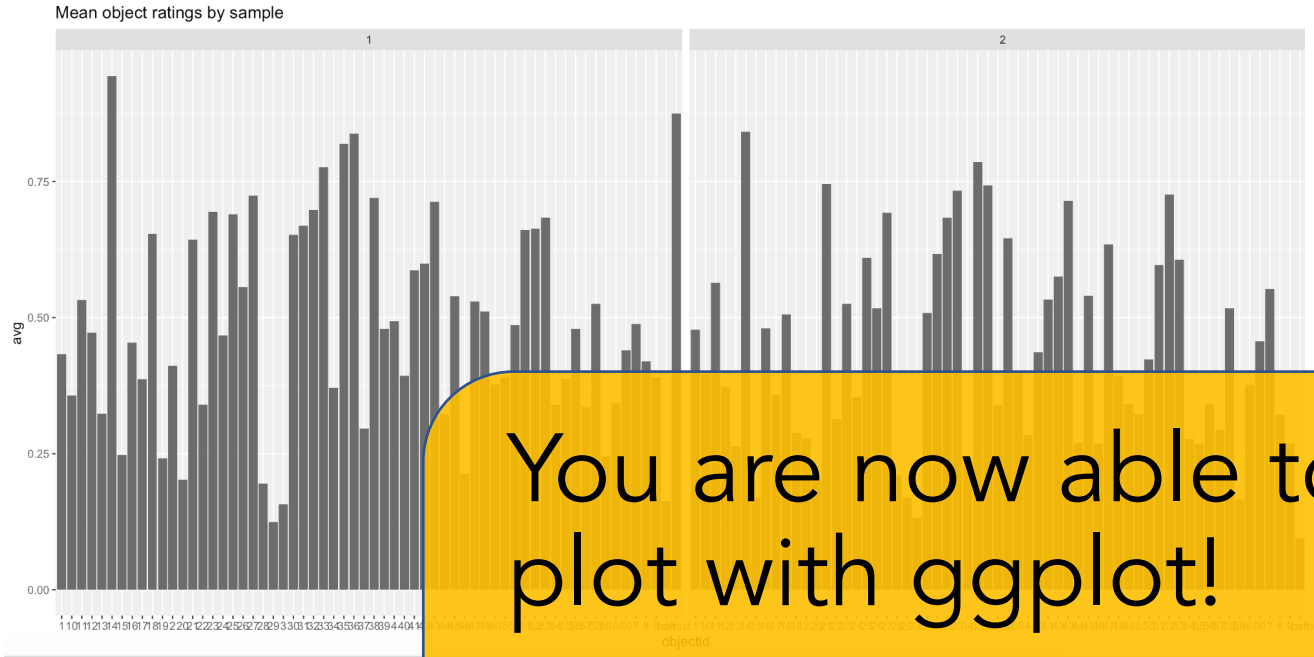
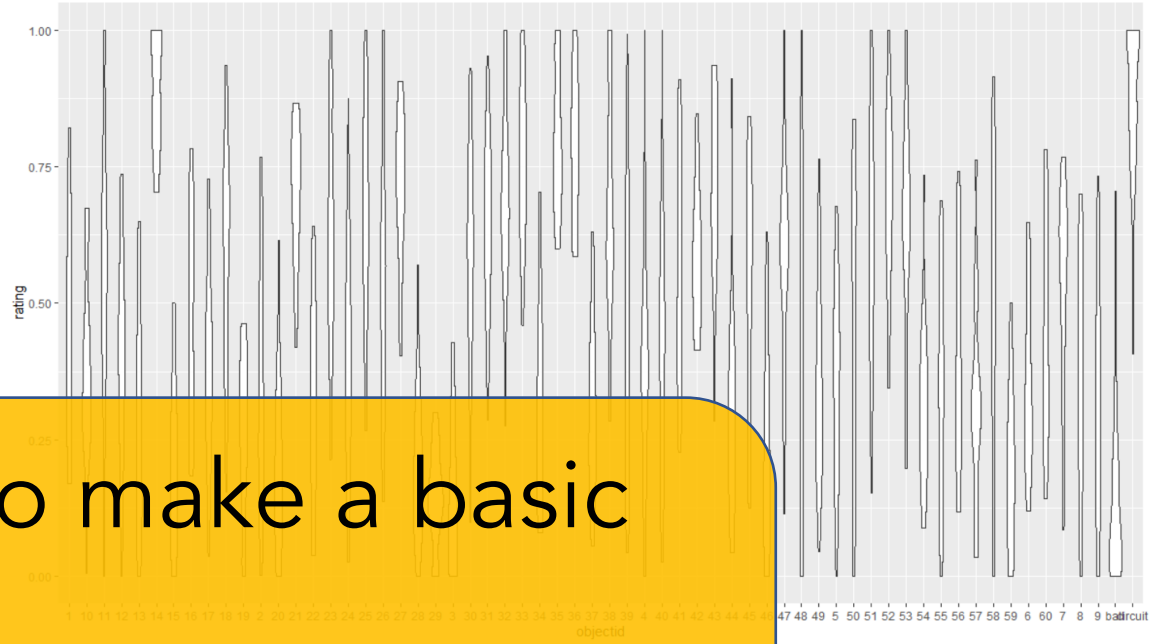
Principles of Visualization

20 September 2021

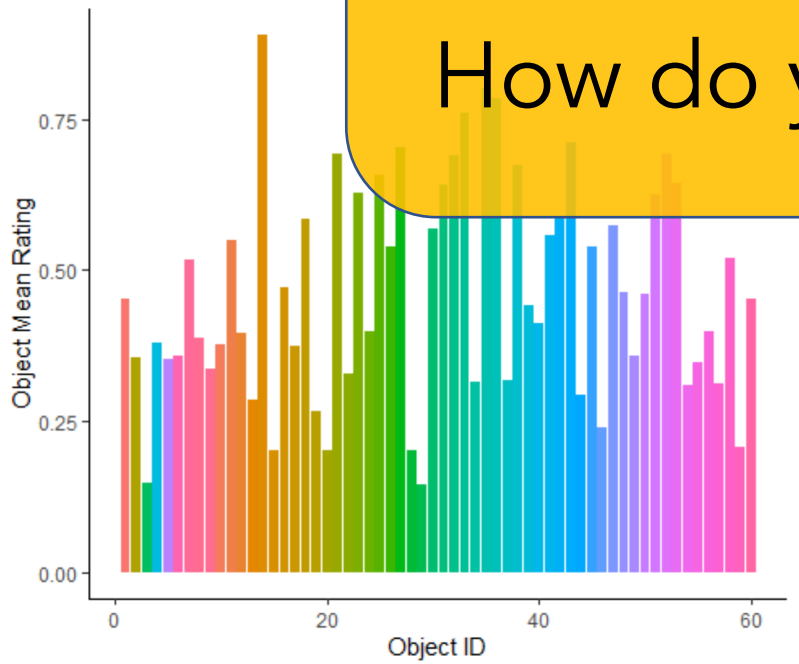
Modern Research Methods

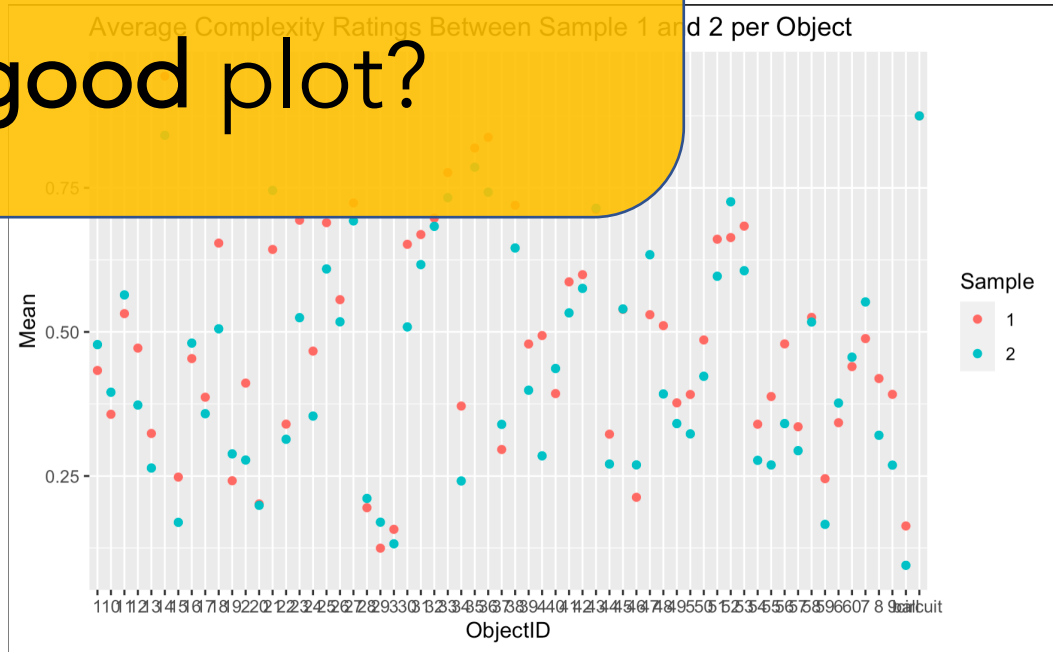




You are now able to make a basic plot with ggplot!

How do you make a **good** plot?





Visualization as **communication**

Visualization as communication

- There is no list of rules for what makes a good visualization
- Design depends on:
  1. the message you want to communicate
  2. who your audience is.
- Your goal is to make it as easy as possible for your audience to understand your message.
- Too much detail/information means your audience might not get the intended message.
- Too little detail and you haven’t communicated what you wanted

Principles of visualization as communication

1. Consider who your audience is
2. Get rid of “chart junk” – maximize info to ink ratio
3. Don’t be deceptive – show the raw data when appropriate
4. Think about human perception to maximize communication.

Consider who your audience is

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-0918-6

Check for updates

Gender stereotypes are reflected in the distributional structure of 25 languages

Molly Lewis[1,2] ✉ and Gary Lupyan[3]

Cultural stereotypes such as the idea that men are more suited for paid work and women are more suited for taking care of the home and family, may contribute to gender imbalances in science, technology, engineering and mathematics (STEM) fields, among other undesirable gender disparities. Might these stereotypes be learned from language? Here we examine whether gender stereotypes are reflected in the large-scale distributional structure of natural language semantics. We measure gender associations embedded in the statistics of 25 languages and relate these to data on an international dataset of psychological gender associations ($N = 656{,}636$). People's implicit gender associations are strongly predicted by gender associations encoded in the statistics of the language they speak. These associations are further related to the extent that languages mark gender in occupation terms (for example, 'waiter'/'waitress'). Our pattern of findings is consistent with the possibility that linguistic associations shape people's implicit judgements.

By the time they are two, children have begun to acquire the gender stereotypes in their culture[1]. These stereotypes can have undesirable effects. For example, in one study, six-year-old girls were less likely than boys to choose activities that were described as being for children 'who are very, very smart' and also less likely to think of themselves as 'brilliant'[2]. Such beliefs may, over time, translate to the observed lower rates of female participation in STEM fields[3–6] and are reflected in large differences in perceived self-efficacy; boys reported having greater ability to understand and explain various scientific findings (independent of actual ability)[6]. Here we attempt to understand where such beliefs a particular gender[9]. In some cases, a subtle turn of phrase can influence children's gender-based generalization[10,11]. For example, Cimpian and Markman found that children were more likely to infer that a novel skill is stereotypical of a gender if the skill is introduced with a generic as opposed to a non-generic subject[10] ("[Girls are/There is a girl who is] really good at a game called 'gorp'"). Such work shows that in certain experimental settings, language can influence stereotype formation. In this study, we investigate whether a similar correspondence between language associations and stereotypes exists in a large corpus of naturalistic text and among an international sample of participants.

SCIENTIFIC AMERICAN

Subscribe

BEHAVIOR

How Dozens of Languages Help Build Gender Stereotypes

Usage patterns shape biases worldwide, whether in Japanese, Persian or English

By Gary Stix on August 3, 2020

Credit: Simone Golob *Getty Images*

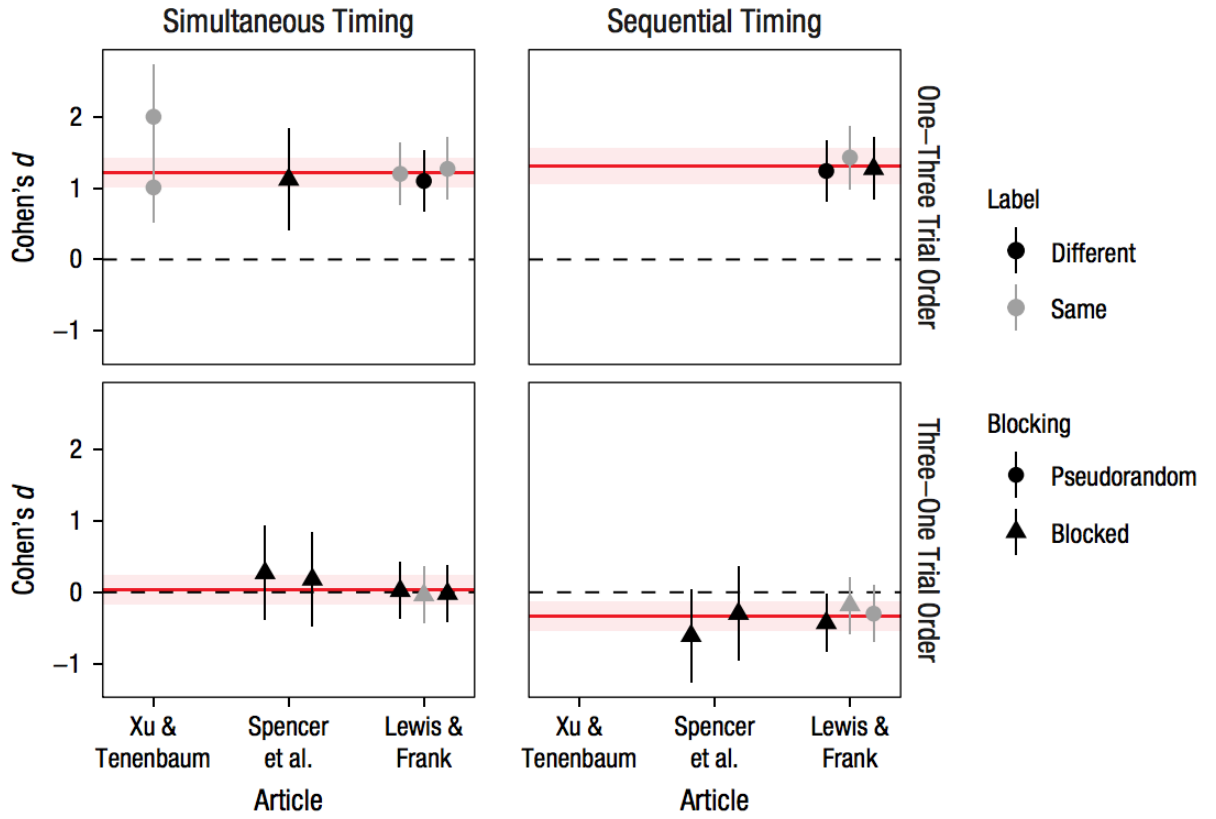
Linguists use machine-learning techniques for mining large text corpora to detect how the structure of a language lends meaning to its words. They work on the assumption that terms that appear in close proximity to one another may have similar connotations: dogs turn up near cats more often than canines appear close to bananas.

This same method of burrowing into texts—more formally called the search for distributional semantics—can also provide a framework for analyzing psychological attitudes, including gender stereotypes that contribute to the underrepresentation of women in scientific and technical fields. Studies in English have shown, for example, that the word "woman" often appears close to "home" and "family," whereas "man" is frequently paired with "job" and "money."

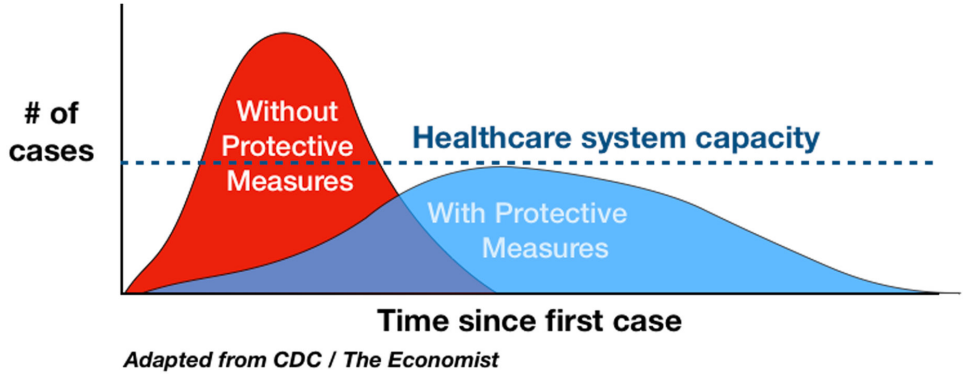
Kinds of Audiences

- You, exploring data
- Newspaper
- Journal Article
- Poster Presentation

- Audiences differ in expertise, how much time they have to view the plot, and motivation

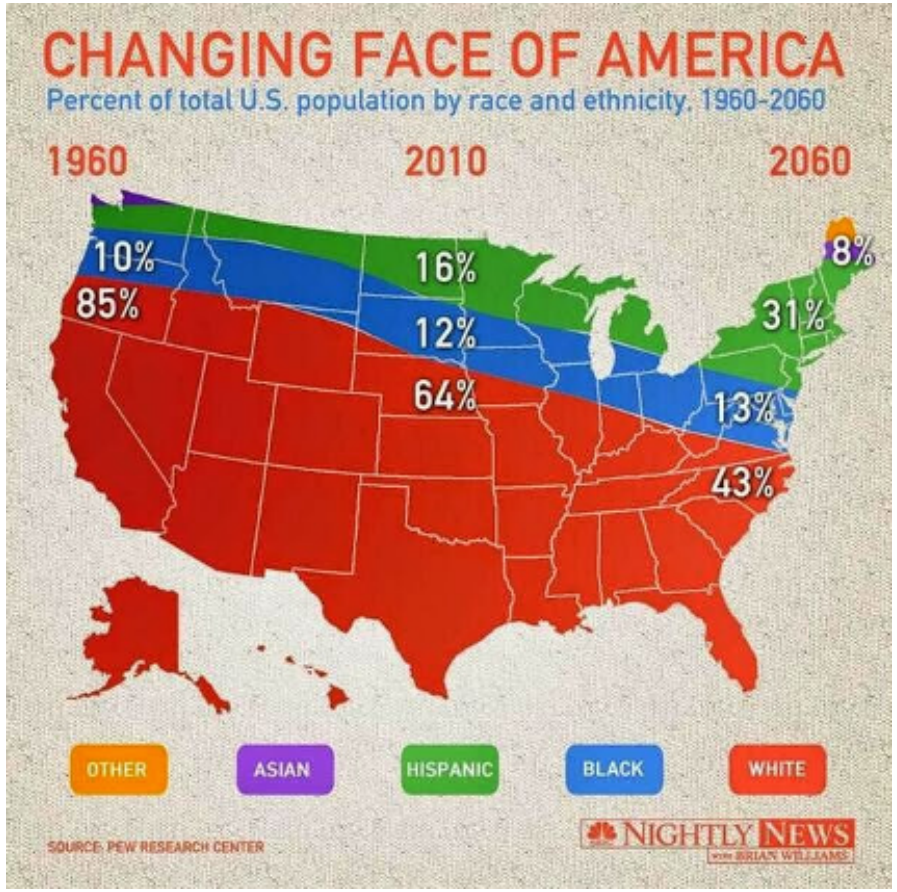


(Lewis & Frank, 2018)



(Source: Nytimes)

Get rid of "chart junk" – maximize info to ink ratio

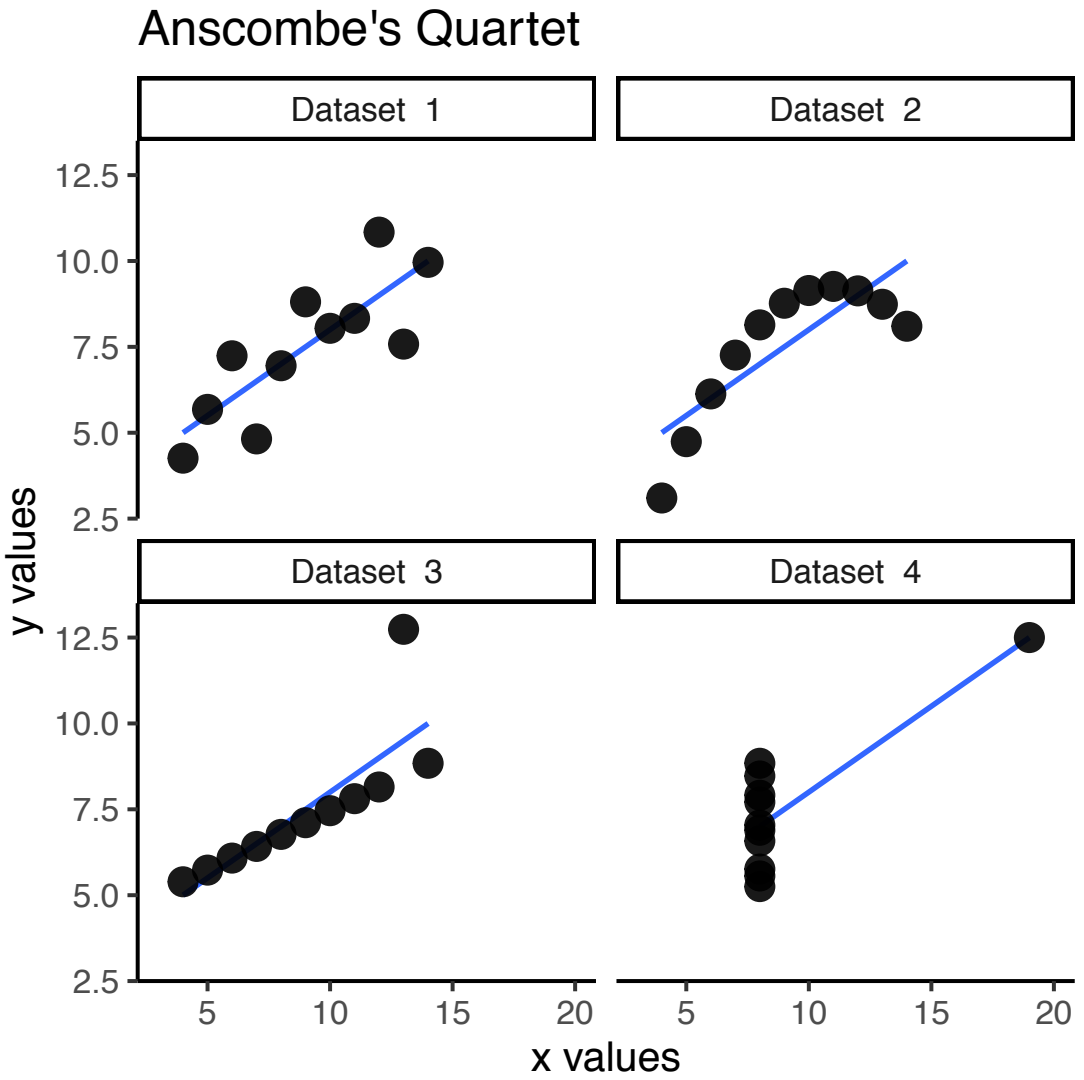


(image source: Healy, 2018)

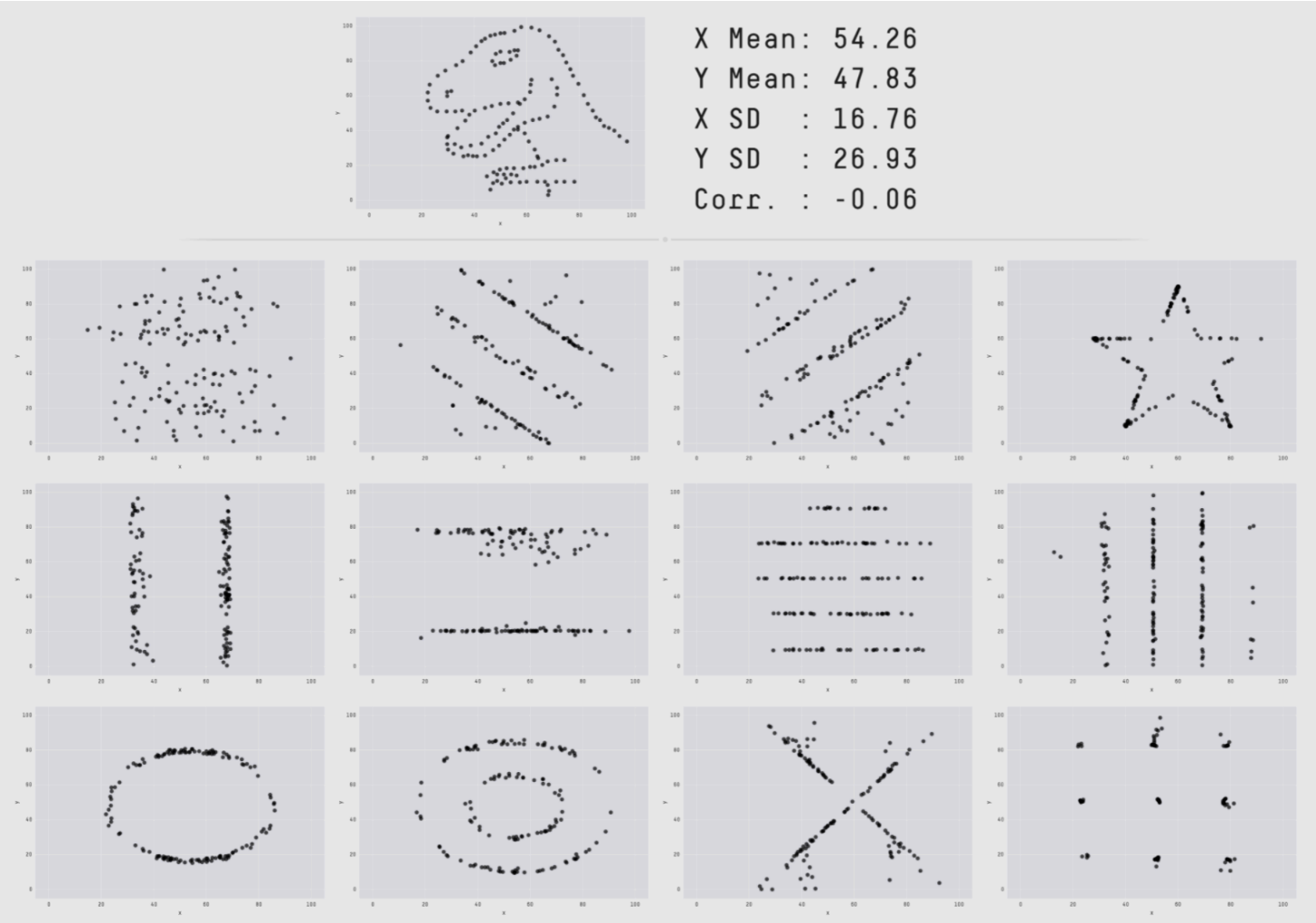
Don’t be deceptive – show the raw data when appropriate

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

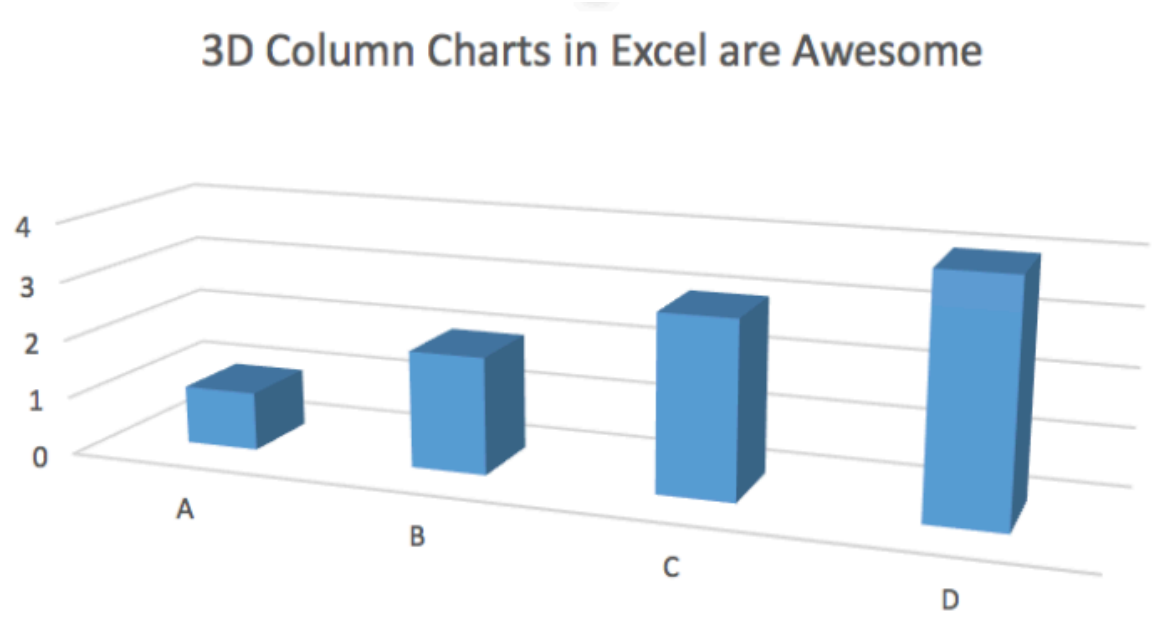


Don’t be deceptive – show the raw data when appropriate

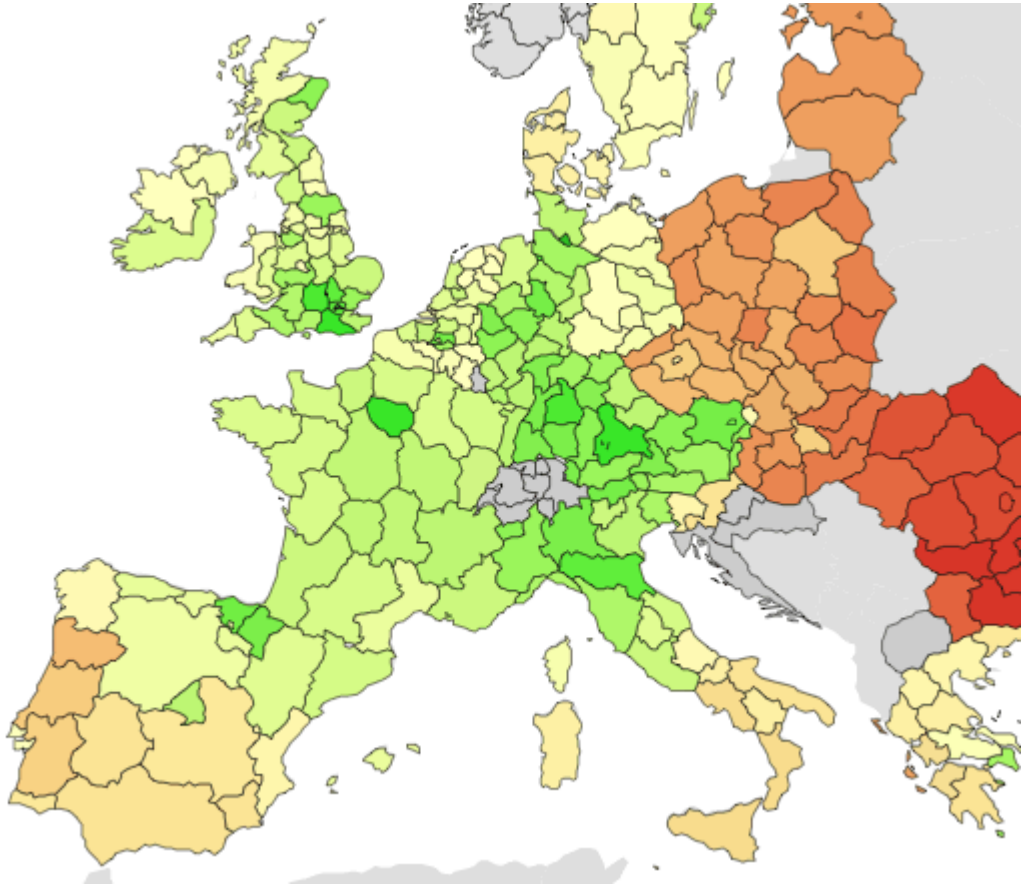


Source: Alberto Cairo

Think about human perception to maximize communication.

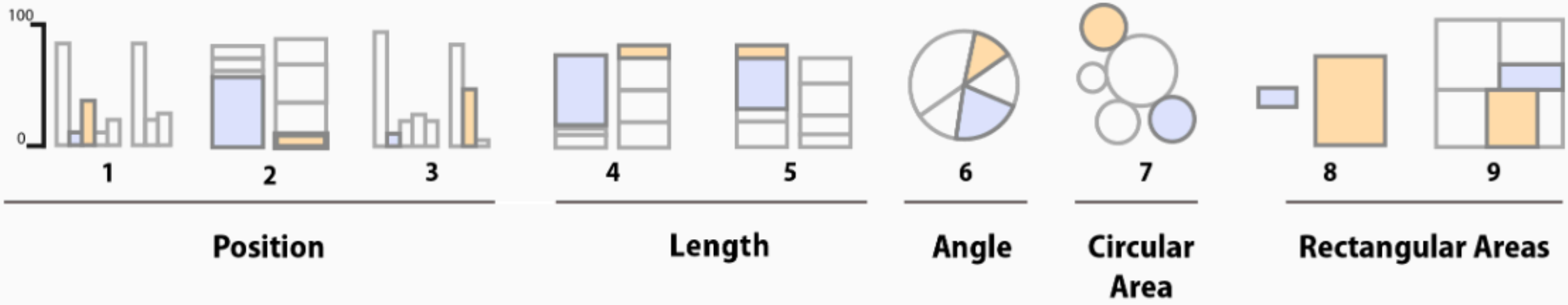


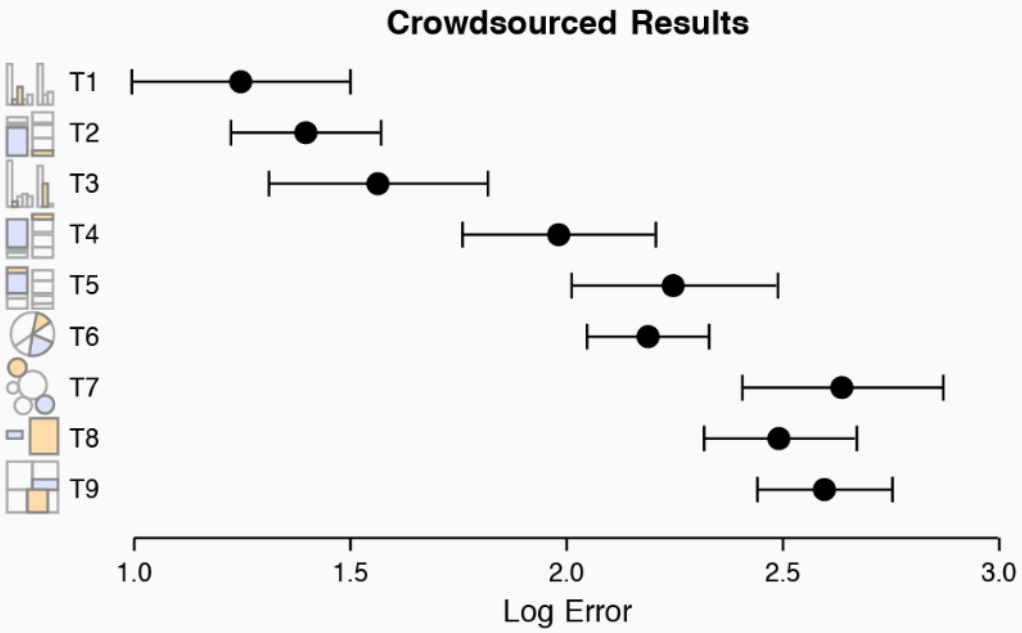
(source: Healy, 2018)



(source: Gregor Aisch)

Some things are easier to perceive than others!





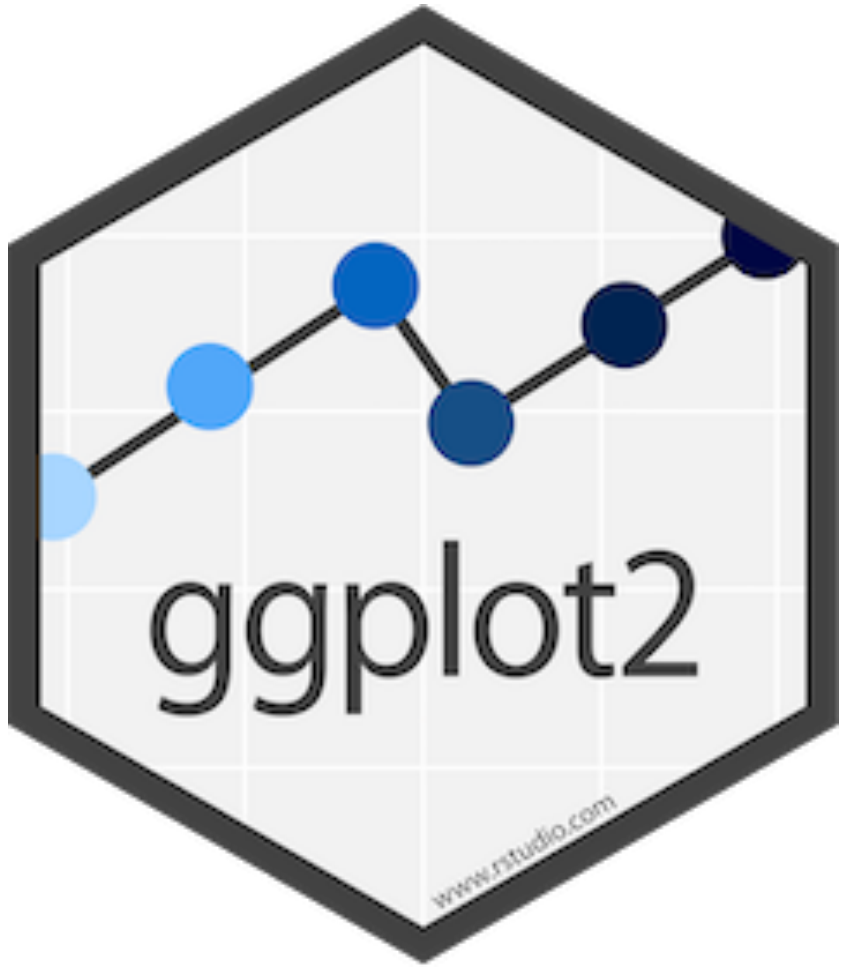
Position and length are easiest for the human perceptual system to distinguish.

Heer & Bostock (2010);
image source: Healy, 2018

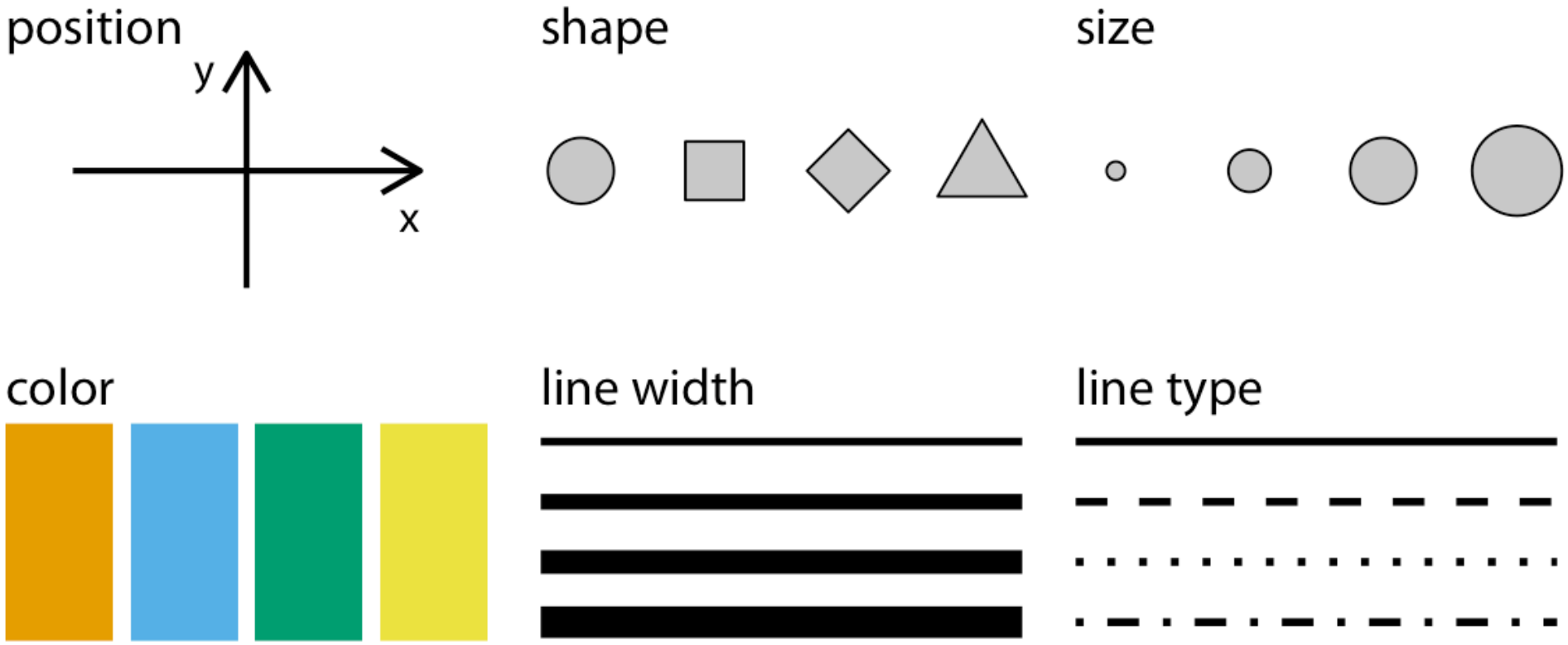
Principles of visualization as communication

1. Consider who your audience is
2. Get rid of “chart junk” – maximize info to ink ratio
3. Don’t be deceptive – show the raw data when appropriate
4. Think about human perception to maximize communication.

Guidelines for implementing principles of visualization as communication



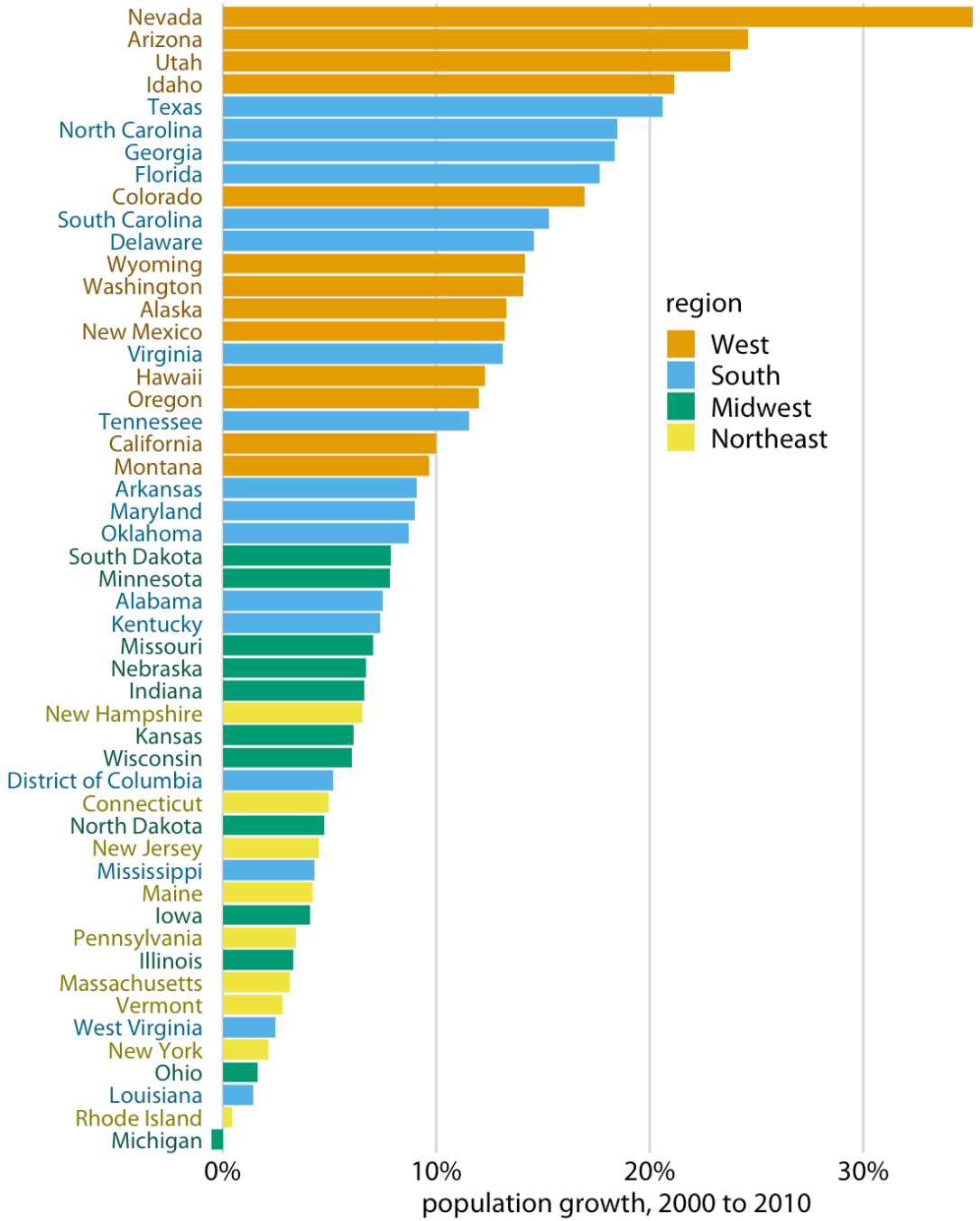
Main channels available in ggplot



(image source: Wilke, *Fundamentals of Data Visualization*)

Color for **qualitative** variables

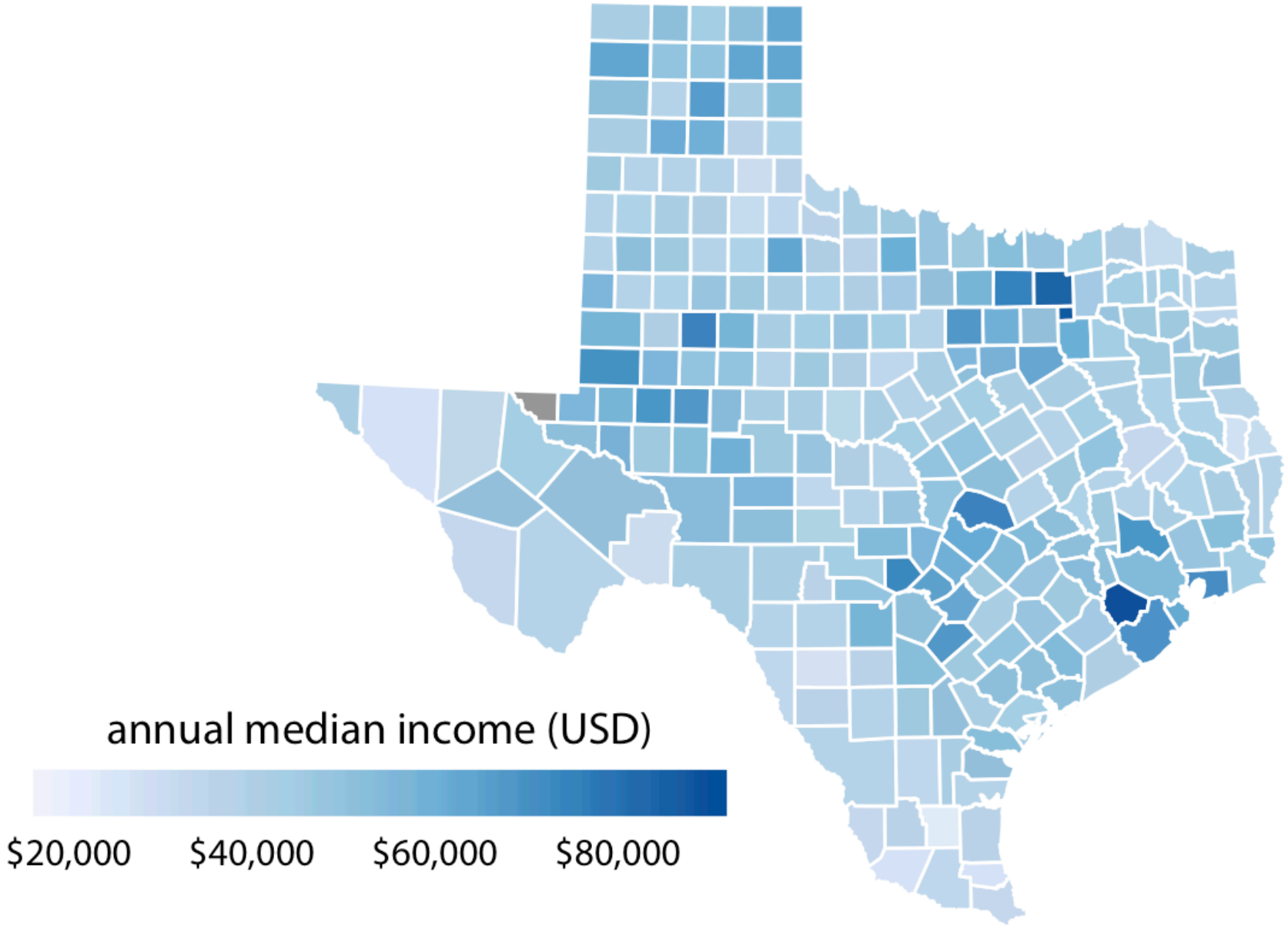
- Color as a tool to distinguish
- Colors clearly distinct from each other
- No one color should stand out relative to the others
- No impression of order



(image source: Wilke, *Fundamentals of Data Visualization*)

Color for **quantitative** variables

- Color indicates which values are larger/smaller
- How distant two values are from each other



(image source: Wilke, *Fundamentals of Data Visualization*)

Working with color in ggplot

Can change the palette of colors used by ggplot.

* = aesthetic

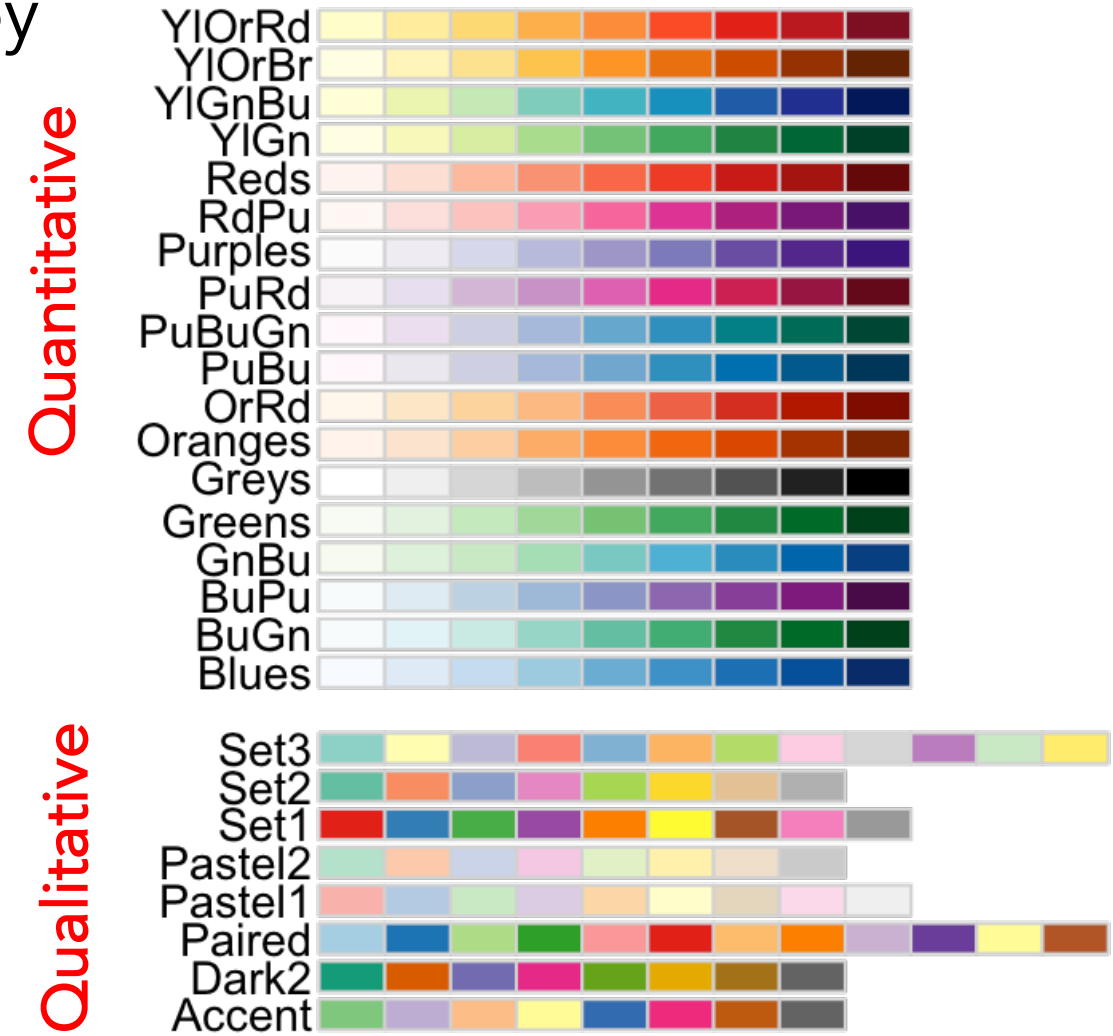
scale_*_brewer(palette = “YlOrRd”)

scale_*_manual(values =
c("#000000", "#E69F00”)

hex color

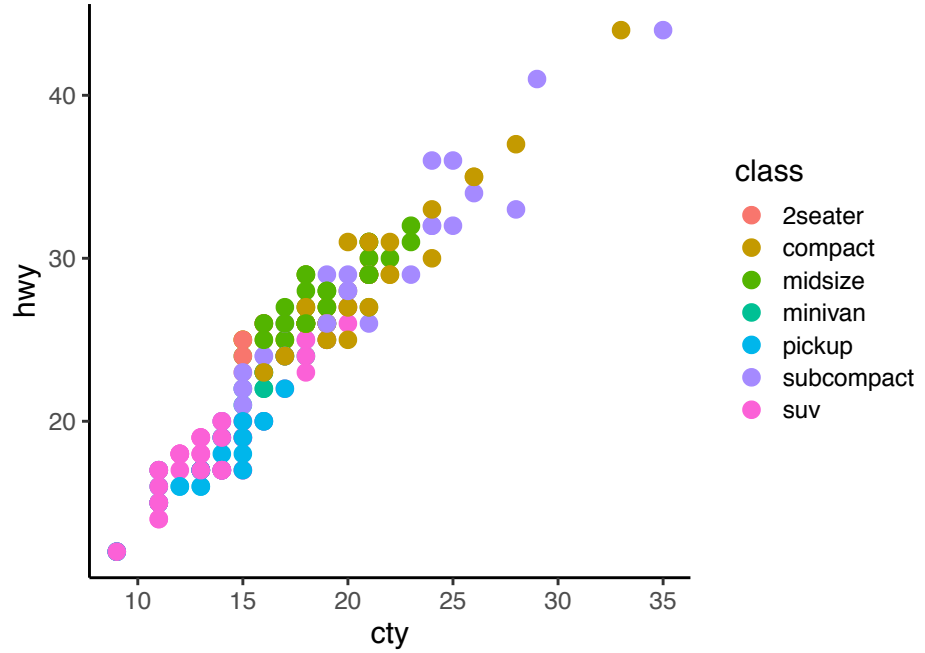
https://htmlcolorcodes.com/

Brewer Color Palette



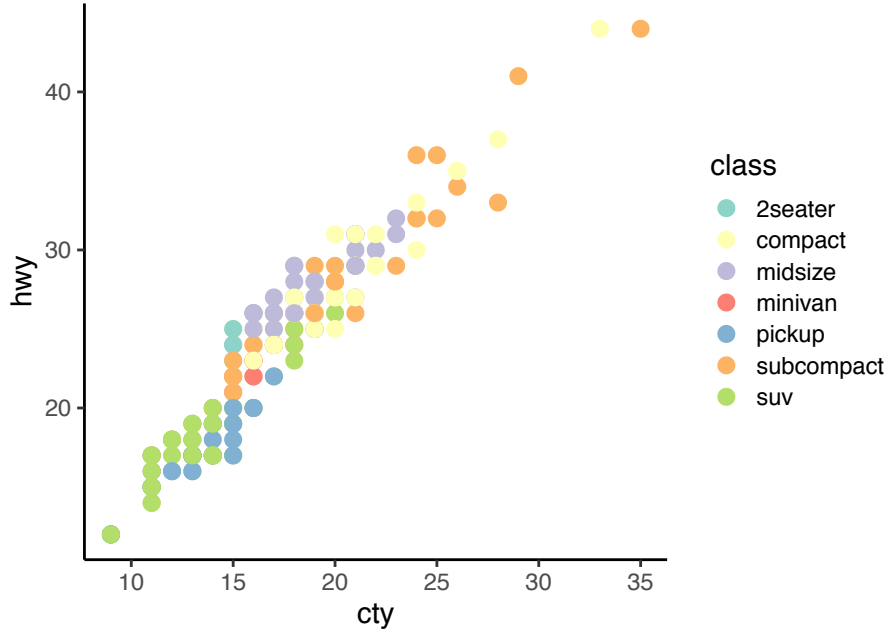
Default

```
ggplot(data = mpg, aes(x = cty, y = hwy)) +
    geom_point(aes(color = class), size = 4) +
     theme_classic(base_size = 16)
```



Brewer palette

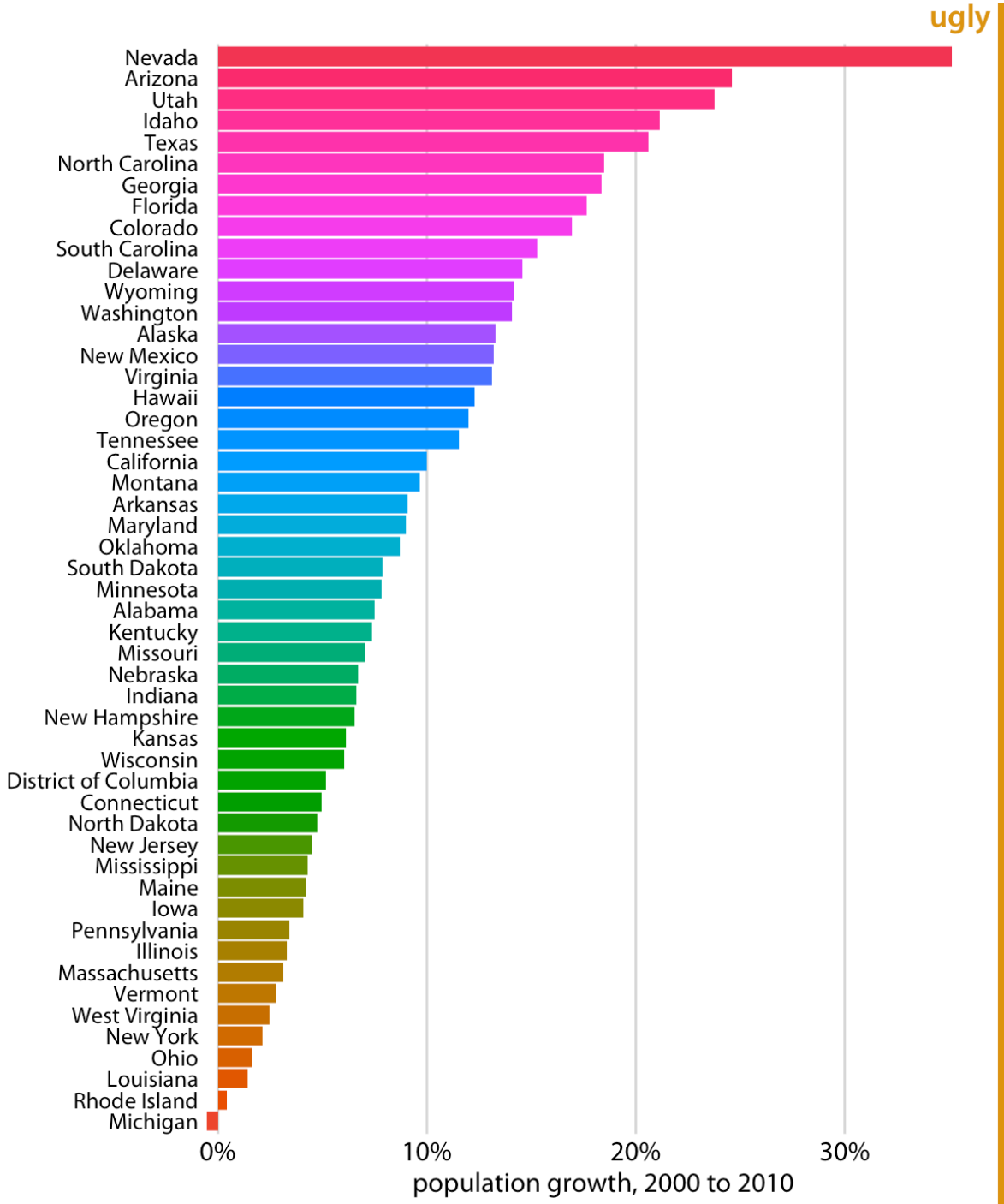
```
ggplot(data = mpg, aes(x = cty, y = hwy)) +
  geom_point(aes(color = class), size = 4) +
  scale_color_brewer(palette = "Set3") +
  theme_classic(base_size = 16)
```



Use colors to communicate something

Avoid using color for the sake of color

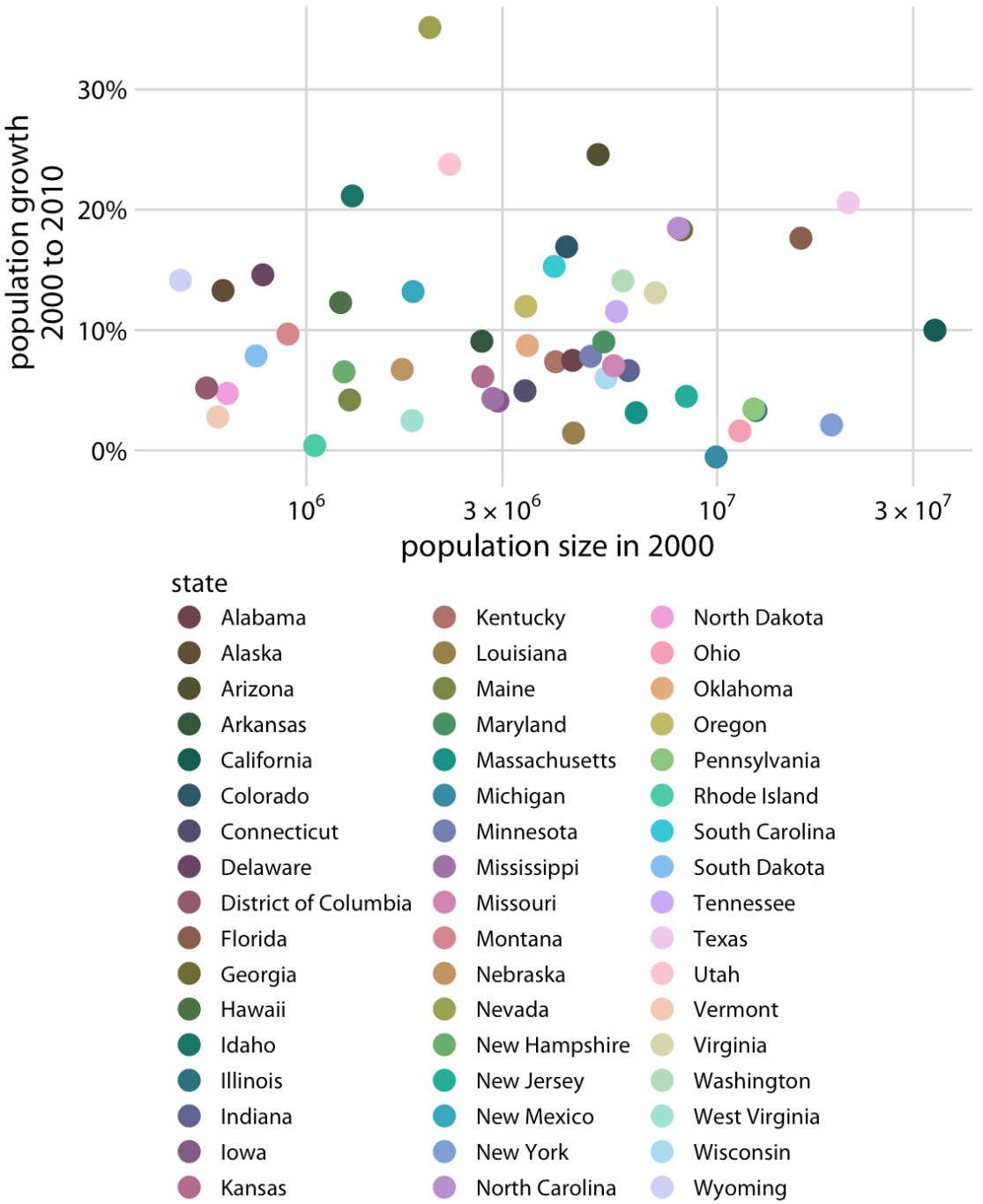
= chart junk!



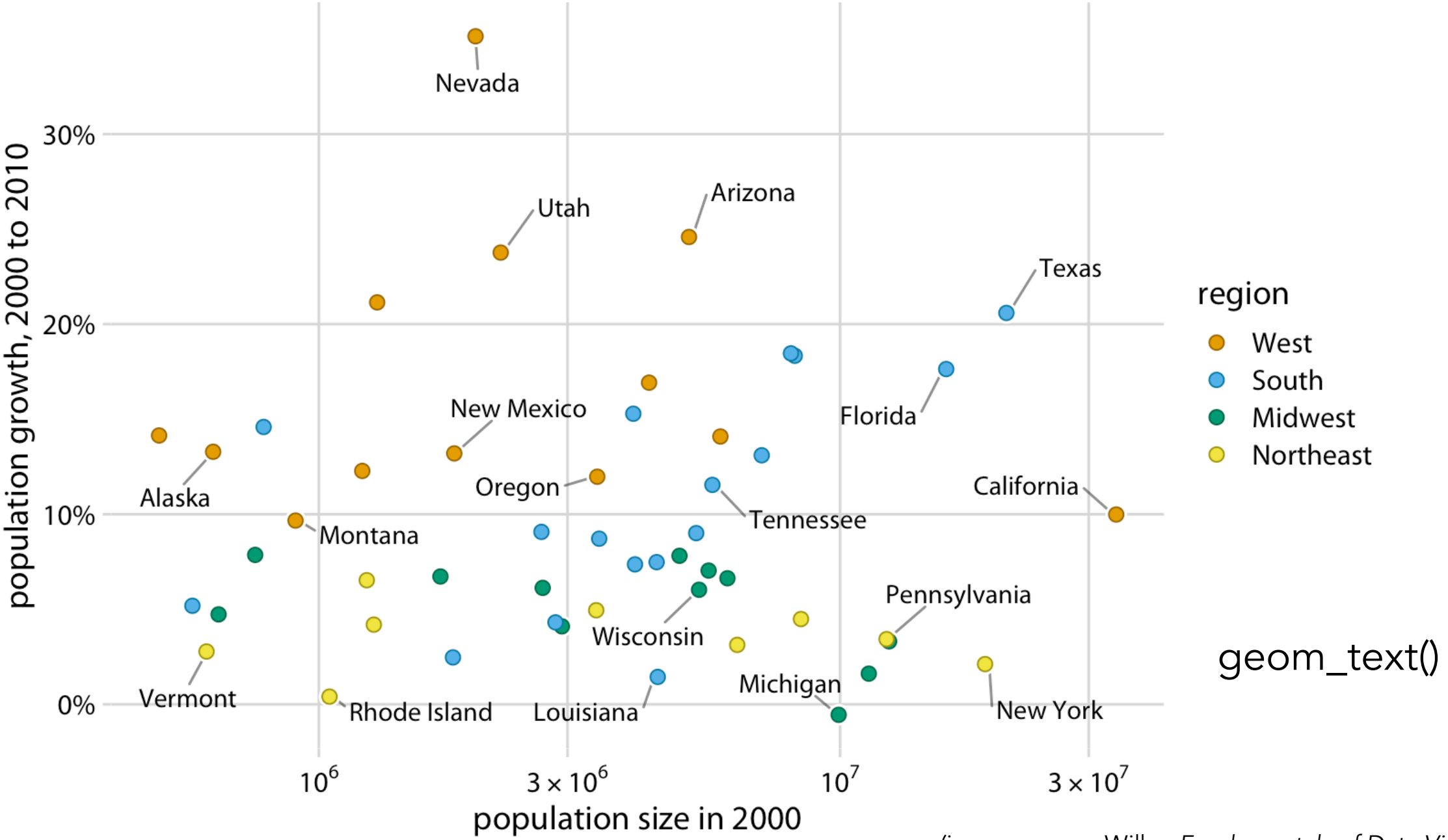
(image source: Wilke, *Fundamentals of Data Visualization*)

Avoid encoding too much/irrelevant information

Use direct labeling to distinguish >8 qualitative variables.



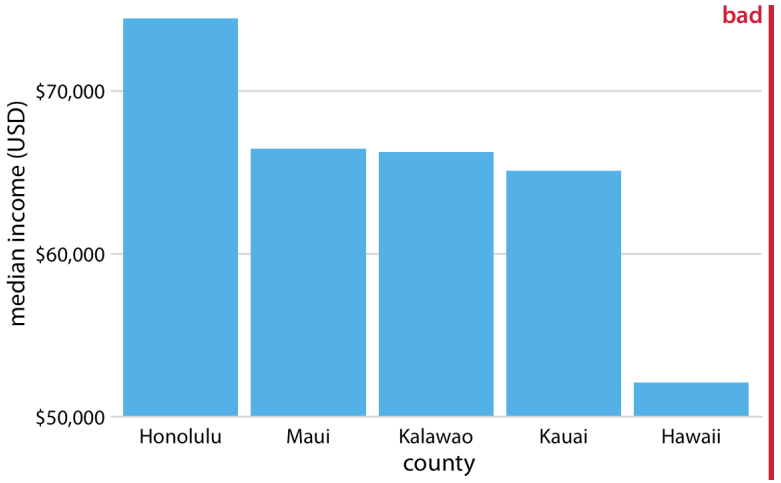
(image source: Wilke, *Fundamentals of Data Visualization*)

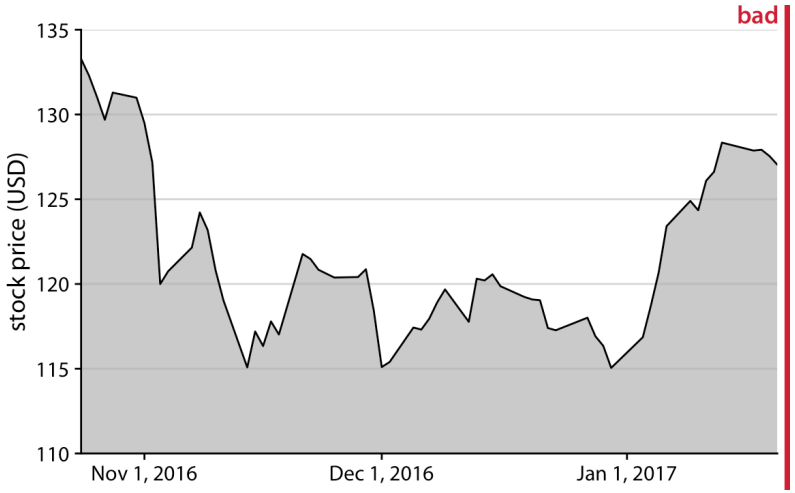


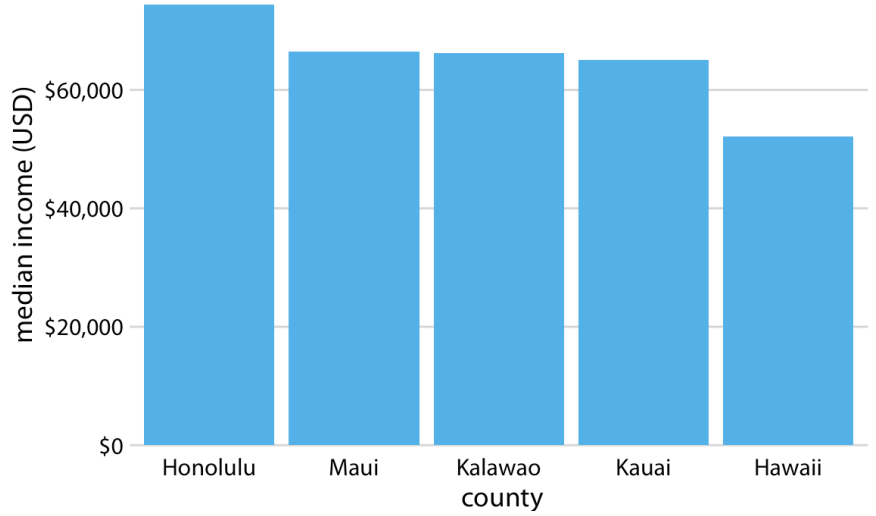
geom_text()

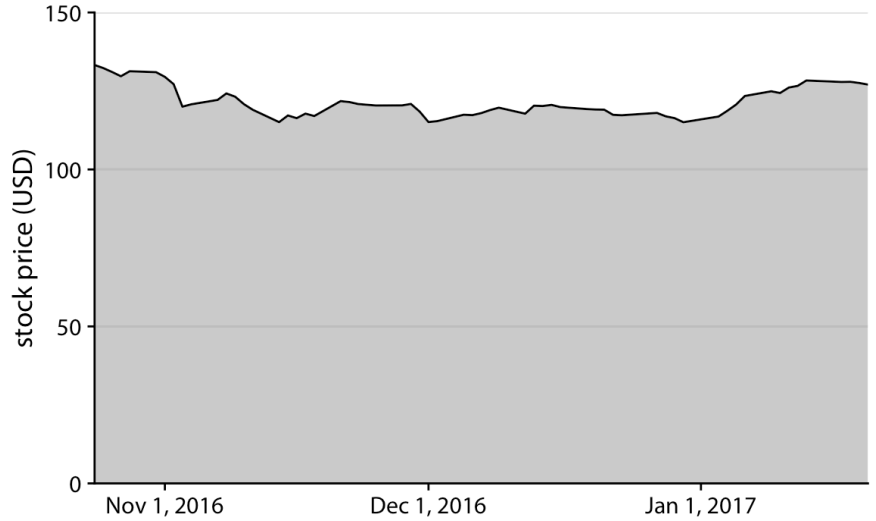
(image source: Wilke, *Fundamentals of Data Visualization*)

Axes should start at 0 (ggplot does this by default)



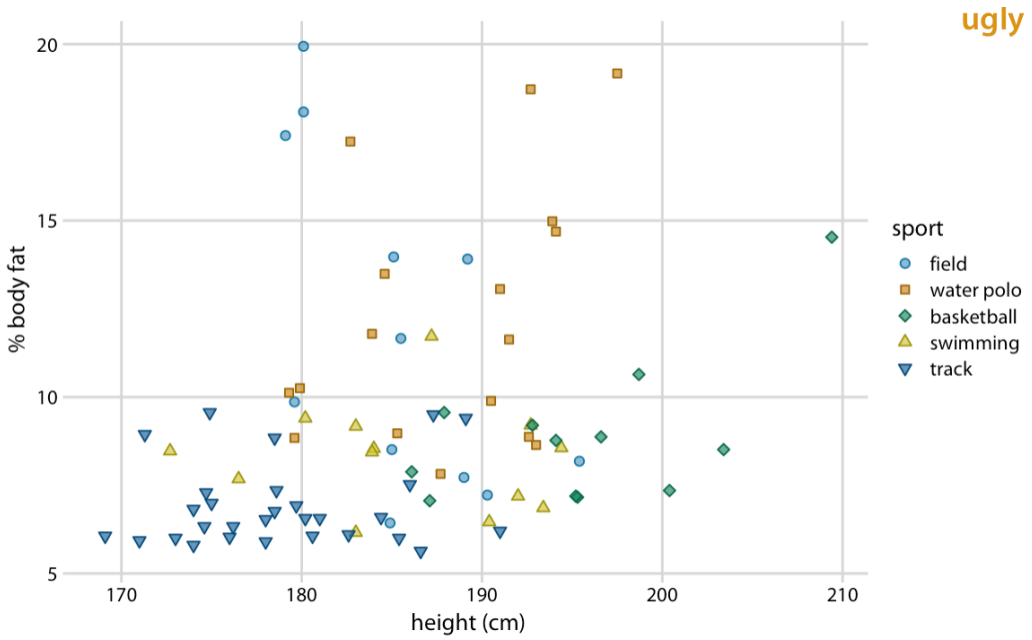


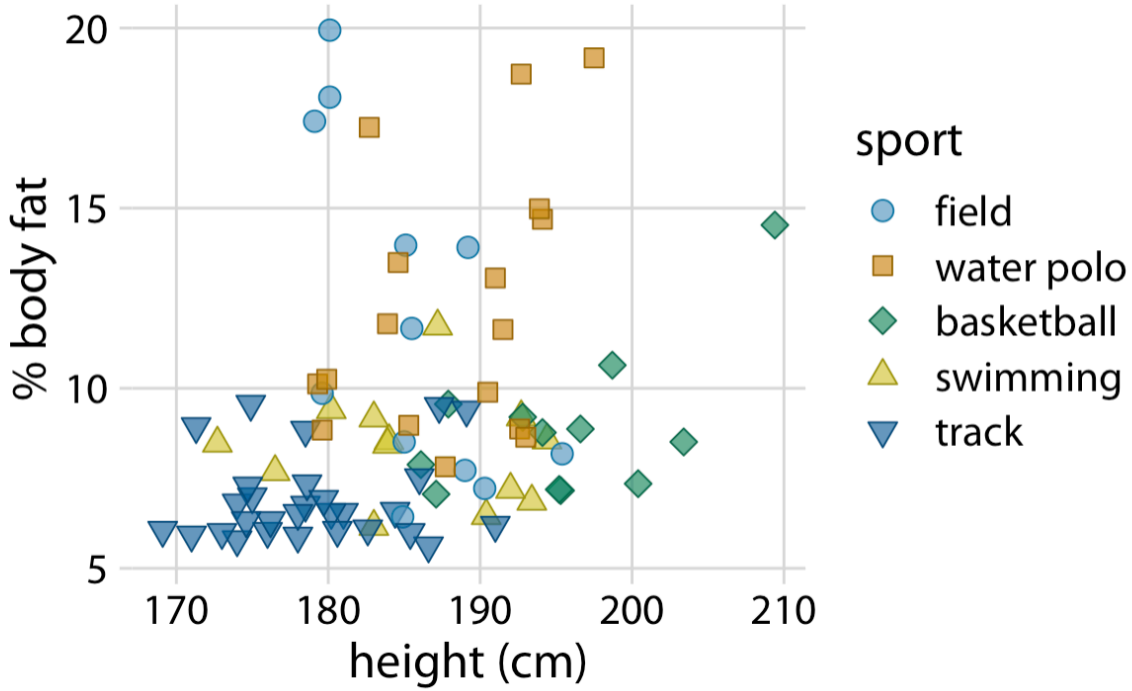




(image source: Wilke, *Fundamentals of Data Visualization*)

Make text readable by increasing font size





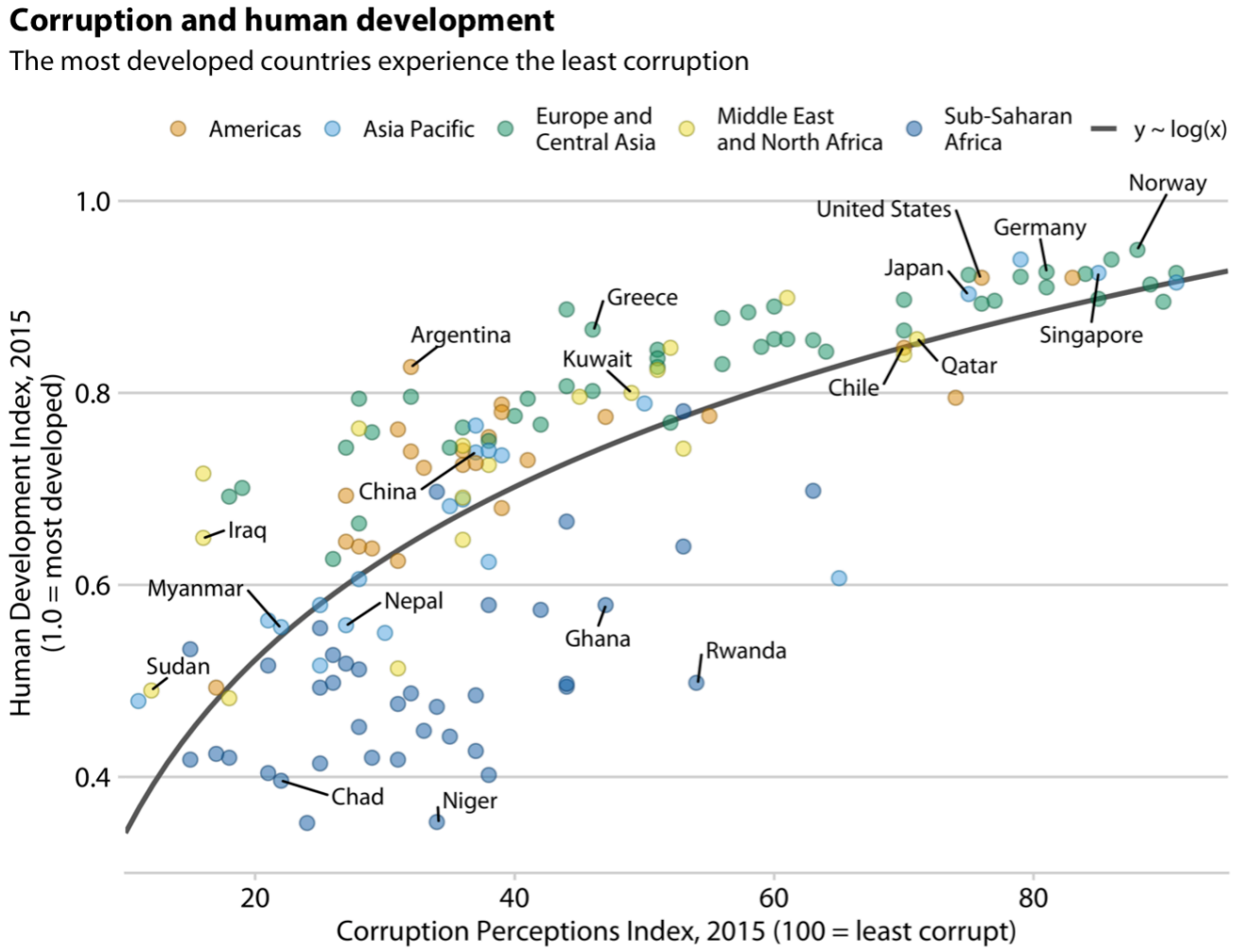
theme_classic(base_size = 16)

(image source: Wilke, *Fundamentals of Data Visualization*)

Add interpretable labels

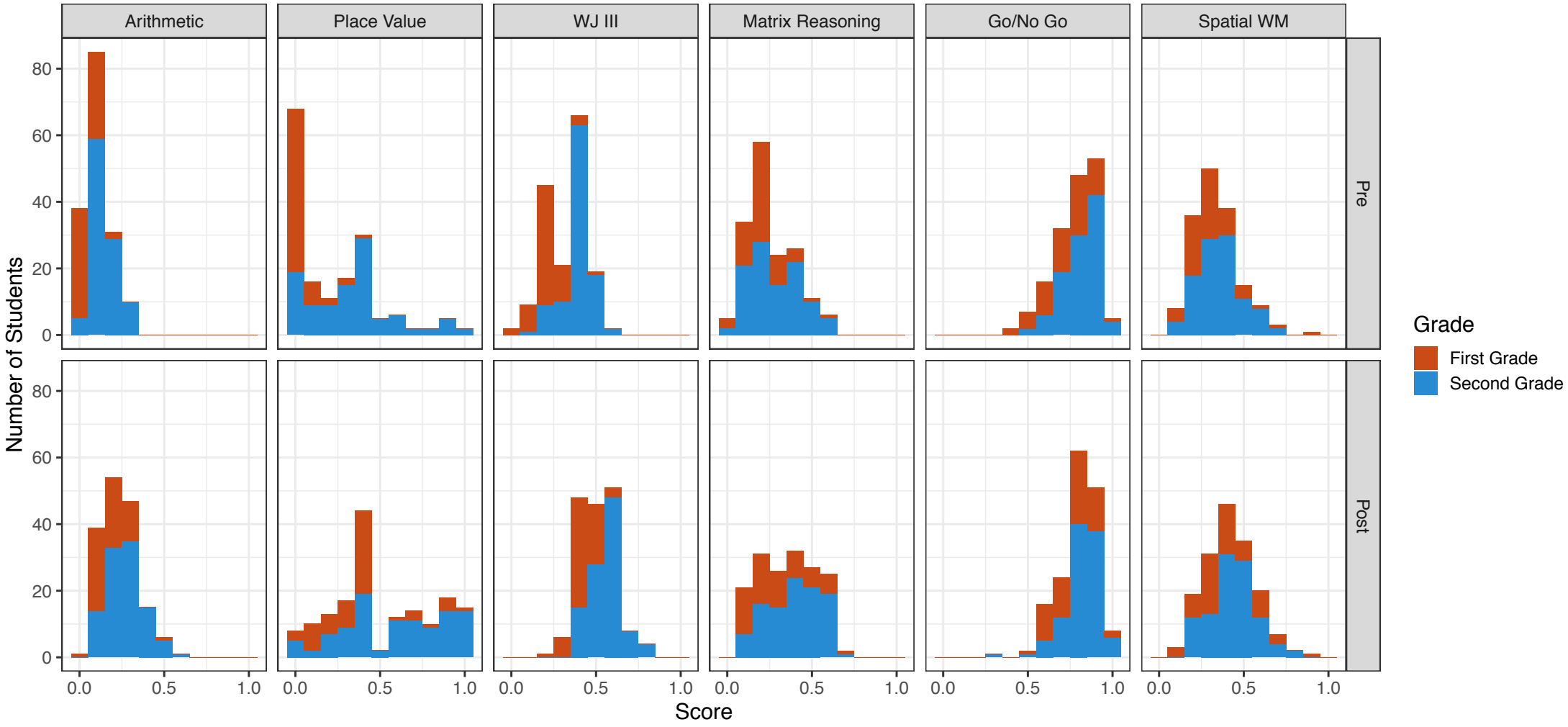
Title and axis labels

```
ggtitle(
    label = "Corruption and human
    development",
    subtitle = "The most developed
    countries experience the least
    corruption"
)


xlab("Corrpution Perceptions
Index, 2015 (100 = least
corrupt)")


ylab("Human Development Index,
2015 (1.0 = most developed)")
```



(image source: Wilke, *Fundamentals of Data Visualization*)

Use “facets” for large data

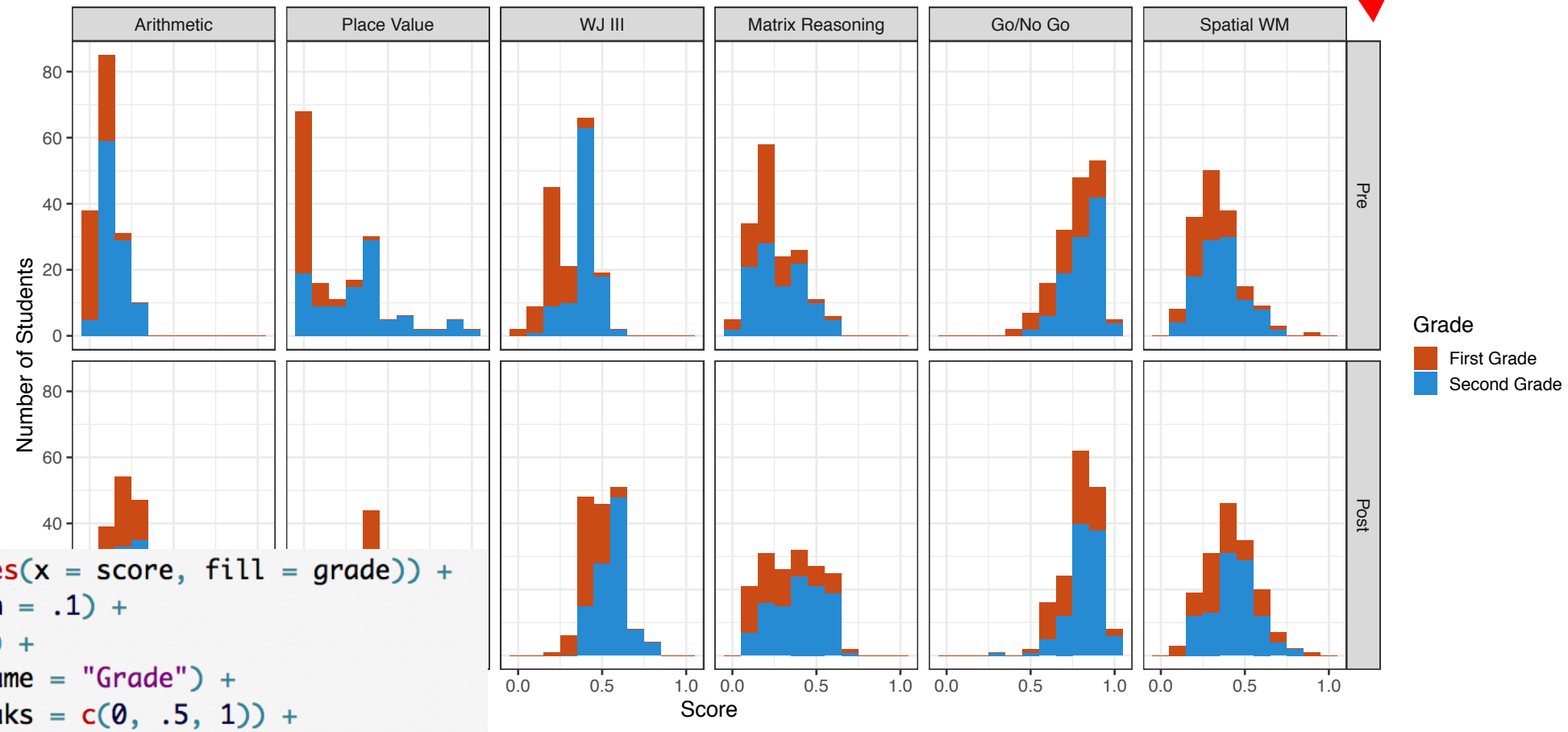


(Barner et al, 2017)

facet_grid()

Task →

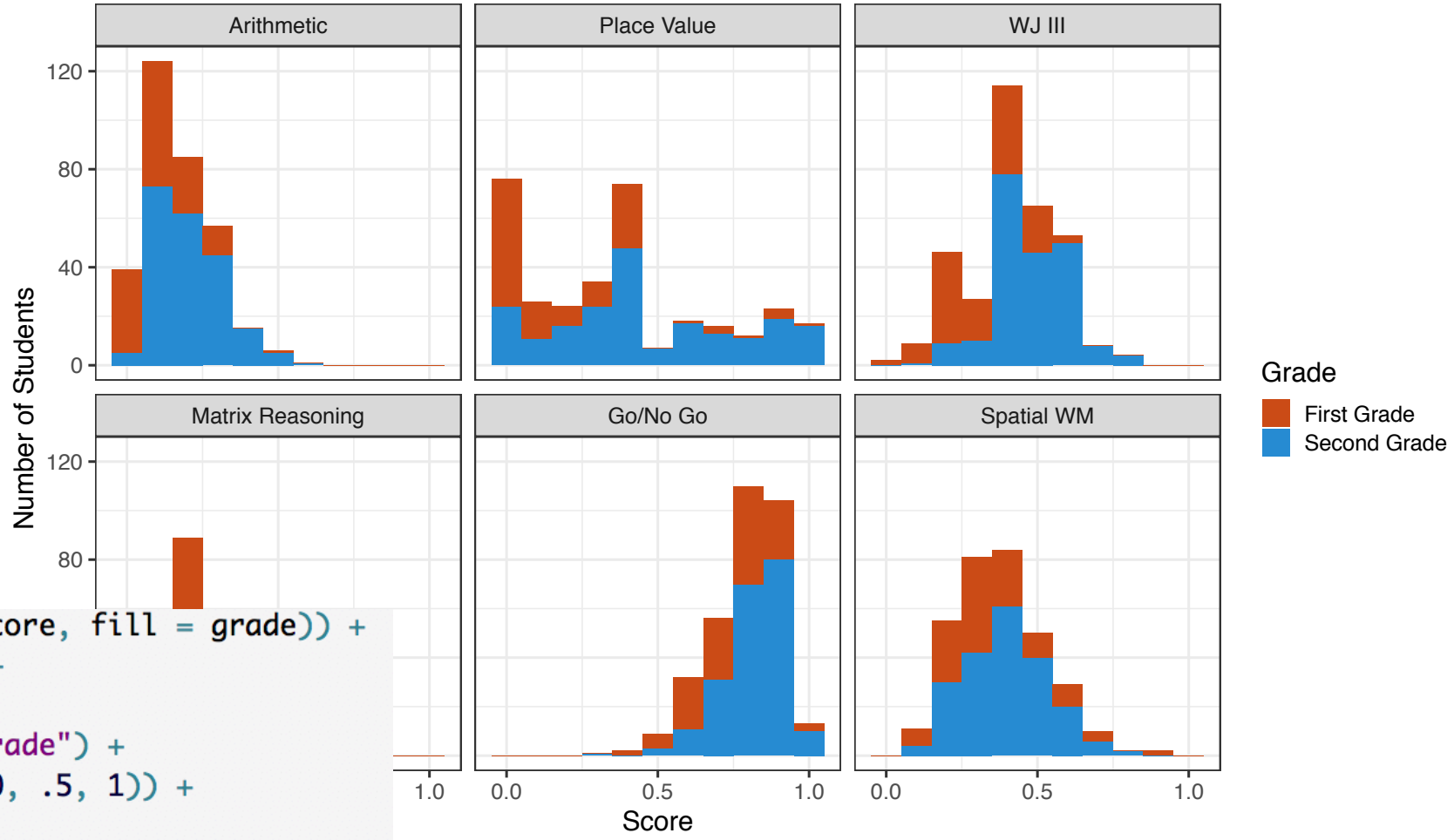
Time ↓




```
ggplot(student_scores, aes(x = score, fill = grade)) +
  geom_histogram(binwidth = .1) +
  facet_grid(time ~ task) +
  scale_fill_solarized(name = "Grade") +
  scale_x_continuous(breaks = c(0, .5, 1)) +
  xlab("Score") +
  ylab("Number of Students") +
  theme_bw(base_size = 16)
```

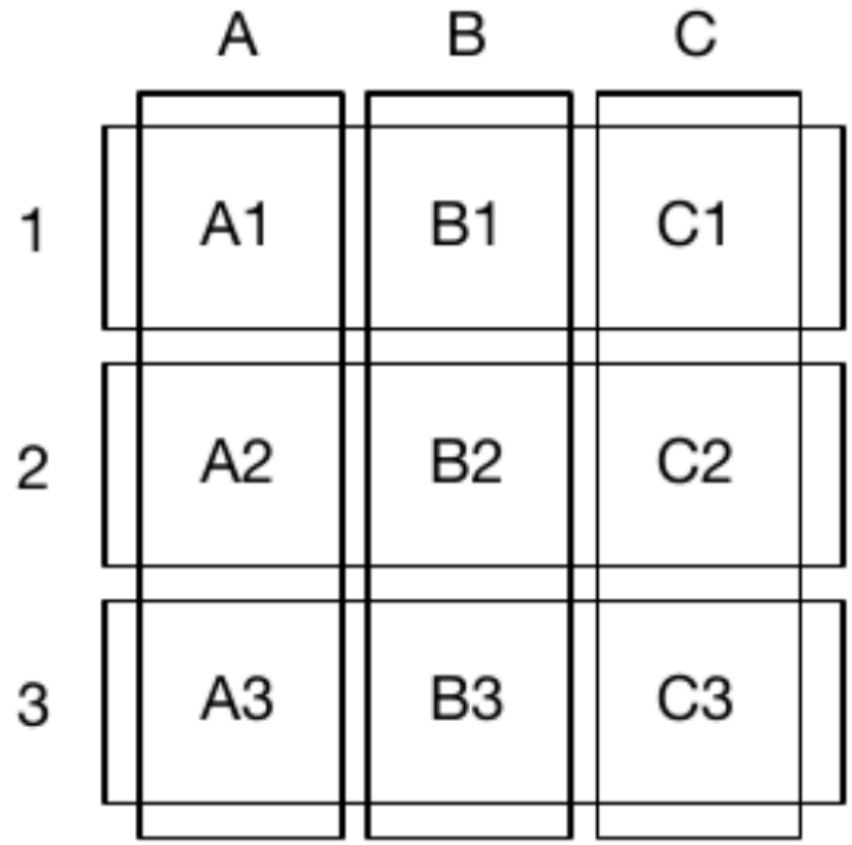
facet_wrap()

Task →

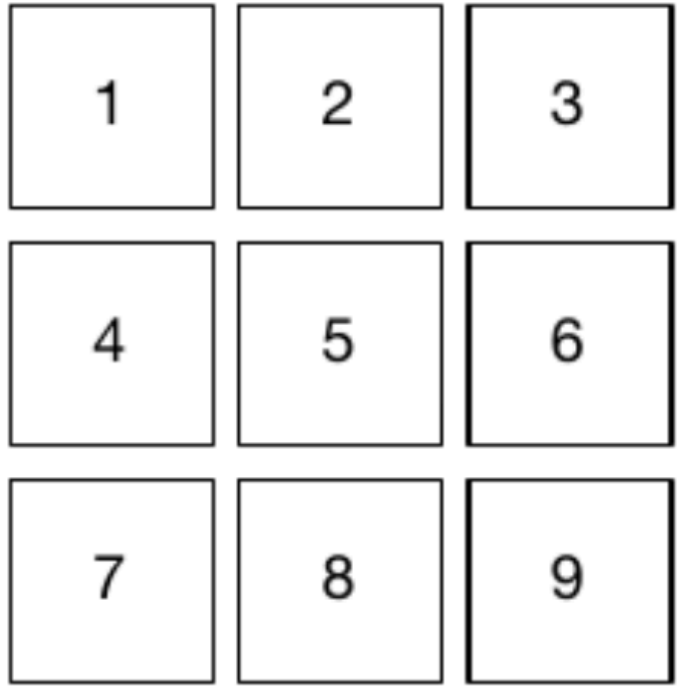



```
ggplot(student_scores, aes(x = score, fill = grade)) +
  geom_histogram(binwidth = .1) +
  facet_wrap(~ task) +
  scale_fill_solarized(name = "Grade") +
  scale_x_continuous(breaks = c(0, .5, 1)) +
  xlab("Score") +
  ylab("Number of Students") +
  theme_bw(base_size = 16)
```

facet_grid() vs. facet_wrap()



facet_grid

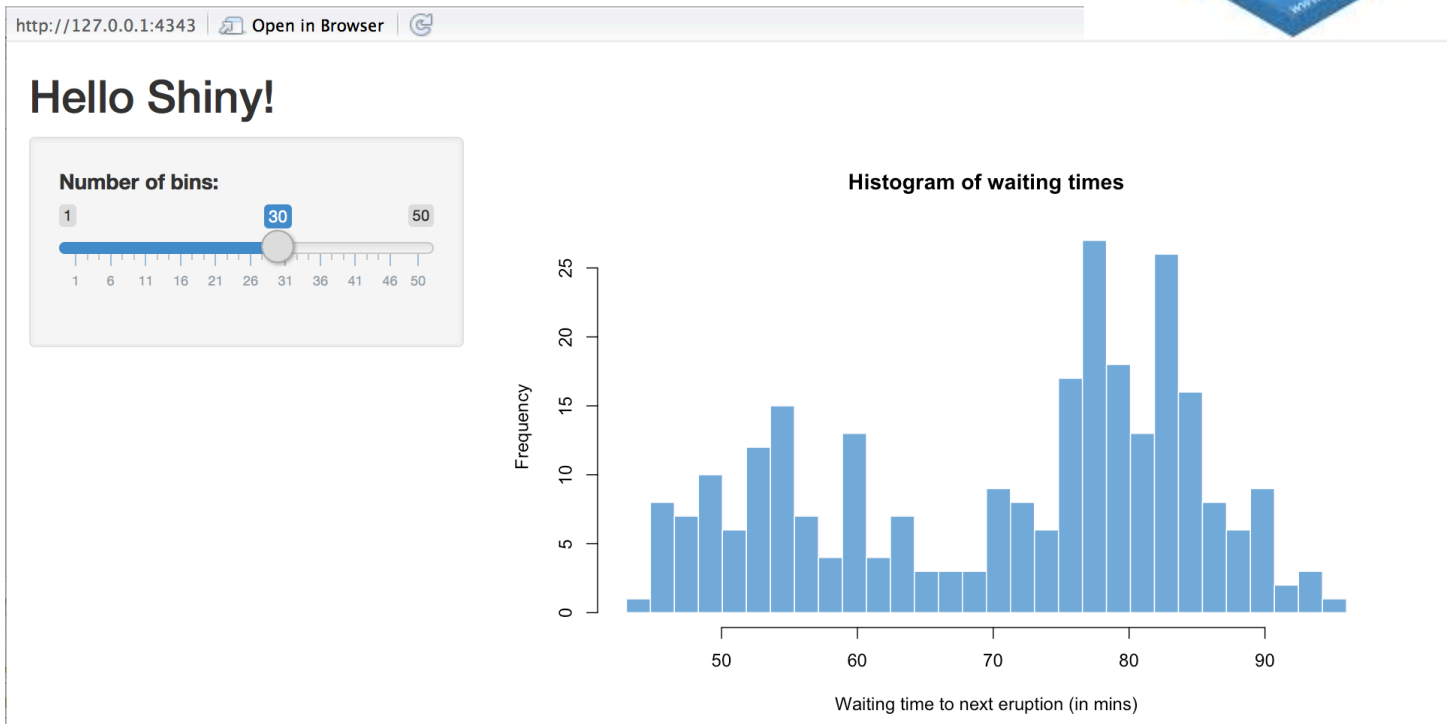


facet_wrap

https://ggplot2-book.org/

Interactive visualizations

- Campaign Donors: https://www.nytimes.com/interactive/2020/02/01/us/politics/democratic-presidential-campaign-donors.html
- Hair styles: https://pudding.cool/2019/11/big-hair/
- Built with Shiny:
  - https://mlewis.shinyapps.io/lnhBrowser/
  - https://mlewis.shinyapps.io/SI_KIDBOOK/



Shiny Tutorial: https://shiny.rstudio.com/tutorial

Guidelines for implementing these principles in ggplot

- Choose a geom that highlights the comparison you want to make, keeping in mind human perception
- Use colors to communicate something
- Axes should start at 0
- Make text readable by increasing font size
- Add interpretable labels
- Use “facets” for large data

Summary of Principles of Visualization

- Think of plotting as communication – you want to maximize the likelihood that your audience gets your message
- ggplot is powerful – you have lots of control!
- ggplot defaults are pretty good, but often you need to make choices/tweak things for your specific plot

Next Time

- Wrapping up some loose tidyverse ends

Office Hours:

Roderick 3:30-5:30pm **today**);

Molly 2:45-4:45pm Wednesday (Porter 208H)

Acknowledgements

Images on slides 13-35 adapted from
http://socviz.co/lookatdata.html (Healy, 2018) and
https://serialmentor.com/dataviz/ (Wilke)