

Principles of Visualization

3 February 2020

Modern Research Methods

Last Time: Plotting with ggplot

What data do you want to plot?



1. Tidy Data

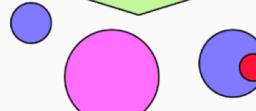
```
p <- ggplot(data = gapminder, ...)
```

gdp	lifexp	pop	continent
340	65	31	Euro
227	51	200	Amer
909	81	80	Euro
126	40	20	Asia

2. Mapping

```
p <- ggplot(data = gapminder,  
             mapping = aes(x = gdp,  
                            y = lifexp, size = pop,  
                            color = continent))
```

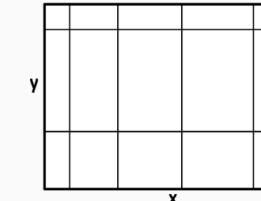
3. Geom



```
p + geom_point()
```

4. Co-ordinates & Scales

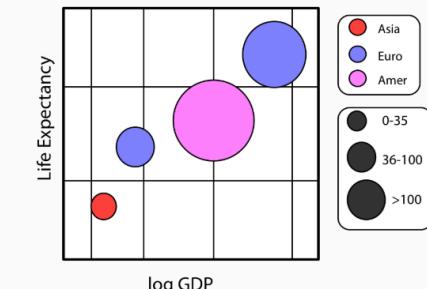
```
p + coord_cartesian() +  
    scale_x_log10()
```



5. Labels & Guides

```
p + labs(x = "log GDP",  
          y = "Life Expectancy",  
          title = "A Gapminder Plot")
```

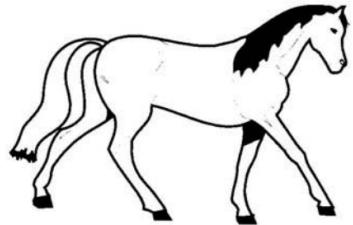
A Gapminder Plot



Last Time: Plotting with ggplot

	word	mean_complexity_rating	complexity_type	word_class	n_chars	n_syllables	n_phonemes	concreteness	familiarity	imageability	frequency
1	a	1.375	low	closed	1	1	1	201	632	217	410.55
2	about	3.9375	high	closed	5	2	4	227	593	225	193.97
3	ache	2.88235294117647	low	open	4	1	3	443	523	443	8.23
4	affirmation	5.83333333333333	high	open	11	4	8	242	332	313	4.25
5	after	2.23076923076923	low	closed	5	2	4	242	575	217	93.85
6	age	2.17647058823529	low	open	3	1	2	390	582	468	36.83
7	aid	1.625	low	open	3	1	2	372	536	413	17.31
8	all	2.66666666666667	low	closed	3	1	2	267	582	332	225.97
9	ally	3.23076923076923	low	open	4	2	3	485	410	453	
10	alphabet	3.57142857142857	high	open	8	3	7	449	493	499	8.23

"The linguistic sign is arbitrary" – Saussure (1916)



horse

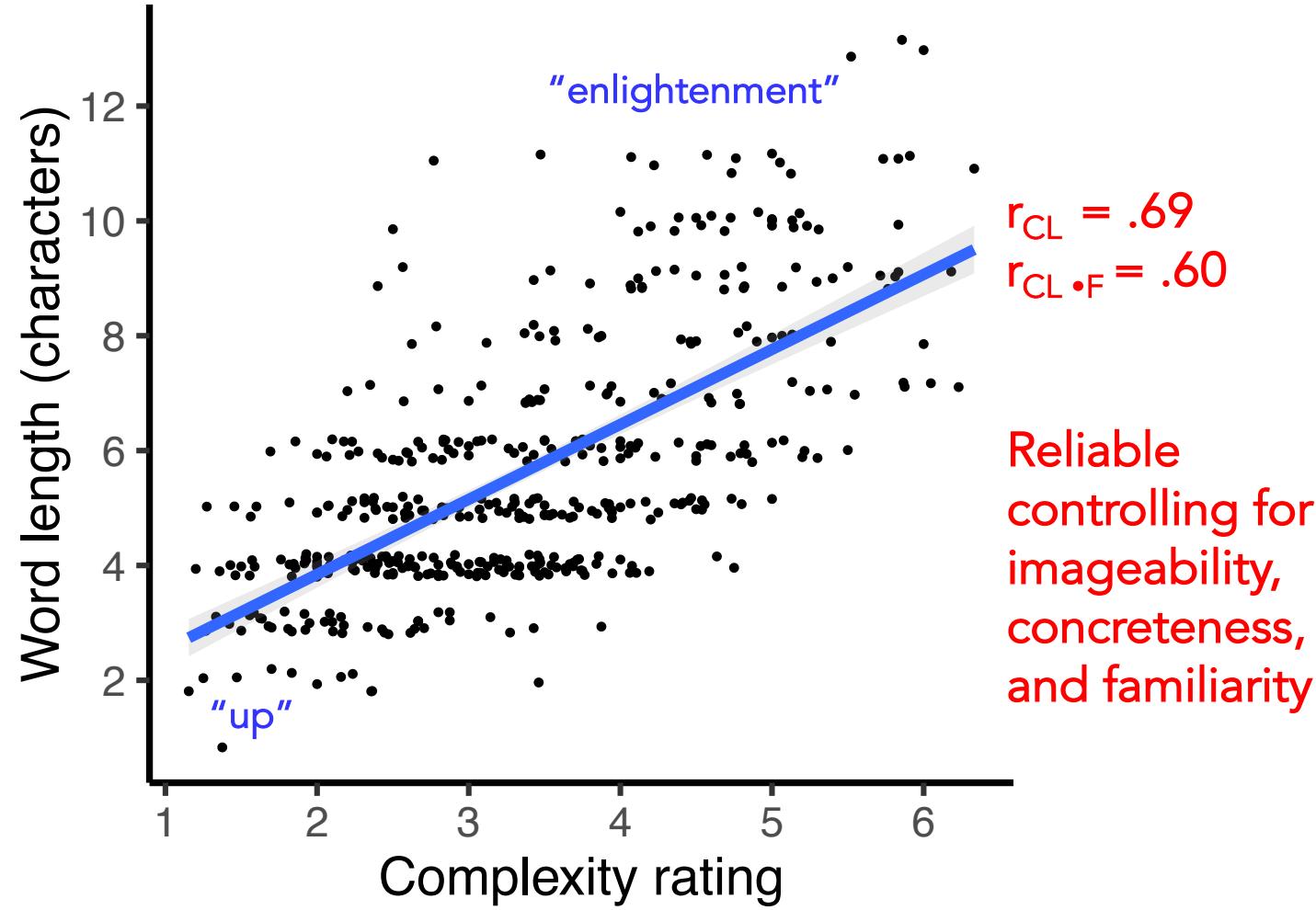
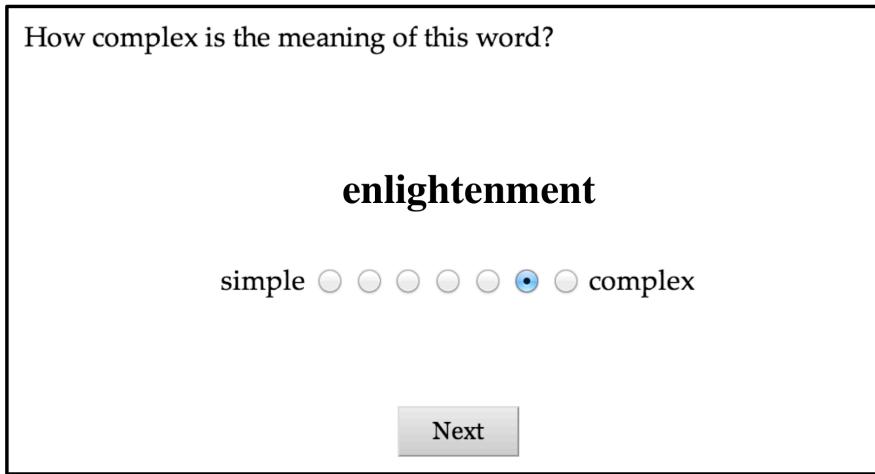
kalë	حصان	اَجِ	at	zaldi	коń	شَاحْ	konj	кон	cavall	kabayo	马	馬	konj
kůň	hest	paard	čevalo		hobune		kabayo		hevonen		cheval		cabalo
Pferd	ପ୍ରାଣୀ	chwal	doki	ଚାଲ	ଘୋଡ଼ା	ନେସ	ଲୋ	hestur		anyinya		kuda	
capall	cavallo			馬	jaran	କୁଦୁର୍ବେ	ସଂଃ	ମଲ୍ଲା	equo	zirgs	arklys	koń	kuda
hoiho	ଘୋଡ଼ା	ଅଦ୍ୟ	ଘୋଡ଼	hest	اَسَب	କୋନ୍ତିର୍ବେ	କାଵାଲୋ	ଛାତା	cal	ଲୋଶାଦ୍ୟ	କୋନ୍ତିର୍ବେ	କୋନ୍ତିର୍ବେ	konj
konj	faras	caballo		farasi	häst	କୁତୀତେ		ରୁରମୁ		ମାଣ୍ଡା	କିନ୍ବ	ଗୁହ୍ରା	ceffyl

However, limits to arbitrariness

(Köhler, 1929; Maurer, et al., 2006; Ramachandran & Hubbard, 2001; Farmer, Christiansen, & Monaghan, 2006; Zipf, 1936; Piantadosi, Tily, & Gibson, 2011)

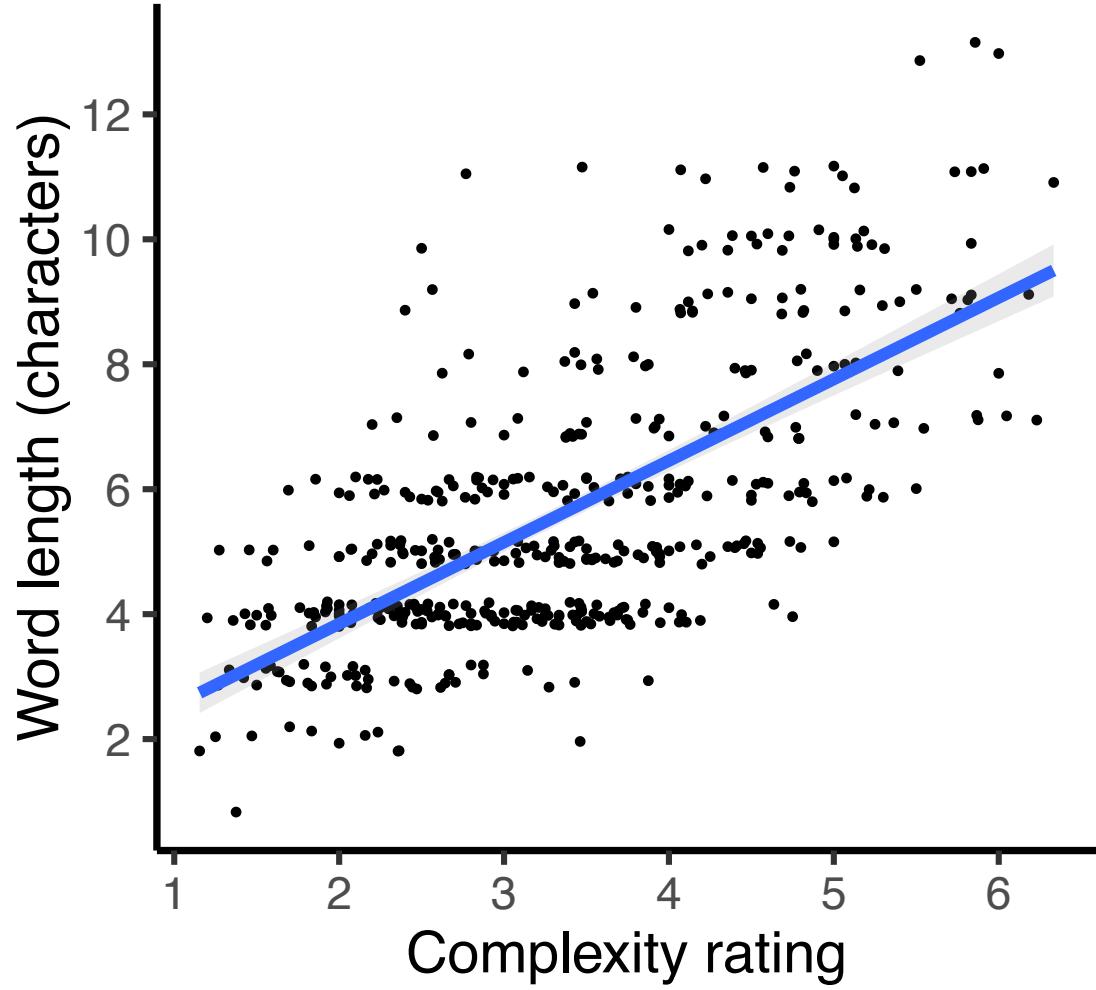
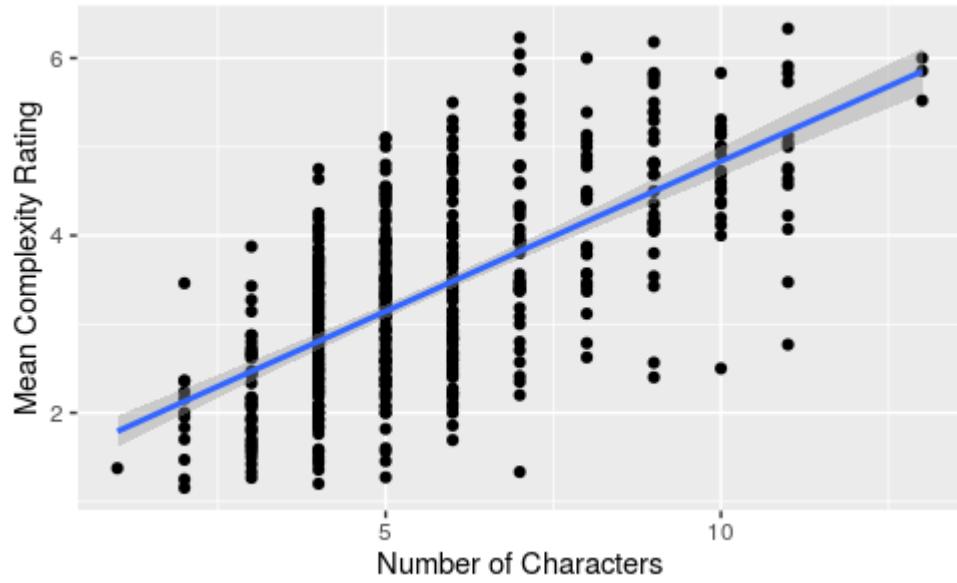
Complexity Bias in natural language

Participants rated 499 words
for conceptual complexity
(Lewis & Frank, 2016)

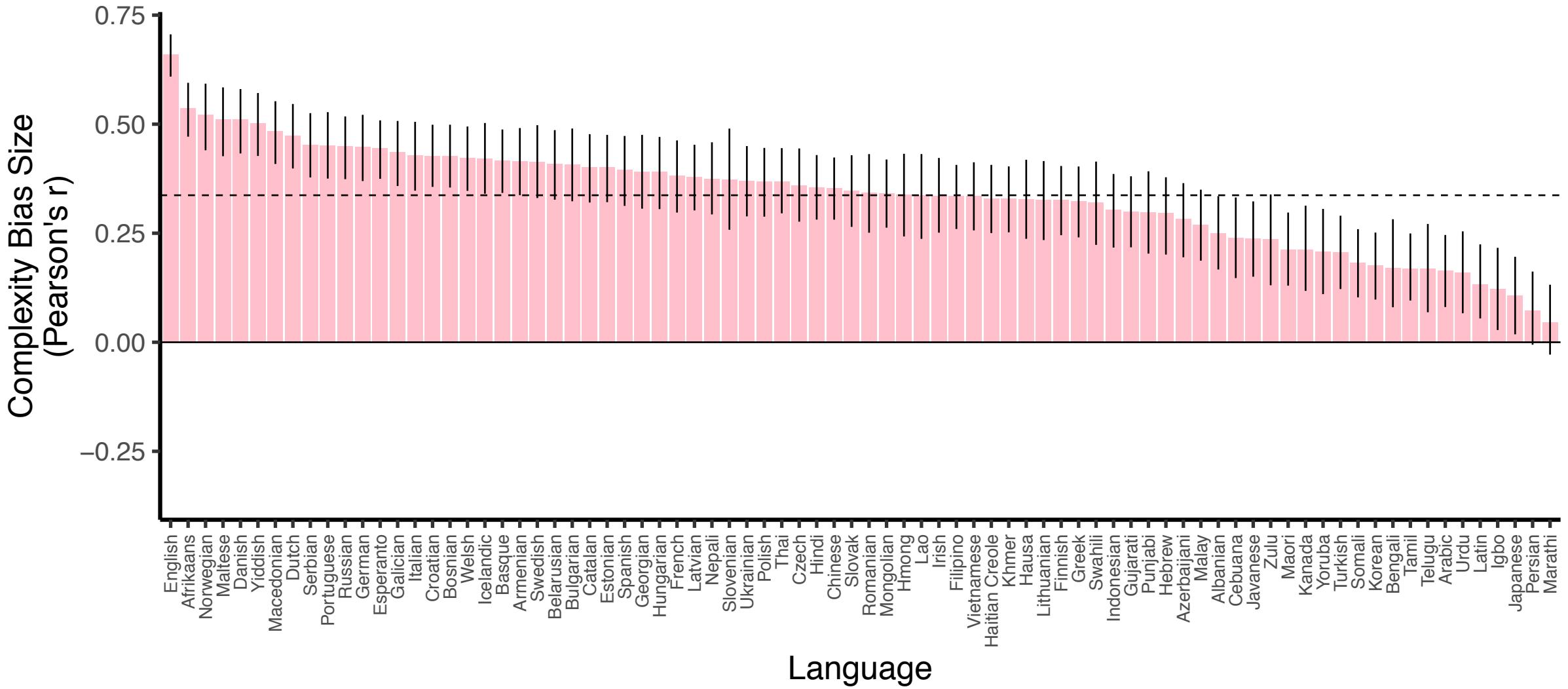


Complexity Bias in natural language

Plot by Nicole Hsing and
Nicole Casey

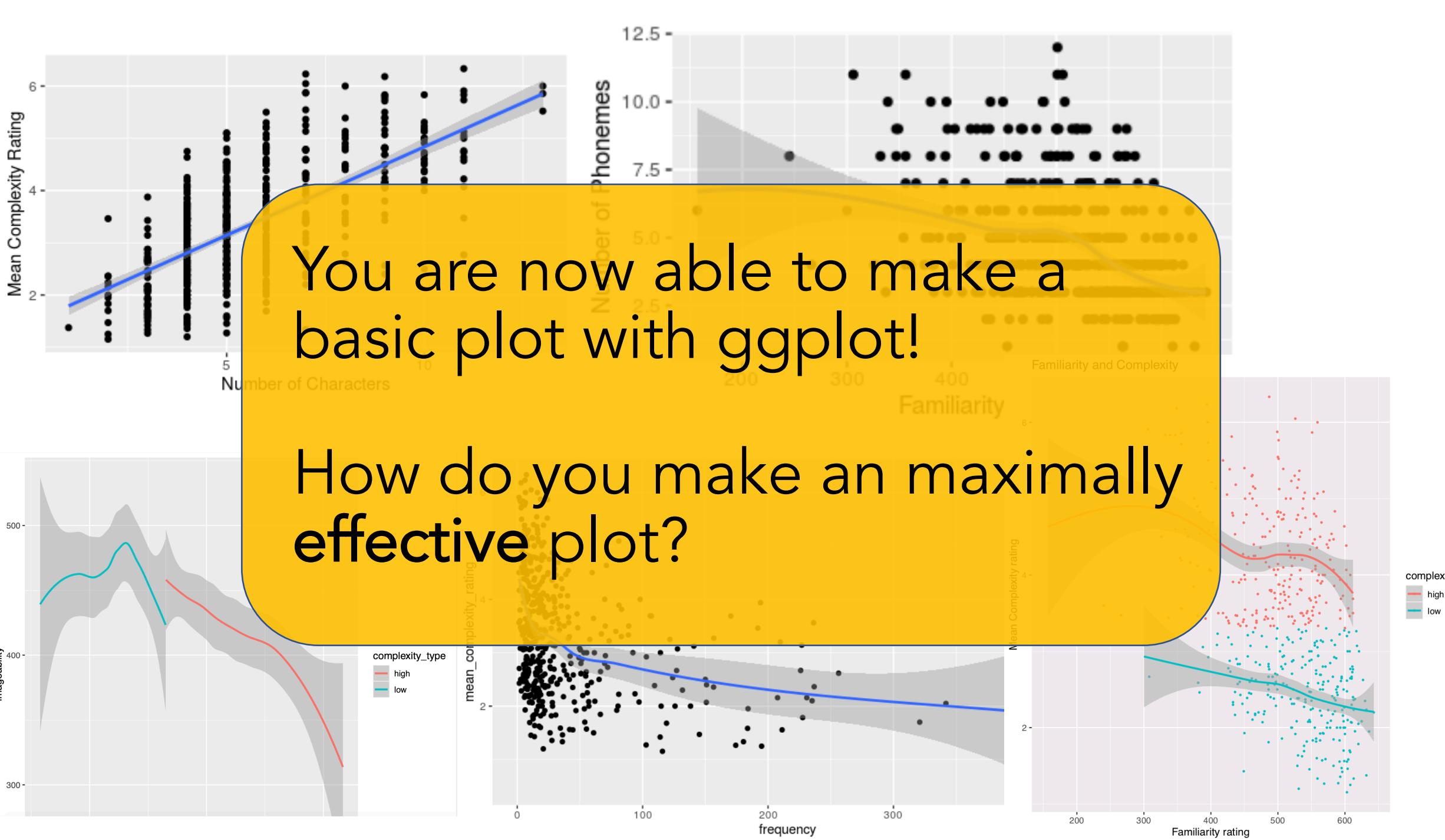


Complexity bias cross-linguistically



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

How do you make an maximally effective plot?



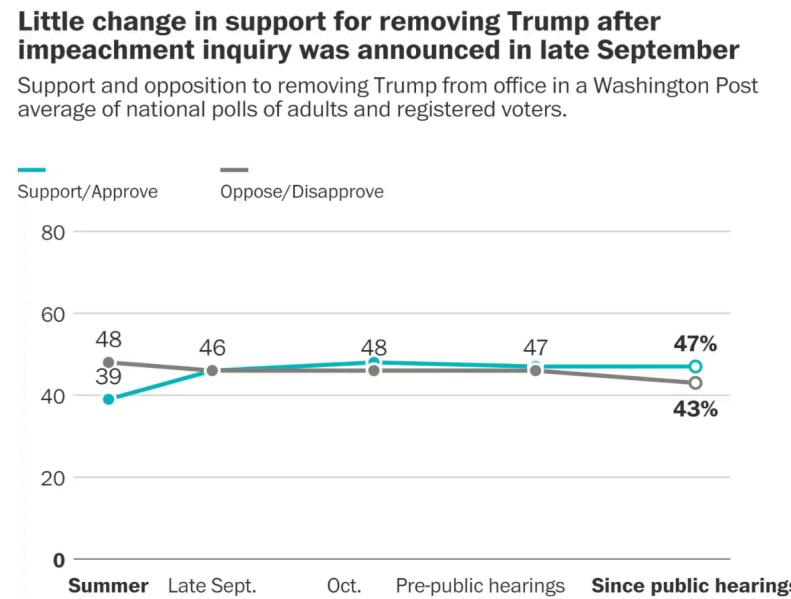
Principles of Data Visualization

- Visualization is **communication**
 - There is no list of rules for what makes a good visualization
 - Design depends on the message you want to communicate
 - And, 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.

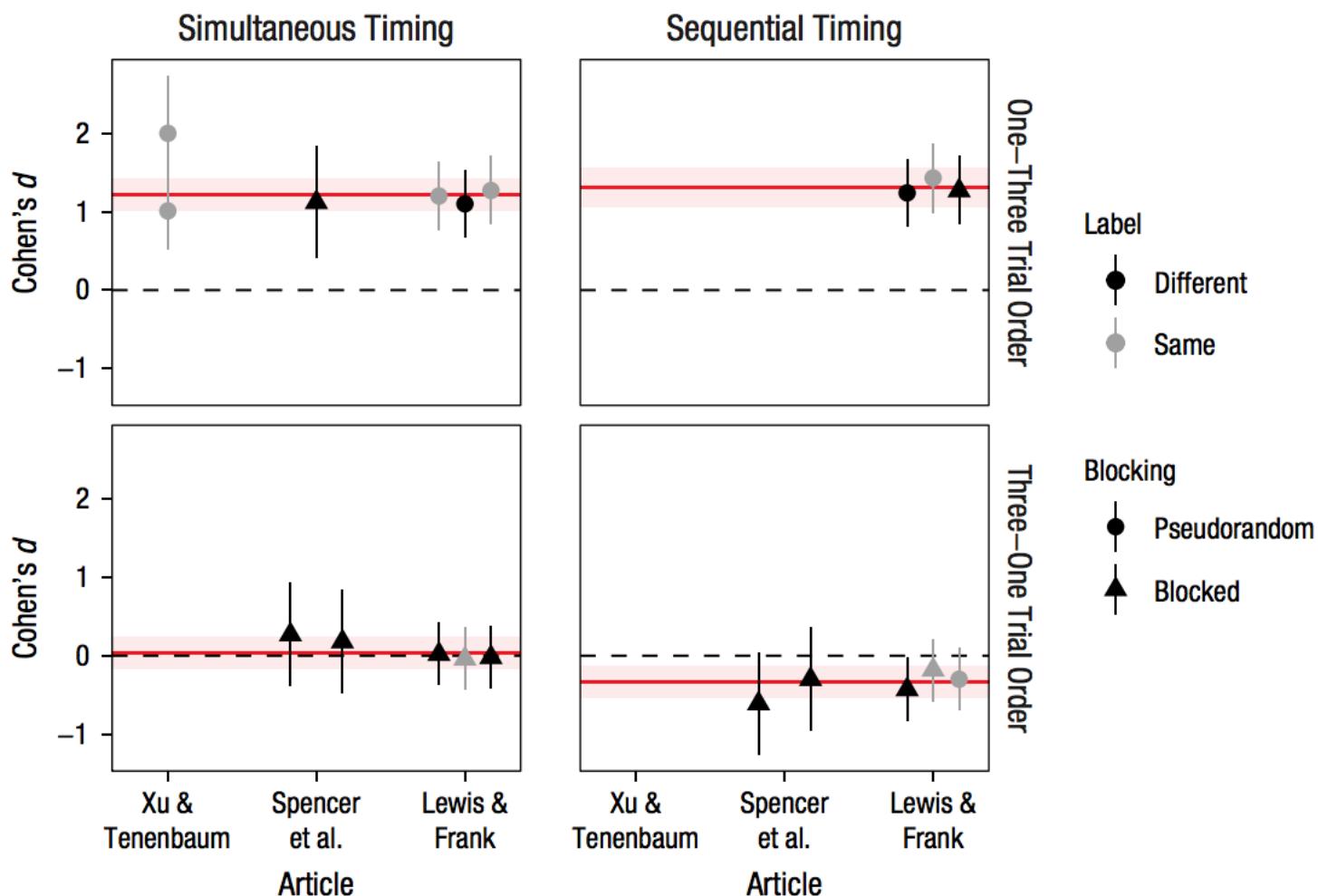


Kinds Audiences

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



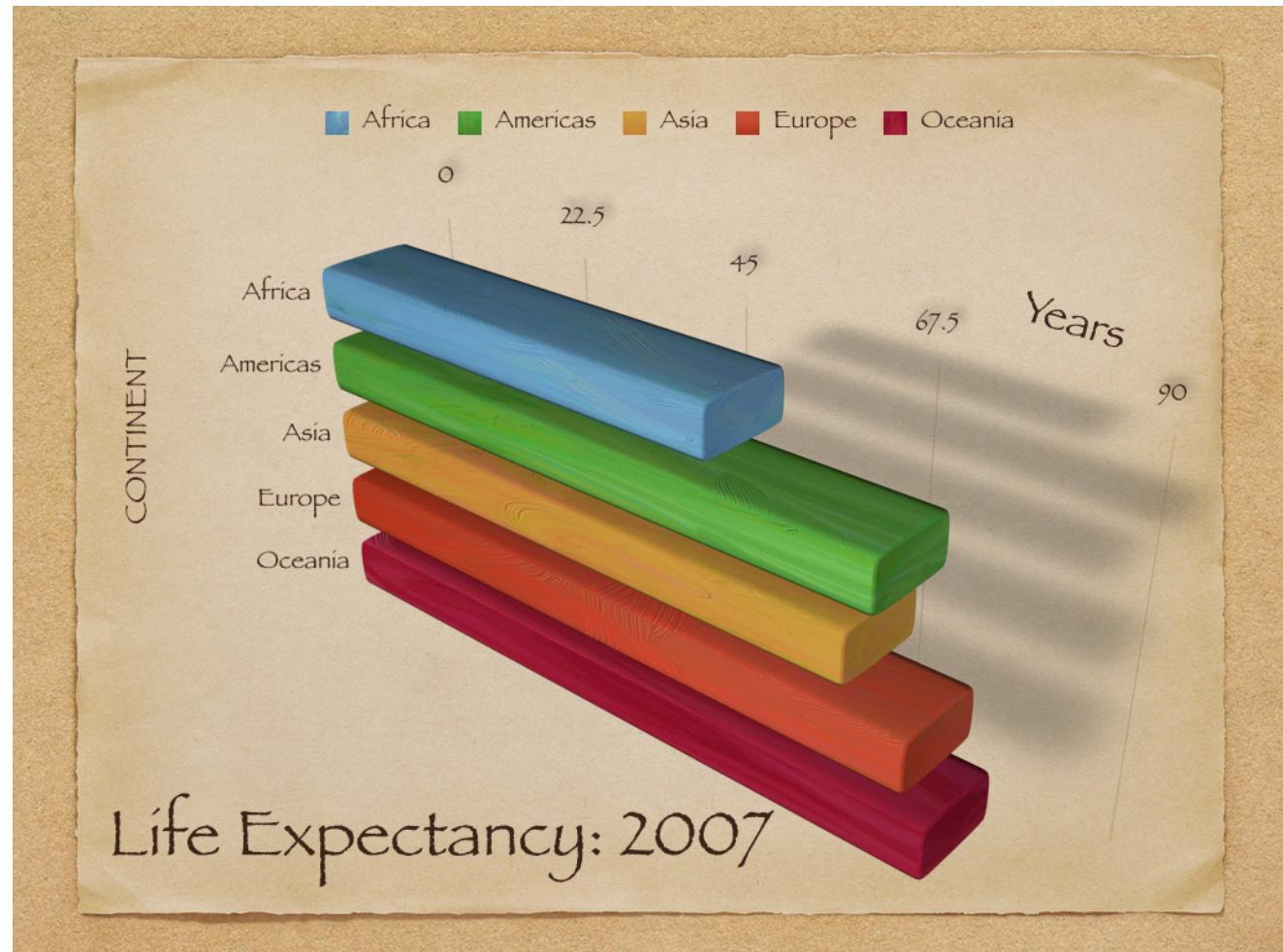
Source: Washington Post polling average of national polls that include language about removing Trump from office.



(Lewis & Frank, 2018)

Some Guidelines

1. Get rid of “chart junk” – maximize info to ink ratio

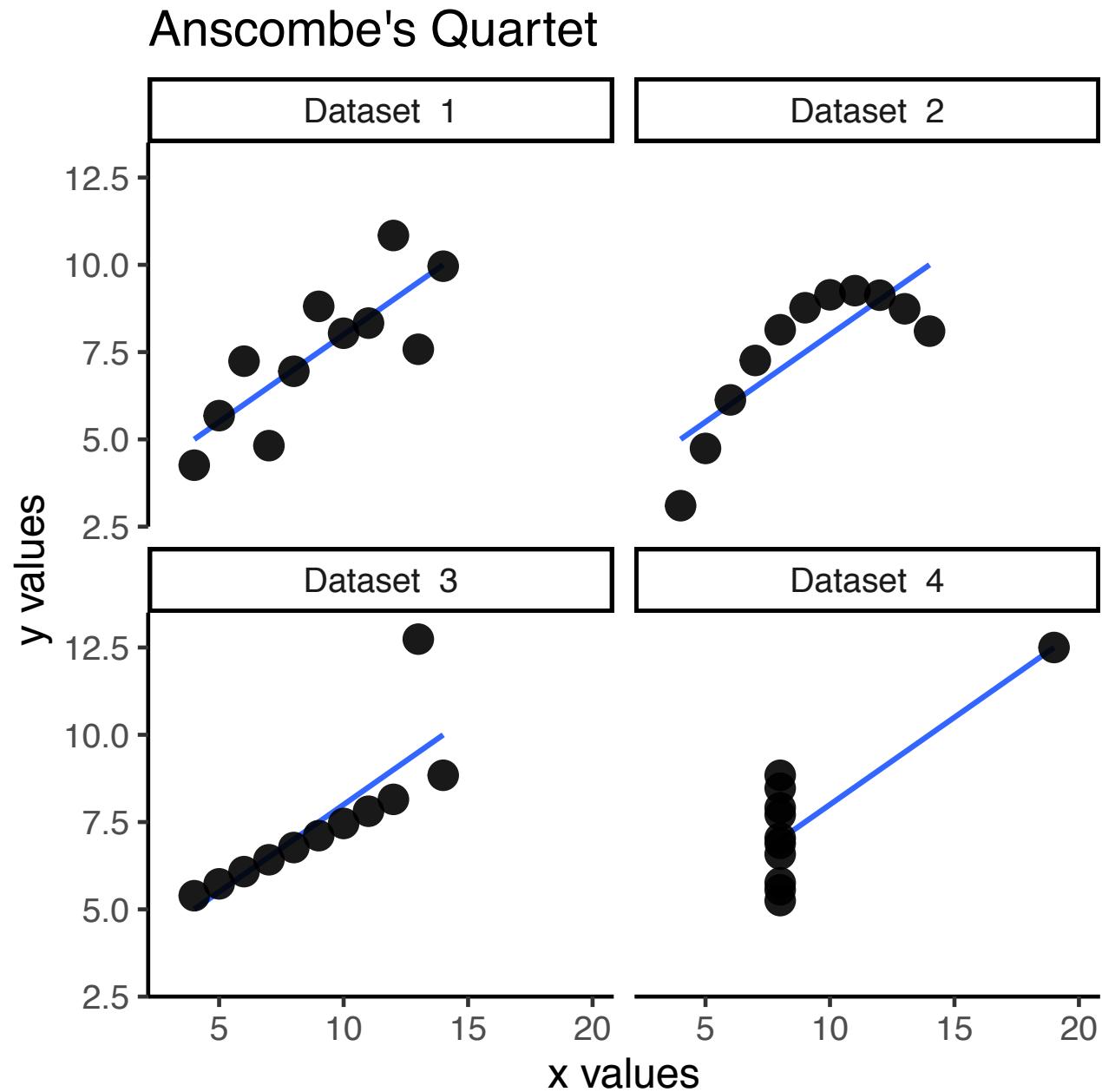


(image source: Healy, 2018)

Some Guidelines

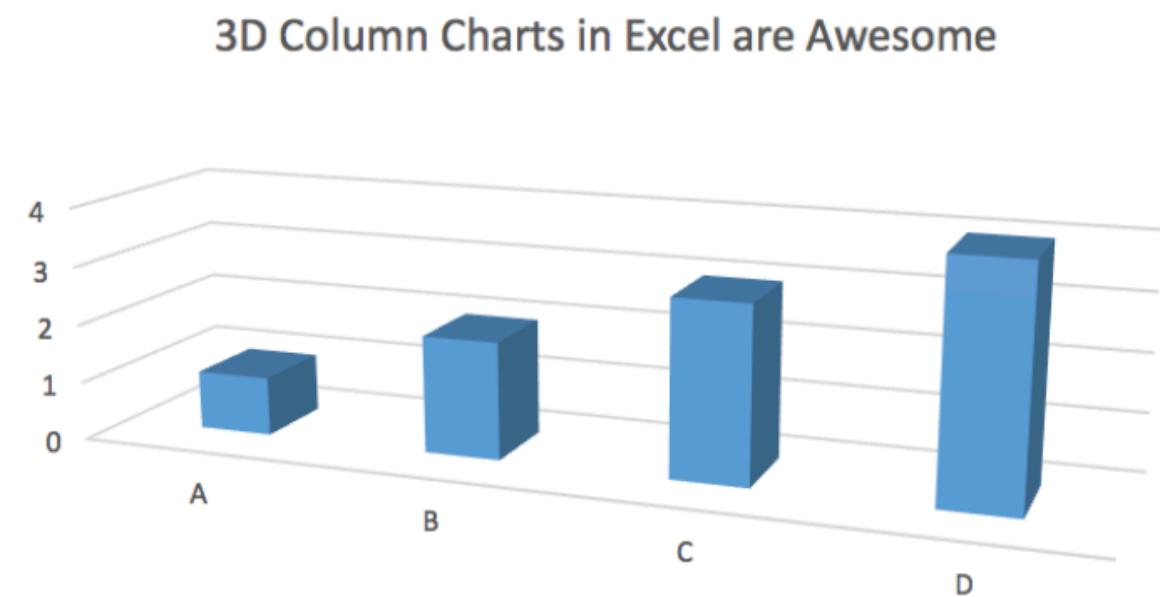
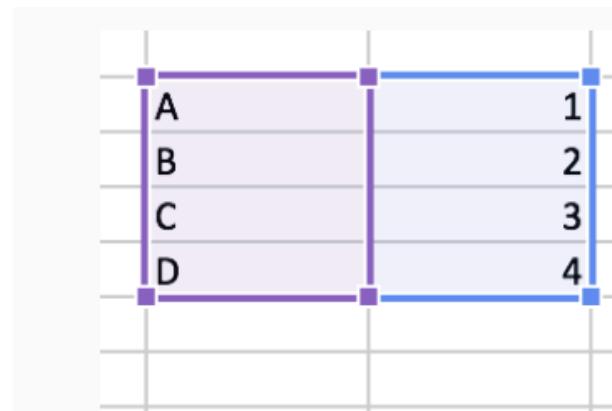
1. Get rid of “chart junk” – maximize info to ink ratio
2. Don’t be deceptive – show the raw data when possible

Anscombe's quartet									
I		II		III		IV			
x	y	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		



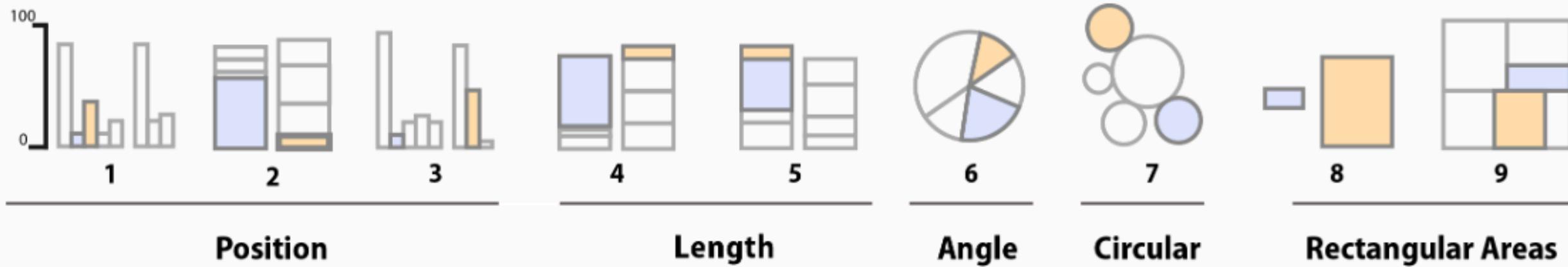
Some Guidelines

1. Get rid of “chart junk” – maximize info to ink ratio
2. Don’t be deceptive – show the raw data when possible
3. Think about human perception.



(image source: Healy, 2018)

Some things are easier to perceive than others!



Position

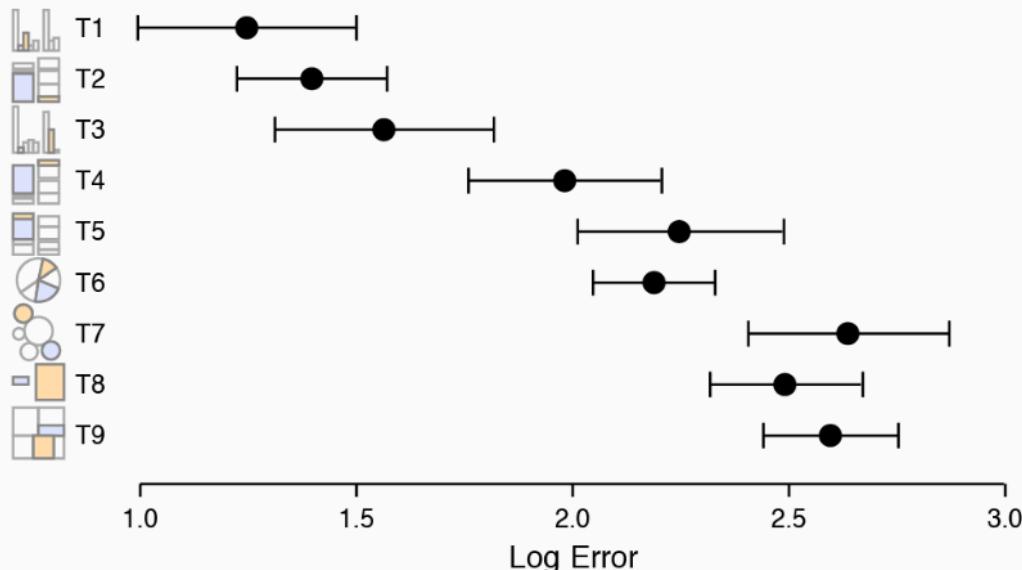
Length

Angle

Circular Area

Rectangular Areas

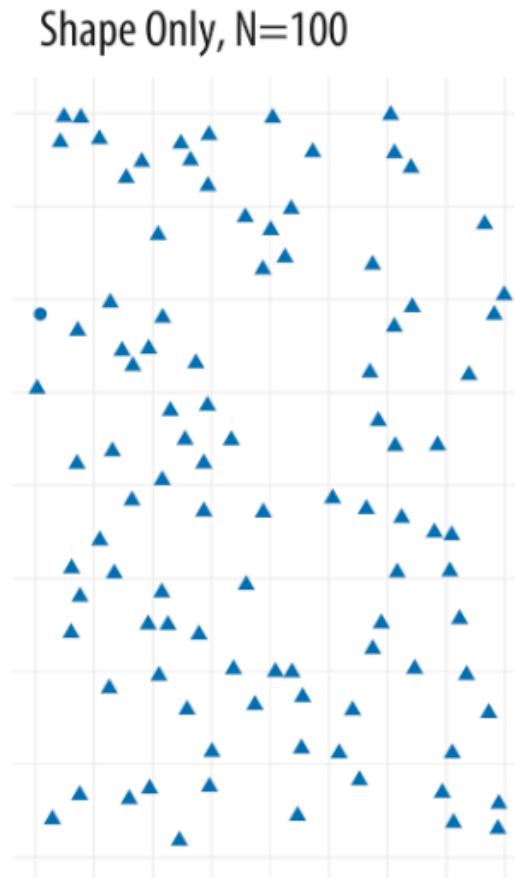
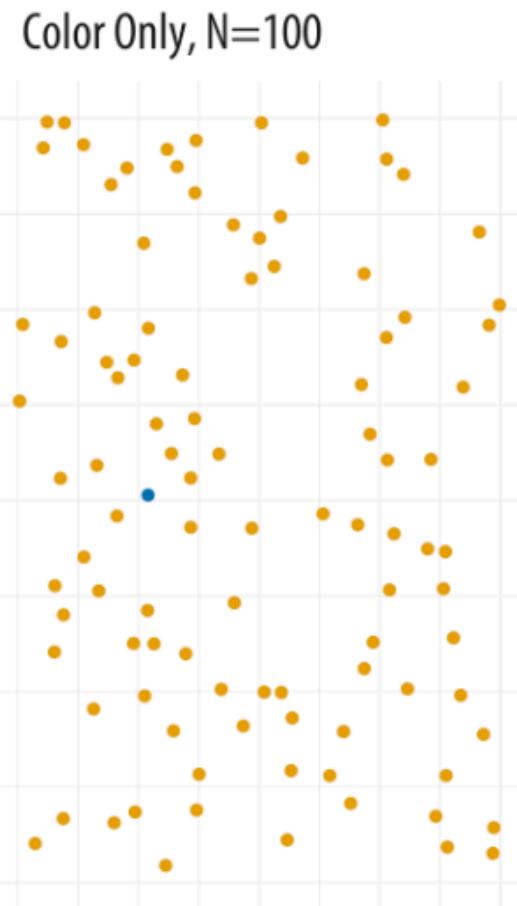
Crowdsourced Results



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

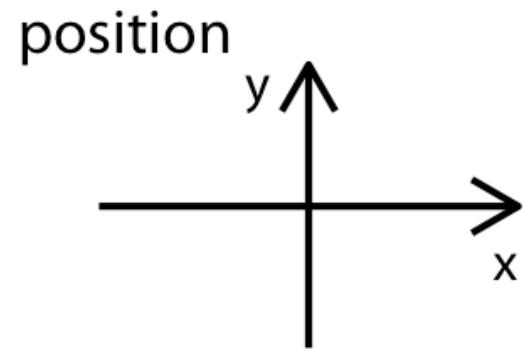
Heer & Bostock (2010);
image source: Healy, 2018

Attentional “pop”



Attentional “pop” is stronger for color than for shape.

Main channels available in ggplot



shape



size



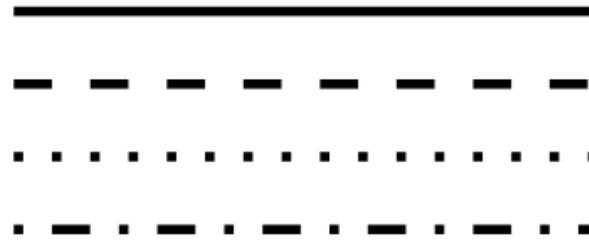
color



line width



line type



Color Channel

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

Okabe Ito

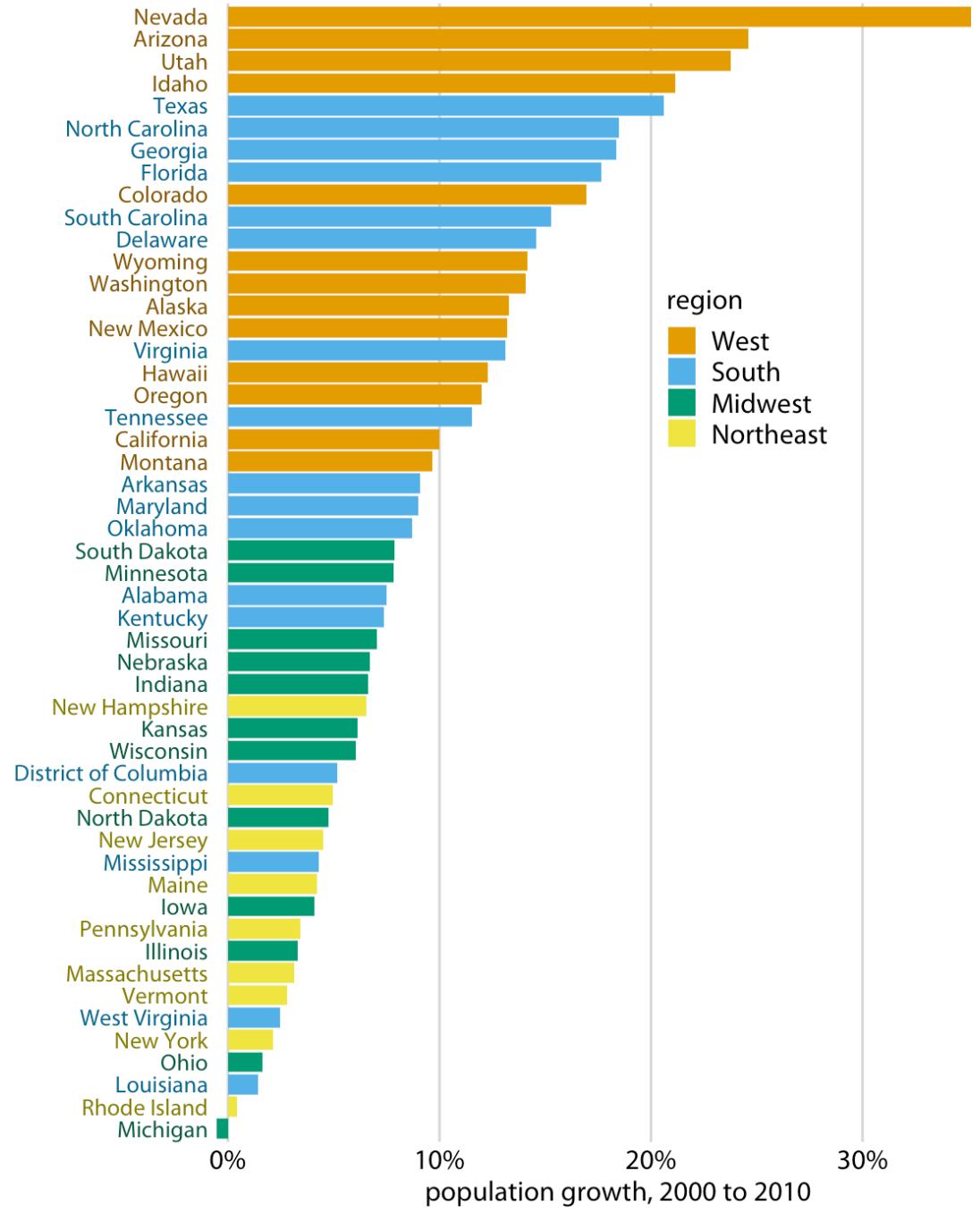


ColorBrewer Dark2



ggplot2 hue





(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

ColorBrewer Blues

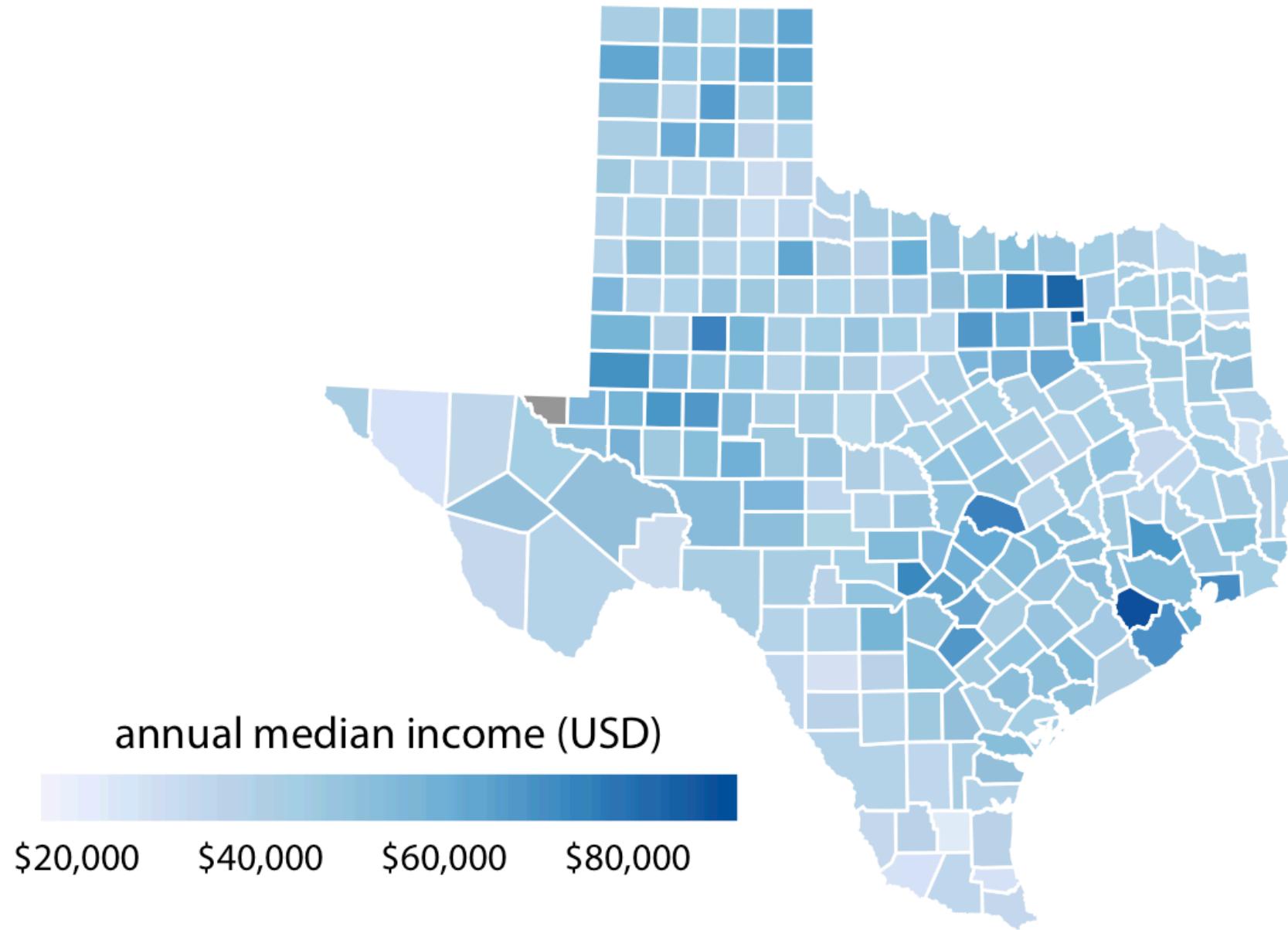


Heat



Viridis





(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/>

Quantitative

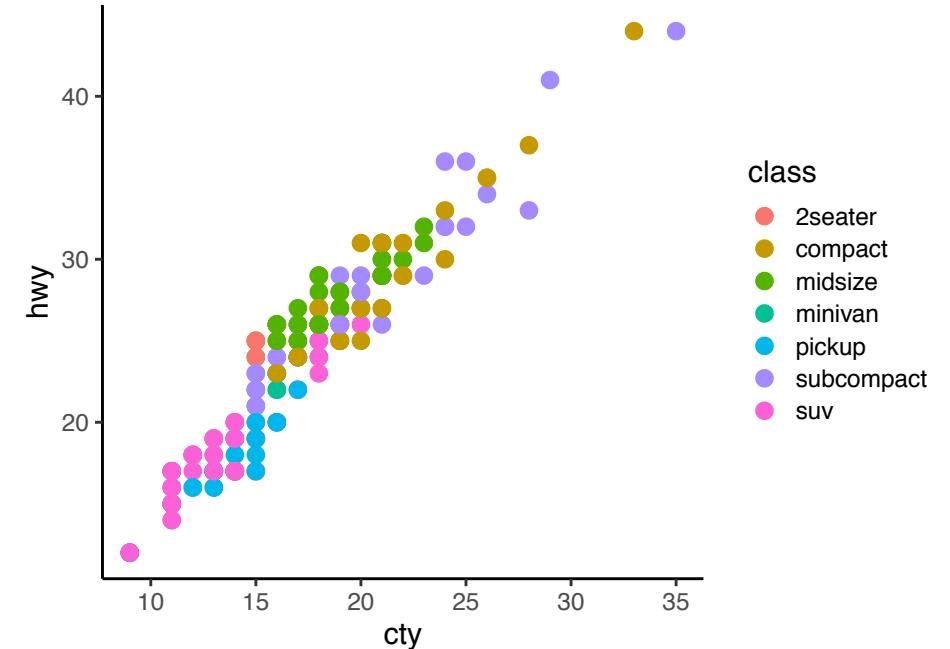
Qualitative

Brewer Color Palette



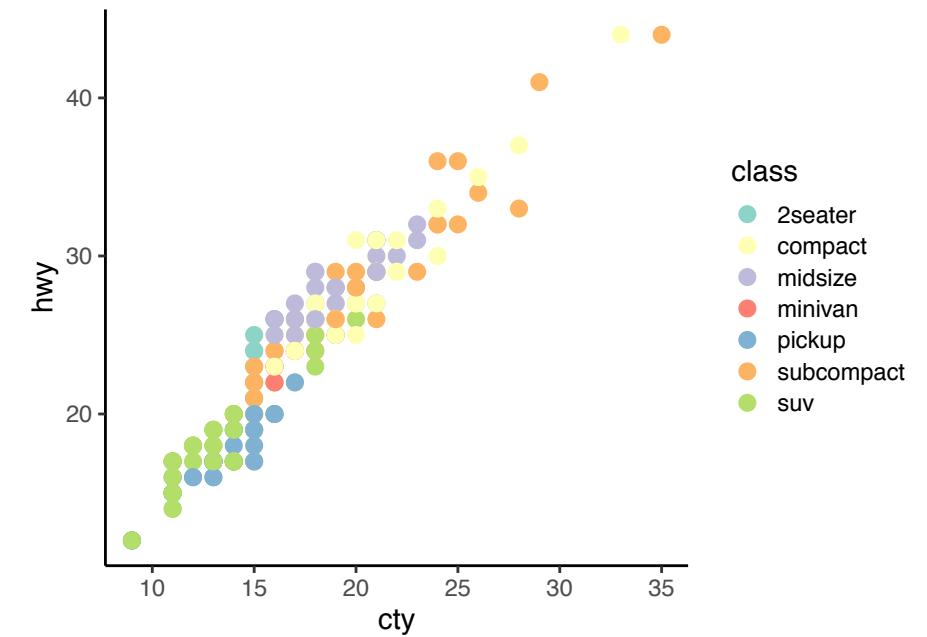
Default

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



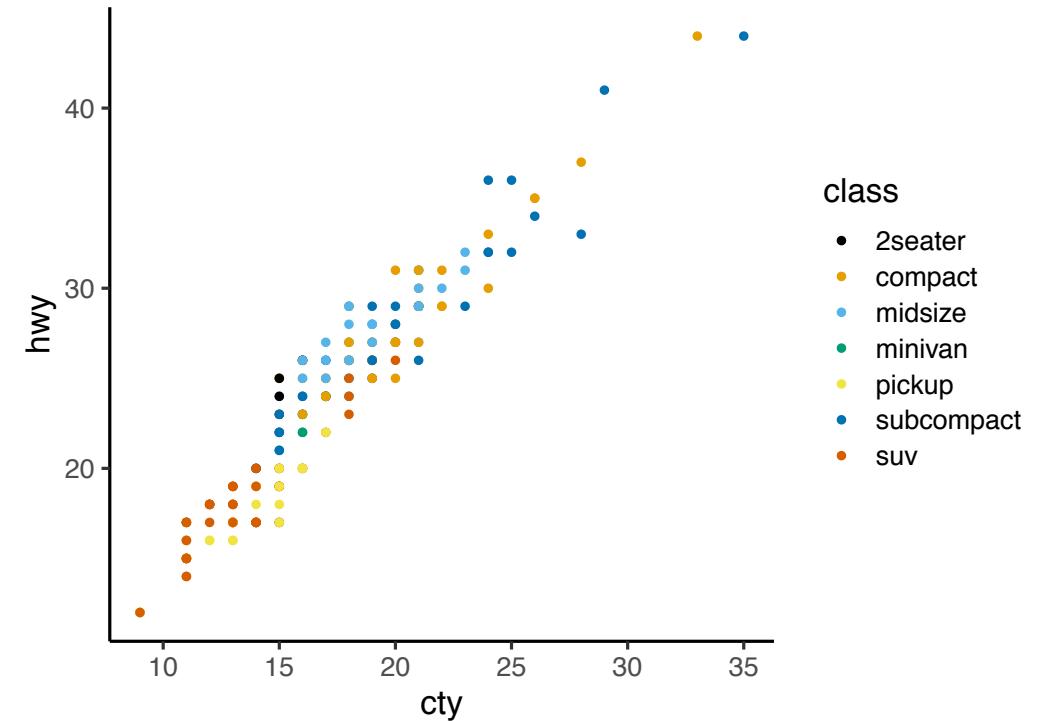
Brewer palette

```
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)
```



Manual palette

```
ggplot(data = mpg, aes(x = cty, y = hwy)) +  
  geom_point(aes(color = class)) +  
  scale_color_manual(values = c("#000000", "#E69F00", "#56B4E9",  
    "#009E73", "#F0E442", "#0072B2",  
    "#D55E00", "#CC79A7")) +  
  theme_classic(base_size = 16)
```

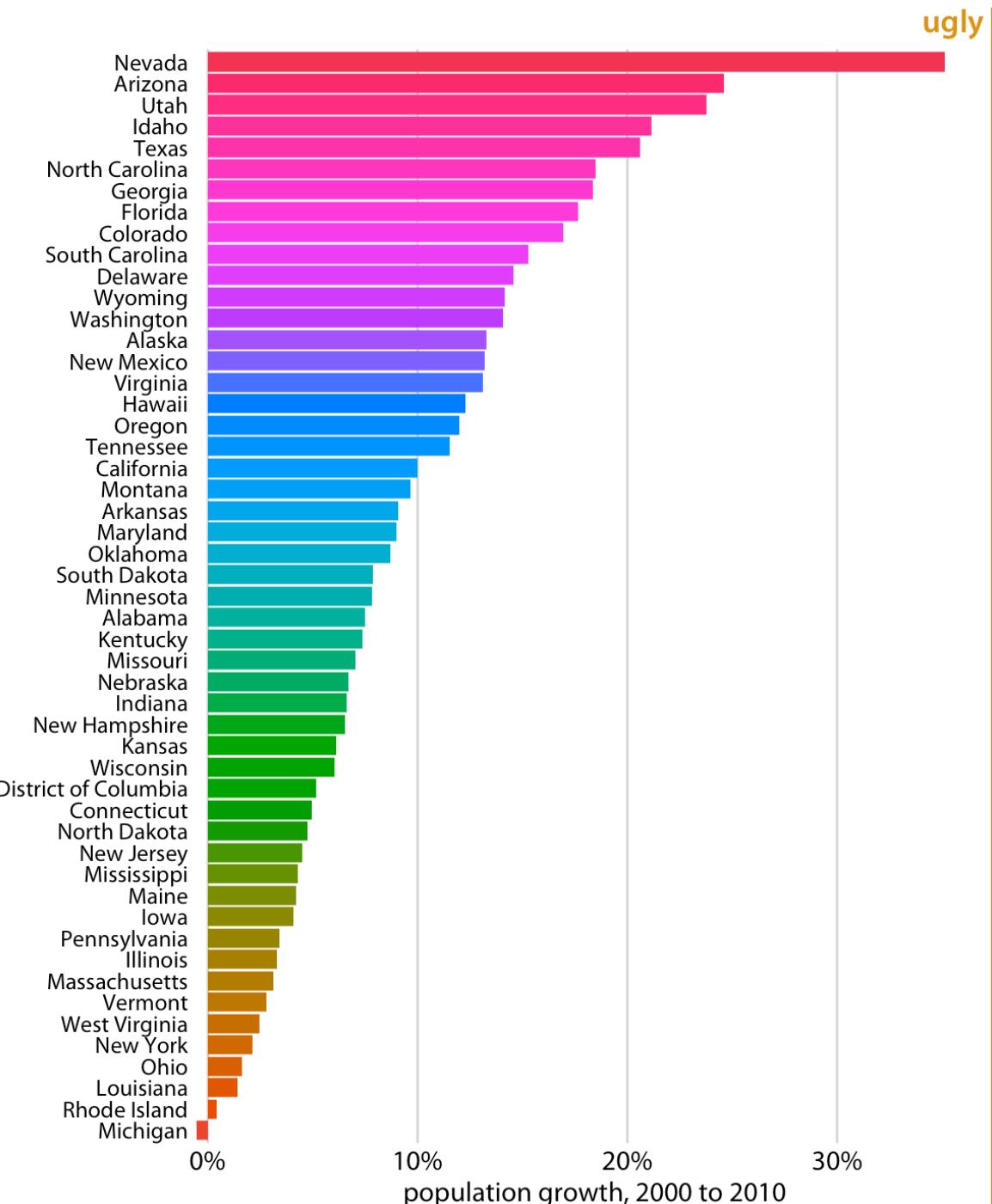


More plotting guidelines

Use colors to communicate something

Avoid using color for the sake of color

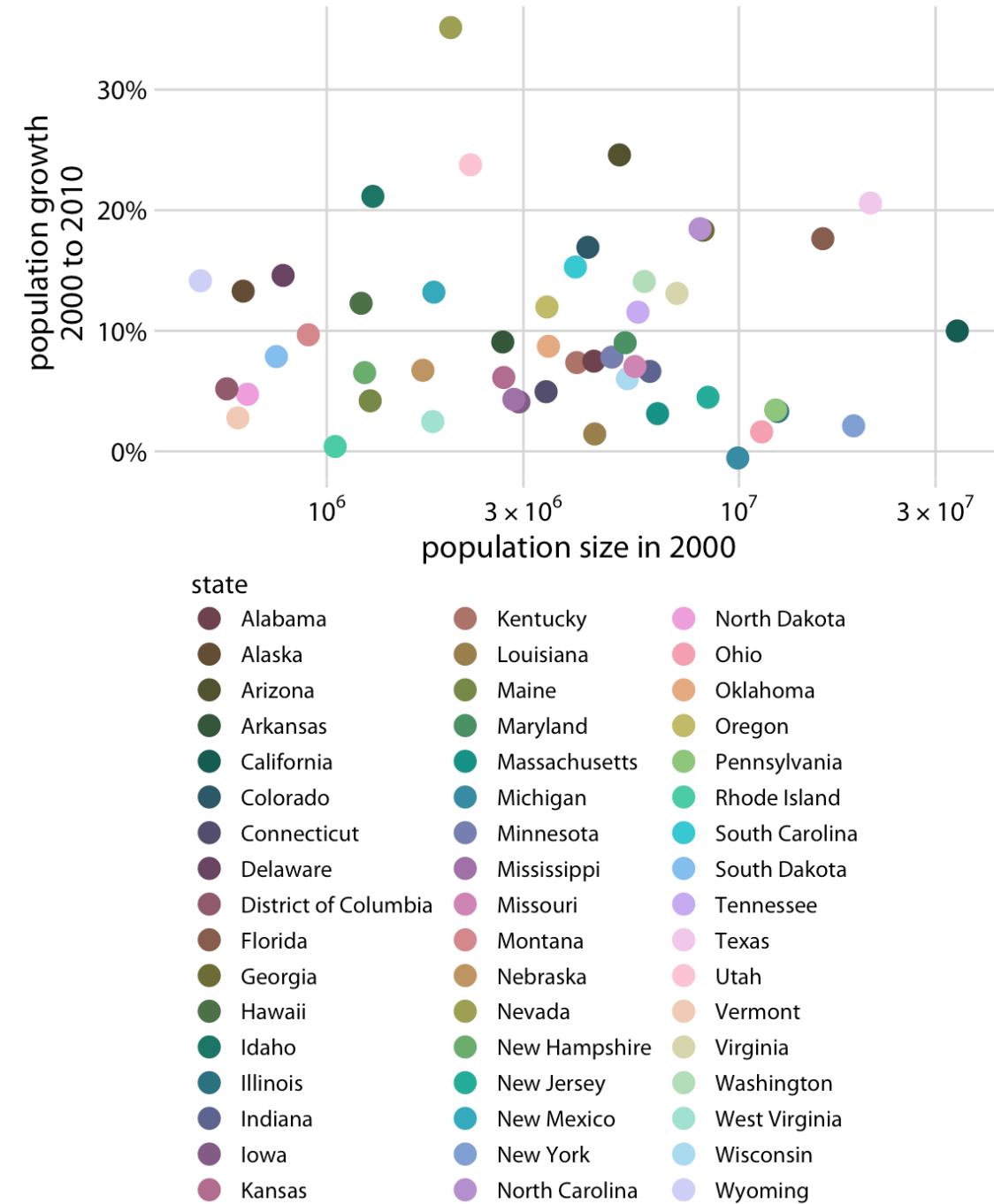
= chart junk!



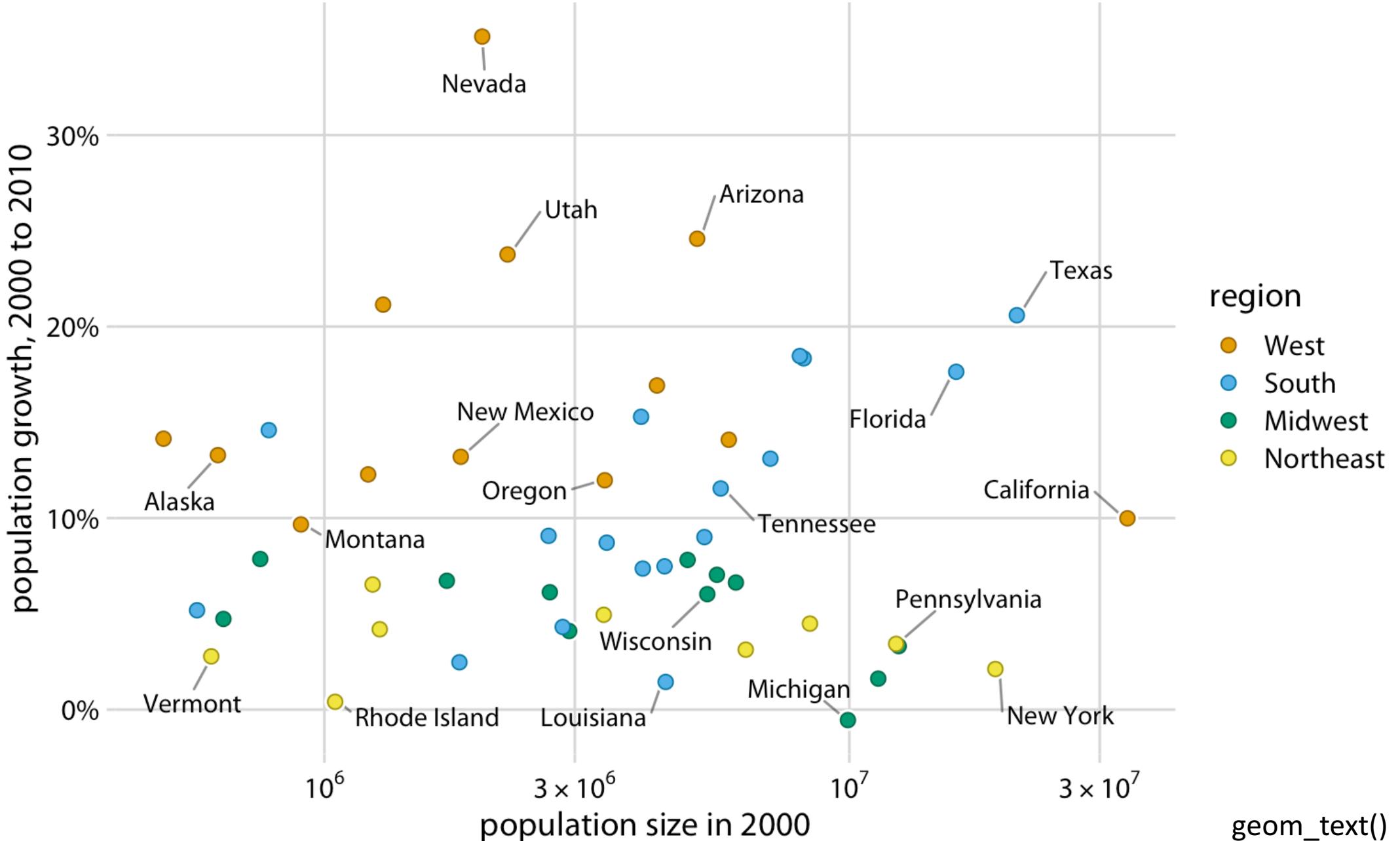
(image source: Wilke, *Fundamentals of Data Visualization*)

Avoid encoding too much/ irrelevant information

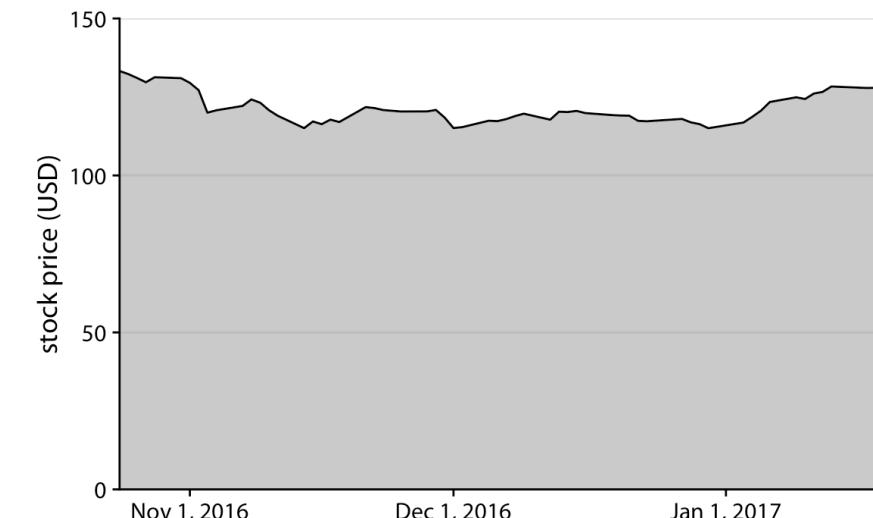
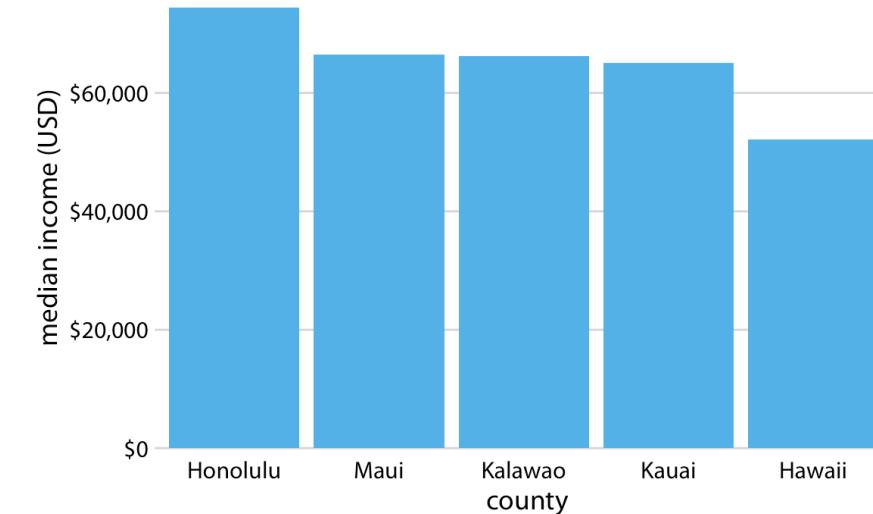
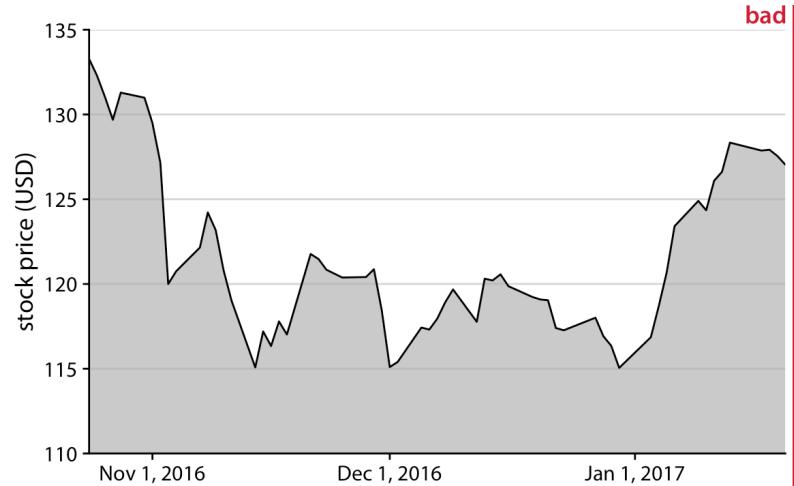
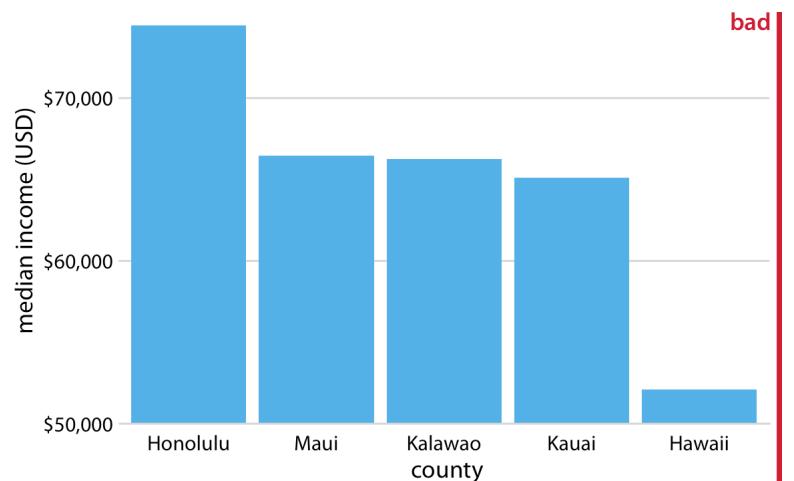
Use direct labeling to distinguish >8 qualitative variables.



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

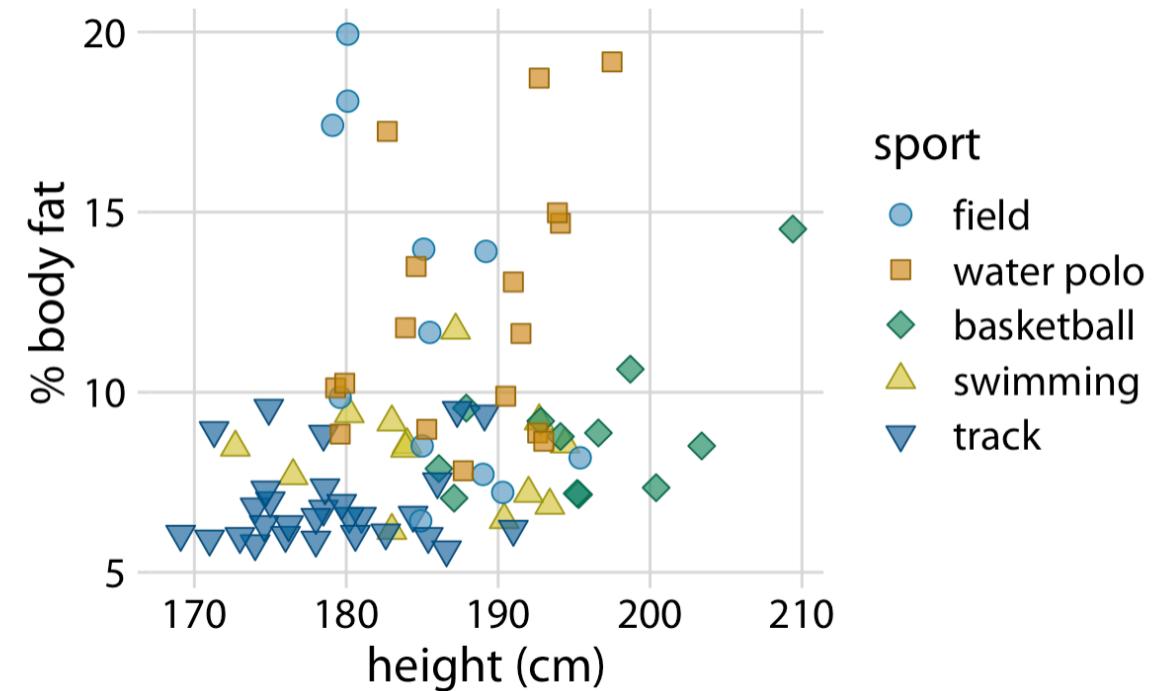
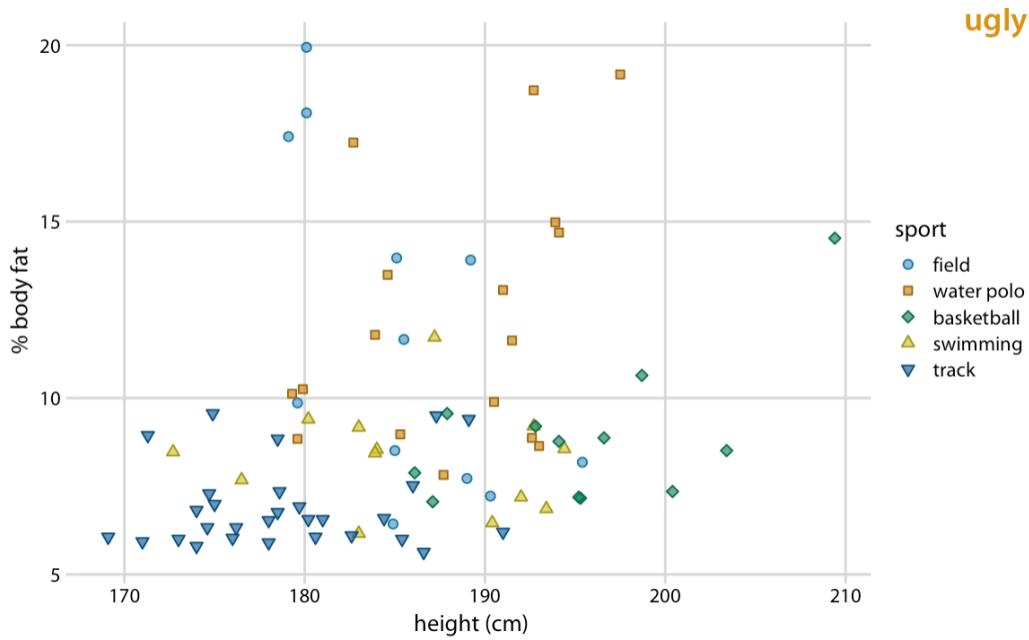


Axes should start at 0 (ggplot does this by default)



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

Make text readable by increasing font size



theme_classic(base_size = 16)

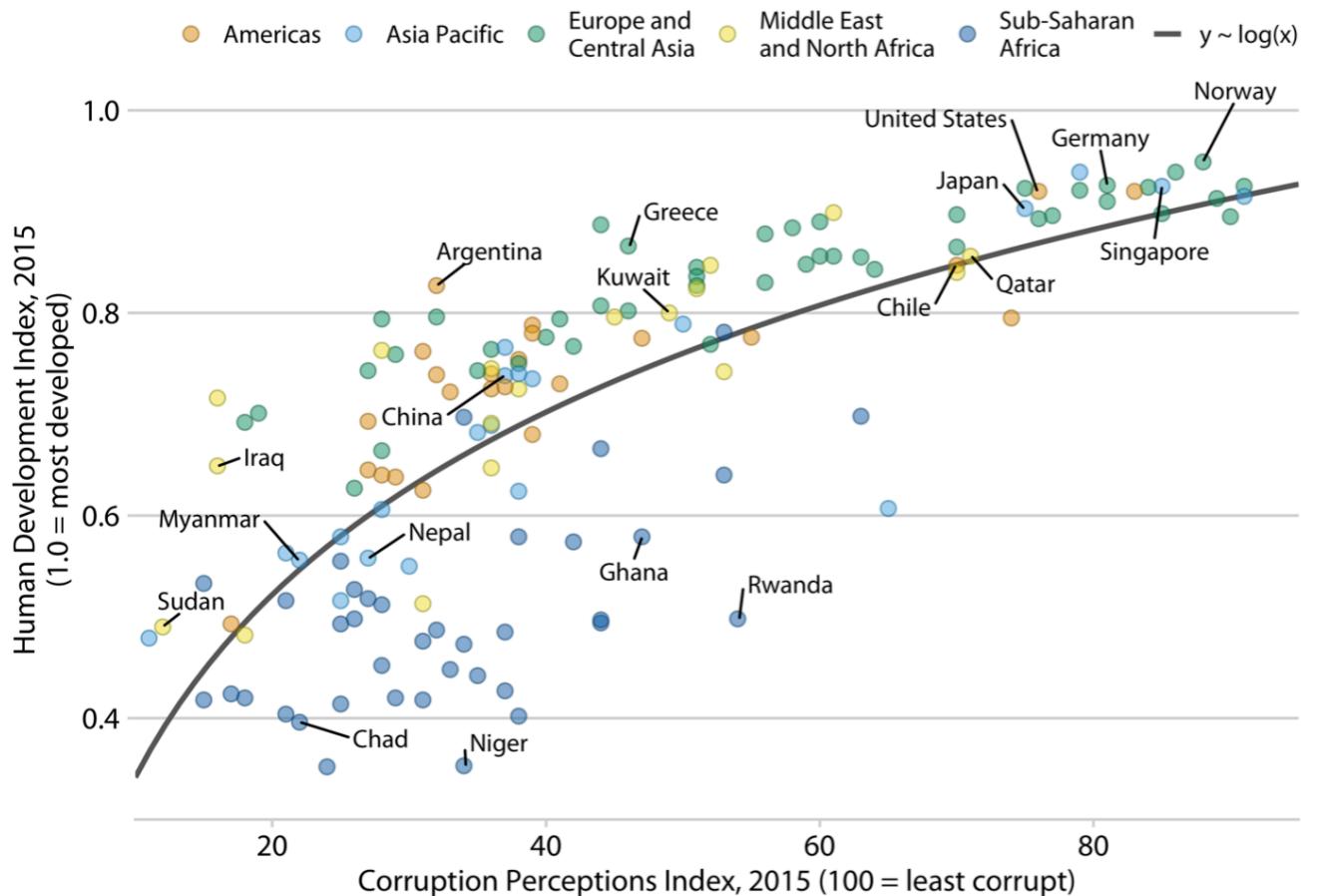
Add interpretable labels

Title and axis labels

```
ggtile(  
  label = "Corruption and human  
  development",  
  subtitle = "The most developed countries  
  experience the least corruption"  
)  
  
xlab("Corruption Perceptions Index, 2015  
(100 = least corrupt)")  
  
ylab("Human Development Index, 2015  
(1.0 = most developed)")
```

Corruption and human development

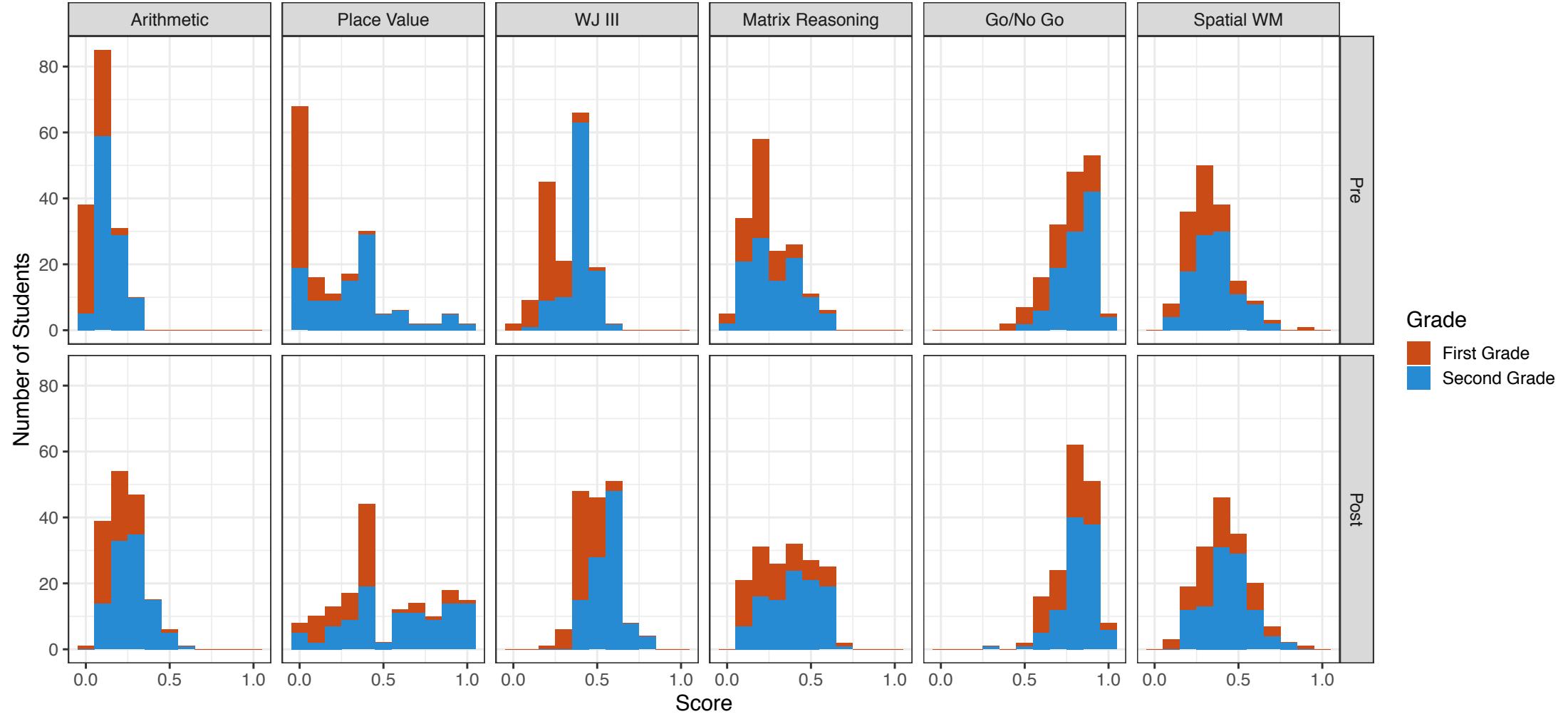
The most developed countries experience the least corruption



Data sources: Transparency International & UN Human Development Report

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

Use “facets” for large data

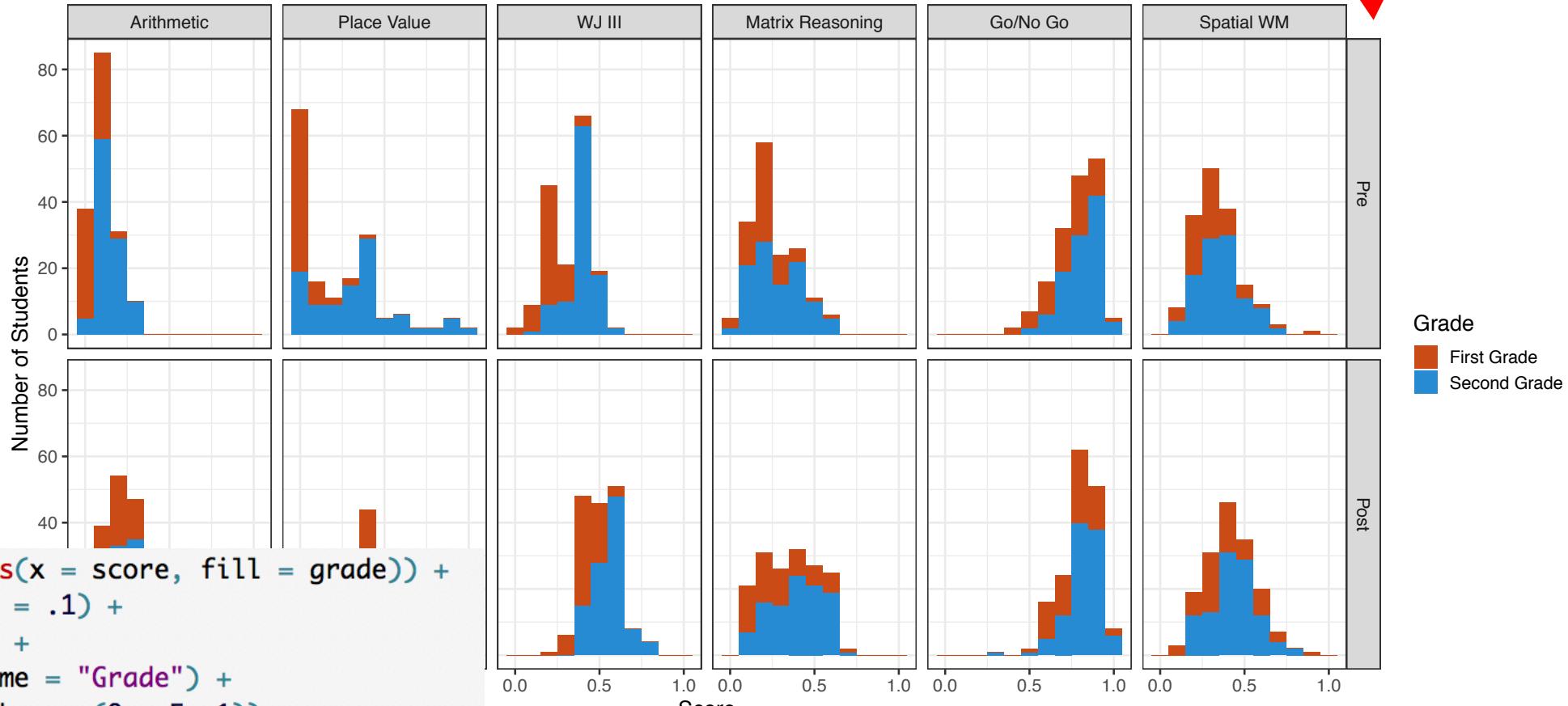


facet_grid()

Task



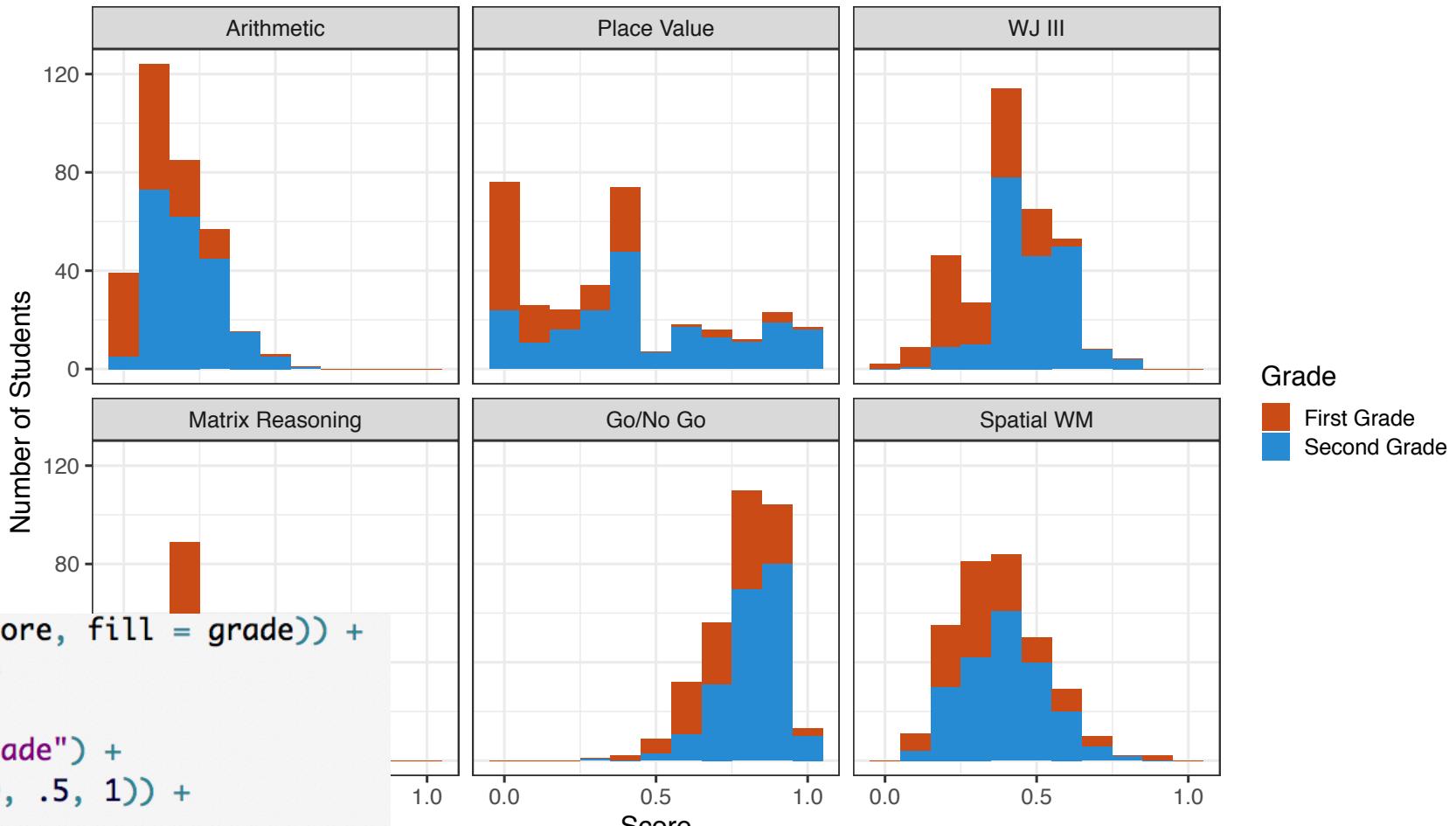
Time



```
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)
```

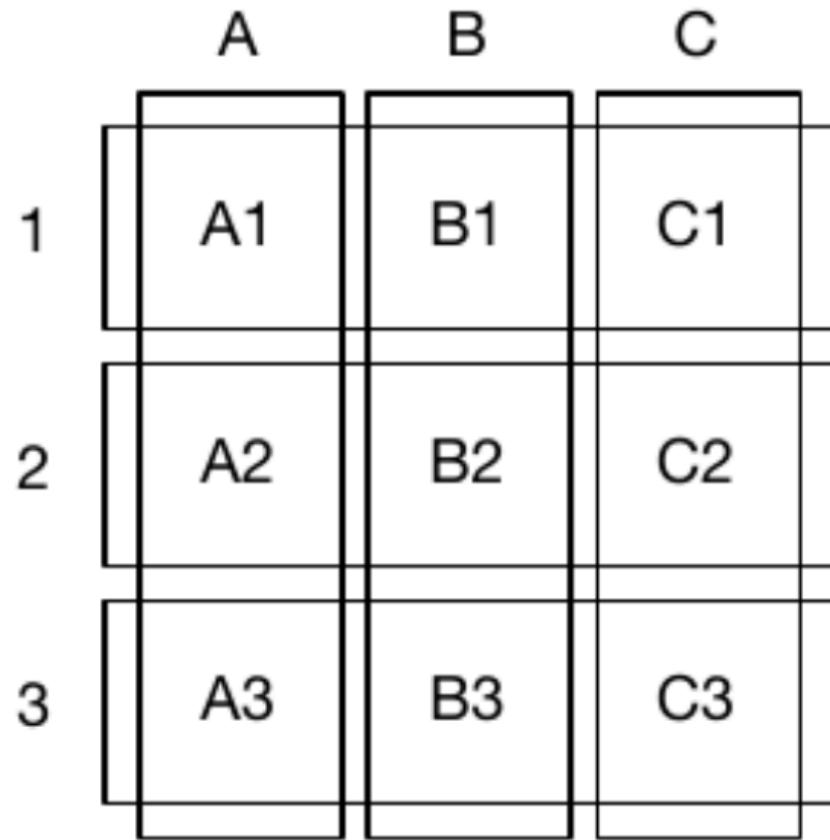
facet_wrap()

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()`



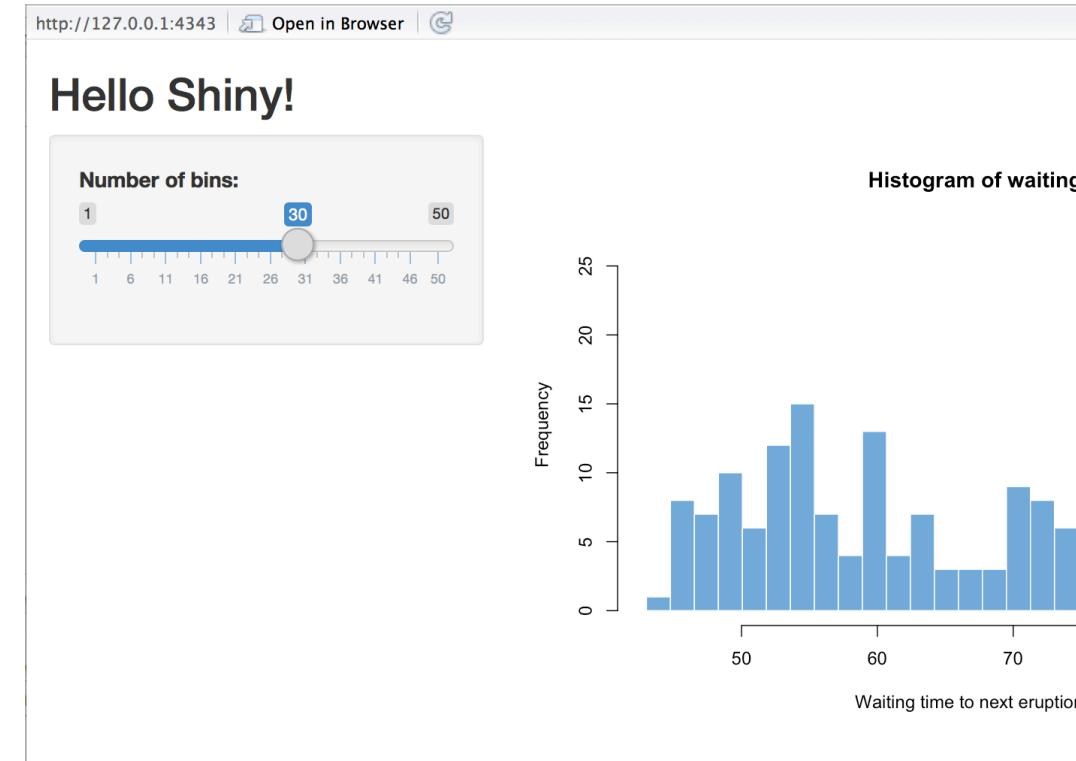
`facet_grid`



`facet_wrap`

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/>
- Build with Shiny:
 - <https://mlewis.shinyapps.io/lNhBrowser/>
 - https://mlewis.shinyapps.io/SI_KIDBOOK/



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

Summary of Principles of Visualization

- Think of plotting as communication – you want to maximize the likelihood that your audience gets your message
- How do you do that?
 - Get rid of “chart junk” – maximize info to ink ratio
 - Don’t be deceptive – show the raw data when possible
 - Think about human perception.
- 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
- Use appropriate color scale and make axes readable

Next Time

- Review of tidyverse verbs (summarize, mutate, group_by, arrange, ggplot, etc.)
- Evaluating plots

Office Hours:

Jaeah 1:00-3pm today (Psych Lounge);

Molly 4:30-6:30pm Wednesday (Porter 223A)

Acknowledgements

Images on slides 11-35 adapted from

<http://socviz.co/lookatdata.html> (Healy, 2018) and

<https://serialmentor.com/dataviz/> (Wilke)