

TEACHING INTRO DATA SCIENCE & ASSESSING LEARNING

PREPARING TO TEACH
JSM 2019

NICHOLAS HORTON

AMHERST COLLEGE



@askdrstats



nicholasjhorton



nhorton@amherst.edu

MINE ÇETINKAYA-RUNDEL

UNIVERSITY OF EDINBURGH + DUKE + RSTUDIO



mine-cetinkaya-rundel

cetinkaya.mine@gmail.com



GAISE 2016

1. Teach statistical thinking.

a. Teach statistics as an investigative process of problem-solving and decision-making.

Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision-making *process* that is fundamental to scientific inquiry and essential for making sound decisions.

b. Give students experience with multivariable thinking.

We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

2. Focus on conceptual understanding.

3. Integrate real data with a context and a purpose.

4. Foster active learning.

5. Use technology to explore concepts and analyze data.

6. Use assessments to improve and evaluate student learning.

① NOT a commonly used subset of tests and intervals and produce them with hand calculations

② Multivariate analysis requires the use of computing

③ NOT use technology that is only applicable in the intro course or that doesn't follow good science principles

④ Data analysis isn't just inference and modeling, it's also data importing, cleaning, preparation, exploration, and visualization

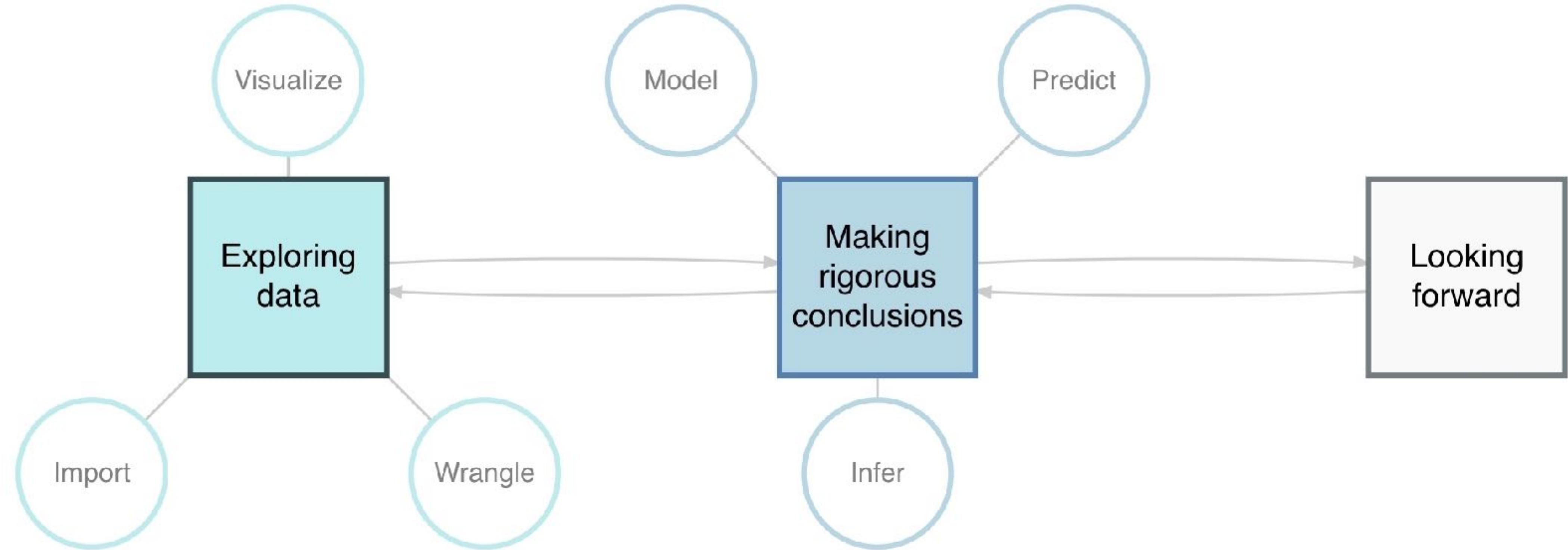
don't start with this

- Exploratory data analysis
- Study design
- Probability
- Random variables
- Central Limit Theorem
- One sample mean HT and CI
- One sample proportion HT and CI
- Two sample mean HT and CI
- Two sample proportion HT and CI
- Chi-square test
- ANOVA
- Simple linear regression

and add all this

- + R
- + R Markdown
- + git / GitHub
- + data scraping
- + iteration
- + working with non-rectangular data
- + interactive visualization

...



Fundamentals of
data & data viz,
confounding variables,
Simpson's paradox
+
R / RStudio,
R Markdown, simple Git

Tidy data, data frames
vs. summary tables,
recoding & transforming,
web scraping & iteration
+
collaboration on GitHub

Building & selecting
models,
visualizing interactions,
prediction & validation,
inference via simulation

Data science ethics,
interactive viz &
reporting, text analysis,
Bayesian inference
+
communication &
dissemination

5

design
principles



Which kitchen would you
rather bake a cake?





Which kitchen would you
rather bake a cake?



cherish

day

one

T

minimize
time spent
on course logistics

maximize
time spent on
creating a data
visualization

don't start like this

- Install R
- Install RStudio
- Install the following packages:
 - rmarkdown
 - tidyverse
 - ...
- Load these packages
- Install git

instead do this

- Go to rstudio.cloud (or some other server based solution)
 - Log in with your ID & pass
- > hello R!





Join Space?

Joining a space gives you access to it and to its contents.

Once you join, admins will be able to see your email address.

Would you like to join this space?

Join Space

Cancel



Welcome to Preparing to Teach

Workspace for the Preparing to Teach pre-JSM 2018 workshop

If you did not intend to join this space, or you later decide you don't want to be a member, just go to the [Members](#) area and click "Leave Space".



All Projects

START

UN Votes

Mine Çetinkaya-Rundel

Created Jul 27, 2019 9:39 AM

New Project



Options



Search Projects



List Projects

 All Shared with everyone Yours

Sort Projects

 By name By date created



File Edit Code View Plots Session Build Debug Profile Tools Help

+ | Go to file/function | Addins ▾

Console Terminal ▾ Jobs ▾

/cloud/project/

```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"  
Copyright (C) 2019 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

> |

Environment History Connections

Import Dataset |

Global Environment ▾

Environment is empty

Files Plots Packages Help Viewer

New Folder | Upload | Delete | Rename | More ▾

Cloud > project

	Name	Size	Modified
	..		
<input type="checkbox"/>	.Rhistory	0 B	Jul 27, 2019, 1:21 PM
<input type="checkbox"/>	project.Rproj	205 B	Jul 27, 2019, 1:21 PM
<input type="checkbox"/>	unvotes.html	1.3 MB	Jul 27, 2019, 9:52 AM
<input type="checkbox"/>	unvotes.Rmd	3.4 KB	Jul 27, 2019, 9:51 AM



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Addins

R 3.6.0

unvotes.Rmd x



```
1 ---  
2 title: "UN Votes"  
3 author: "Mine Çetinkaya-Rundel"  
4 date: "`r Sys.Date()`"  
5 output:  
6   html_document:  
7     toc: true  
8     toc_float: true  
9 ---  
10  
11 ## Analysis  
12  
13 Let's take a look at the voting history of countries in the United Nations  
14 General Assembly. We will be using data from the unvotes package.  
15 Additionally, we will make use of the tidyverse and lubridate packages  
16 for the analysis, and the DT package for interactive display of tabular  
17 output  
1:1 # UN Votes
```

Insert



Run



Stop



Reset



Save



Print



Copy



Paste



Find



Replace



Delete



Edit



Format



Align



Font



Size



Color



Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



Font Weight



Font Size



Font Color



Font Background Color



Font Style



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

R 3.6.0

unvotes.Rmd

```
1 ---  
2 title: "UN Votes"  
3 author: "Mine Çetinkaya-Rundel"  
4 date: `r Sys.Date()`  
5 output:  
6   html_document:  
7     toc: true  
8     toc_float: true  
9 ---  
10 ## Analysis  
11 Let's take a look at the voting history of countries in t  
12 General Assembly. We will be using data from the __unvote  
13 Additionally, we will make use of the __tidyverse__ and __  
14 tidyverse__ and __DT__ for the analysis, and the __DT__ package for interactive  
15 output  
1:1 # UN Votes
```

Console Terminal R Markdown Jobs

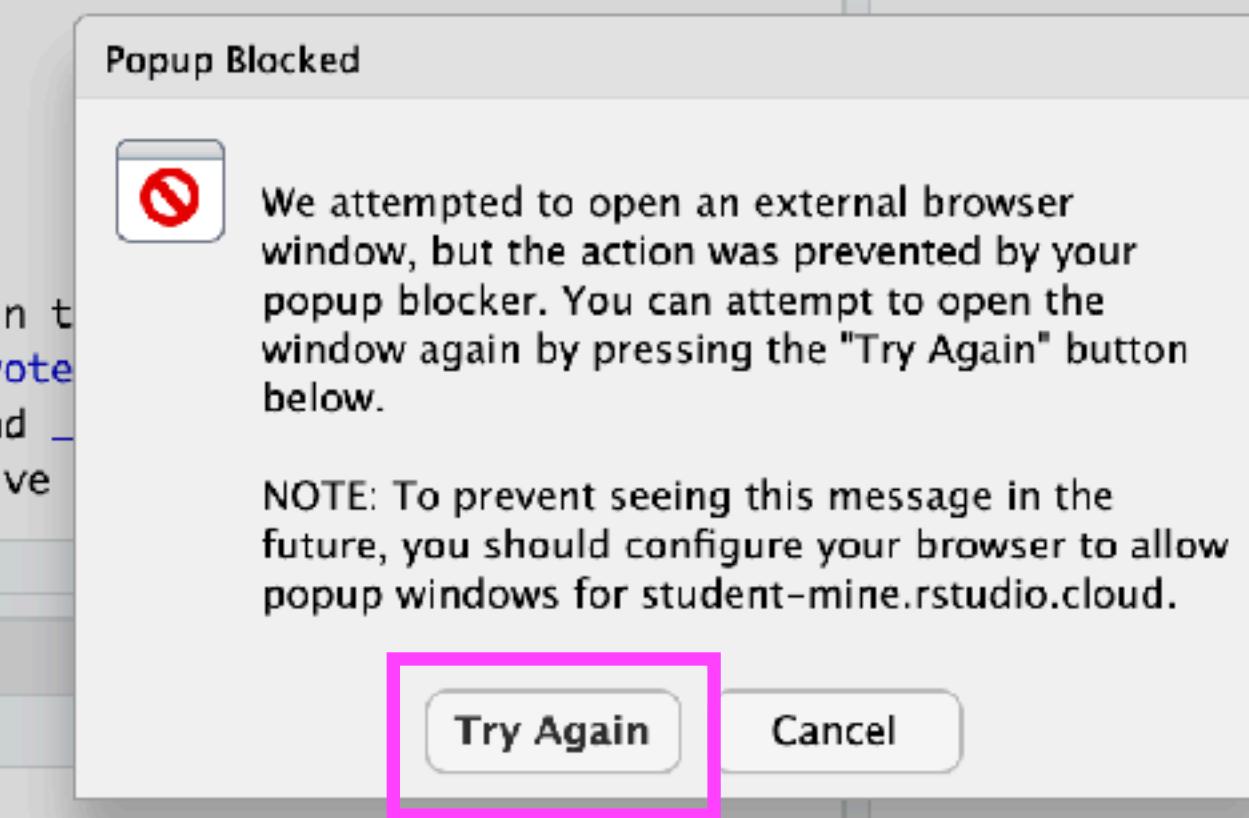
/cloud/project/ ↗

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |



Environment History Connections

Insert Run Import Dataset

Global Environment

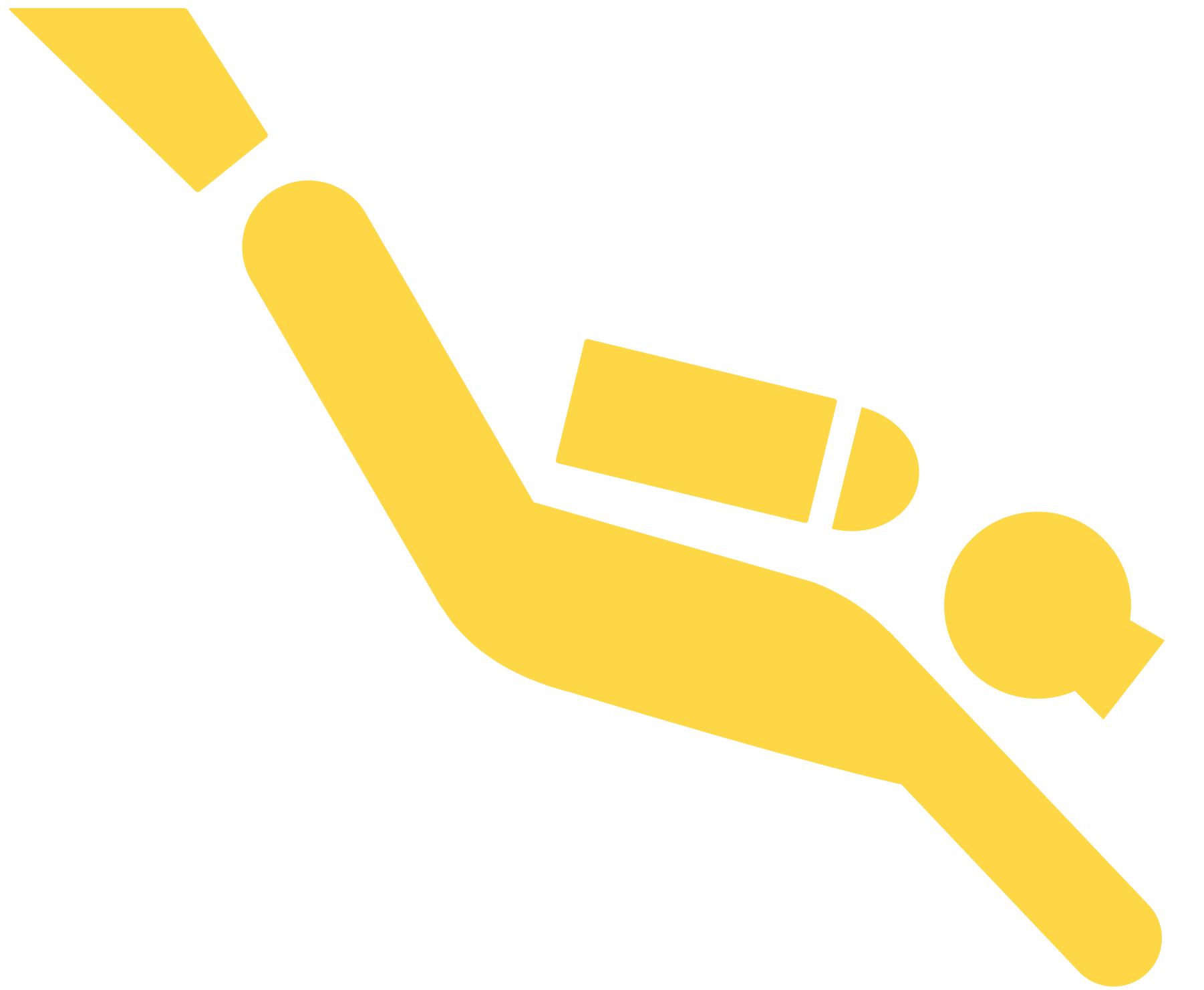
Environment is empty

Pages Help Viewer

Upload Delete Rename More

ct

	Size	Modified
	0 B	Jul 27, 2019, 1:21 PM
obj	205 B	Jul 27, 2019, 1:21 PM
nl	1.3 MB	Jul 27, 2019, 1:22 PM
id	3.4 KB	Jul 27, 2019, 9:51 AM



- ▶ Go to bit.ly/ptt-rscloud and create an account to join the RStudio Cloud workspace for this workshop.
- ▶ Start the assignment called UN Votes.
- ▶ Open the R Markdown document called un-votes.Rmd, knit the document, view the result.
- ▶ Then, change “Turkey” to another country, and knit again.

Resources:

Computing in the classroom

- ▶ **RStudio Cloud in the Classroom** (Mine Çetinkaya-Rundel)
 - ▶ Nitty-gritty of setting up your course on RStudio Cloud
 - ▶ Video and slides: <https://resources.rstudio.com/webinars/rstudio-cloud-in-the-classroom>
- ▶ **Infrastructure and tools for teaching computing throughout the statistical curriculum** (Mine Çetinkaya-Rundel and Colin Rundel)
 - ▶ Overview of cloud computing resources for teaching
 - ▶ Part of the Practical Data Science for Stats collection
 - ▶ <https://peerj.com/preprints/3181/>
- ▶ **JSM 2019 - BoF: Teaching (with) R** (Mine Doğucu)
 - ▶ Tue, 7/30/2019, 12:30 PM - 1:30 PM, CC-SocietyTables Registration Promenade

minimize
time spent
on course logistics

maximize
time spent on
creating a data
visualization

show examples of
data in the wild



JULIA SILGE

BLOG ABOUT RESUME

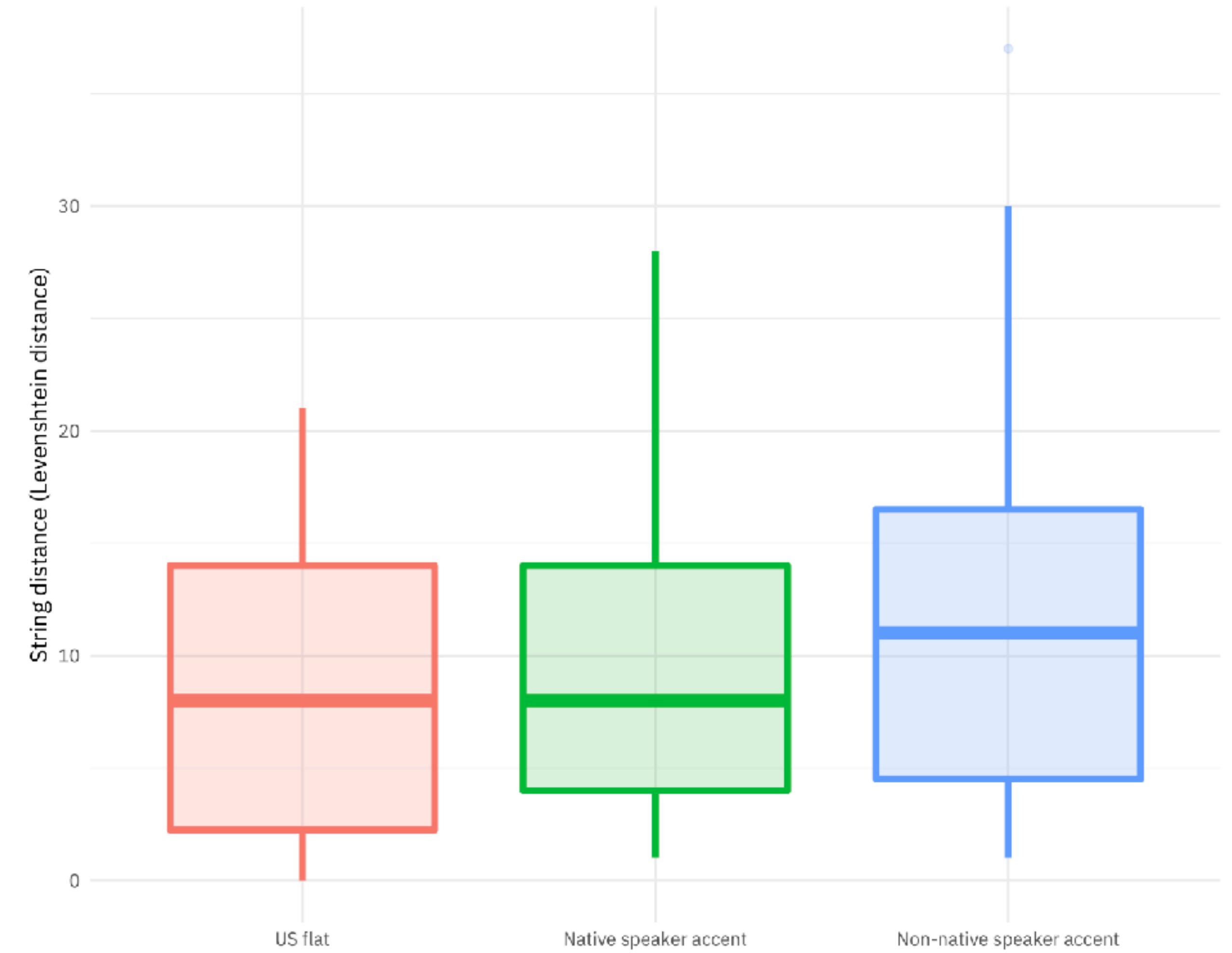
AMAZON ALEXA AND ACCENTED ENGLISH

Jul 19, 2018 · 6 minute read · rstats

Earlier this spring, one of my data science friends here in SLC got in contact

How well does Alexa understand different accents?

Speech with non-native accents is converted to text with the lowest accuracy





David Robinson

Chief Data Scientist at
DataCamp, works in R and
Python.

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:

Donald J. Tr
Good luck #
#OpeningCe
pic.twitter.c

27,391 Likes

Aug 5, 2016 at 8:59 PM

Todd Vaziri @tvaziri

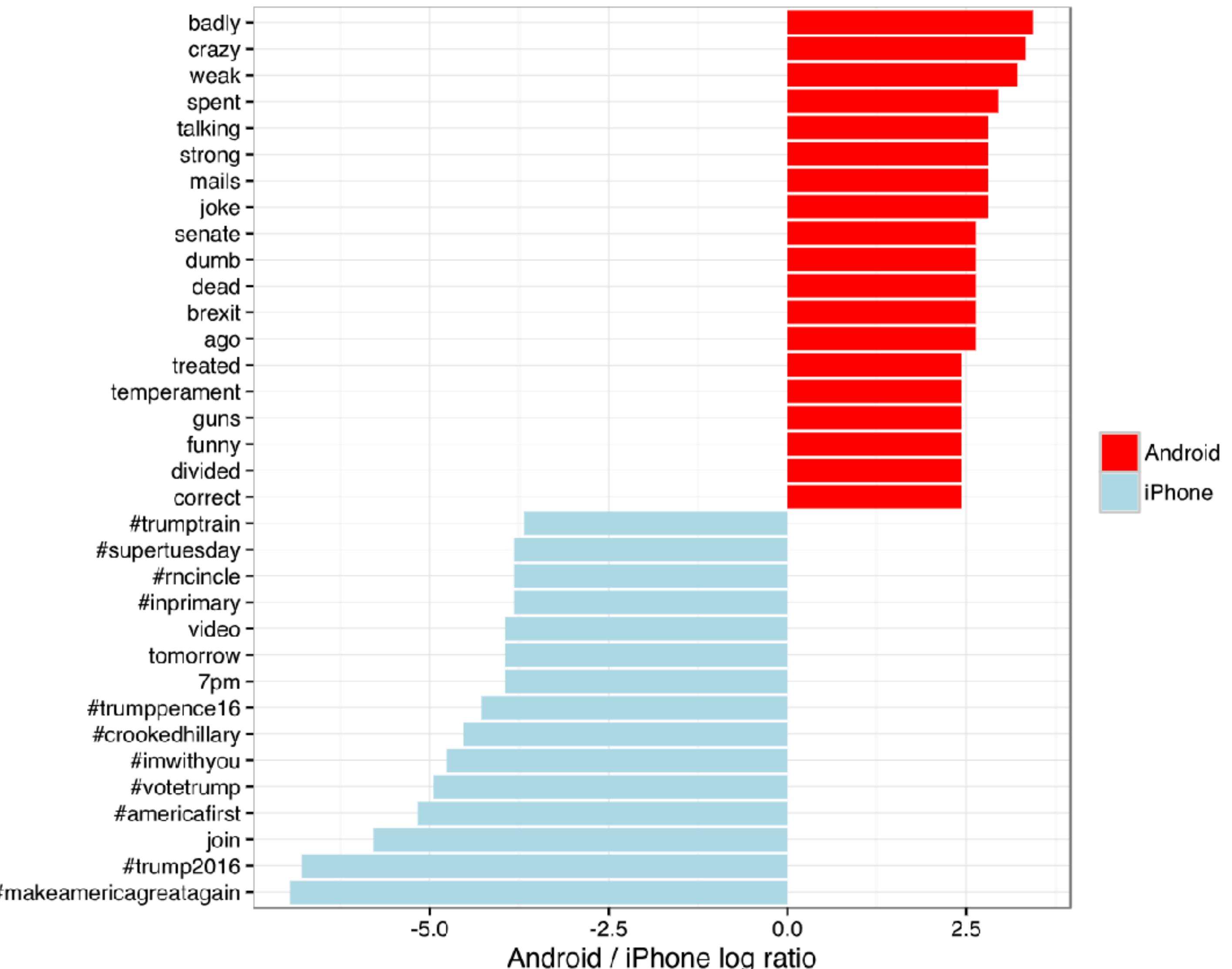
Every non-hyperbolic tweet is from iPhone (his staff).

Every hyperbolic tweet is from Android (from him).

12:20 PM - Aug 6, 2016

14.1K 10.2K people are talking about this

Which are the words most likely to be from Android and most likely from iPhone?



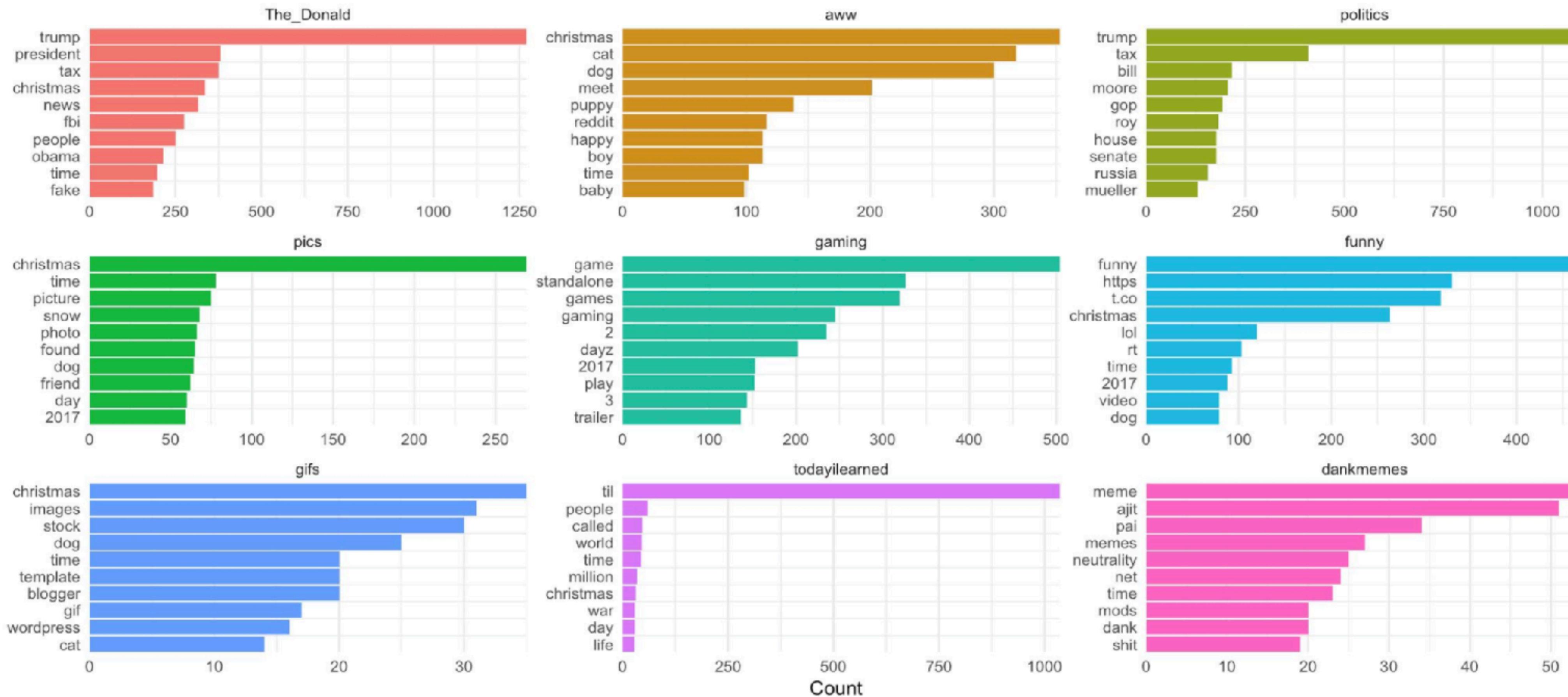
How to Succeed on Reddit

Former intro DS students!!

Team InterstellR

Most frequent words within popular subreddits

in December 2017



How to Succeed on Reddit

Team InterstellR

Modeling Popularity

- Multivariate linear model
- Target: score
- Predictors: subreddit, sentiment, dog_cat, [text features], ...
- Stepwise selection by AIC
- $R^2 = 0.177$

Docs	Terms									
	1	12	2	2017	amp	christmas	game	https	time	world
7hbf0d	1	0	1	0	1	0	0	0	0	0
7hnto4	0	1	1	1	0	0	1	0	0	0
7iiku8	1	1	0	1	0	0	1	0	0	0
7ioafs	1	0	0	0	1	0	0	0	1	0
7ixrdt	0	0	1	1	0	0	1	0	0	0
7jvb0s	1	1	1	0	0	1	0	0	0	0
7kp1v0	0	0	0	1	0	0	1	0	0	0
7l5l52	0	0	0	1	0	0	1	0	0	0
7m1umi	0	0	0	1	0	0	0	0	0	0
7mguty	1	1	1	0	0	0	0	0	0	0

How to Succeed on Reddit

— Team InterstellR —

Modeling Popularity

- Multivariate linear model
- Target: score
- Predictors: subreddit, sentiment, dog_cat, [text features], ...
- Stepwise selection by AIC
- $R^2 = 0.177$

Docs	Terms									
	1	12	2	2017	amp	christmas	game	https	time	world
7hbf0d	1	0	1	0	1	0	0	0	0	0
7hnto4	0	1	1	0	0	1	0	0	0	0
7iiku8	1	1	0	1	0	0	1	0	0	0
7ioafs	1	0	0	0	1	0	0	0	1	0
7ixrdt	0	0	1	1	0	0	1	0	0	0
7jvb0s	1	1	1	0	0	1	0	0	0	0
7kplv0	0	0	0	1	0	0	1	0	0	0
7l5l52	0	0	0	1	0	0	1	0	0	0
7m1umi	0	0	0	1	0	0	0	0	0	0
7mgutu	1	1	0	0	0	0	0	0	0	0

Conclusions

1. Be negative
2. Dogs and cats are both good choices
3. Post on /r/gifs
4. Don't talk about December, games, and don't ask questions
5. Do talk about home and news
6. Don't use a linear model to predict Reddit post scores!

Ingredients

For the Cake:

16 ounces plain or **toasted sugar** (about 2 1/4 cups; 455g)

4 1/2 teaspoons baking powder

2 teaspoons (8g) Diamond Crystal kosher salt; for table salt, use about half as much by volume or the same weight

8 ounces unsalted butter (16 tablespoons; 225g), soft but cool, about 60°F (16°C)

3 large eggs, brought to about 65°F (18°C)

1/2 ounce vanilla extract (about 1 tablespoon; 15g)

16 ounces whole milk (about 2 cups; 455g), brought to about 65°F (18°C)

16 ounces all-purpose flour (about 3 1/2 cups, spooned; 455g)

Directions

1. **For the Cake:** Adjust oven rack to lower-middle position and preheat to 350°F (180°C). Lightly grease three 8-inch anodized aluminum cake pans and line with parchment (explanation and tutorial [here](#)). If you don't have three pans, it's okay to bake the cakes in stages, the batter will keep at room temperature until needed.
2. In the bowl of a stand mixer fitted with the paddle attachment, combine sugar, baking powder, salt, and butter. Mix on low speed to roughly incorporate, then increase to medium and beat until fluffy and light, about 5 minutes. About halfway through, pause to scrape the bowl and beater with a flexible spatula.
3. With the mixer still running, add the eggs one at a time, letting each fully incorporate before adding the next, then dribble in the vanilla. Reduce speed to low and sprinkle in about 1/3 of the flour, then drizzle in 1/3 of the milk. Repeat with remaining flour and milk, working in thirds as before.
4. Scrape the bowl and beater with a flexible spatula, and resume mixing on medium speed for about 3 seconds to ensure everything is well combined. The batter should look creamy and thick, registering between 65 and 68°F (18 and 20°C) on a digital thermometer. (Significant
5. Fold batter once or twice from the bottom up with a flexible spatula, then divide evenly between prepared cake pans (about 20 ounces or 565g if you have a scale). Stagger pans together on the oven rack, and bake until puffed, firm, and pale gold, about 32 minutes. If your oven has very uneven heat, pause to rotate the pans after about 20 minutes. Alternatively, bake two layers at once and finish the third when they're done.
6. Cool cakes directly in their pans for 1 hour, then run a butter knife around the edges to loosen. Invert onto a wire rack, peel off the parchment, and return cakes right-side-up (covered in plastic, the cakes can be left at room temperature for a few hours). Prepare the buttercream.

How do you prefer your cake recipes? Words only, or words & pictures?



Ingredients

For the Cake:

16 ounces plain or [toasted sugar](#) (about 2 1/4 cups; 455g)

4 1/2 teaspoons baking powder

2 teaspoons (8g) Diamond Crystal kosher salt; for table salt, use about half as much by volume or the same weight

8 ounces unsalted butter (16 tablespoons; 225g), soft but cool, about 60°F (16°C)

3 large eggs, brought to about 65°F (18°C)

1/2 ounce vanilla extract (about 1 tablespoon; 15g)

16 ounces whole milk (about 2 cups; 455g), brought to about 65°F (18°C)

16 ounces all-purpose flour (about 3 1/2 cups, spooned; 455g)

Directions

1. **For the Cake:** Adjust oven rack to lower-middle position and preheat to 350°F (180°C). Lightly grease three 8-inch anodized aluminum cake pans and line with parchment (explanation and tutorial [here](#)). If you don't have three pans, it's okay to bake the cakes in stages, the batter will keep at room temperature until needed.
2. In the bowl of a stand mixer fitted with the paddle attachment, combine sugar, baking powder, salt, and butter. Mix on low speed to roughly incorporate, then increase to medium and beat until fluffy and light, about 5 minutes. About halfway through, pause to scrape the bowl and beater with a flexible spatula.
3. With the mixer still running, add the eggs one at a time, letting each fully incorporate before adding the next, then dribble in the vanilla. Reduce speed to low and sprinkle in about 1/3 of the flour, then drizzle in 1/3 of the milk. Repeat with remaining flour and milk, working in thirds as before.
4. Scrape the bowl and beater with a flexible spatula, and resume mixing on medium speed for about 3 seconds to ensure everything is well combined. The batter should look creamy and thick, registering between 65 and 68°F (18 and 20°C) on a digital thermometer. (Significant
5. Fold batter once or twice from the bottom up with a flexible spatula, then divide evenly between prepared cake pans (about 20 ounces or 565g if you have a scale). Stagger pans together on the oven rack, and bake until puffed, firm, and pale gold, about 32 minutes. If your oven has very uneven heat, pause to rotate the pans after about 20 minutes. Alternatively, bake two layers at once and finish the third when they're done.
6. Cool cakes directly in their pans for 1 hour, then run a butter knife around the edges to loosen. Invert onto a wire rack, peel off the parchment, and return cakes right-side-up (covered in plastic, the cakes can be left at room temperature for a few hours). Prepare the buttercream.

How do you prefer your cake recipes? Words only, or words & pictures?



start

with

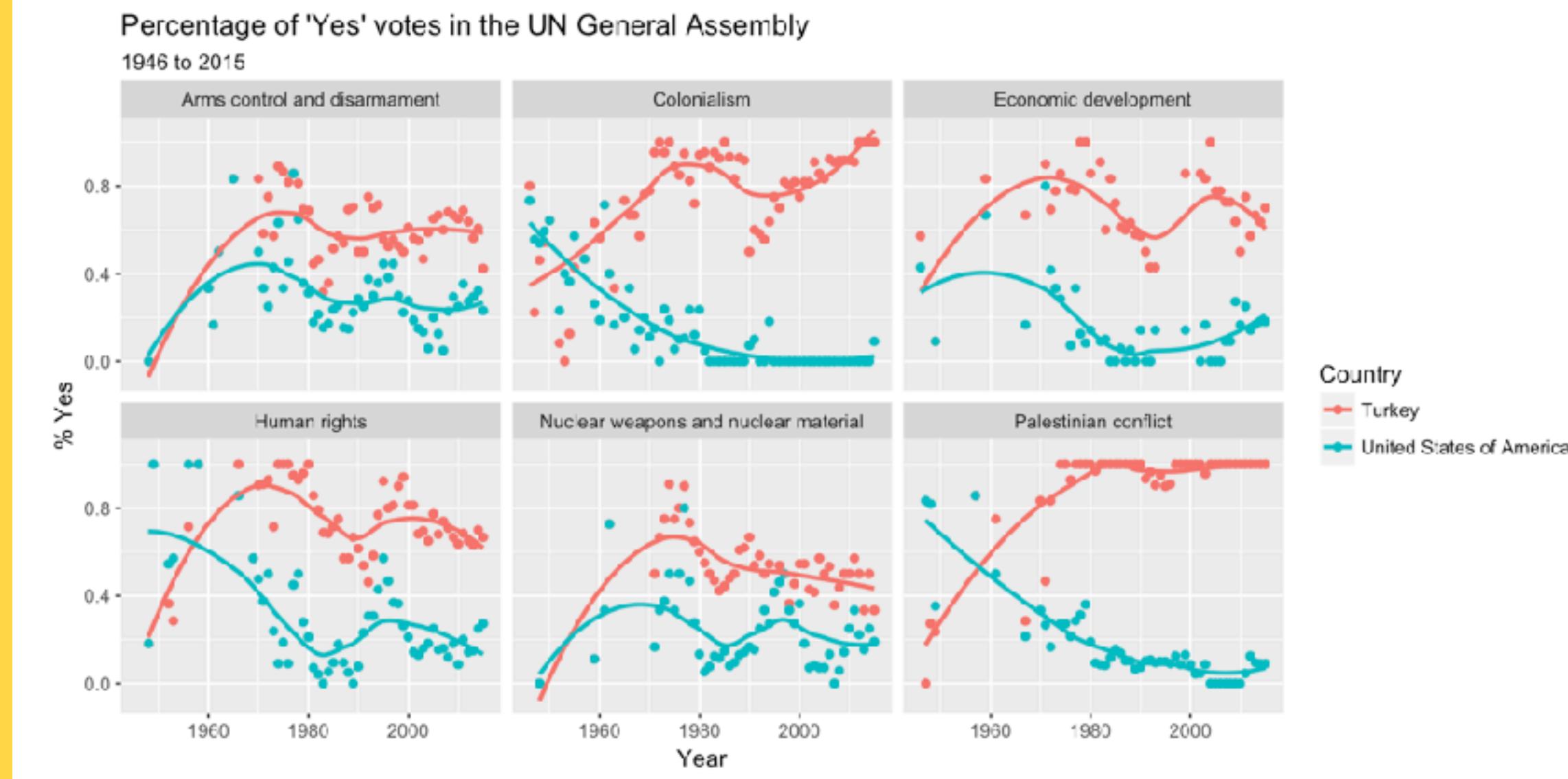
cake



don't start with this

```
class(mtcars$mpg)  
#> [1] "numeric"  
  
mean(mtcars$mpg)  
#> [1] 20.09062  
  
median(mtcars$mpg)  
#> [1] 19.2  
  
sd(mtcars$mpg)  
#> [1] 6.026948
```

instead do this



WHY 🍰 = 📊?

more likely for
students to have
intuition
coming in

easier for students
to catch their own
mistakes

who doesn't
like a good
piece of cake
visualization?

Resources:

Data visualization

- ▶ **Data Science Through Data Visualization in the Intro Course** (Stacey Hancock)
 - ▶ Invited EPoster Session
 - ▶ Sun, 7/28/2019, 8:30 PM - 10:30 PM, CC-Hall C
- ▶ **BoF: Visualizing Visualization** (Jo Hardin)
 - ▶ Mon, 7/29/2019, 12:30 PM - 1:30 PM, CC-SocietyTables Registration Promenade

Which motivates you more to learn how to cook: perfectly chopped onions or ratatouille?



Which motivates you more to learn how to cook: perfectly chopped onions or ratatouille?



skip

baby

steps

3

skip baby steps in
your examples

non-trivial examples can be motivating,
but need to avoid !

How to draw an owl

1.



2.



1. Draw some circles

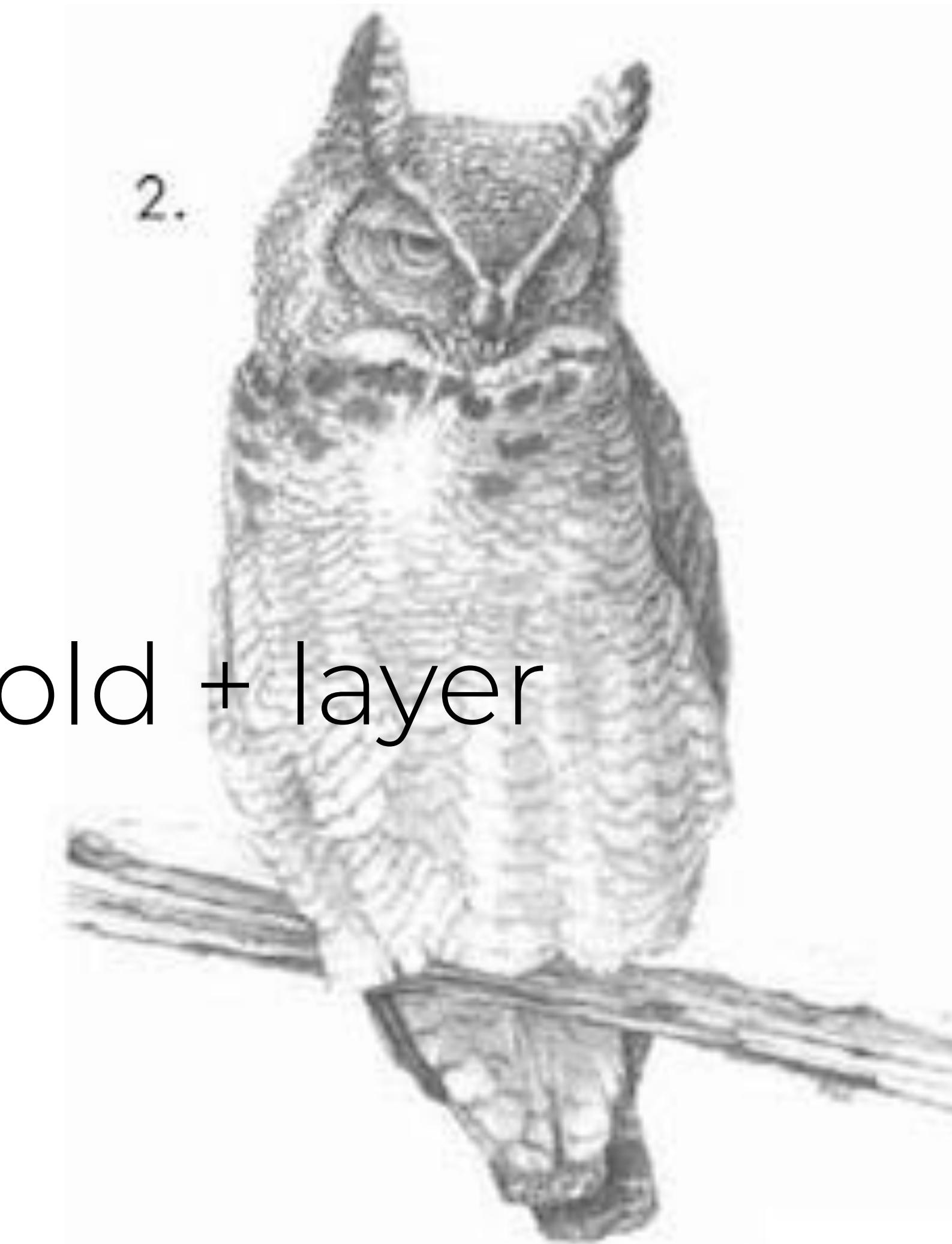
2. Draw the rest of the  owl

How to draw an owl

1.



2.

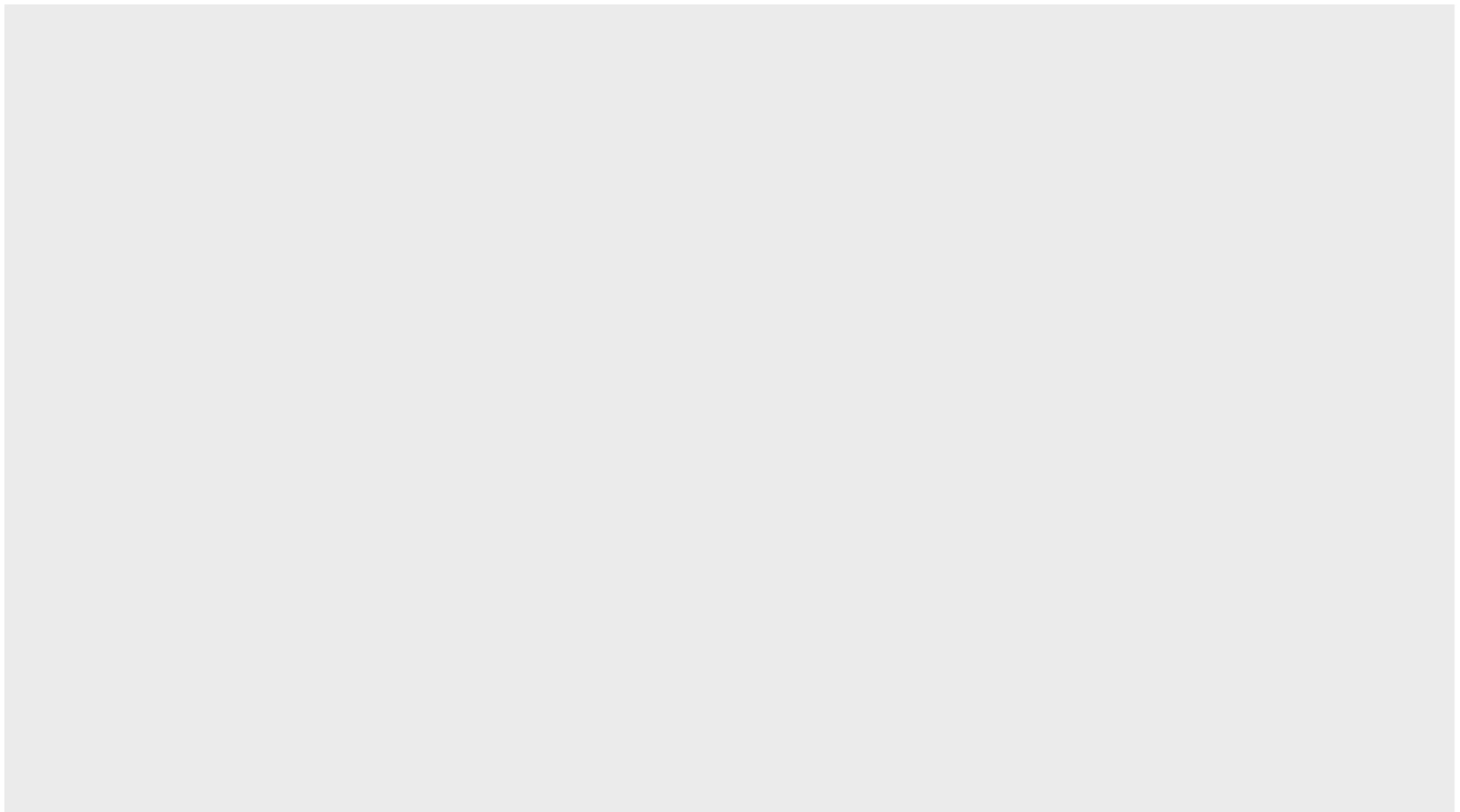


scaffold + layer

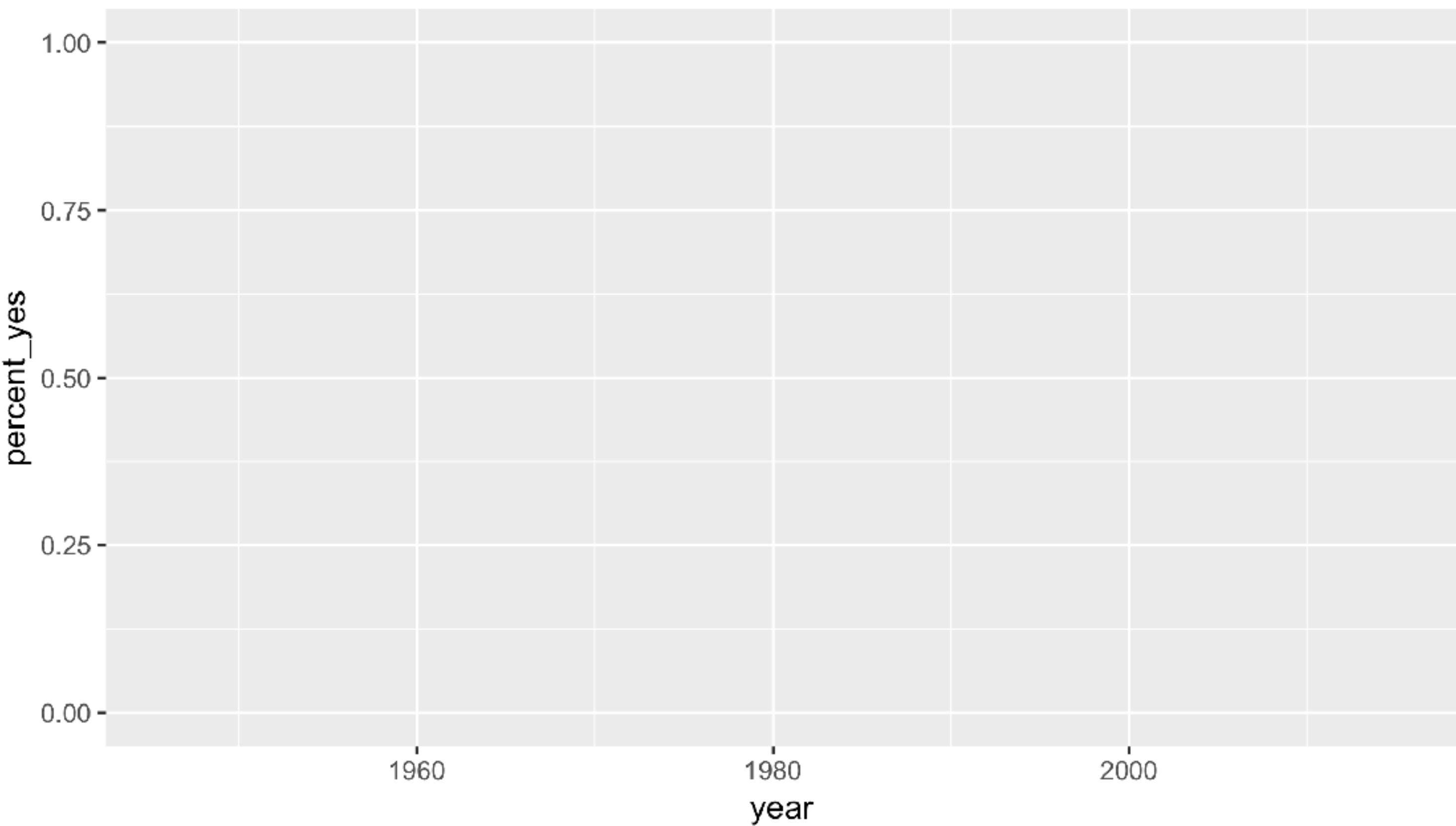
1. Draw some circles

2. Draw the rest of the @#\$% owl

```
ggplot(data = un_votes_joined)
```



```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes))
```



```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes))
```

function(arguments)

often a verb

what to apply that
verb to

```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes))
```

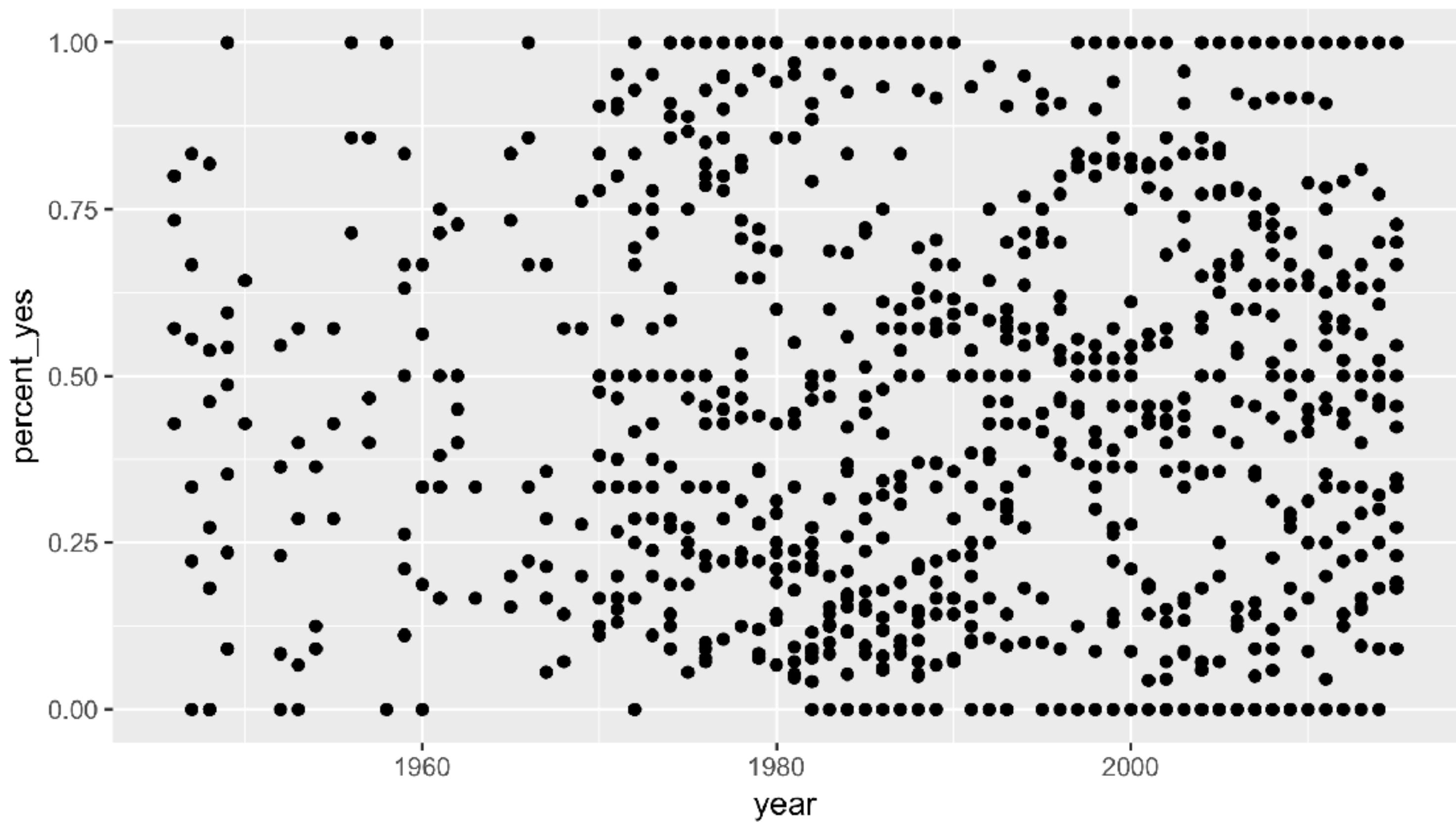
rows =
observations

	country	year	issue	votes	percent_yes
1	Turkey	1946	Colonialism	15	0.80000000
2	Turkey	1946	Economic development	7	0.57142857
3	Turkey	1947	Colonialism	9	0.22222222
4	Turkey	1947	Palestinian conflict	6	0.00000000
5	Turkey	1948	Arms control and disarmament	8	0.00000000
6	Turkey	1948	Colonialism	13	0.46153846
7	Turkey	1948	Human rights	11	0.18181818
8	Turkey	1948	Nuclear weapons and nuclear material	7	0.00000000
9	Turkey	1948	Palestinian conflict	11	0.27272727
10	Turkey	1949	Colonialism	35	0.54285714
11	Turkey	1949	Economic development	11	0.09090909
12	Turkey	1949	Palestinian conflict	17	0.23529412
13	Turkey	1950	Colonialism	14	0.64285714
14	Turkey	1952	Colonialism	12	0.08333333
15	Turkey	1952	Human rights	11	0.36363636
16	Turkey	1953	Colonialism	9	0.00000000
17	Turkey	1953	Human rights	7	0.28571429
18	Turkey	1954	Colonialism	8	0.12500000

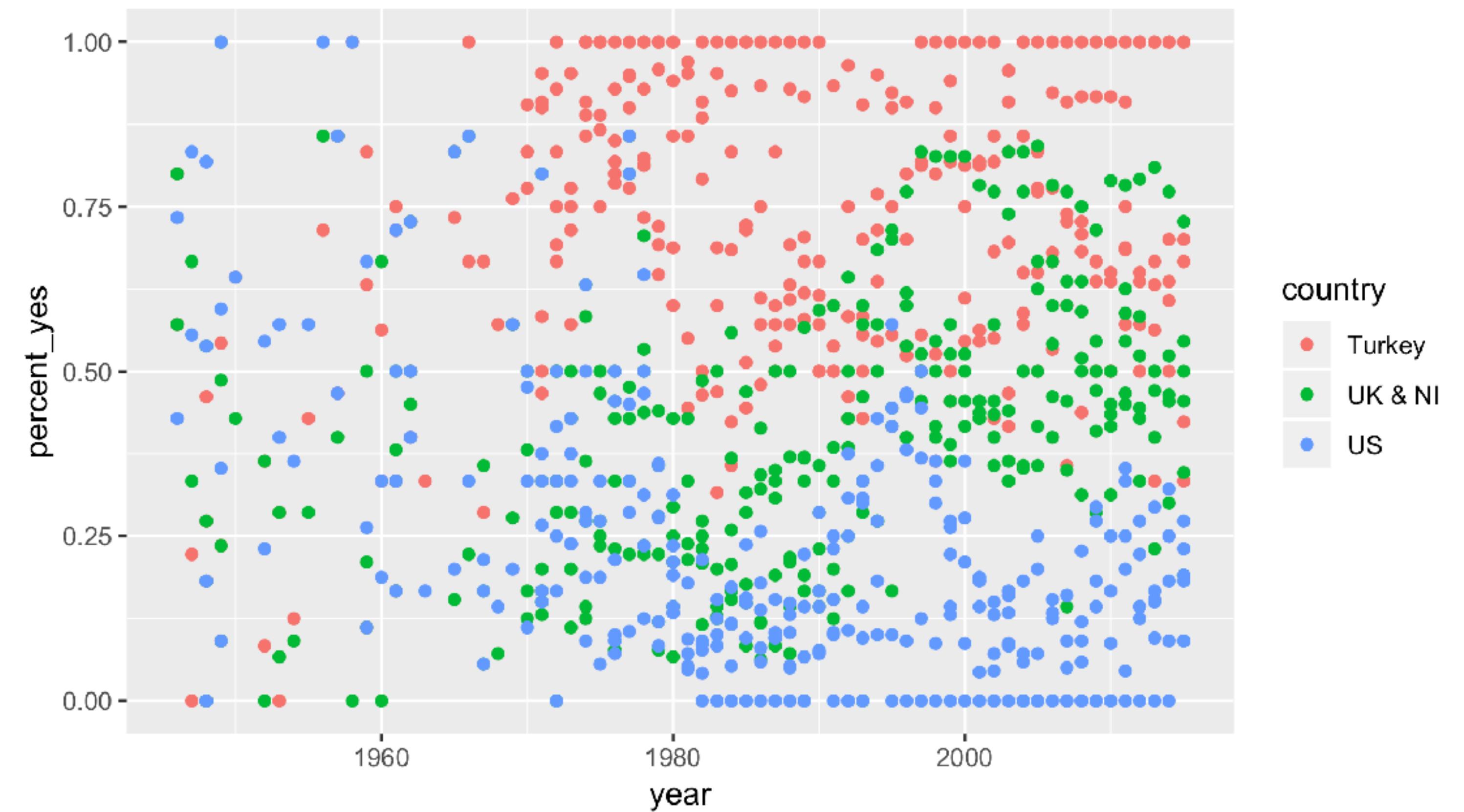
"tidy"
data frame

columns =
variables

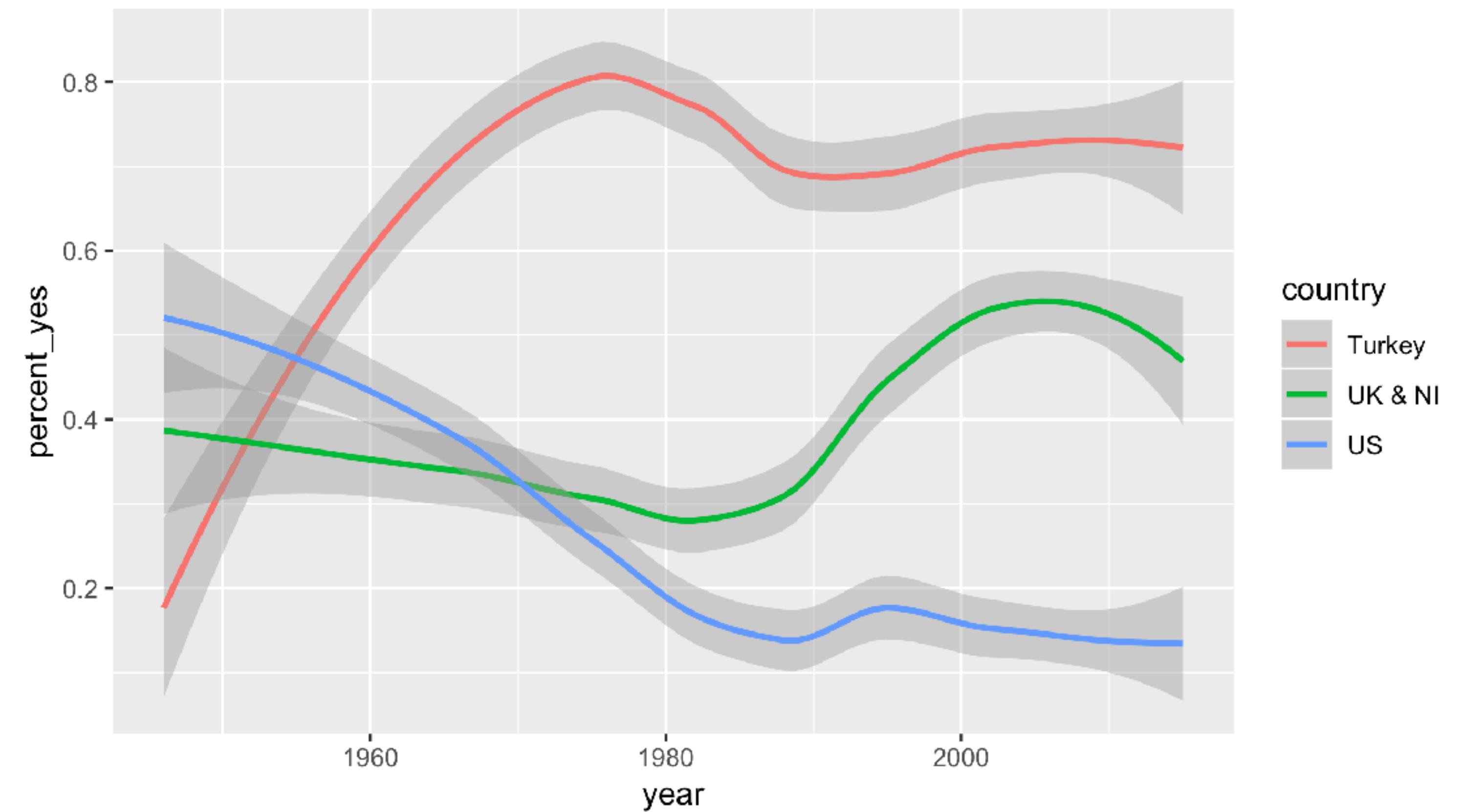
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes)) +  
  geom_point()
```



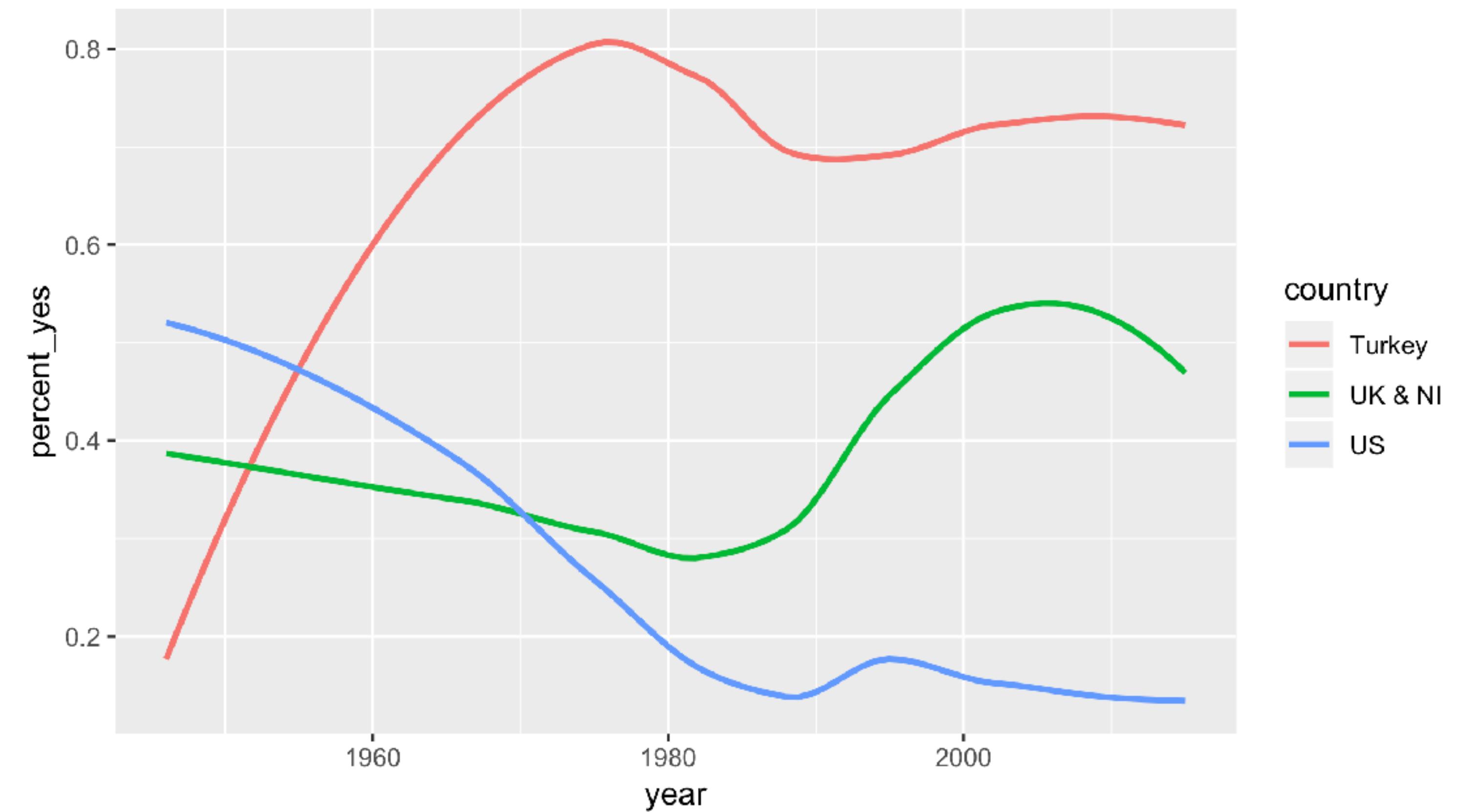
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes, color = country)) +  
geom_point()
```



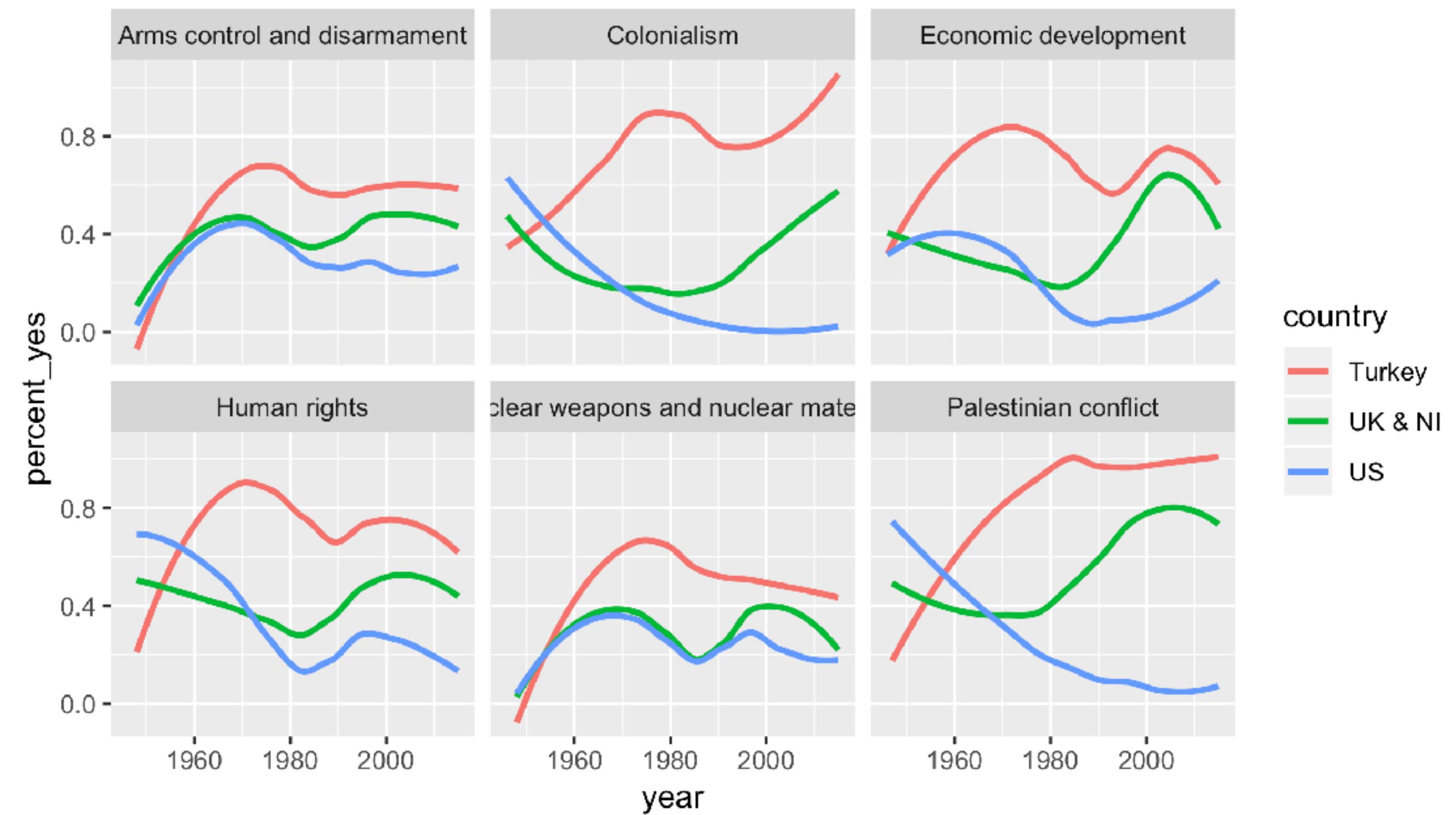
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes, color = country)) +  
  geom_smooth(method = "loess")
```



```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes, color = country)) +  
  geom_smooth(method = "loess", se = FALSE)
```



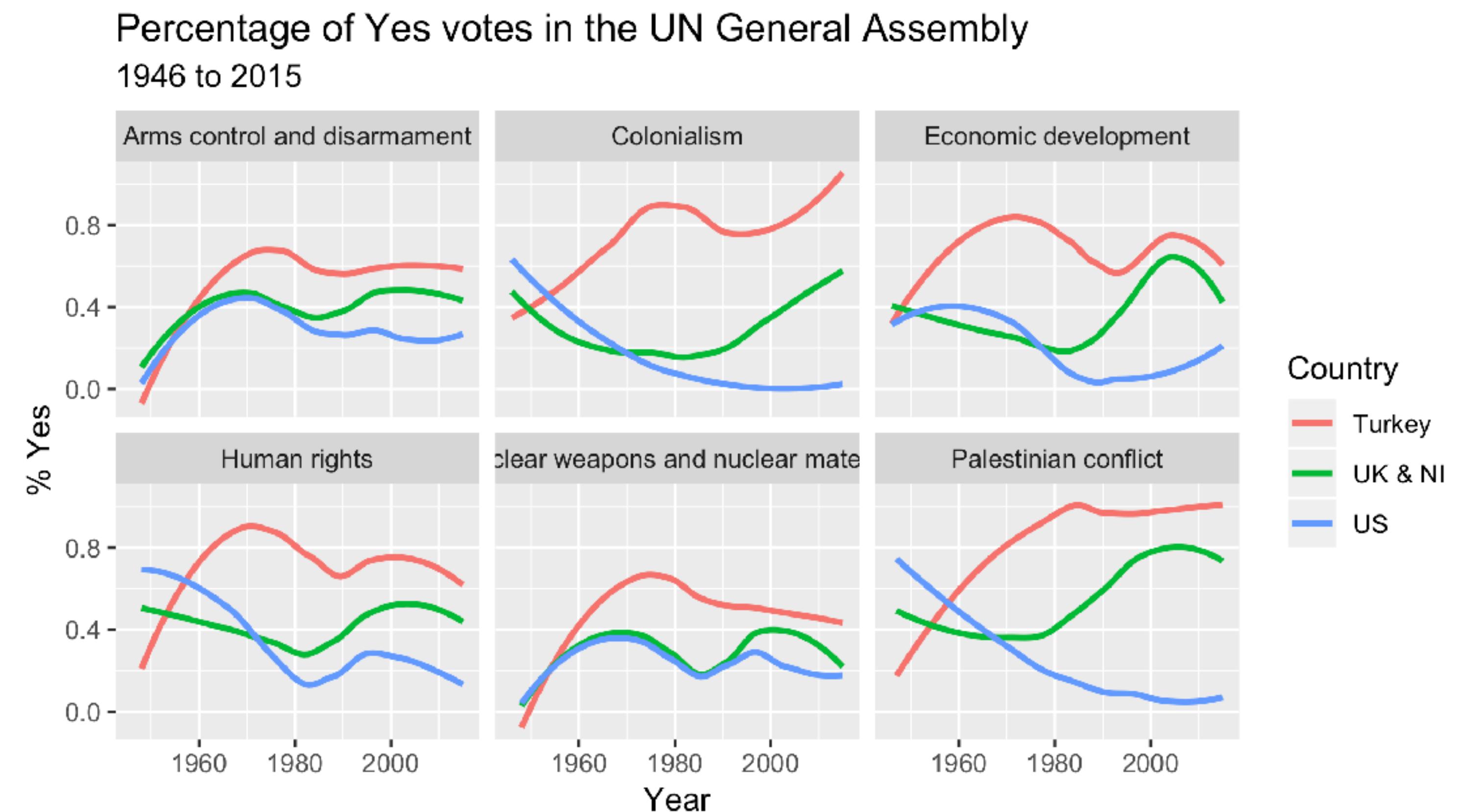
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes, color = country)) +  
  geom_smooth(method = "loess", se = FALSE) +  
  facet_wrap(~ issue)
```



```

ggplot(data = un_votes_joined,
       mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ issue) +
  labs(
    title = "Percentage of 'Yes' votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
)

```



skip baby steps in
your examples

re-insert baby steps
into assessment,
especially formative
assessment

Visualizing data

Data visualization with ggplot2

The data: Star Wars

Scatterplots

Setting aesthetic features

Faceting your visualizations

Data types

Univariate analysis

Start Over

Scatterplots

How can we visualize the relationship between characters' heights and masses? Following the structure of the `ggplot` function that we laid out earlier, we pass `starwars` to the `data` argument, and map `height` and `mass` to the `x` and `y` `aes` thetics, respectively. Then, we specify on the next layer that we would like the data points to be represented by points with `geom_point`.

Fill in the blanks below to create the scatterplot.

Code

Start Over

Solution

Run Code

Submit Answer

```
1 ggplot(data = ___, mapping = aes(x = ___, y = ___)) +  
2   ---  
3
```

Notice the warning that tells us that 28 of the observations have not been graphed, which means that some of the necessary information (height and mass) was missing for those rows.

Your turn!

How would you describe the relationship between height and weight?

- positive and nonlinear
- positive and linear
- negative and nonlinear
- negative and linear

Submit Answer

How many outliers does the graph show?

- 0
- 1
- 2

Submit Answer

Resources:

Curriculum design

- ▶ **JSM 2019 - Building Bridges for Data Science Education**
 - ▶ Panel with statisticians & computer scientists
 - ▶ Tue, July 30, 2019 : 2:00 PM to 3:50 PM, CC-603
- ▶ **JSM 2019 - Data Science: a Three Ring Circus or a Big Tent?**
(Jennifer Bryan and Hadley Wickham)
 - ▶ Response to Donoho's 50 Years of Data Science paper
 - ▶ Mon, July 29, 2019 : 2:00 PM to 3:50 PM, CC-301
- ▶ **JSM 2019 - Teaching Data Science: R, Git, and the Undergraduate Curriculum**
 - ▶ Mon, 7/31/2019, 10:30 AM - 12:20 PM, CC-302
- ▶ **Curriculum Guidelines for Undergraduate Programs in Data Science**
 - ▶ <https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf>



Which is more likely to appeal to someone who has never tried broccoli?





Which is more likely to appeal to someone who has never tried broccoli?



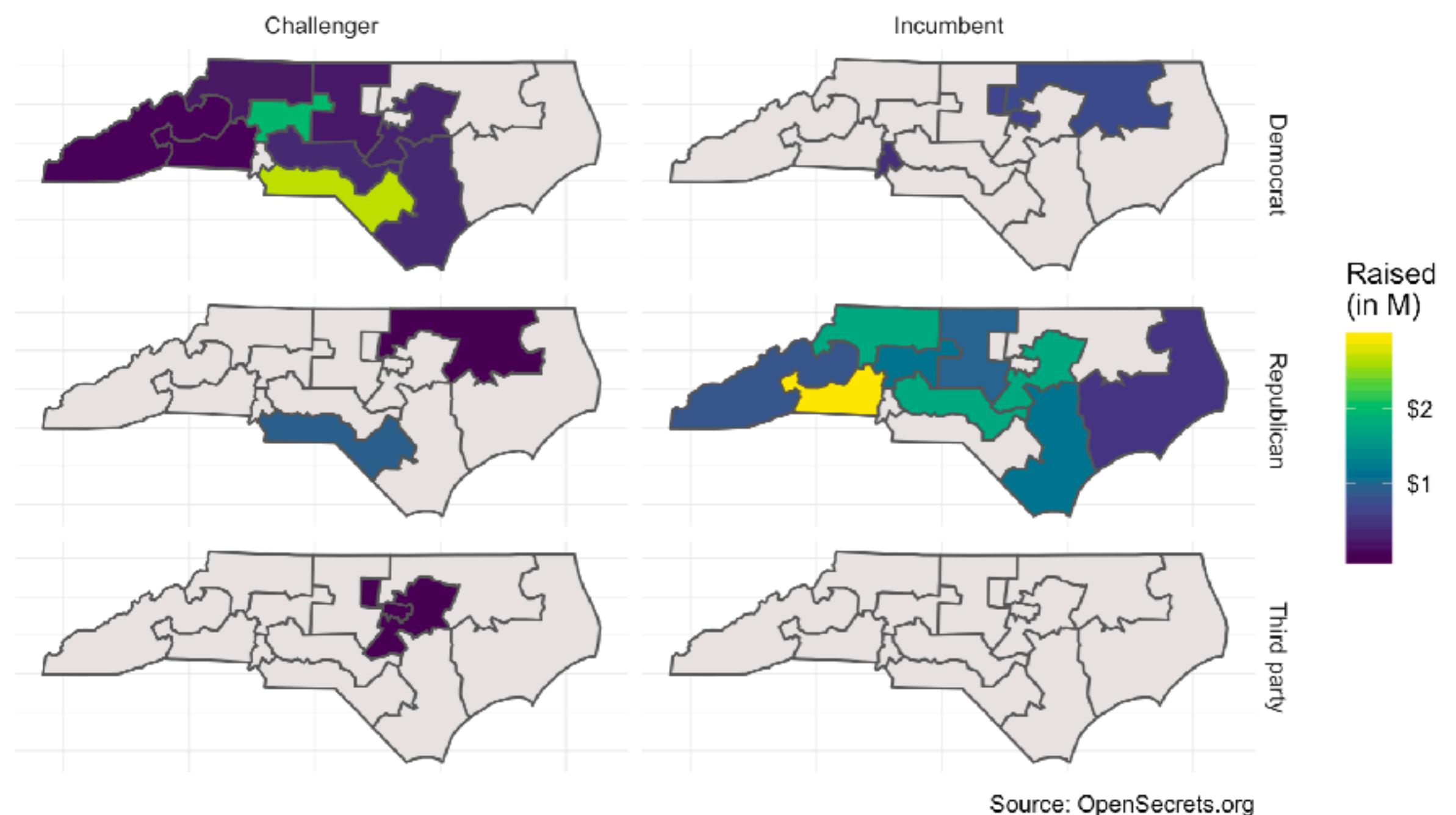
hide
the
veggies



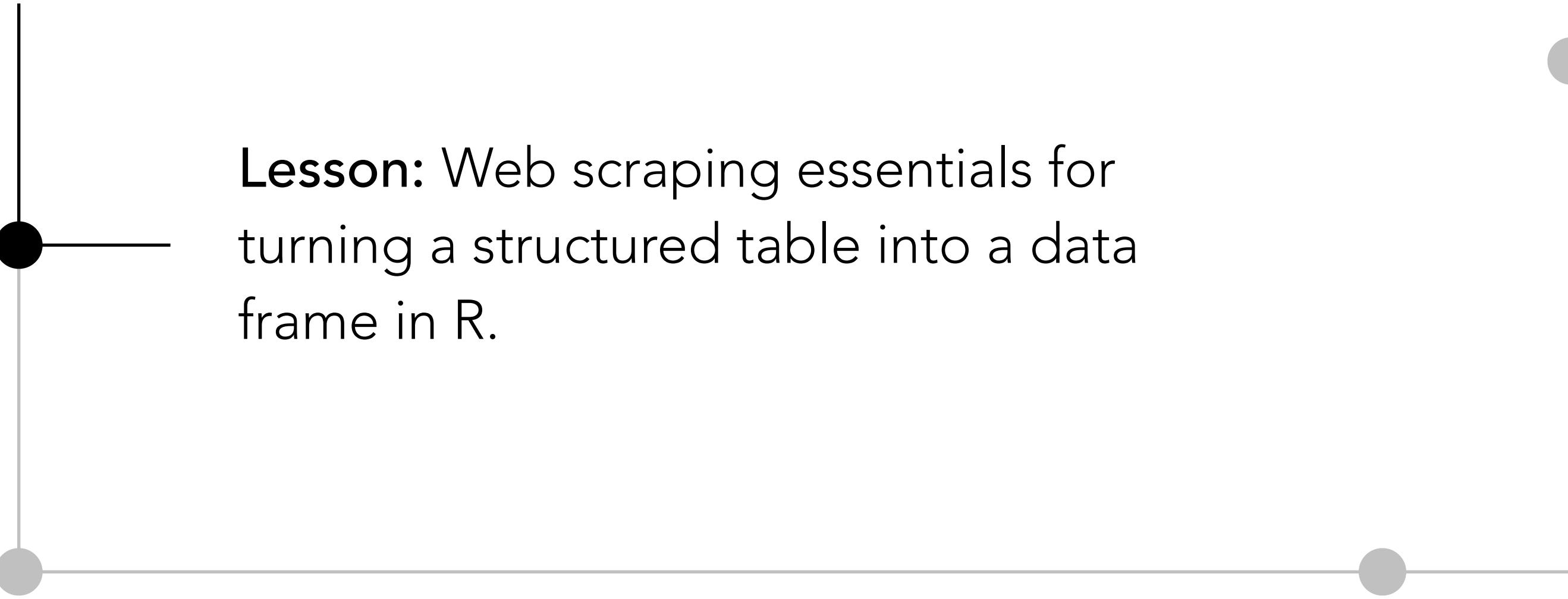
- ▶ Today we go from this to that



Political contributions for 2018 NC Congressional Races
as of 9/30/2018



- ▶ And do so in a way that is easy to replicate for another state



Lesson: Web scraping essentials for
turning a structured table into a data
frame in R.

Lesson: Web scraping essentials for turning a structured table into a data frame in R.

Ex 1: Scrape the table off the web and save as a data frame.

Candidate	Raised	Spent	Cash on Hand	Last Report
G K Butterfield (D) • Incumbent	\$714,219	\$797,700	\$560,416	10/17/2018
Roger Allison (R)	\$28,314	\$27,817	\$497	10/17/2018



#	candidate_info	raised	spent	cash_on_hand	last_report	race
1	G K Butterfield (D) • Incumbent	714219	797700	560416	2018-10-17	North Carolina District 01
2	Roger Allison (R)	28314	27817	497	2018-10-17	North Carolina District 01

Lesson: Web scraping essentials for turning a structured table into a data frame in R.

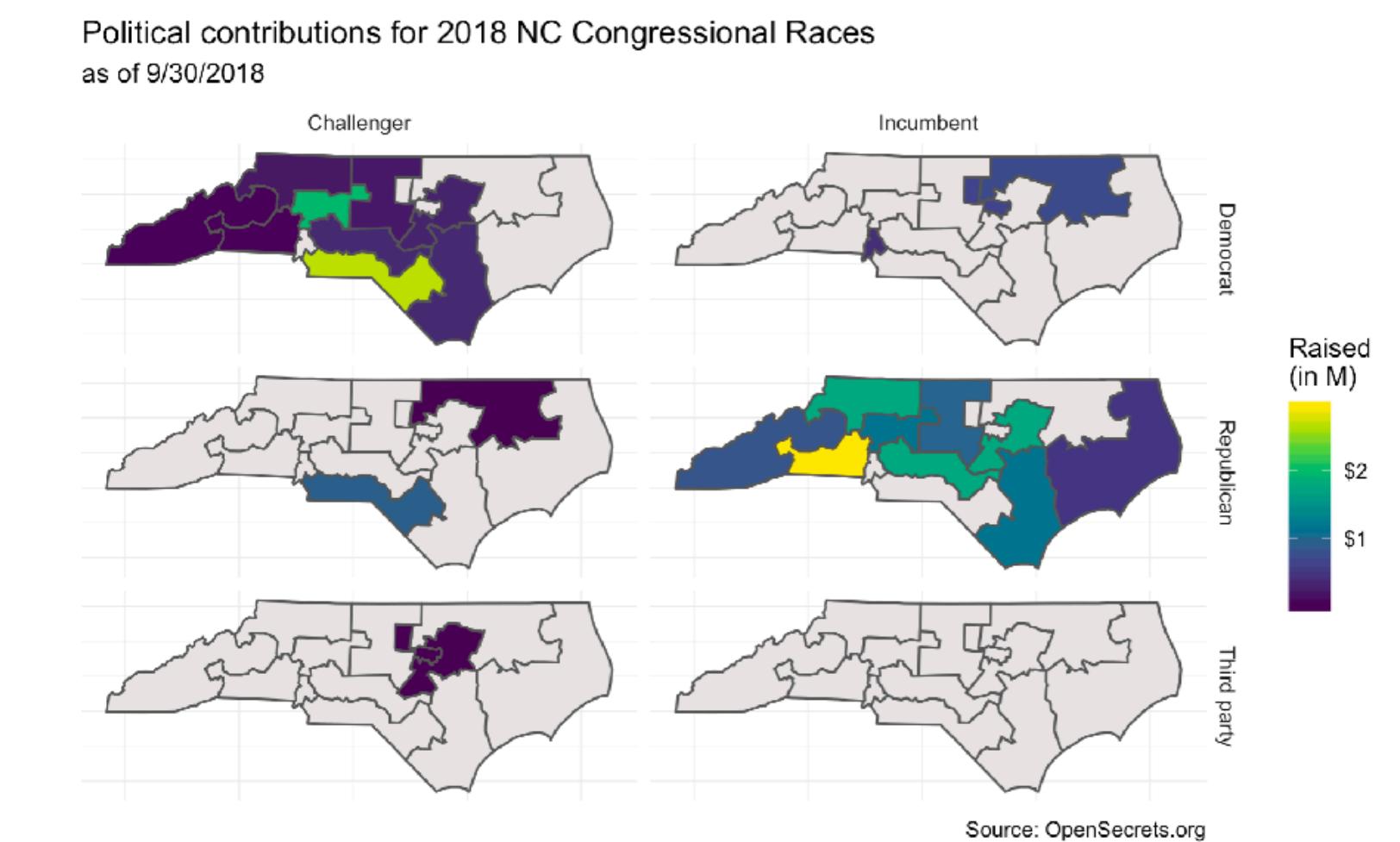
Ex 1: Scrape the table off the web and save as a data frame.

Candidate	Raised	Spent	Cash on Hand	Last Report
G K Butterfield (D) • Incumbent	\$714,219	\$797,700	\$560,416	10/17/2018
Roger Allison (R)	\$28,314	\$27,817	\$497	10/17/2018

↓

#	candidate_info	raised	spent	cash_on_hand	last_report	race
1	G K Butterfield (D) • Incumbent	714219	797700	560416	2018-10-17	North Carolina District 01
2	Roger Allison (R)	28314	27817	497	2018-10-17	North Carolina District 01

Ex 2: What other information do we need represented as variables to make this figure?



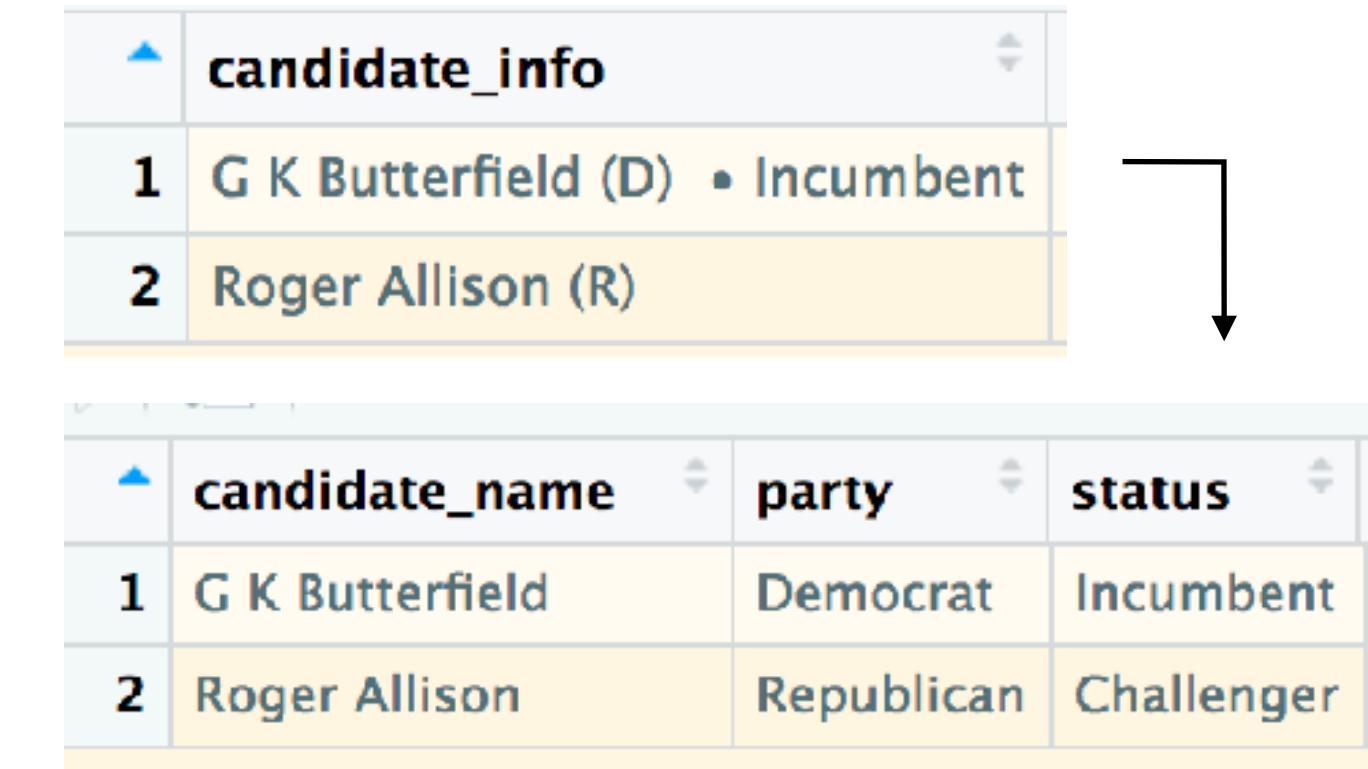
Lesson: Web scraping essentials for turning a structured table into a data frame in R.

Ex 1: Scrape the table off the web and save as a data frame.

Candidate	Raised	Spent	Cash on Hand	Last Report
G K Butterfield (D) • Incumbent	\$714,219	\$797,700	\$560,416	10/17/2018
Roger Allison (R)	\$28,314	\$27,817	\$497	10/17/2018

candidate_info	raised	spent	cash_on_hand	last_report	race
1 G K Butterfield (D) • Incumbent	714219	797700	560416	2018-10-17	North Carolina District 01
2 Roger Allison (R)	28314	27817	497	2018-10-17	North Carolina District 01

Lesson: “Just enough” regex

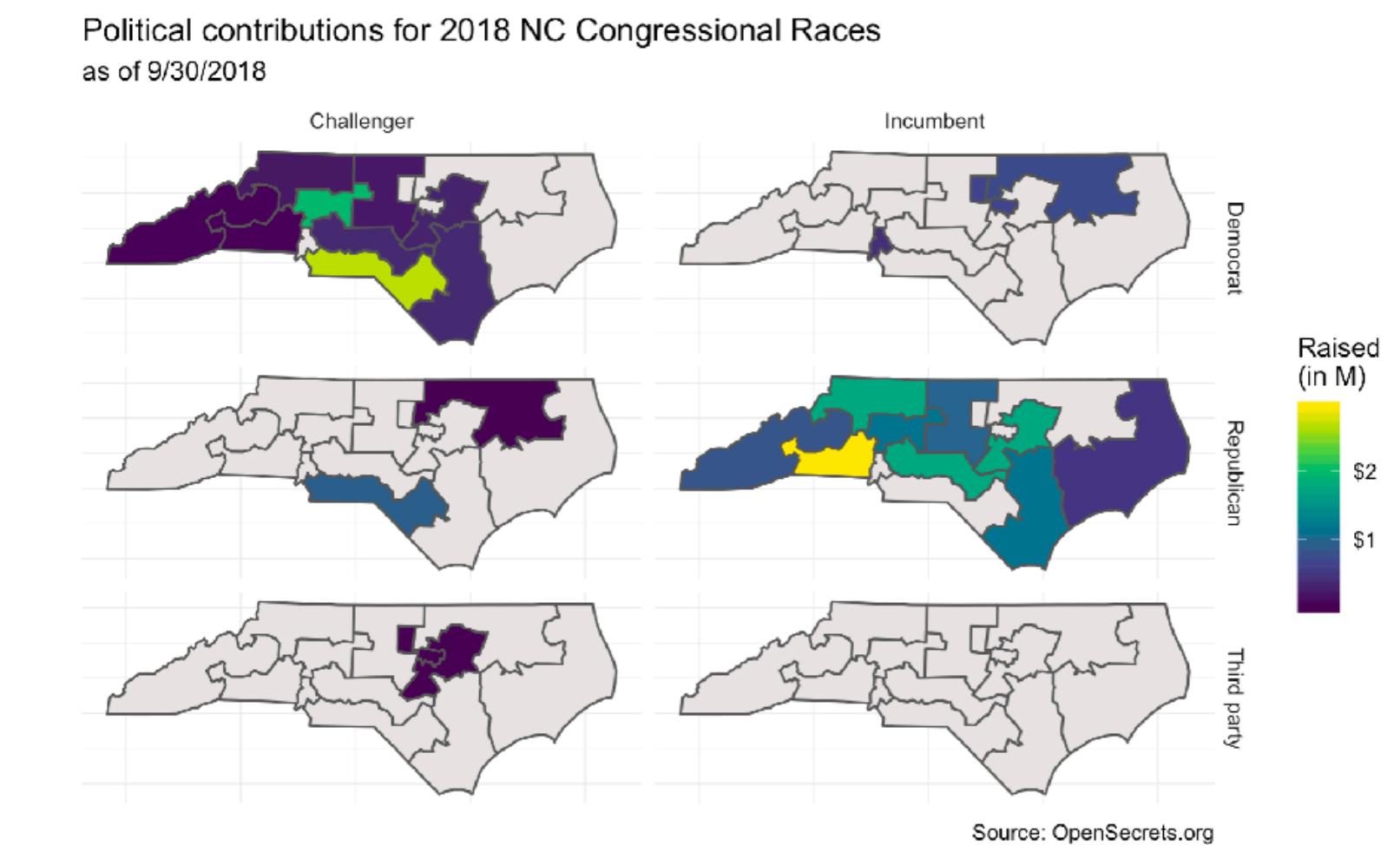


The diagram illustrates the process of cleaning up a raw data frame. At the top, a table titled "candidate_info" contains two rows: "G K Butterfield (D) • Incumbent" and "Roger Allison (R)". An arrow points down to a second table titled "candidate_name", which has columns "candidate_name", "party", and "status". This table also contains two rows: "G K Butterfield" (Democrat, Incumbent) and "Roger Allison" (Republican, Challenger). The "party" and "status" columns are highlighted in yellow, indicating they were derived from the original table using regular expressions.

candidate_info			
1	G K Butterfield (D) • Incumbent		
2	Roger Allison (R)		

candidate_name	party	status
1 G K Butterfield	Democrat	Incumbent
2 Roger Allison	Republican	Challenger

Ex 2: What other information do we need represented as variables to make this figure?





If you are already taking a baking class, which will be easier to venture on to?





If you are already taking a baking class, which will be easier to venture on to?



leverage
the
ecosystem



leverage the
ecosystem
for your students

- ▶ Estimate the difference between the average evaluation score of male and female faculty.

	score	rank	ethnicity	gender	bty_avg
1	<dbl>	<chr>	<chr>	<chr>	<dbl>
2	4.7	tenure track	minority	female	5
3	4.1	tenure track	minority	female	5
4	3.9	tenure track	minority	female	5
5	4.8	tenure track	minority	female	5
6	4.6	tenured	not minority	male	3
7	4.3	tenured	not minority	male	3
8	2.8	tenured	not minority	male	3
9	4.1	tenured	not minority	male	3.33
10	3.4	tenured	not minority	male	3.33
...	4.5	tenured	not minority	female	3.17
463
	4.1	tenure track	minority	female	5.33

concise, but foreign

```
t.test(evals$score ~ evals$gender)

# Welch Two Sample t-test

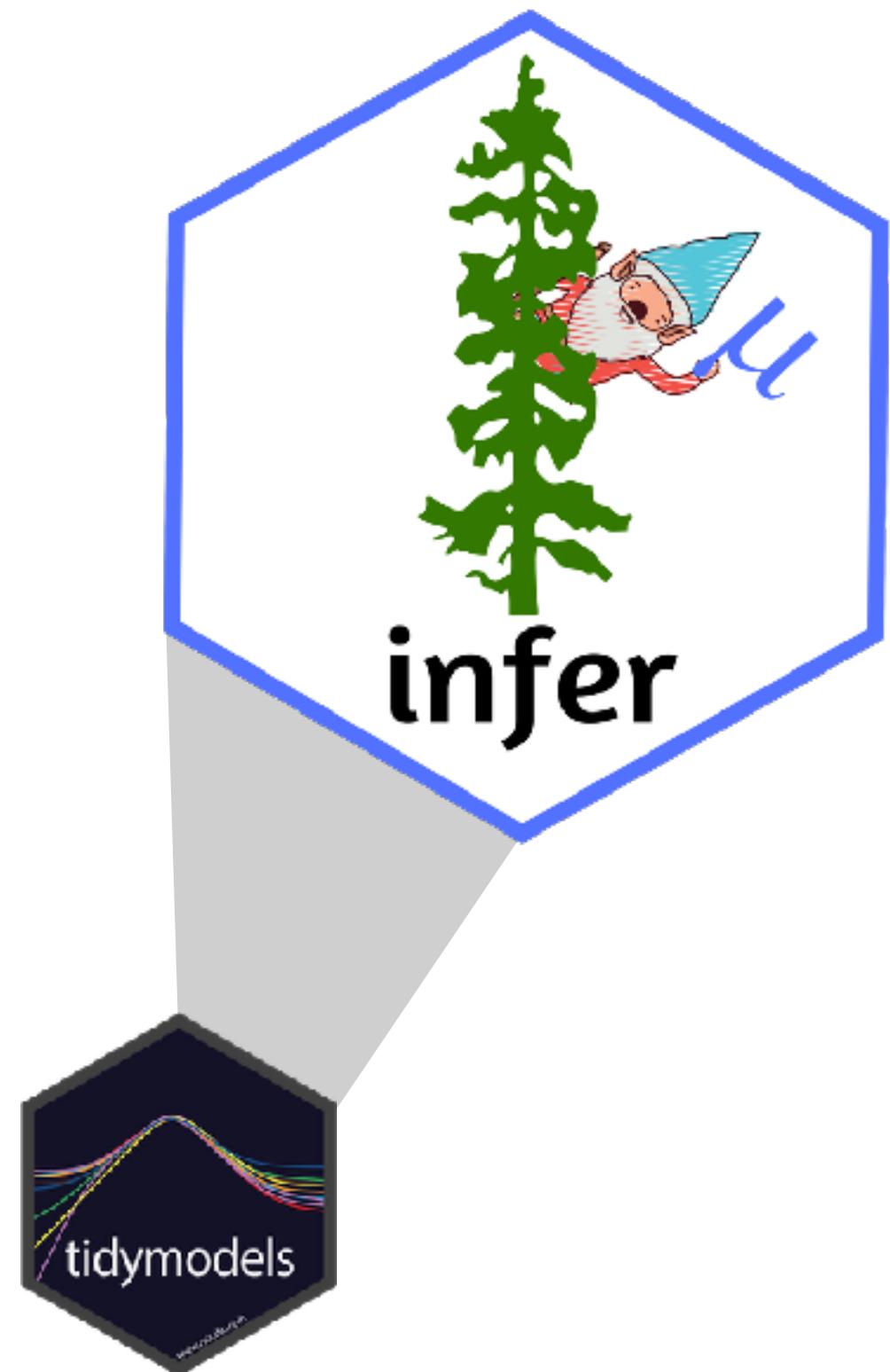
# data: evals$score by evals$gender
# t = -2.7507, df = 398.7, p-value = 0.006218
# alternative hypothesis: true difference in
# means is not equal to 0
# 95 percent confidence interval:
# -0.24264375 -0.04037194
# sample estimates:
# mean in group female    mean in group male
#                 4.092821                4.234328
```

verbose, but familiar

```
library(tidyverse)
library(infer)

evals %>%
  specify(score ~ gender) %>%
  generate(reps = 15000,
            type = "bootstrap") %>%
  calculate(stat = "diff in means",
            order = c("male", "female")) %>%
  summarise(
    l = quantile(stat, 0.025),
    u = quantile(stat, 0.975)
  )

#      l      u
# 0.0410 0.243
```



infer

The objective of this package is to perform statistical inference using an expressive statistical grammar that coheres with the `tidyverse` design framework.

Now part of the `tidymodels` suite of modeling packages.

```
library(tidyverse)  
library(infer)
```

```
evals %>%
```

start with data

```
library(tidyverse)  
library(infer)
```

```
evals %>%  
  specify(score ~ gender)
```

specify the model

```
library(tidyverse)  
library(infer)
```

```
evals %>%  
  specify(score ~ gender) %>%  
  generate(reps = 15000, type = "bootstrap")
```

generate bootstrap samples

```
library(tidyverse)  
library(infer)
```

```
evals %>%  
  specify(score ~ gender) %>%  
  generate(reps = 15000, type = "bootstrap") %>%  
  calculate(stat = "diff in means", order = c("male", "female"))
```

calculate sample statistics

```
library(tidyverse)
library(infer)

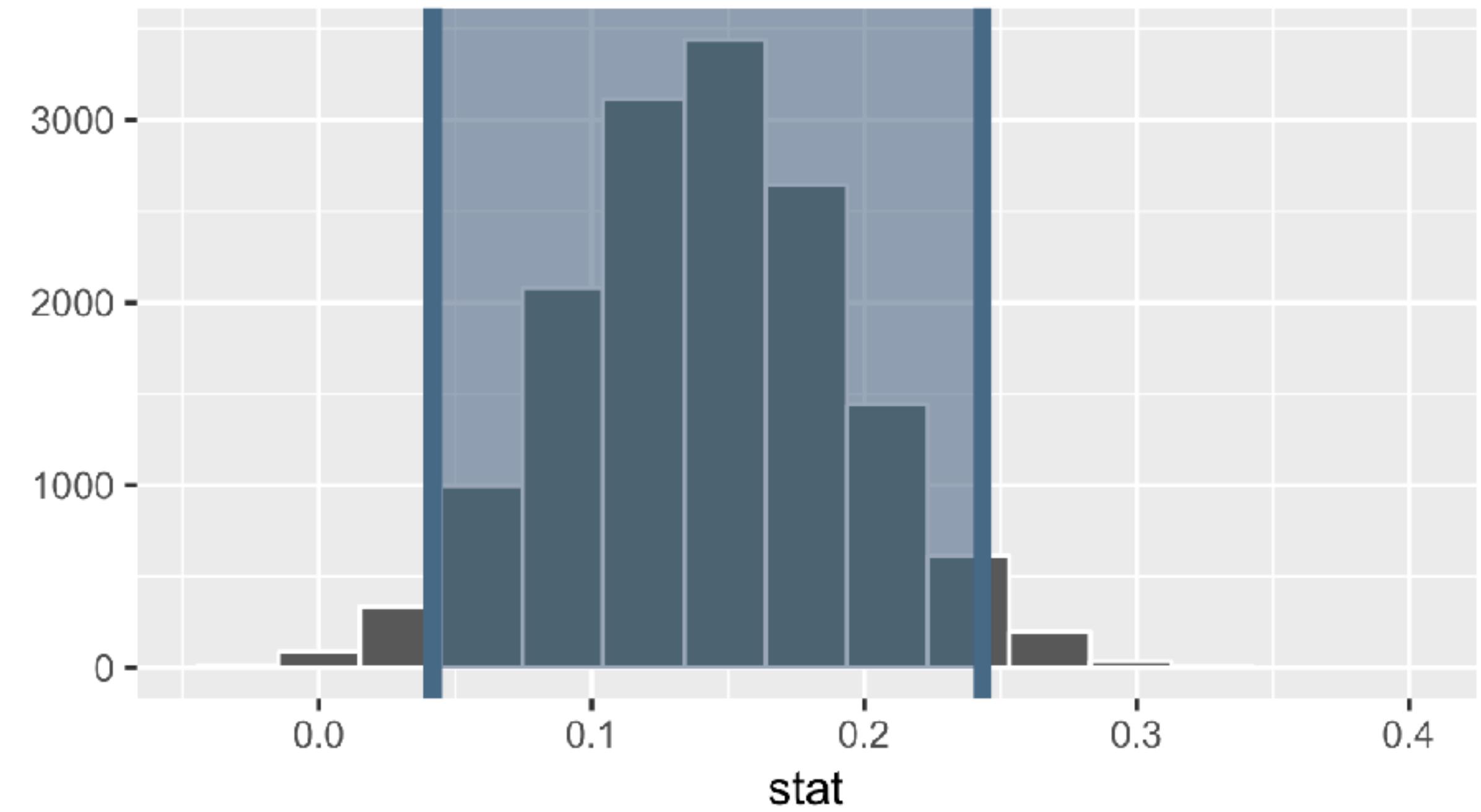
evals %>%
  specify(score ~ gender) %>%
  generate(reps = 15000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("male", "female")) %>%
  summarise(l = quantile(stat, 0.025), u = quantile(stat, 0.975))
```

summarise CI bounds

```
library(tidyverse)
library(infer)

evals %>%
  specify(score ~ gender) %>%
  generate(reps = 15000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("male", "female")) %>%
  summarise(l = quantile(stat, 0.025), u = quantile(stat, 0.975))
```

```
#      l     u
# 0.0410 0.243
```



leverage the
ecosystem
for your students

leverage the
ecosystem
for yourself

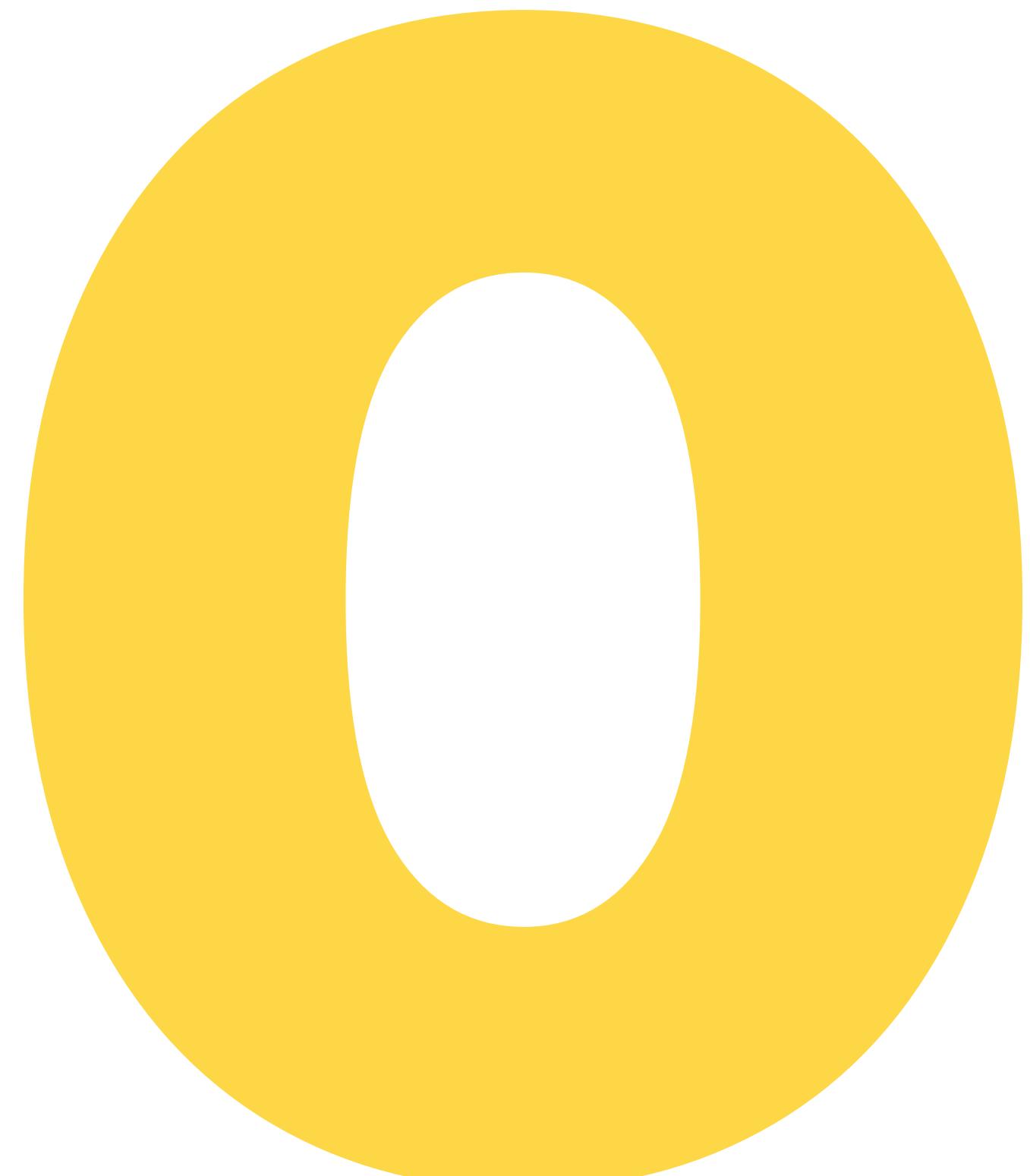
student + instructor



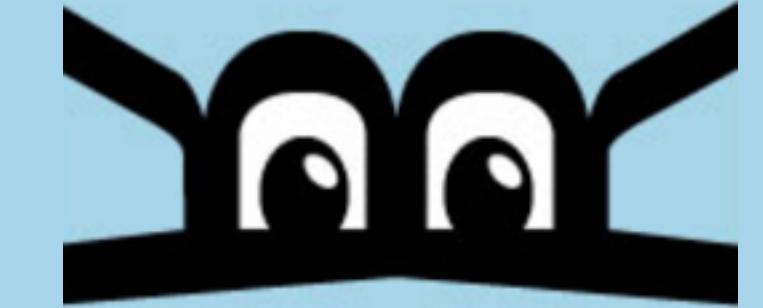
instructor



**you're
not
alone**

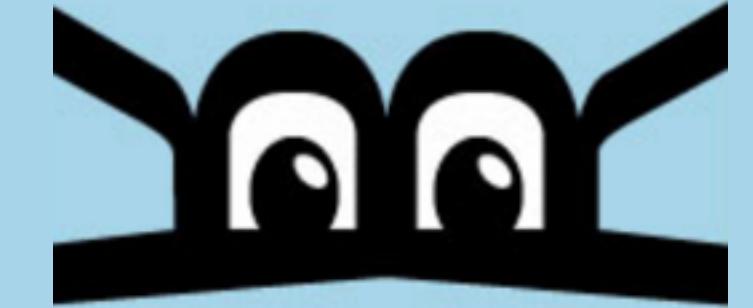


Three questions that keep me up at night...



- 1 What should my students learn?
- 2 How will my students learn best?
- 3 What tools will enhance my students' learning?

Three questions that keep me up at night...

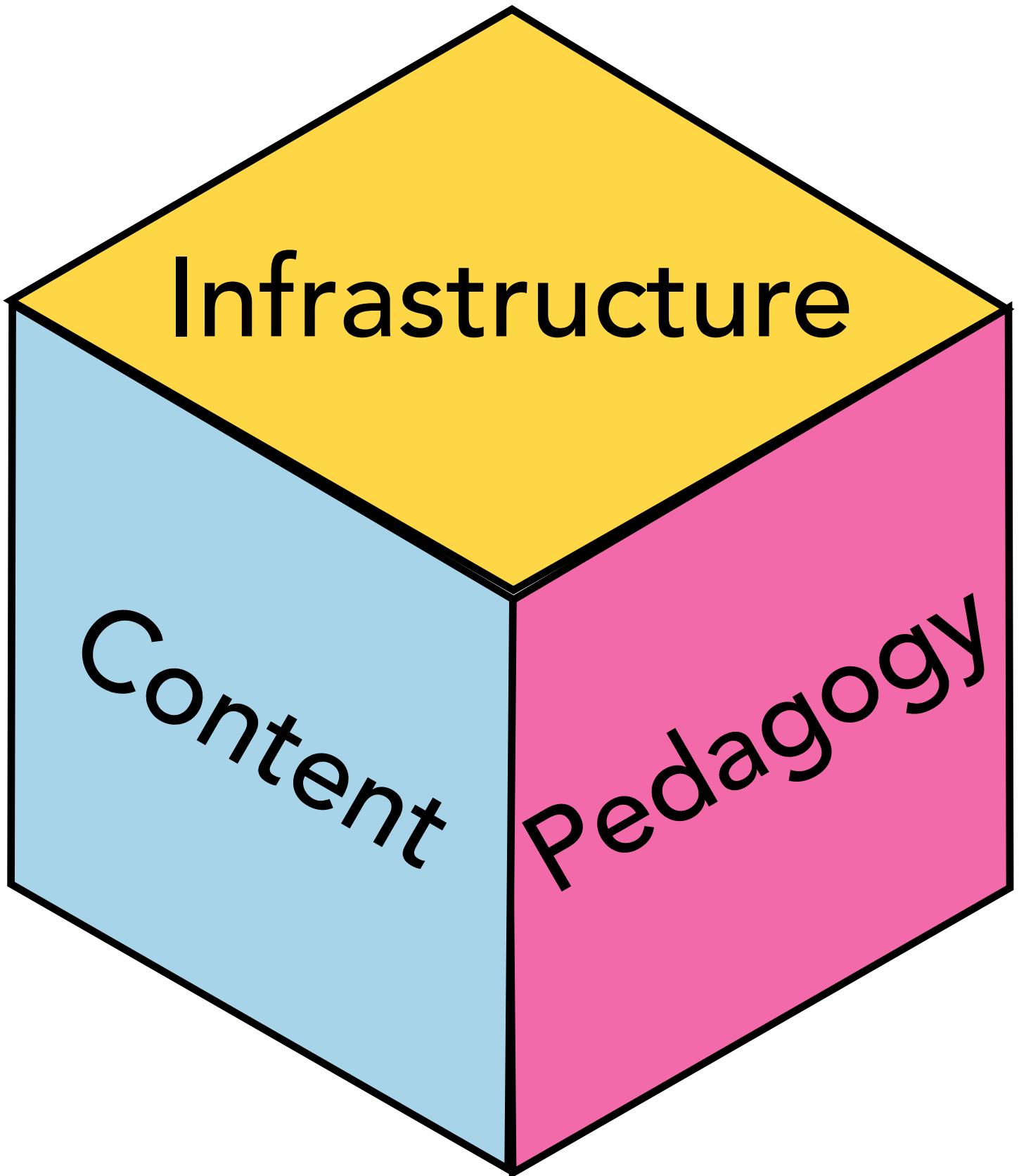


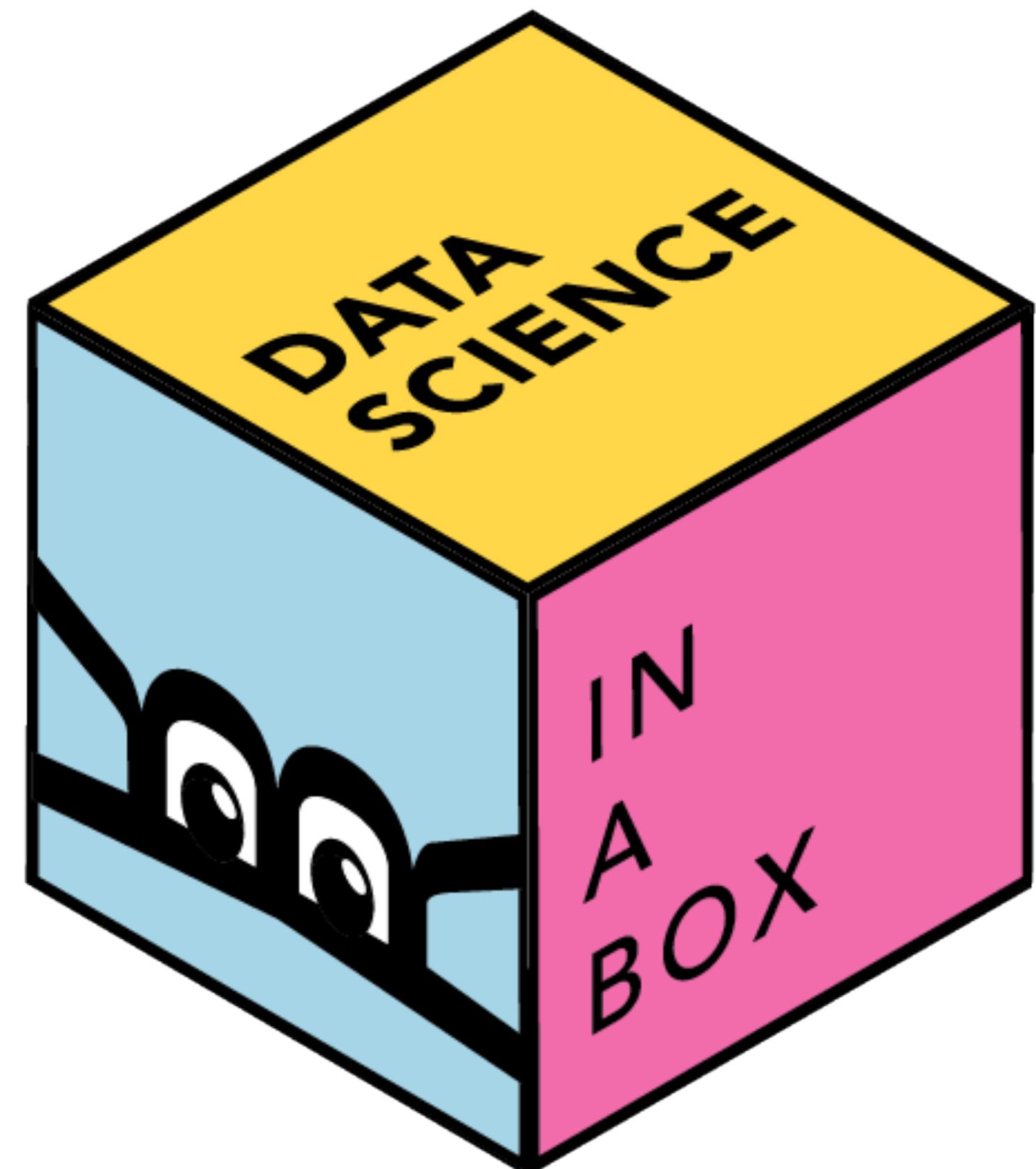
- | | | |
|-----------------------|---|--|
| Content | 1 | What should my students learn? |
| Pedagogy | 2 | How will my students learn best? |
| Infrastructure | 3 | What tools will enhance my students' learning? |

Infrastructure

Content

Pedagogy





The screenshot shows the 'Topics' page of the Data Science in a Box website. At the top left is a logo of a 3D cube labeled 'DATA SCIENCE IN A BOX'. Below it is a search bar and a sidebar with links like 'Hello #dsbox', 'Overview', 'Design principles', 'Topics' (which is currently selected), 'Tech stack', 'Community', 'Course content', 'Infrastructure', and 'Pedagogy'. The main content area has a heading 'Topics' and a sub-section 'Exploring data' which includes 'Visualize', 'Model', 'Infer', and 'Unit 2 - Making rigorous conclusions'. A large pink arrow points diagonally upwards from the bottom left towards the GitHub repository.

The screenshot shows the GitHub repository page for 'rstudio-education / datascience-box'. It displays basic repository statistics: 299 commits, 1 branch, 0 releases, and a master branch. The repository is described as a 'Data Science Course in a Box' with a link to <https://datasciencebox.org/>. The code tab is selected. A large yellow arrow points diagonally upwards from the bottom right towards the Data Science in a Box website.



27
slide
decks



10
application
exercises



10
computing
labs



6
homework
assignments



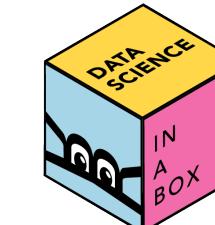
2
take-home
exams



1
open-ended
project



(10)
interactive
tutorials

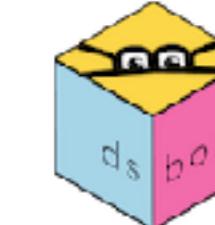


website

datasciencebox.org



repository



package
`dsbox`



Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.



The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

TEACHING INTRO DATA SCIENCE & ASSESSING LEARNING

PREPARING TO TEACH
JSM 2019

NICHOLAS HORTON

AMHERST COLLEGE



@askdrstats



nicholasjhorton



nhorton@amherst.edu

MINE ÇETINKAYA-RUNDEL

UNIVERSITY OF EDINBURGH + DUKE + RSTUDIO



mine-cetinkaya-rundel

cetinkaya.mine@gmail.com

