

01

# curriculum design

teaching  
data  
science



Imagine you're new to baking,  
and you're in a baking class.  
I'm going to present two  
options for starting the class.  
Which one gives you **better**  
**sense** of the final product?

# Pineapple and Coconut Sandwich cake



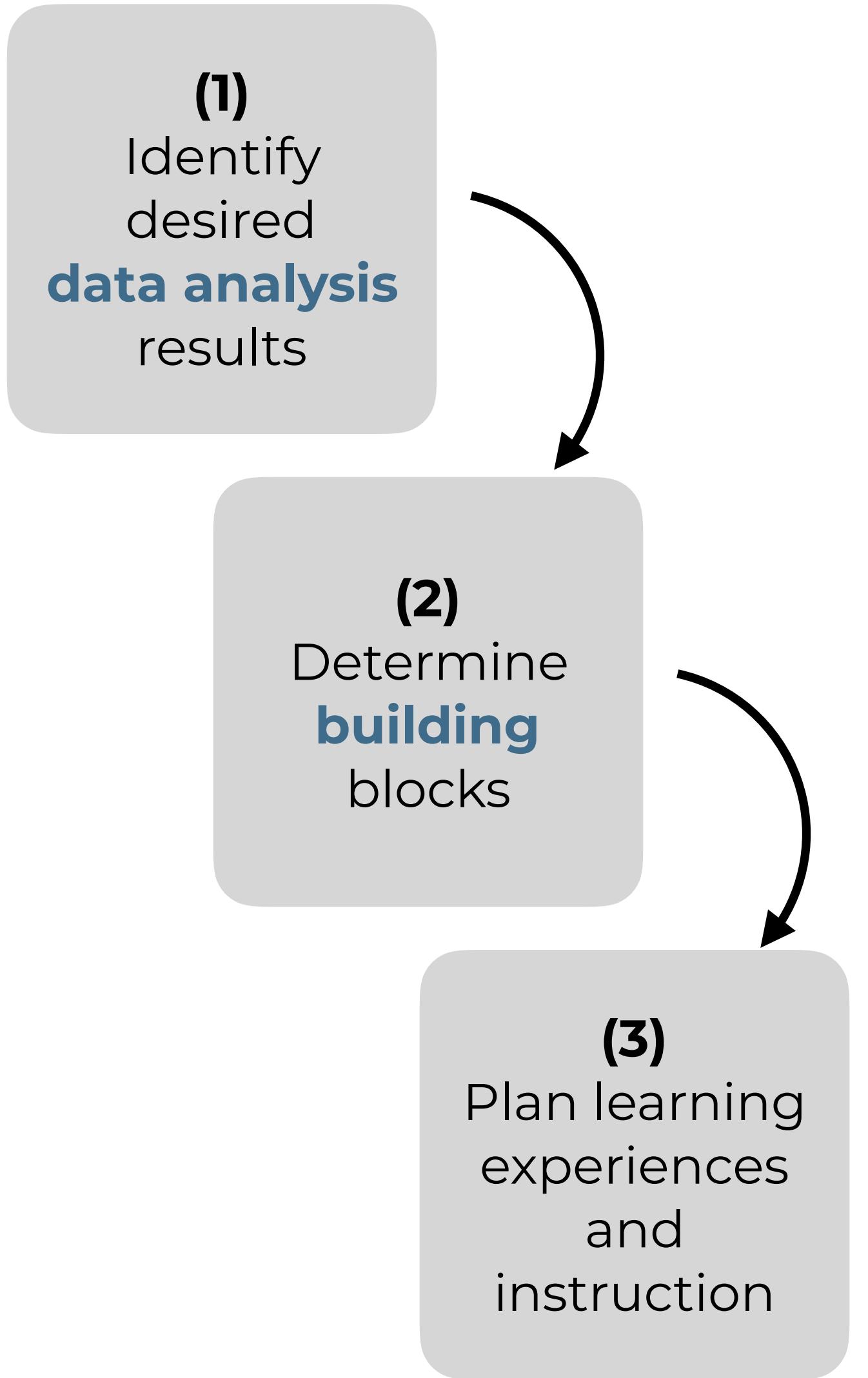
# Pineapple and coconut sandwich cake



# design foundations



# 1 backwards design



# 2 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.

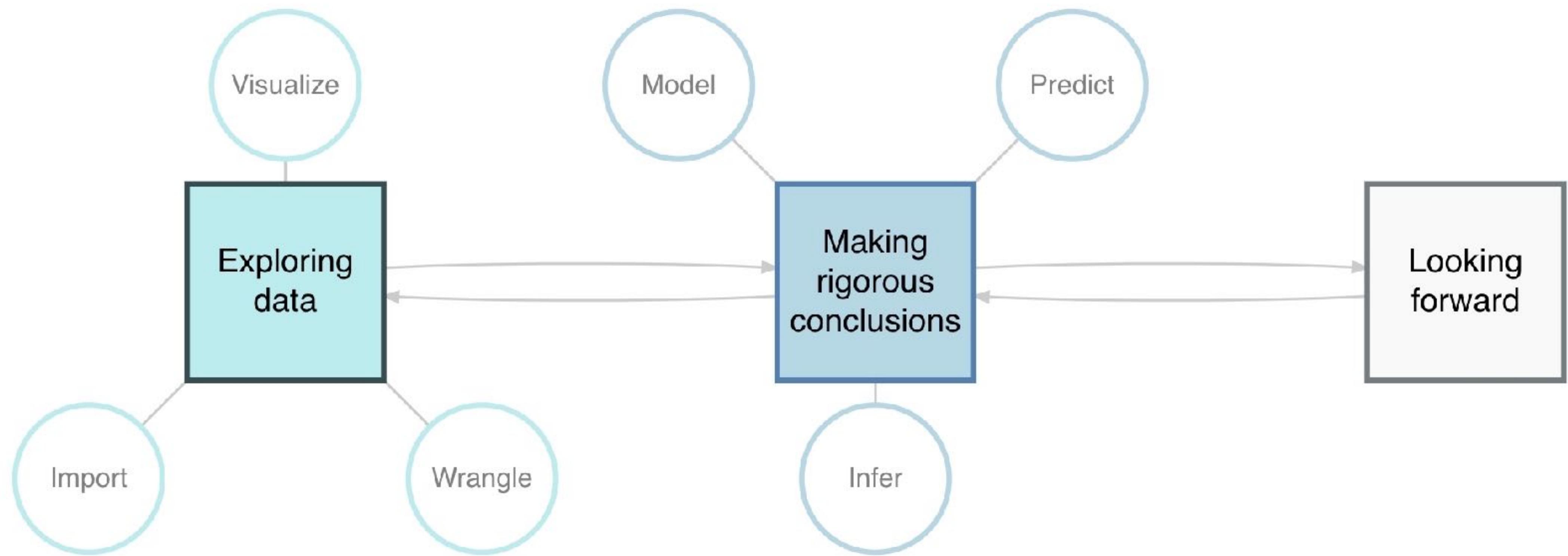
## 6. Use assessments to improve and evaluate student learning.

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 Not just inference & modeling, also data importing, cleaning, preparation, exploration, & 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 and transforming,  
web scraping and 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

# design principles



# Q

Which kitchen would you  
rather bake a cake?

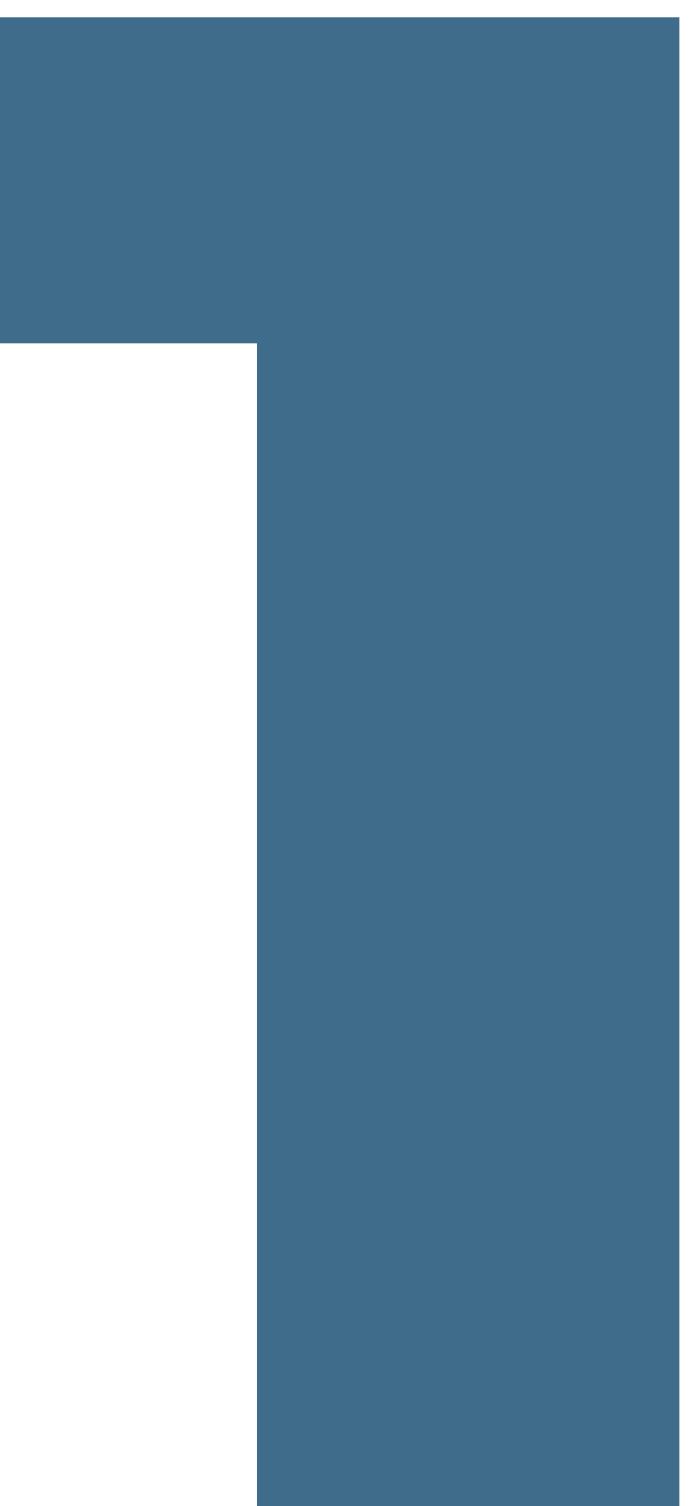


# Q

Which kitchen would you  
rather bake a cake?



cherish  
day  
one





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

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



Data

Analysis

References

Appendix

## UN Votes

Mine Çetinkaya-Rundel

2018-09-26

Let's take a look at the voting history of countries in the United Nations General Assembly. We will be using data from the `unvotes` package. Additionally, we will make use of the `tidyverse` and `lubridate` packages for the analysis, and the `DT` package for interactive display of tabular output.

### Data

The `unvotes` package provides three datasets we can work with: `un_roll_calls`, `un_roll_call_issues`, and `un_votes`. Each of these datasets contains a variable called `roid`, the roll call id, which can be used as a unique identifier to join them with each other.

- The `un_votes` dataset provides information on the voting history of the United Nations General Assembly. It contains one row for each country-vote pair.

```
un_votes
```

```
## # A tibble: 738,764 x 4
##   roid country      country_code vote
##   <int> <chr>        <chr>     <fct>
## 1 3 United States of America US      yes
## 2 3 Canada          CA      no
## 3 3 Cuba            CU      yes
## 4 3 Haiti           HT      yes
## 5 3 Dominican Republic DO      yes
## 6 3 Mexico          MX      yes
## 7 3 Guatemala       GT      yes
## 8 3 Honduras         HN      yes
## 9 3 El Salvador      SV      yes
## 10 3 Nicaragua        NI     yes
## # ... with 738,754 more rows
```

- The `un_roll_calls` dataset contains information on each roll call vote of the United Nations General Assembly.

```
un_roll_calls
```

```
## # A tibble: 5,429 x 9
##   roid session importantvote date      unres amend para short descr
##   <int> <dbl> <dbl> <date>    <chr> <dbl> <dbl> <dbl> <dbl>
## 1 3     1     1     0 1946-01-01 8/1/66    1     0 AMEN_ TO ADD_
## 2 4     1     1     0 1946-01-02 8/1/79    0     0 SECU_ TO ADD_
## 3 5     1     1     0 1946-01-04 8/1/98    0     0 VOTI_ "TO AD_
```

# Q

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 (15 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



# Q

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 (15 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



**start  
with  
cake**



Ingredients	Directions
<b>For the Cake:</b>	
16 ounces plain or <b>toasted sugar</b> (about: 2 1/4 cups; 455g)	1. <b>For the Cake:</b> 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 <a href="#">here</a> ). 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.
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 (15 tablespoons; 225g), soft but cool, about 60°F (16°C)	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 large eggs, brought to about 65°F (18°C)	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.
1/2 ounce vanilla extract (about 1 tablespoon; 15g)	
16 ounces whole milk (about 2 cups; 455g), brought to about 65°F (18°C)	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
16 ounces all-purpose flour (about: 3 1/2 cups, spooned; 455g)	



- Declare the following variables
- Then, determine the class of each variable

```
# Declare variables
```

```
x <- 8
```

```
y <- "monkey"
```

```
z <- FALSE
```

```
class(x)
```

```
#> [1] "numeric"
```

```
class(y)
```

```
#> [1] "character"
```

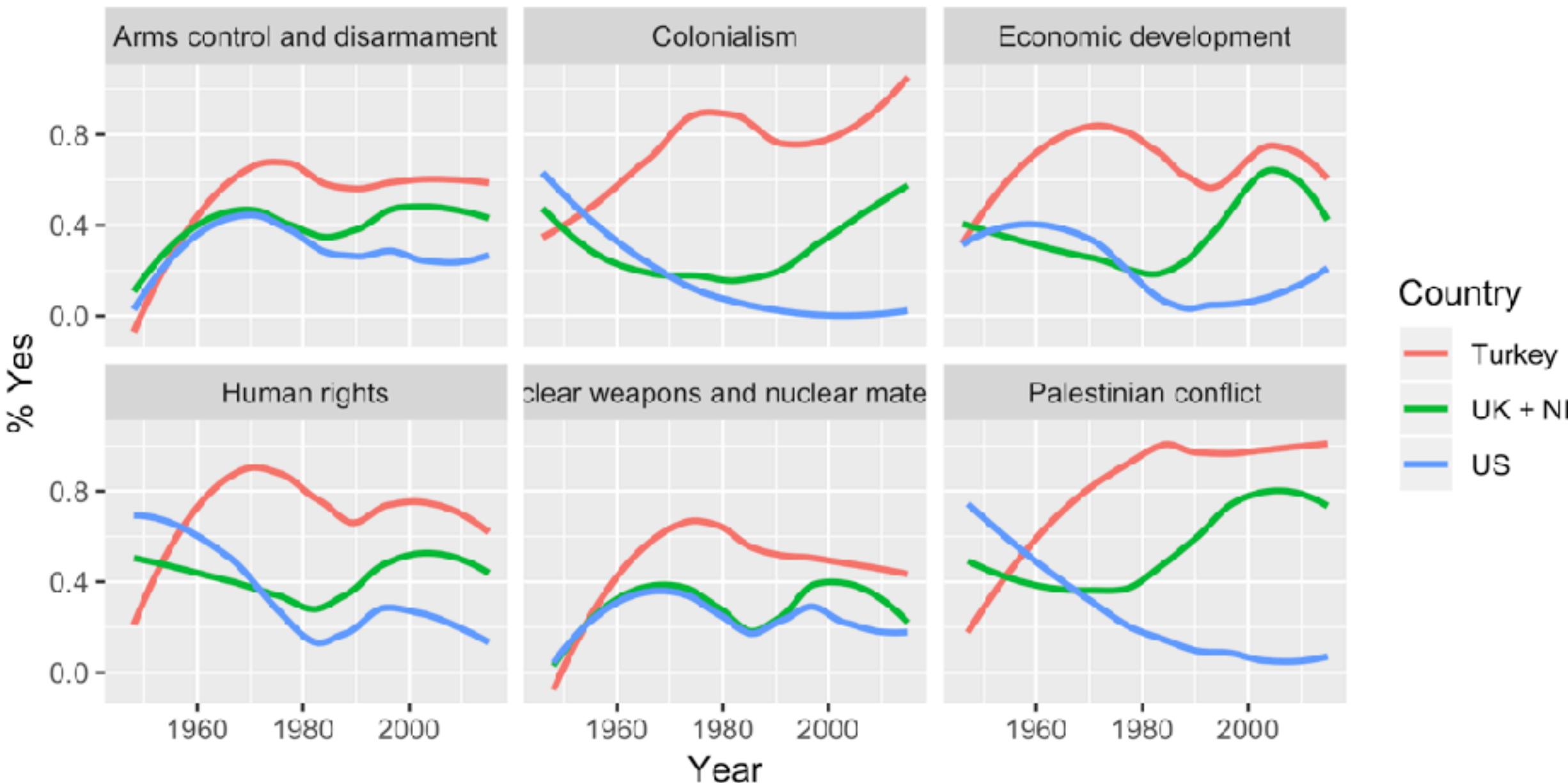


[bit.ly/teach-as-latinR](http://bit.ly/teach-as-latinR)

```
#> [1] "logical"
```

- Open today's demo project
- Knit the document and discuss the results with your neighbor

Percentage of Yes votes in the UN General Assembly  
1946 to 2015



- Then, change Turkey to a different country, and plot again

with great examples,  
comes a great amount of code...

but let's focus on the task at hand...

- Open today's demo project
- Knit the document and discuss the results with your neighbor
- Then, change Turkey to a different country, and plot again

```
un_votes %>%  
  filter(country %in% c("UK & NI", "US", "Turkey")) %>%  
  inner_join(un_roll_calls, by = "rcid") %>%  
  inner_join(un_roll_call_issues, by = "rcid") %>%  
  group_by(country, year = year(date), issue) %>%  
  summarize(  
    votes = n(),  
    percent_yes = mean(vote == "yes")  
  ) %>%  
  filter(votes > 5) %>% # only use records where there are more than 5 votes  
  ggplot(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"  
)
```

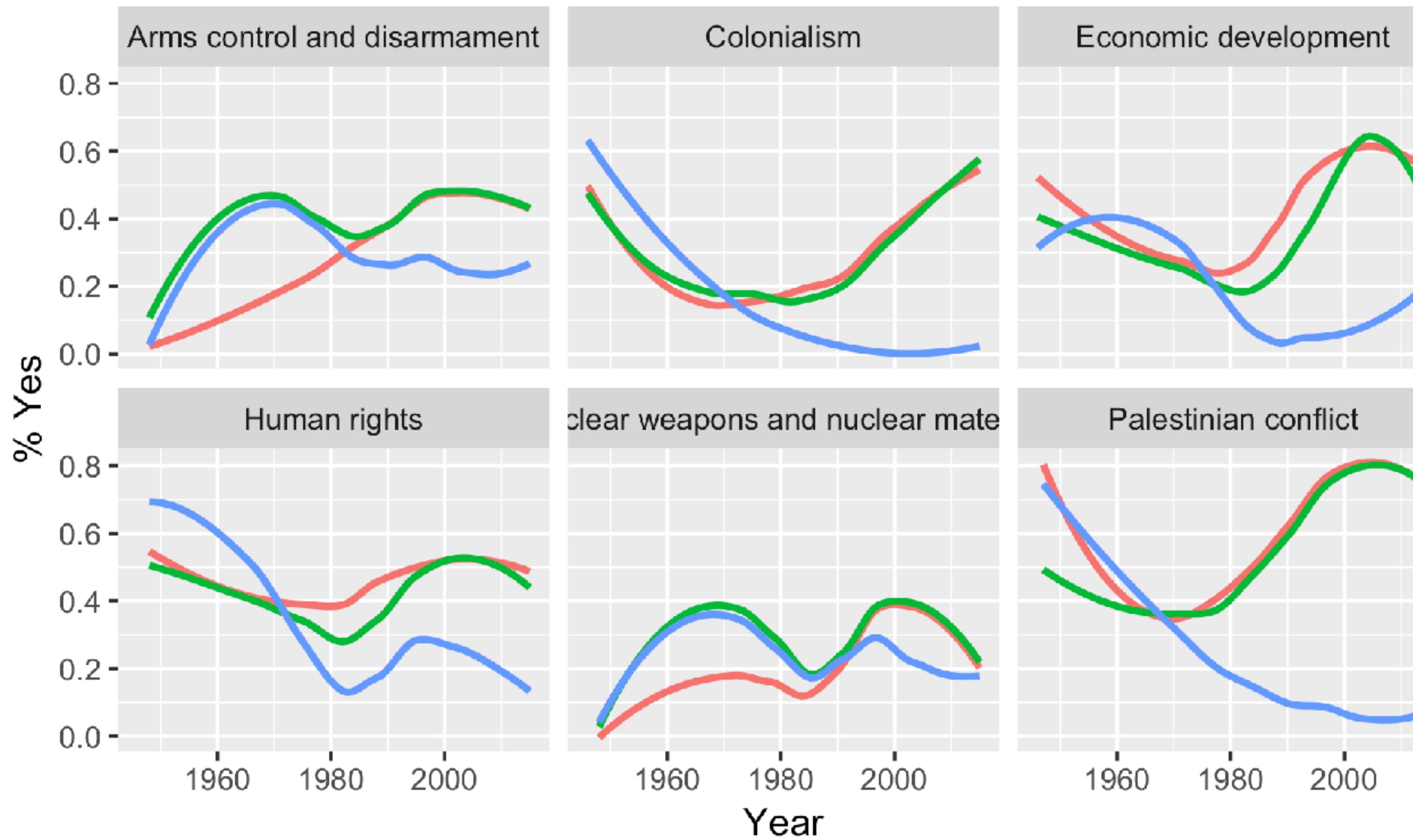
```
un_votes %>%
  filter(country %in% c("UK & NI", "US", "Turkey")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
  ) %>%
  filter(votes > 5) %>% # only use records where there are more than 5 votes
  ggplot(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"
  )

```

```
un_votes %>%
  filter(country %in% c("UK & NI", "US", "France")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
  ) %>%
  filter(votes > 5) %>% # only use records where there are more than 5 votes
  ggplot(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"
  )

```

# Percentage of Yes votes in the UN General Assembly 1946 to 2015



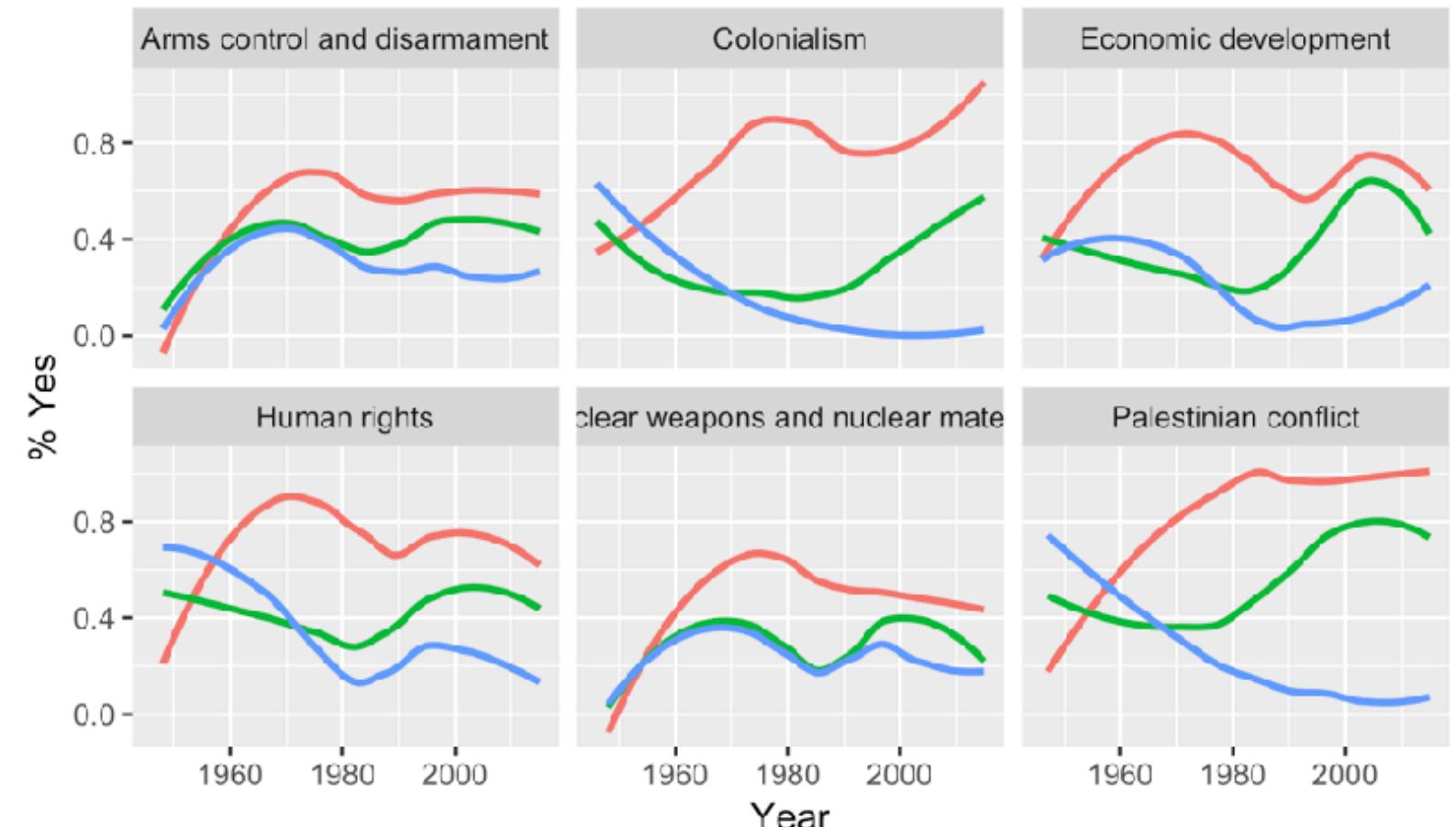
Country

- France
- UK + NI
- US

Country

- Turkey
- UK + NI
- US

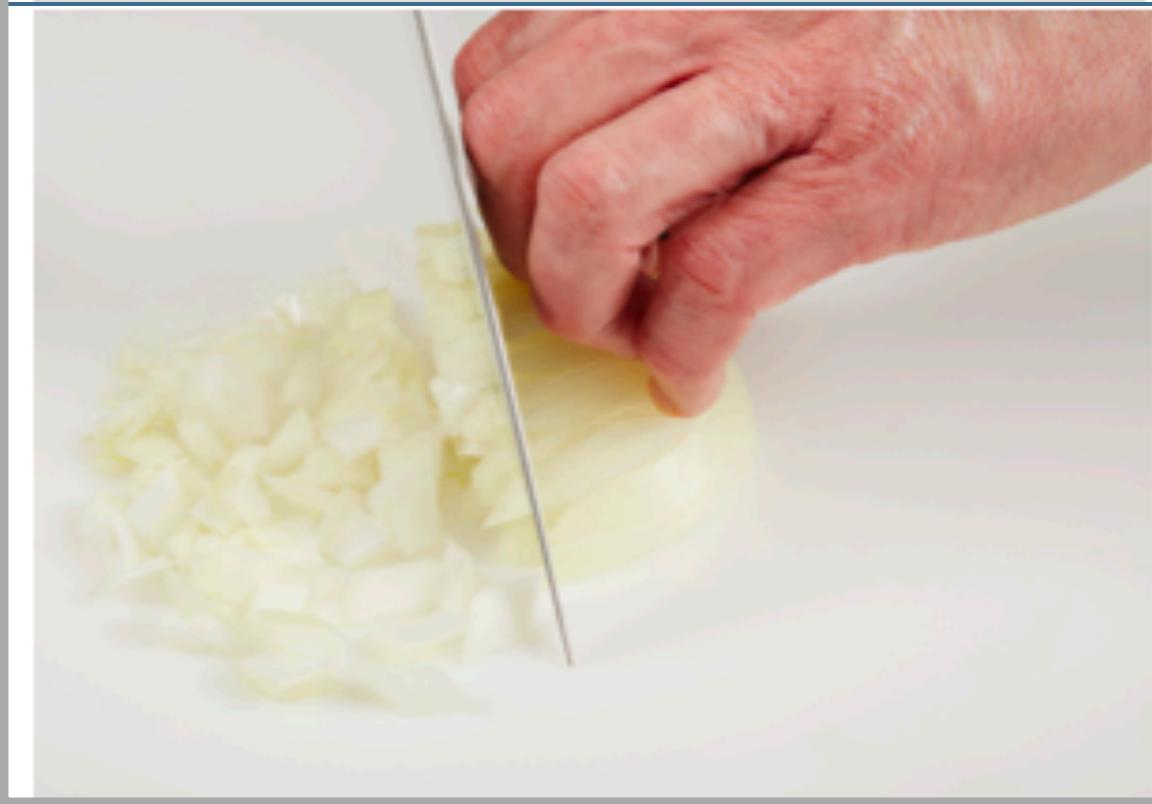
Percentage of Yes votes in the UN General Assembly  
1946 to 2015



[bit.ly/teach-ds-latinR](http://bit.ly/teach-ds-latinR)

Q

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



# Q

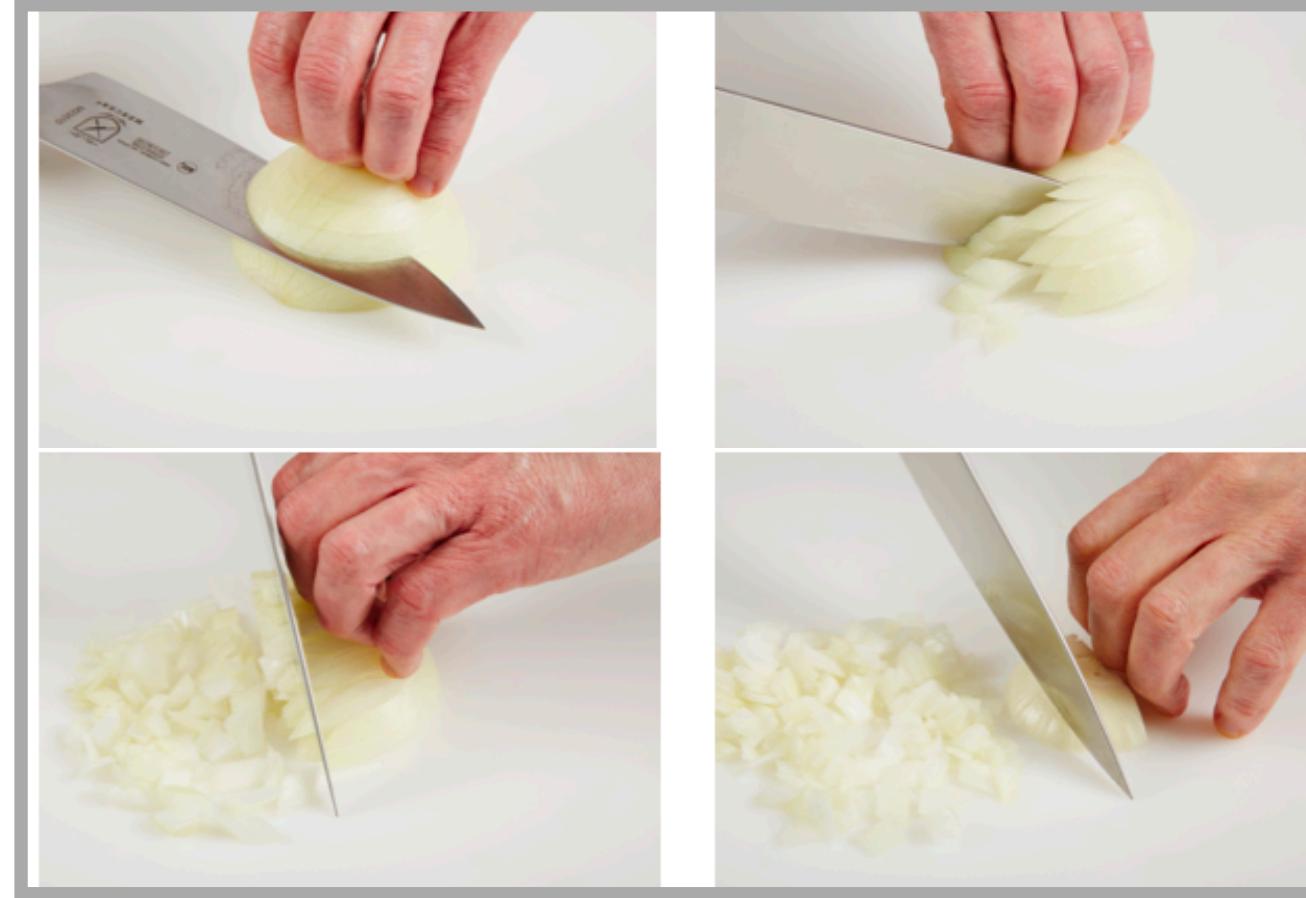


Which motivates you more to learn how to cook: perfectly chopped onions or ratatouille?

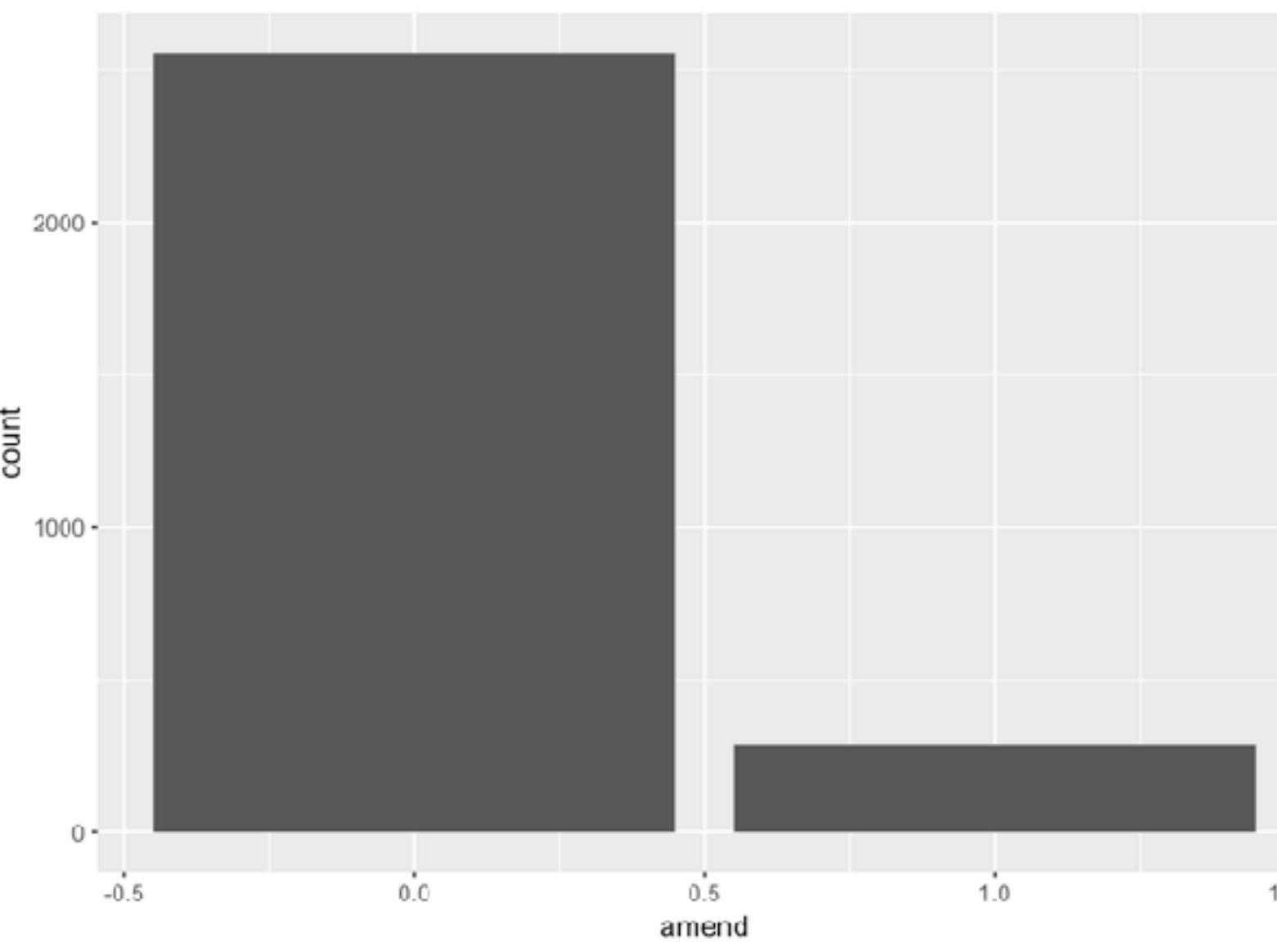


**skip  
baby  
steps**

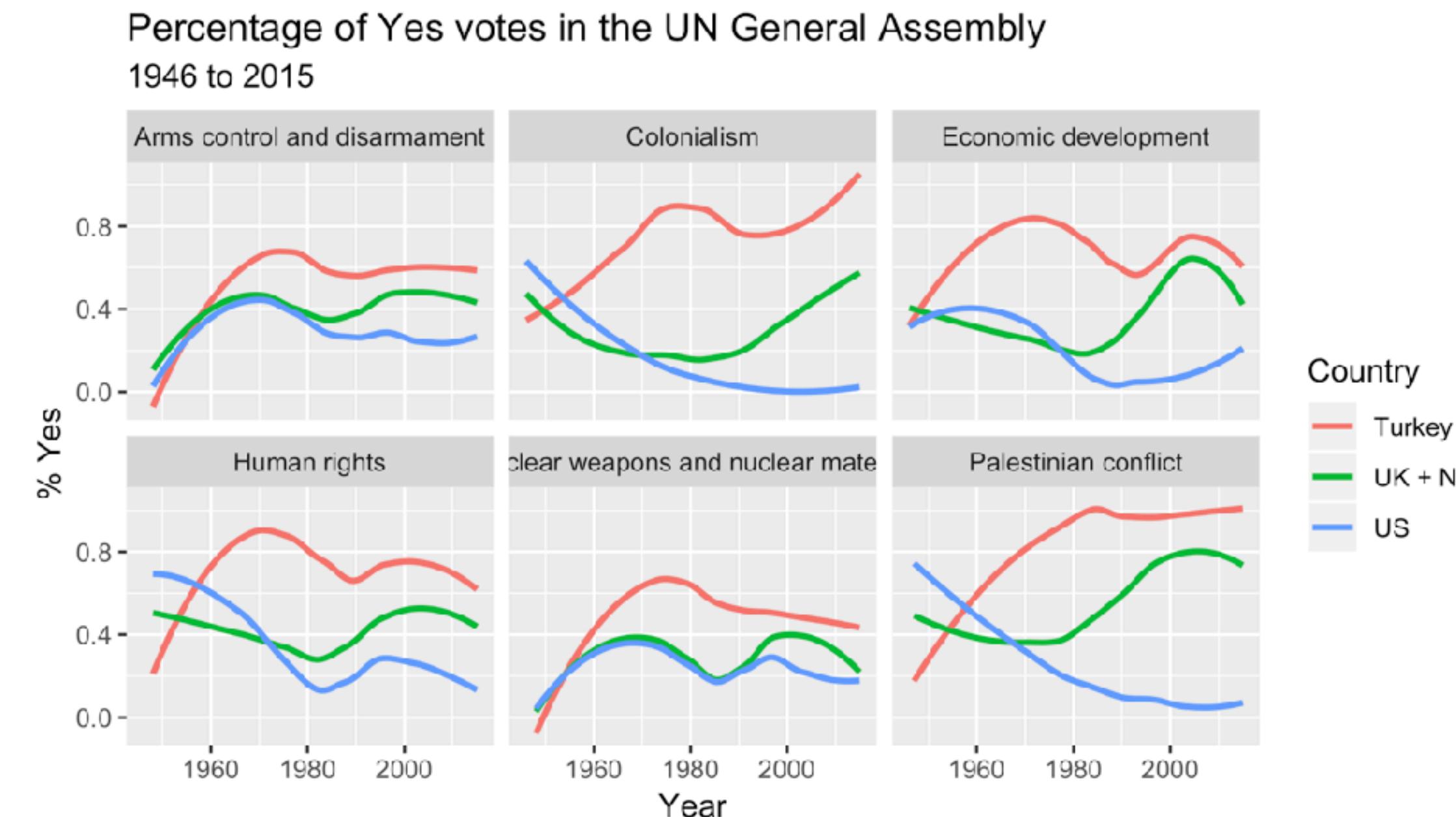
**z  
s**



Create a visualization displaying whether the vote was on an amendment.



Create a visualization displaying how US, UK, and Turkey voted over the years on issues of arms control and disarmament, colonialism, economic development, human rights, nuclear weapons, and Palestinian conflict.



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.



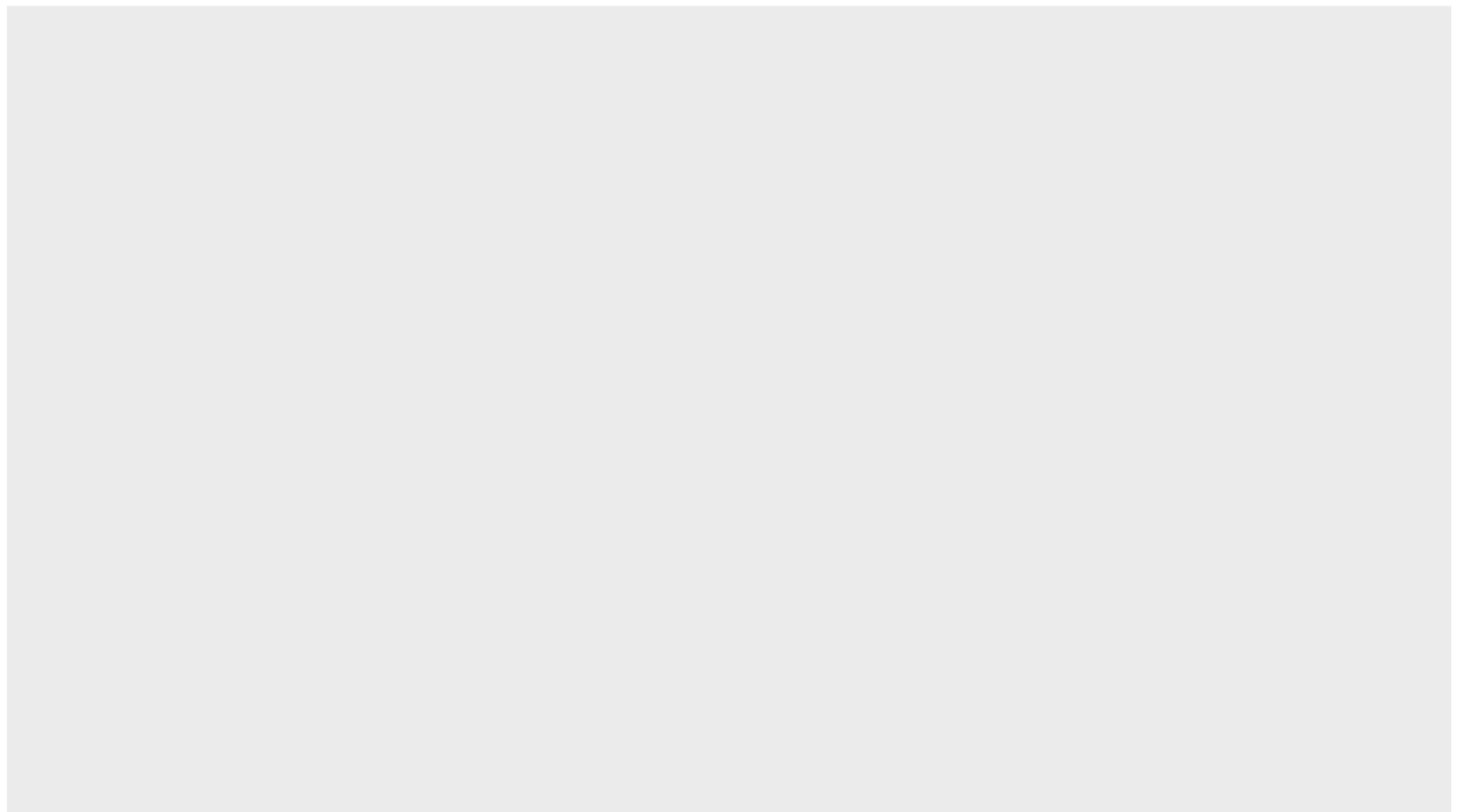
scaffold + layer

1. Draw some circles

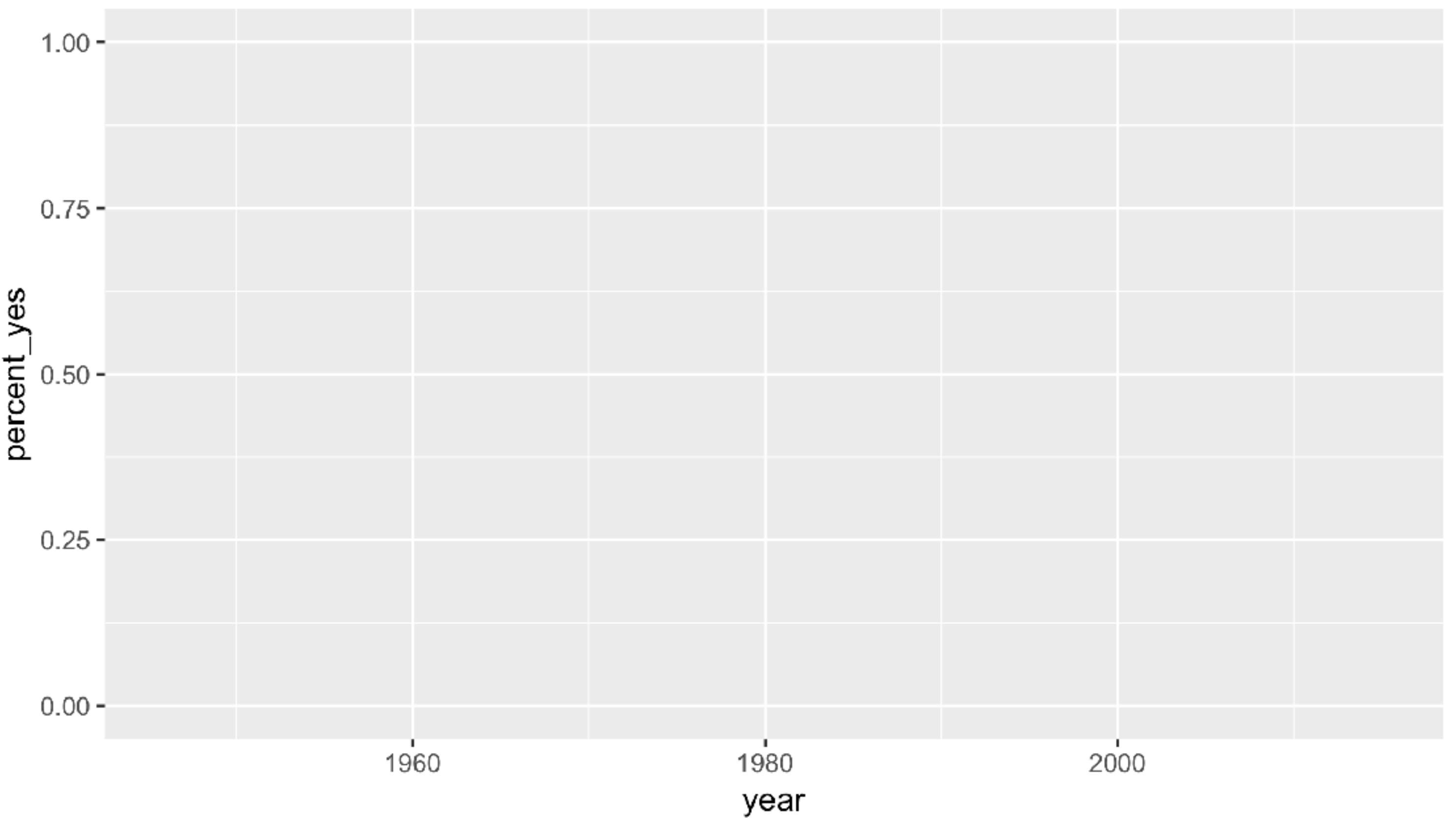
2. Draw the rest of the owl

@#\$% owl

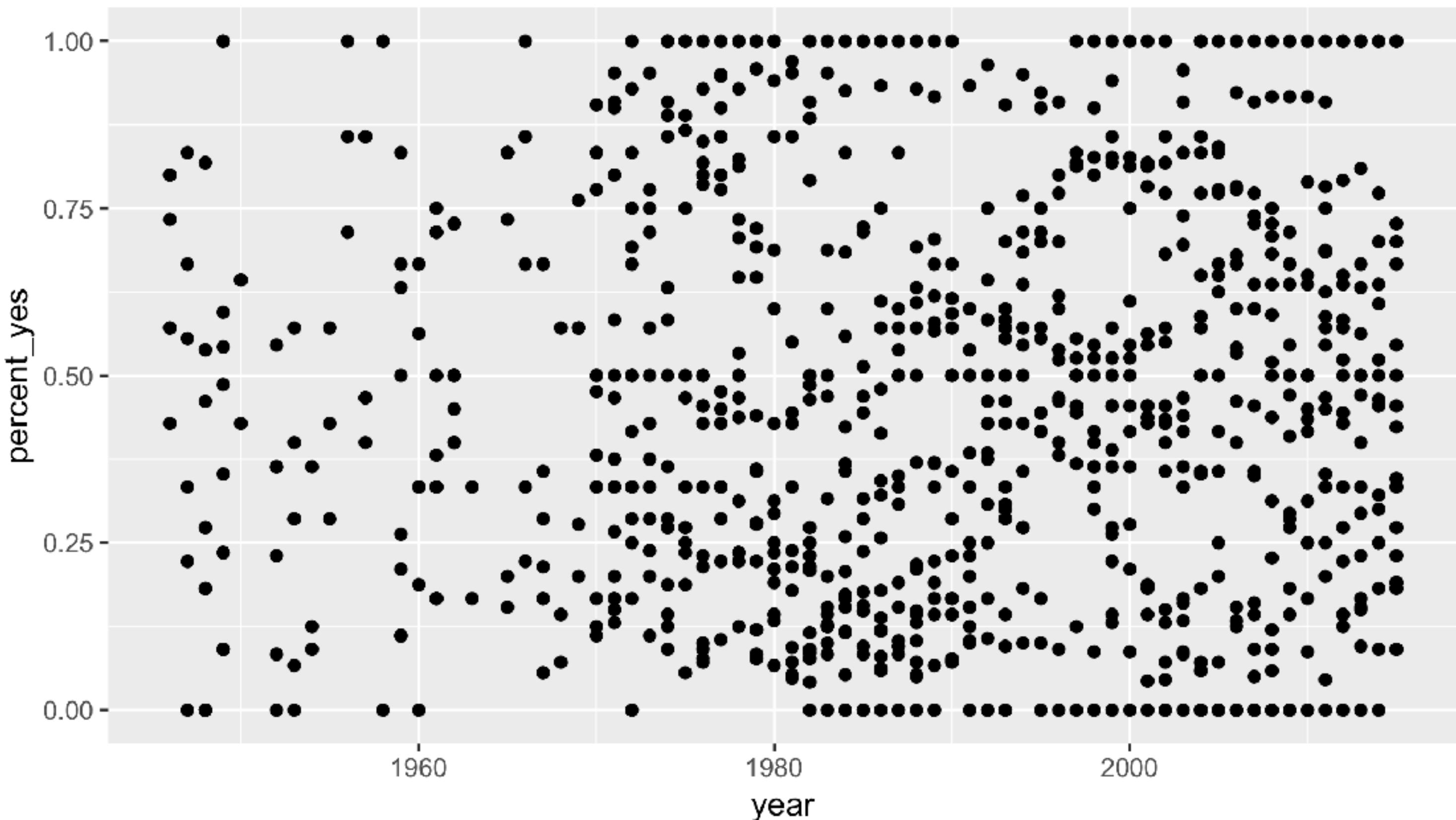
```
ggplot(data = un_votes_joined)
```



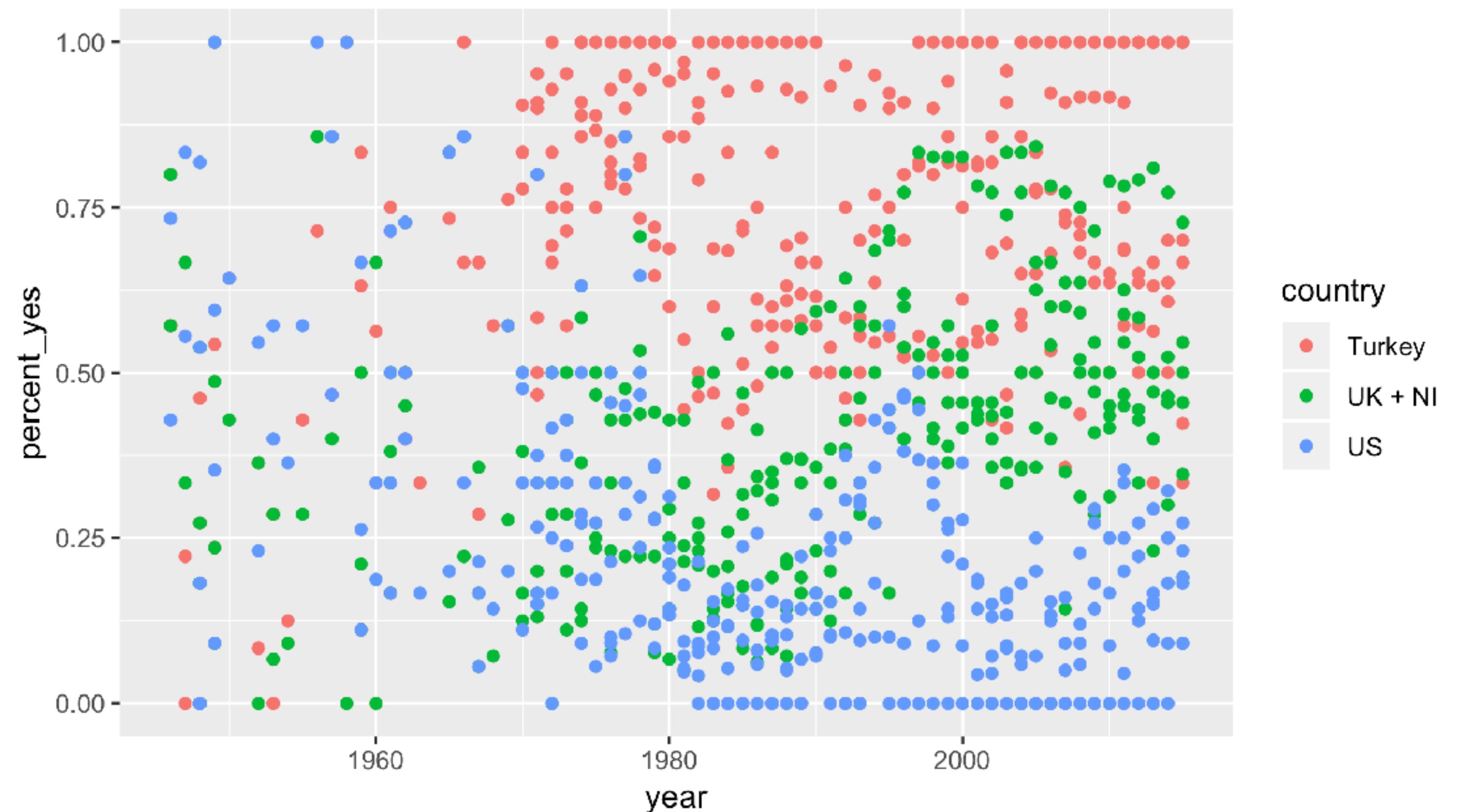
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes))
```



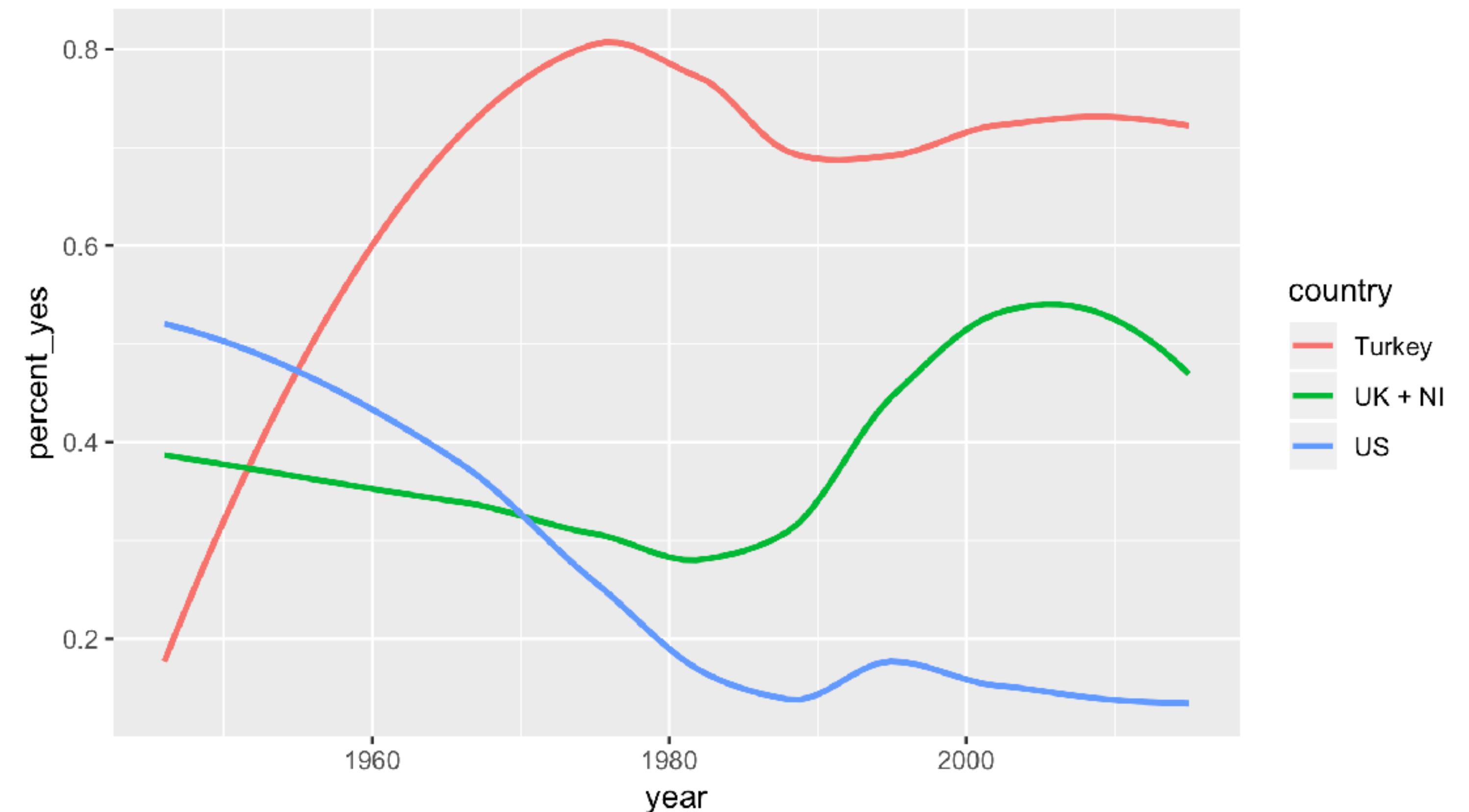
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes)) +  
  geom_point()
```



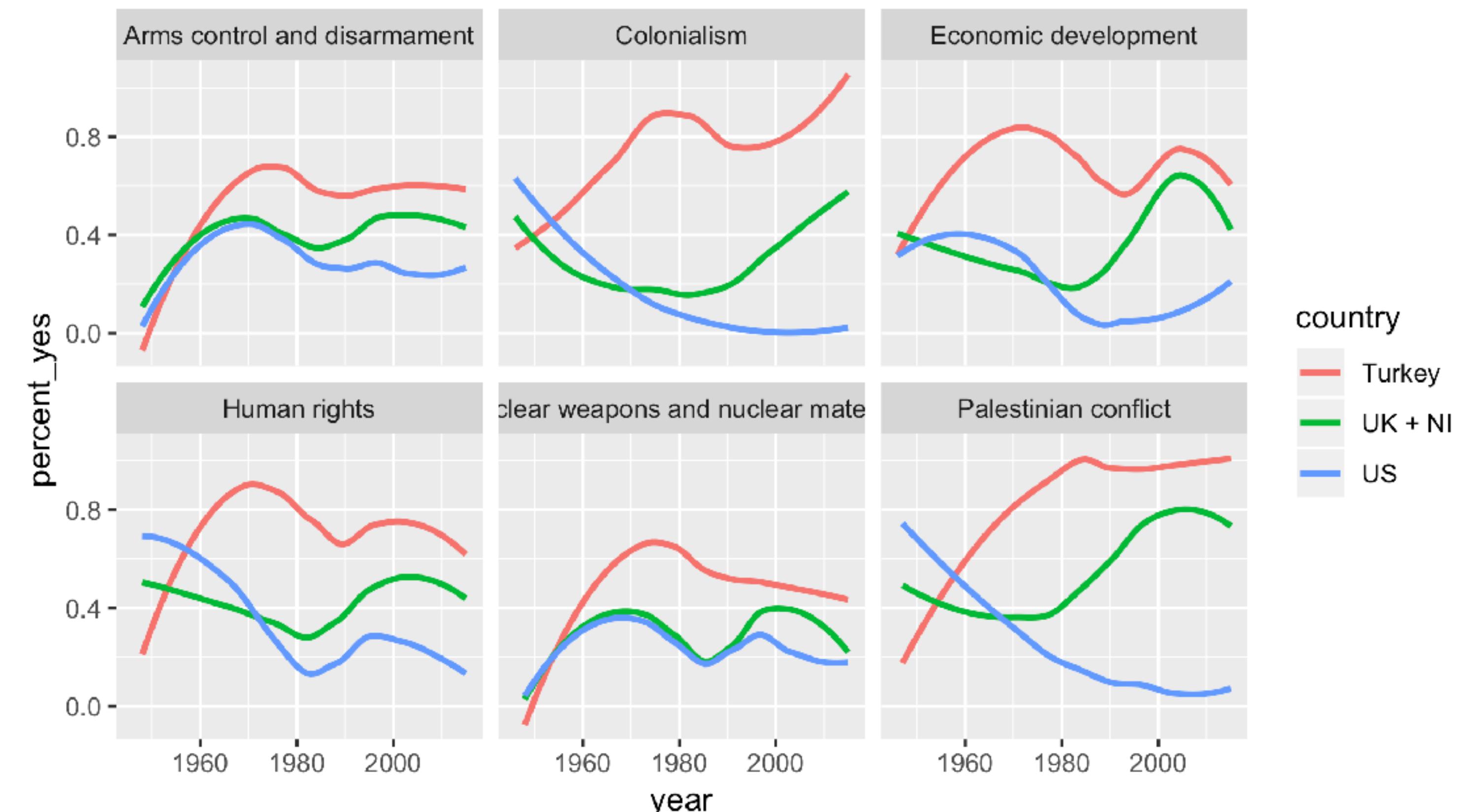
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes, color = country)) +  
geom_point()
```



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



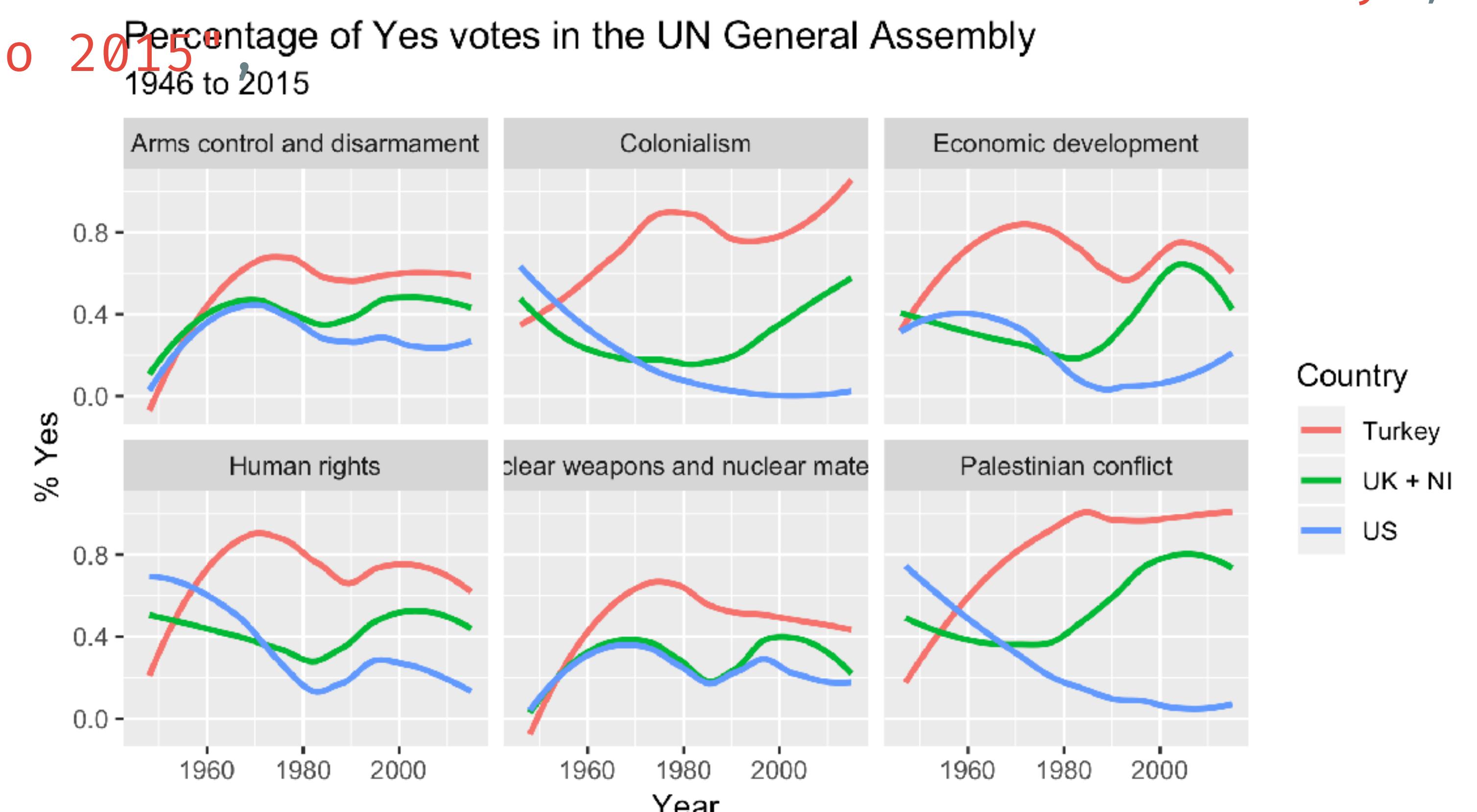
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes, color = country)) +  
  geom_smooth(method = "loess", se = FALSE) +  
  facet_wrap(~ issue)
```



```

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"
)

```



# re-insert ~~skip~~ baby steps

## 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

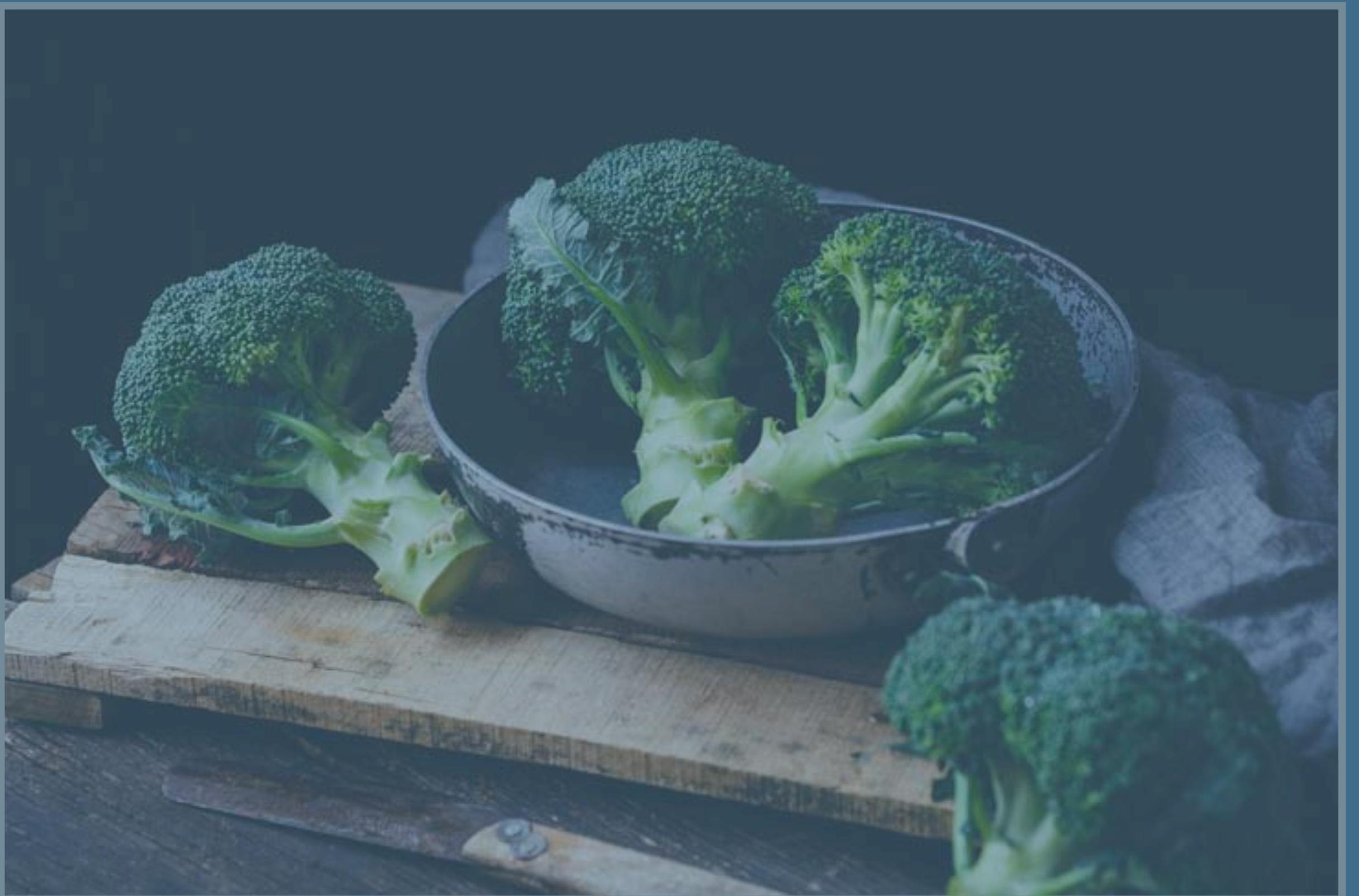
# Q

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



# Q

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



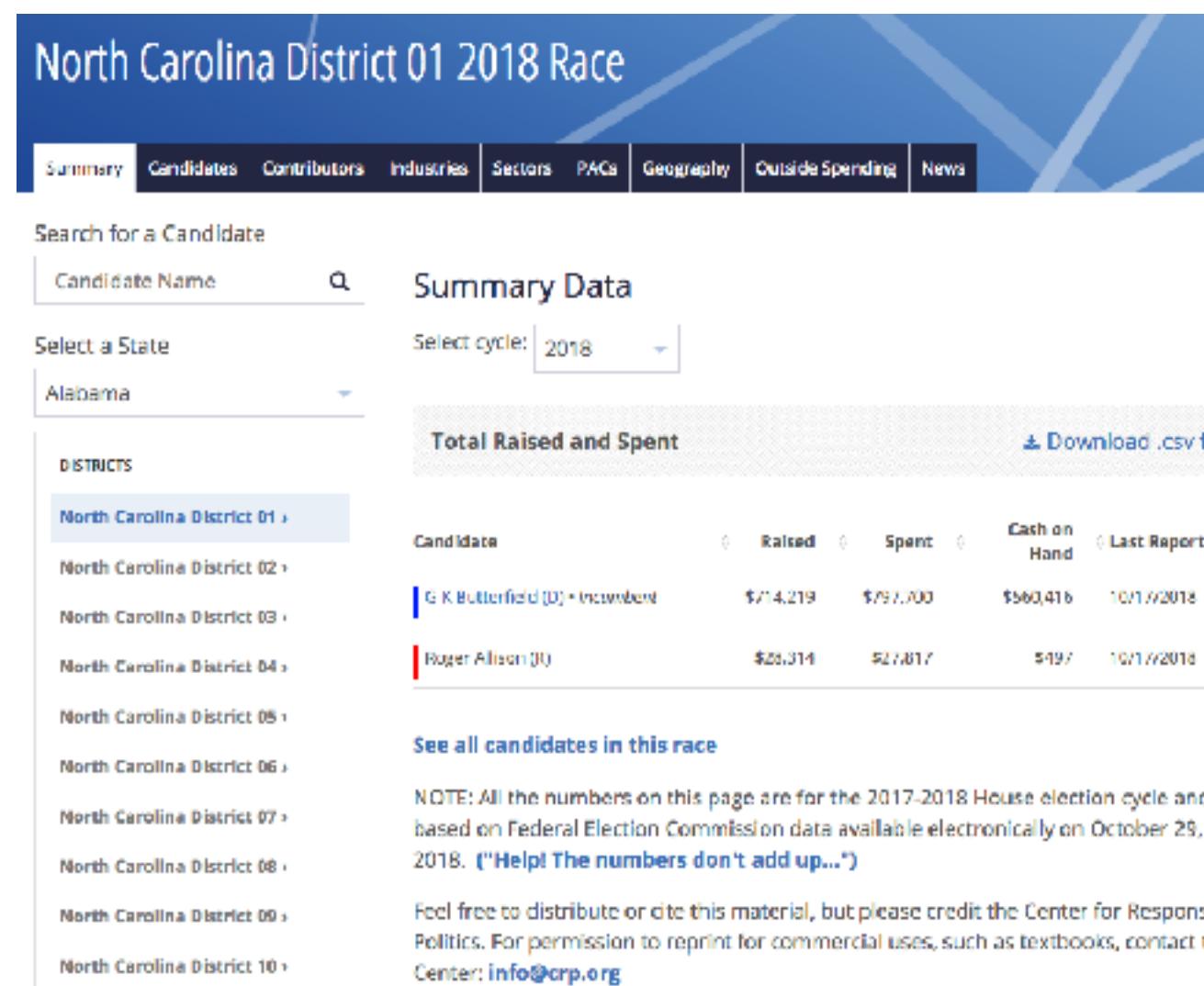
hide  
the  
veggies





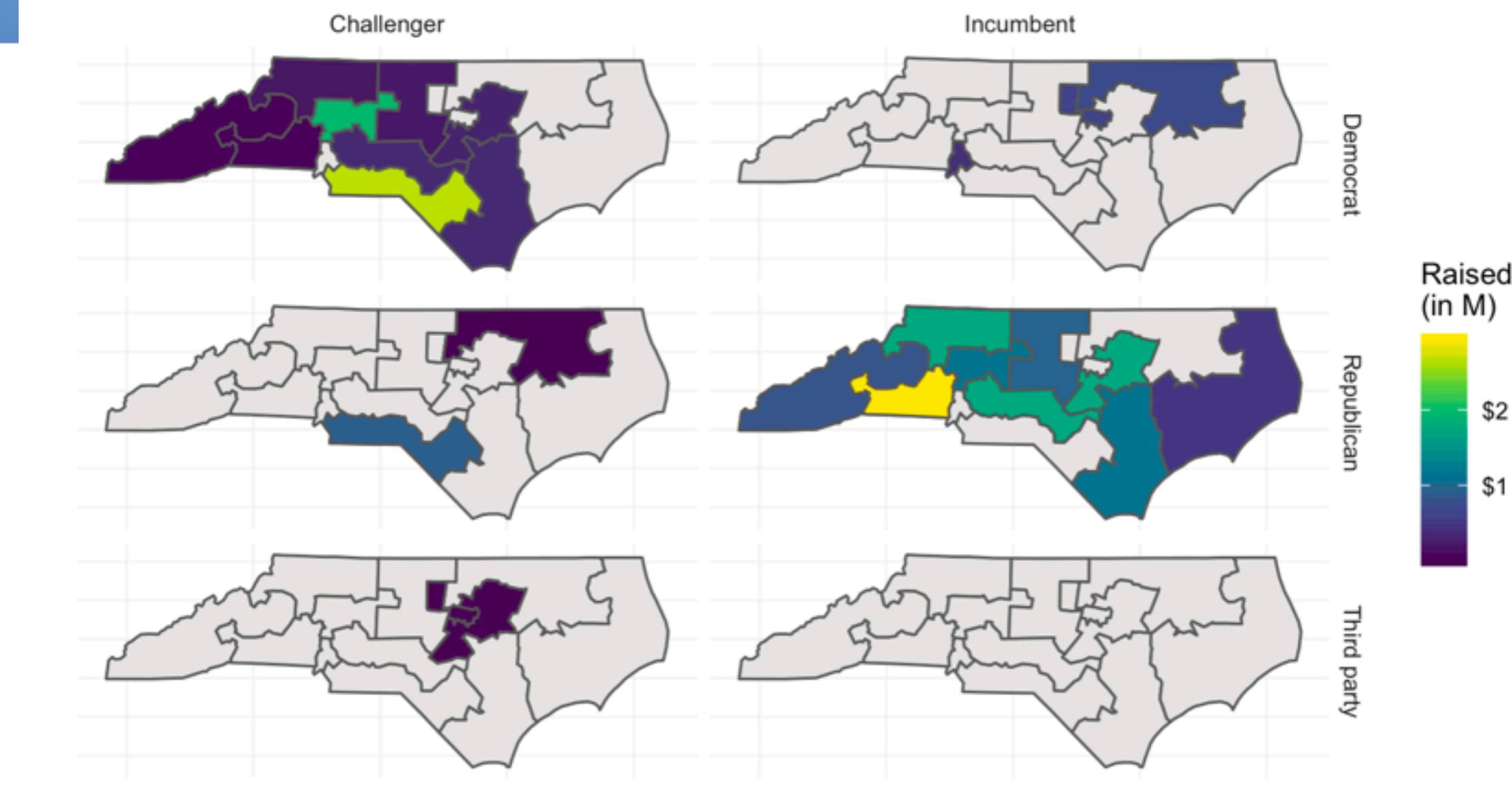
- Topic: Web scraping
- Tools:
  - **rvest**
  - regular expressions

- Today we start with this:



- and end with this:

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



Source: OpenSecrets.org

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

students will encounter lots of new  
challenges along the way —  
let that happen,  
and then provide a solution

- **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.

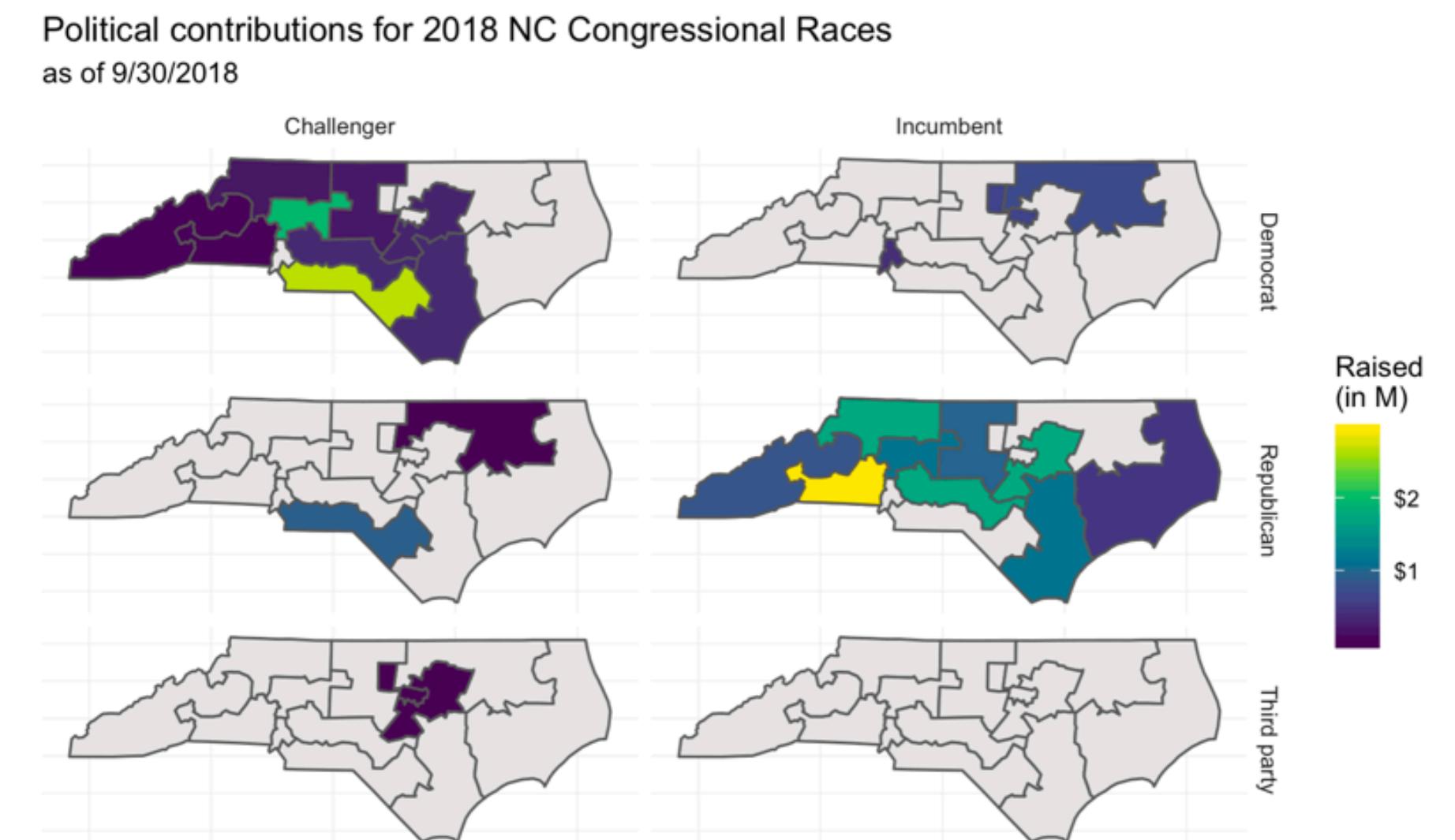
- **Ex 1:** Scrape the table off the web and save as a data frame.

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 in the data to obtain the desired facets?



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

□ **Lesson:** “Just enough” string parsing and regular expressions to go from

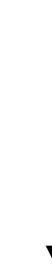
candidate_info	
1	G K Butterfield (D) • Incumbent
2	Roger Allison (R)

to

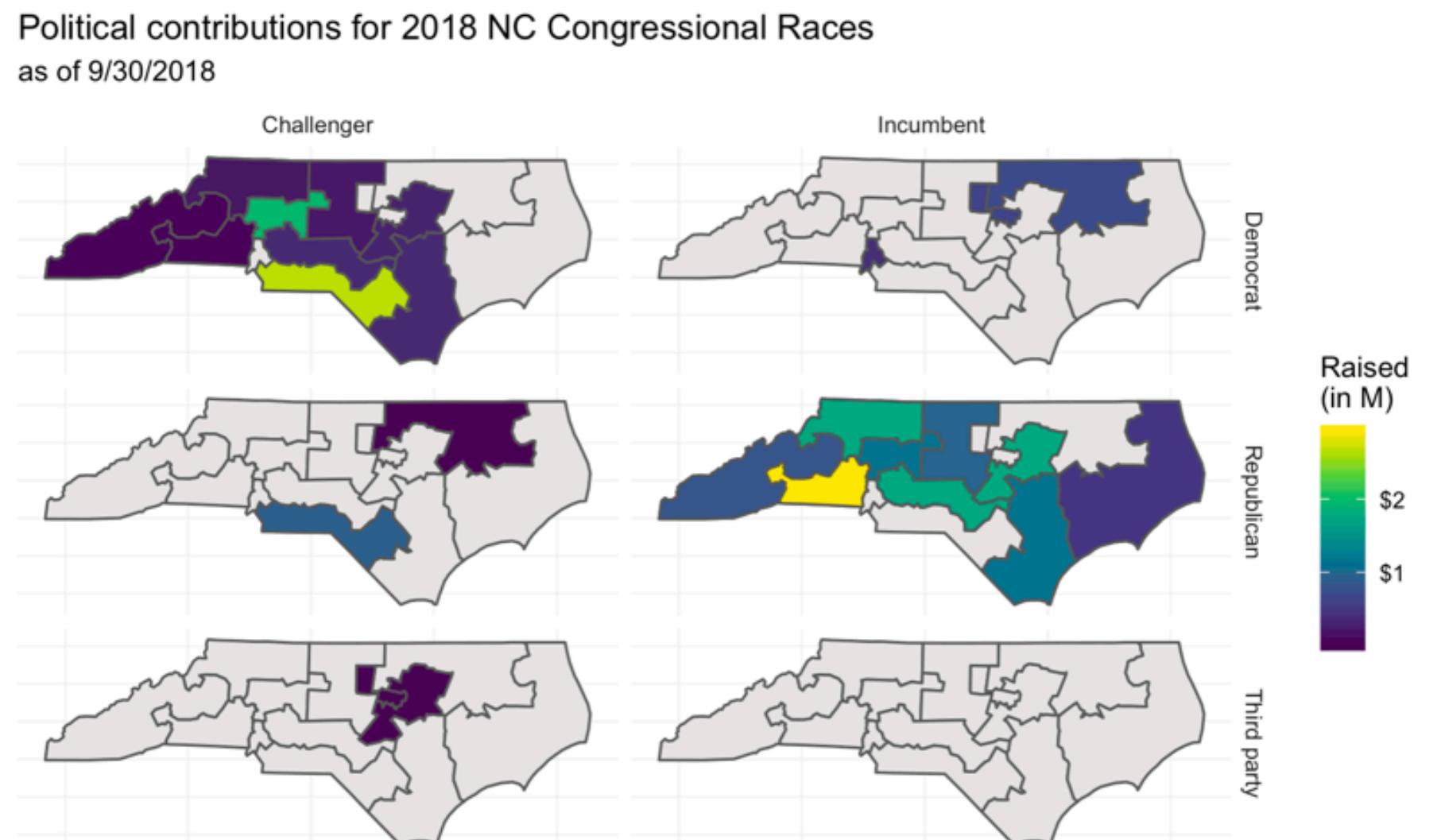
candidate_name			
	party	status	
1	Democrat	Incumbent	G K Butterfield
2	Republican	Challenger	Roger Allison

- **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



- **Ex 2:** What other information do we need represented as variables in the data to obtain the desired facets?



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



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

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



```
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
```



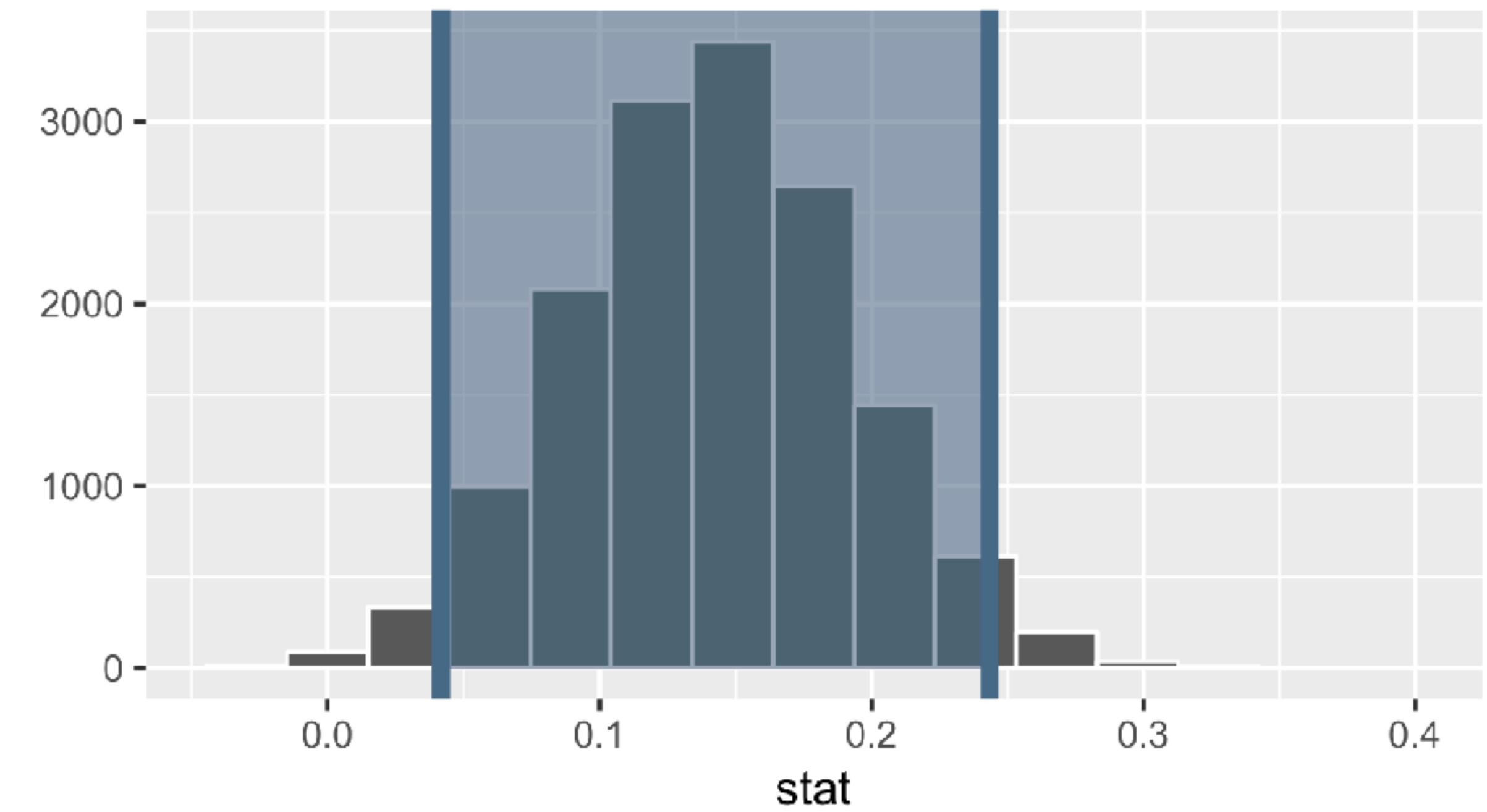
```
library(tidyverse)  
library(infer)  
  
evals %>%  
  specify(score ~ gender) %>%  
  generate(reps = 15000,  
            type = "bootstrap") %>%  
  calculate(stat = "diff in means",  
            order = c("male", "female")) %>%  
  summarise(  
    l = quantile(stat, 0.025),  
    u = quantile(stat,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

```
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, 0.0
```

#	l	u
#	0.0410	0.243



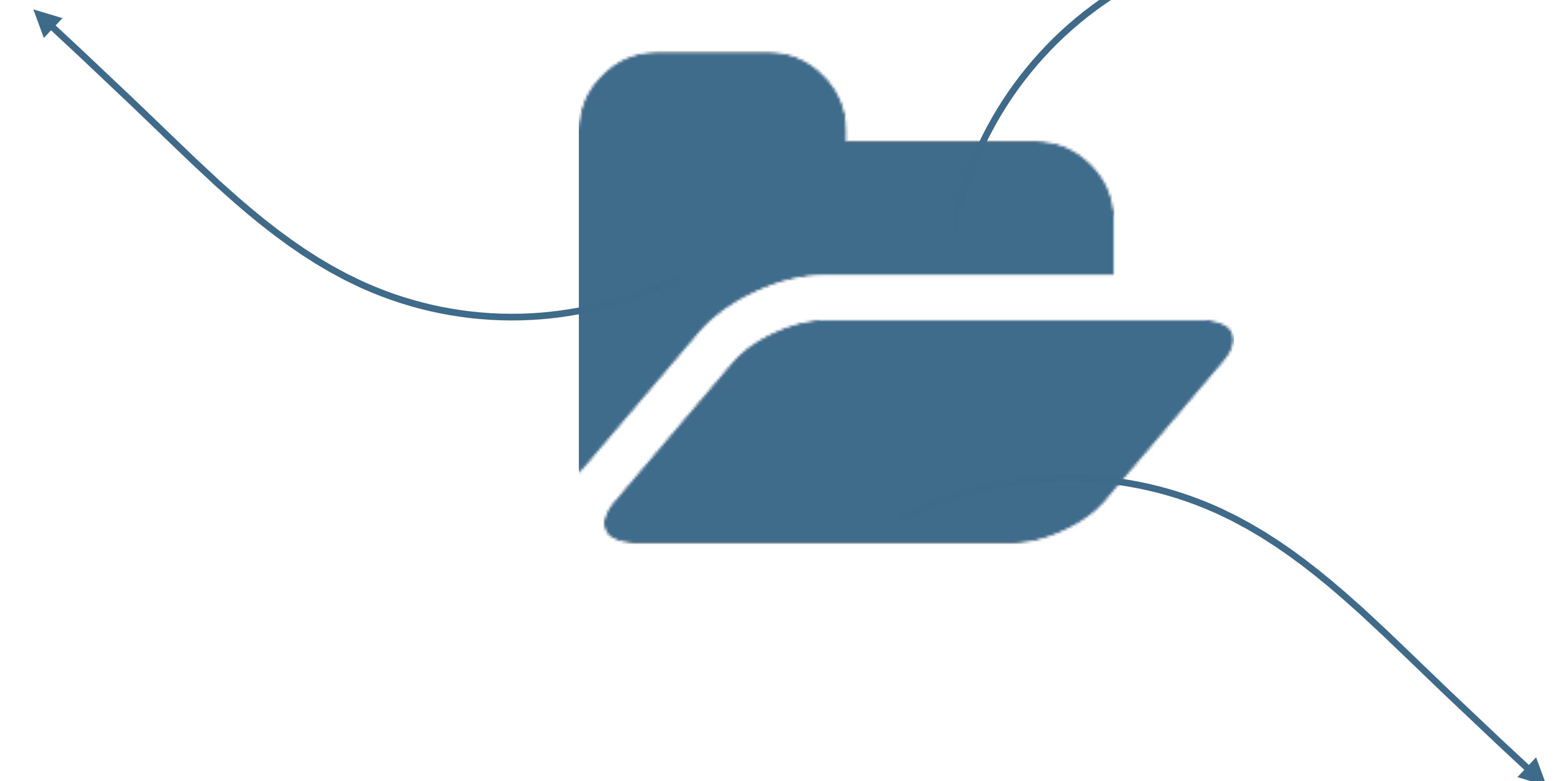
# goals

scalable



open

validated



 Search...

Hello #dsbox

Overview

Philosophy

Topics

Tech stack

Community

Course content

Infrastructure

Pedagogy

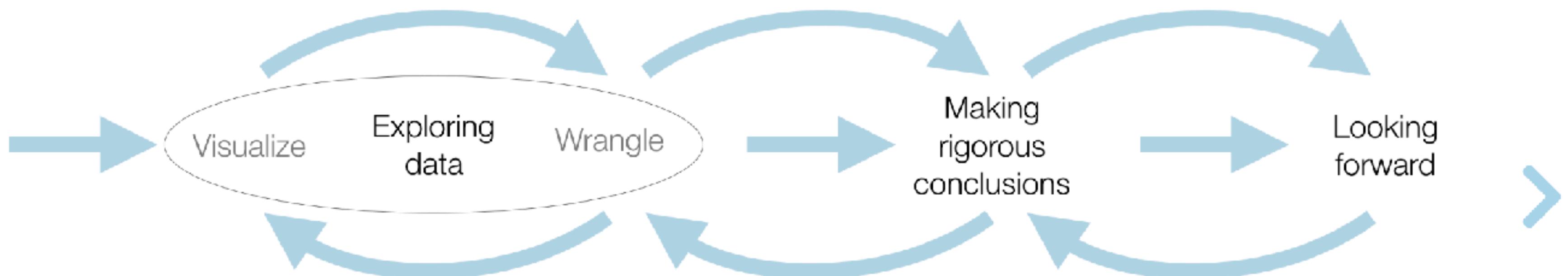
rstd.io/teach-ds-latinR

Data Science in a Box &gt; Hello #dsbox &gt; Topics



# Topics

The course content is organized in three units:



**Unit 1 - Exploring data:** This unit focuses on data visualization and data wrangling. Specifically we cover fundamentals of data and data visualization, confounding variables, and Simpson's paradox as well as the concept of tidy data, data import, data cleaning, and data curation. We end the unit with web scraping and introduce the idea of iteration in preparation for the next unit. Also in this unit students are introduced to the toolkit: R, RStudio, R Markdown, Git, GitHub, etc.

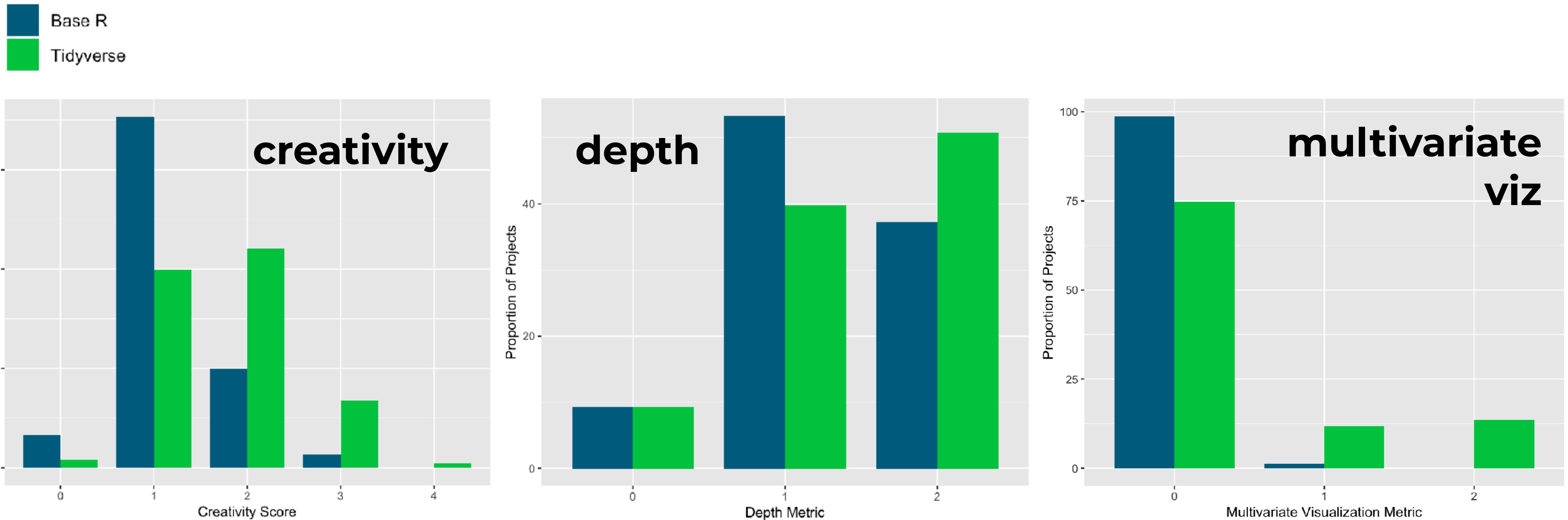
**Unit 2 - Making rigorous conclusions:** In this part we introduce modeling and statistical inference for making data based conclusions. We discuss building, interpreting, and selecting models, visualizing interaction effects, and prediction and model validity. Statistical inference is introduced from a simulation based perspective, and the Central Limit Theorem is discussed very briefly to lay the foundation for future coursework in statistics.

**Unit 3 - Looking forward:** In the last unit we present a series of modules such as interactive reporting and visualization with Shiny, text analysis, and Bayesian inference. These are independent modules that instructors can choose to include in their introductory data science curriculum depending on how much time they have left in the semester.

open

**Retrospective study** of 205 open ended student projects

- on **creativity**, **depth** and the complexity of **multivariate visualizations**
- compared across students who learned R using **base R** syntax vs. **tidyverse**





1. Formative assessments
2. Automated grading
3. Calibrated peer review

Let them eat cake (first)! \*

\* You can tell them  
all about the  
ingredients later!



@minebocek



mine-cetinkaya-rundel



cetinkaya.mine@gmail.com

