

Let them  
eat cake  
(first)!



@minebocek



mine-cetinkaya-rundel



cetinkaya.mine@gmail.com



[bit.ly/let-eat-cake](http://bit.ly/let-eat-cake)



# Backward design

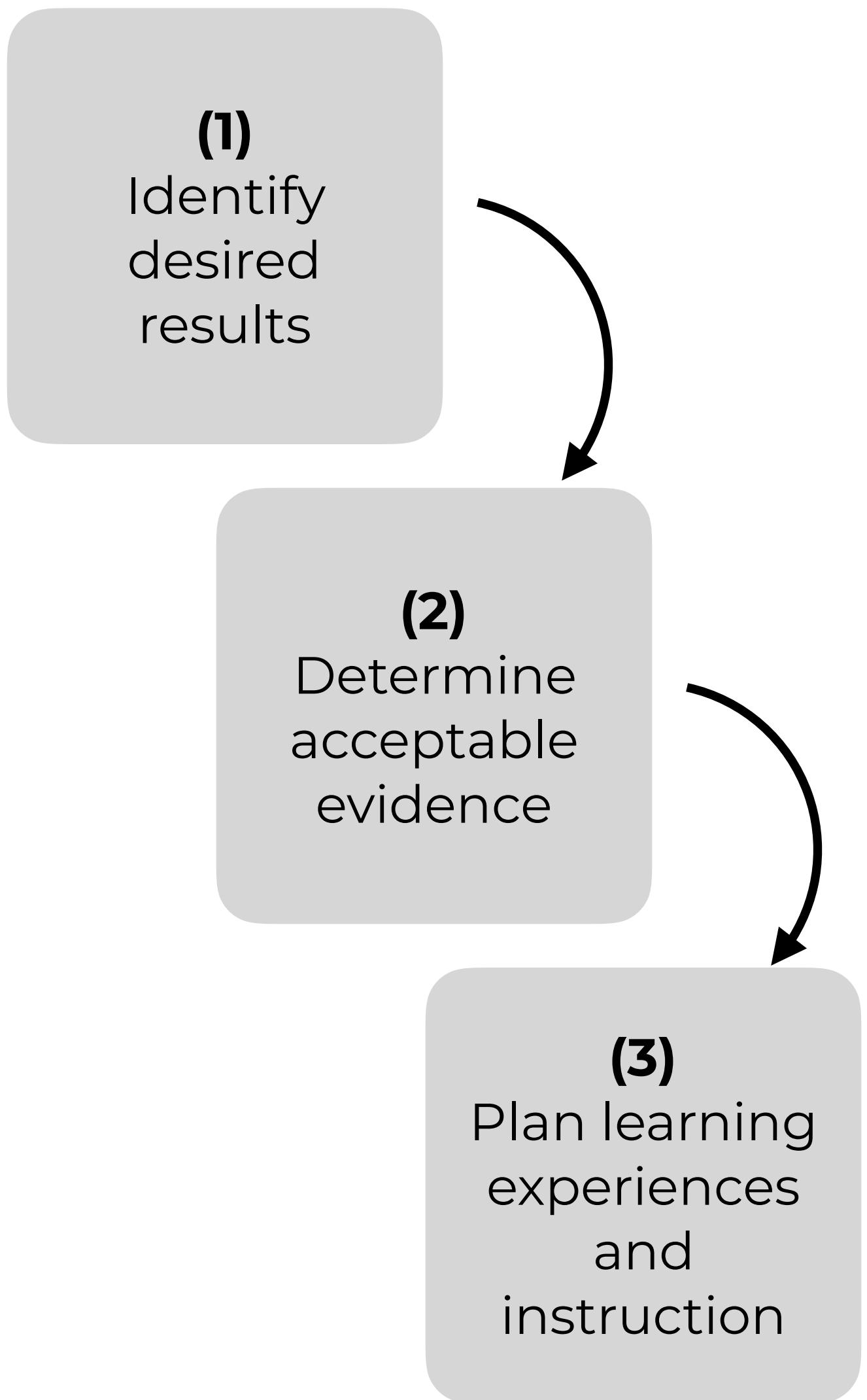
set goals for educational curriculum before choosing instructional methods + forms of assessment



analogous to travel planning - itinerary deliberately designed to meet cultural goals, not purposeless tour of all major sites in a foreign country



Wiggins, Grant P., Grant Wiggins, and Jay McTighe. *Understanding by design*. Ascd, 2005.

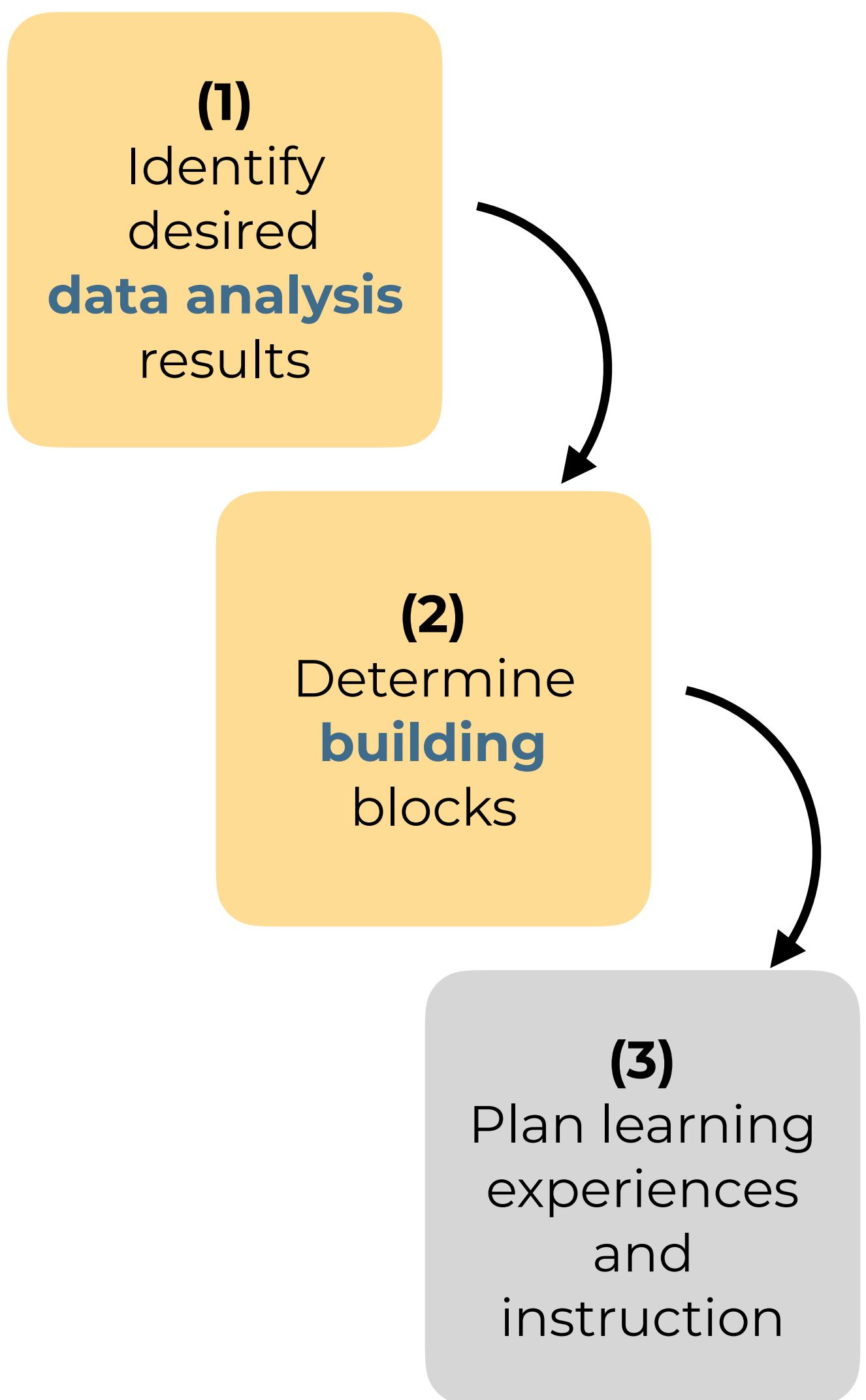


# Designing backwards

students are first exposed to results and findings of a data analysis



and then learn the building blocks of the methods and techniques used along the way



# Context

assumes  
no  
background



focuses on  
EDA +  
modeling &  
inference +  
modern  
computing



requires  
reproducibility



emphasizes  
collaboration +  
effective  
communi-  
cation



uses R as the  
statistical  
programming  
language





Which of the following four descriptions give you a **better sense** of the final product?

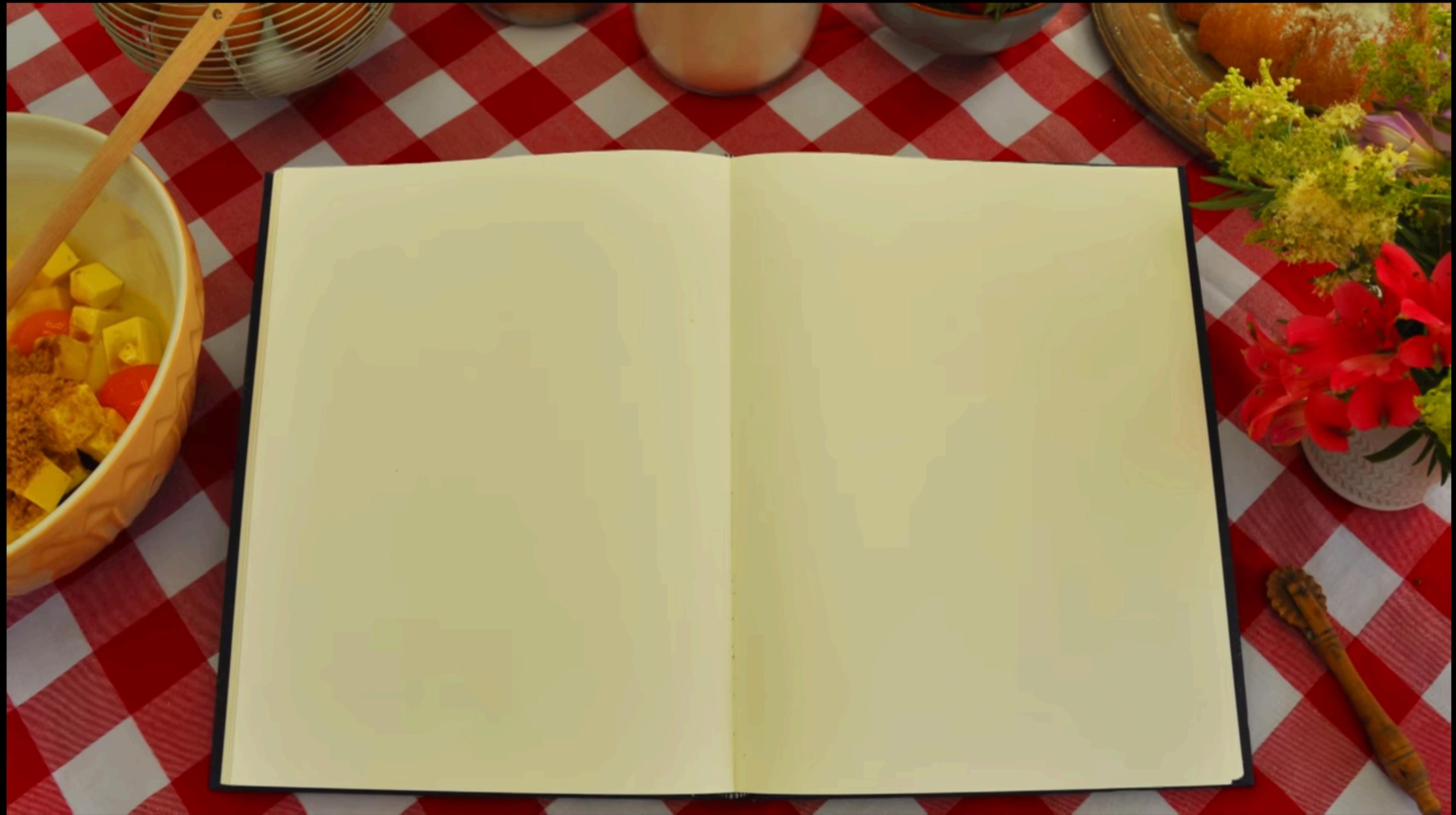
# Pineapple and coconut sandwich cake

# Pineapple and Coconut sandwich cake



# Pineapple and Coconut sandwich cake





**(a)** Pineapple and coconut sandwich cake



**(c)** **<with audio>**



**(d)**

Toasted  
Coconut  
Flakes

Pineapple  
'Flower'



**start  
with  
cake**



ex1.

visualization



Which of the following two examples is more likely to be **interesting** for a wide range of students?

**(a)**

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

```
# Declare variables
```

```
x ← 8  
y ← "monkey"  
z ← FALSE
```

```
# Check class of x  
class(x)
```

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

```
# Check class of y  
class(y)
```

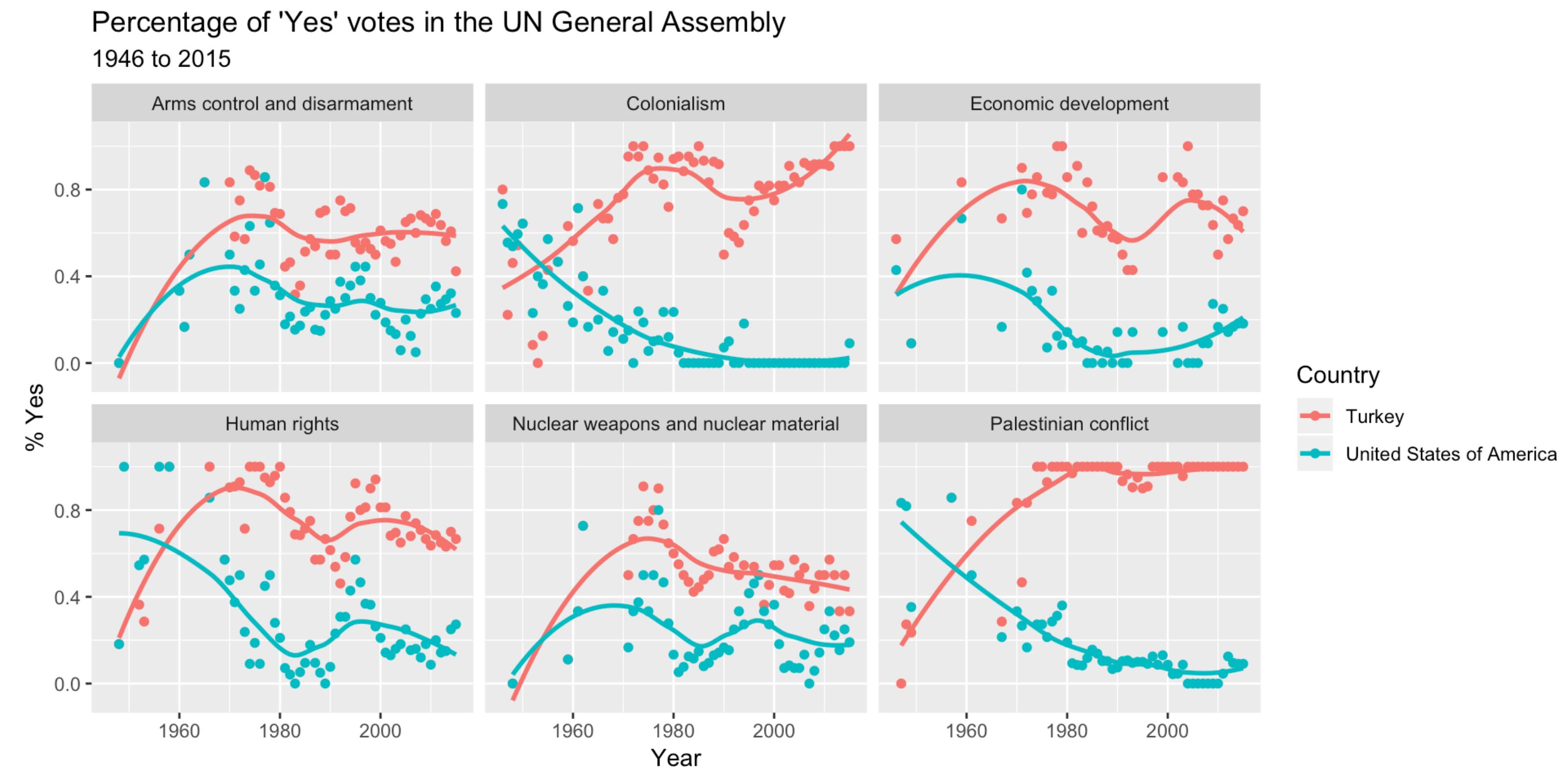
```
#> [1] "character"
```

```
# Check class of z  
class(z)
```

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

**(b)**

- 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



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

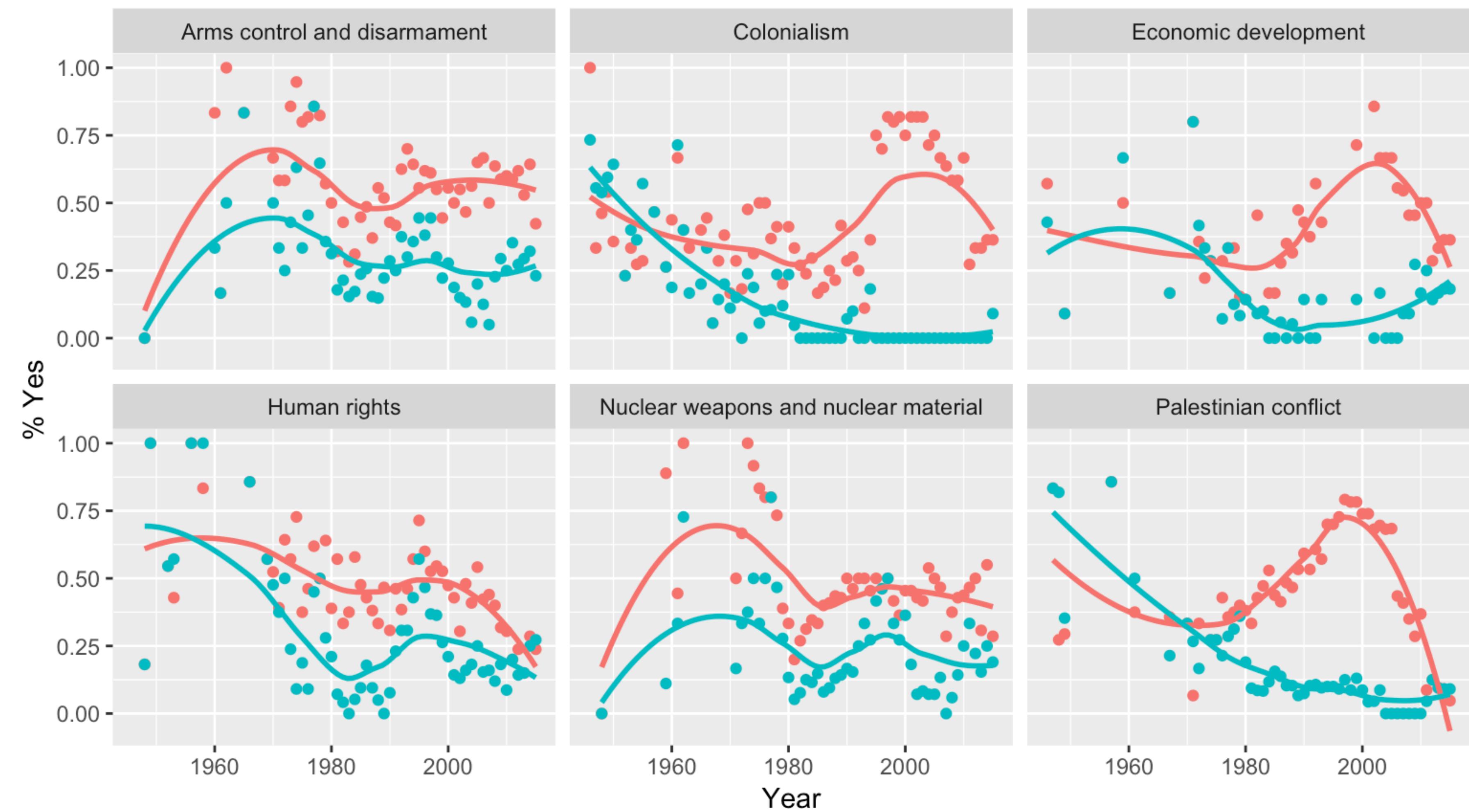
```
un_votes %>%  
  filter(country %in% c("United States of America", "Turkey")) %>%  
  inner_join(un_roll_calls, by = "rcid") %>%  
  inner_join(un_roll_call_issues, by = "rcid") %>%  
  group_by(country, year = year(date), issue) %>%  
  summarize(  
    votes = n(),  
    percent_yes = mean(vote == "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_point() +  
    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("United States of America", "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_point() +  
  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("United States of America", "Canada")) %>%  
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_point() +  
  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

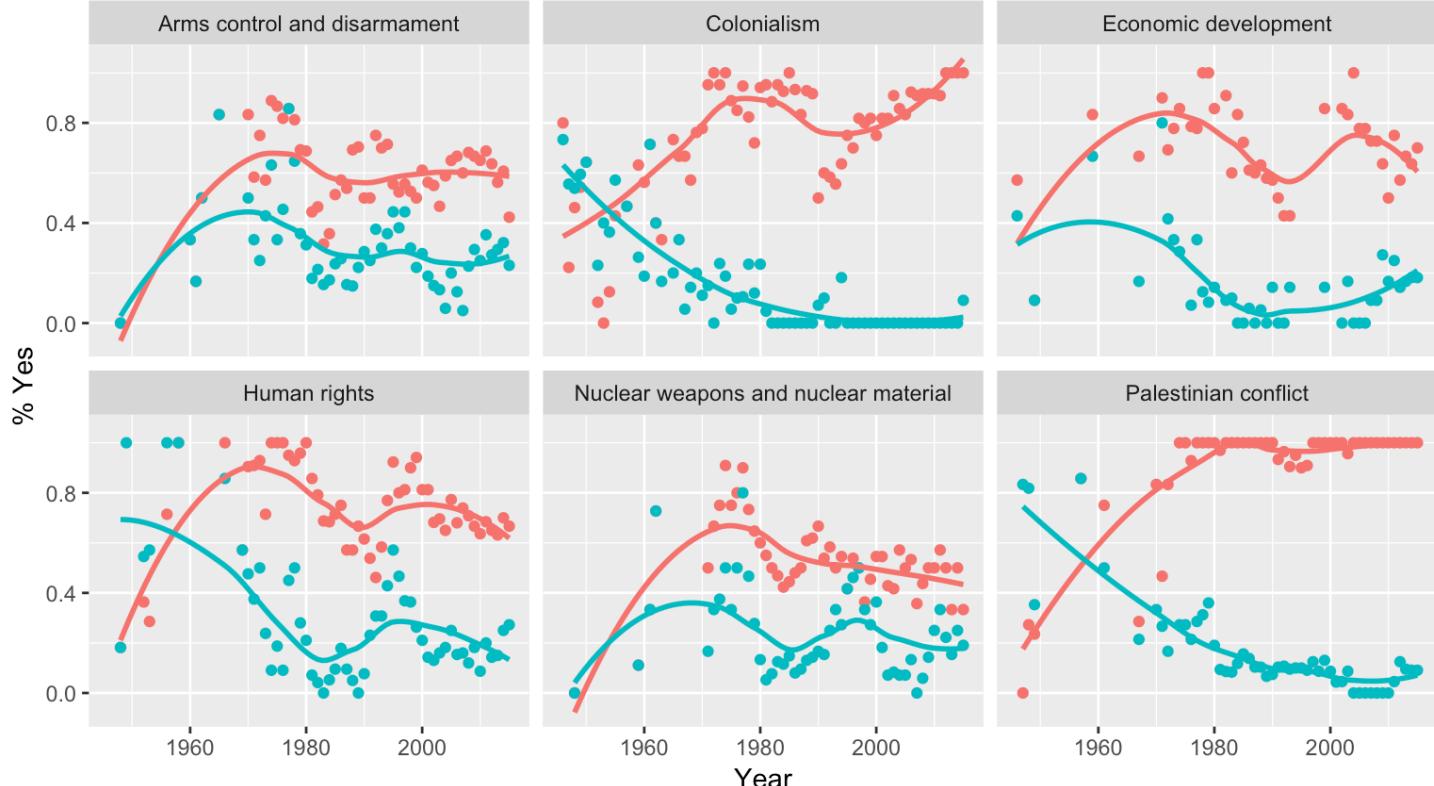
1946 to 2015



Country

- Canada
- United States of America

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



# why 🍰 = 📊?

more likely for students to have intuition for interpretations coming in



easier for them to catch their own mistakes



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



# ex: Introduction to R for Data Science

Microsoft Professional Program Certificate in Data Science

## Course Syllabus

### Section 1: Introduction to Basics

Take your first steps with R. Discover the basic data types in R and assign your first variable.

### Section 2: Vectors

Analyze gambling behaviour using vectors. Create, name and select elements from vectors.

### Section 3: Matrices

Learn how to work with matrices in R. Do basic computations with them and demonstrate your knowledge by analyzing the Star Wars box office figures.

### Section 4: Factors

R stores categorical data in factors. Learn how to create, subset and compare categorical data.

### Section 5: Data Frames

When working R, you'll probably deal with Data Frames all the time. Therefore, you need to know how to create one, select the most interesting parts of it, and order them.

### Section 6: Lists

Lists allow you to store components of different types. Section 6 will show you how to deal with lists.

### Section 7: Basic Graphics

Discover R's packages to do graphics and create your own data visualizations.

# ex: Data Science Specialization

Johns Hopkins University

JOHNS HOPKINS Data Science Specialization Enroll Starts Sep 27

About How It Works **Courses** Instructors Enrollment Options FAQ

1  
COURSE

## The Data Scientist's Toolbox

★★★★★ 4.5 16,022 ratings • 3,325 reviews

In this course you will get an introduction to the main tools and ideas in the data scientist's toolbox. The course gives an overview of the data, questions, and tools that data analysts and data scientists work with. There are two components to this course. The first is a c... [MORE](#)

2  
COURSE

## R Programming

★★★★★ 4.6 12,076 ratings • 2,558 reviews

In this course you will learn how to program in R and how to use R for effective data analysis. You will learn how to install and configure software necessary for a statistical programming environment and describe generic programming language concepts as they are i... [MORE](#)

3  
COURSE

## Getting and Cleaning Data

★★★★★ 4.6 5,178 ratings • 829 reviews

Before you can work with data you have to get some. This course will cover the basic ways that data can be obtained. The course will cover obtaining data from the web, from APIs, from databases and from colleagues in various formats. It will also cover the basics of data ... [MORE](#)

4  
COURSE

## Exploratory Data Analysis

★★★★★ 4.7 3,957 ratings • 591 reviews

This course covers the essential exploratory techniques for summarizing data. These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data. We will cover in detail the plotting systems in R as well as some of the basic principles of constructing data graphics. We will also cover some of the common multivariate statistical techniques used to visualize high-dimensional data. [LESS](#)

1  
SECTION

25 hours to complete

## Week 1: Background, Getting Started, and Nuts & Bolts

This week covers the basics to get you started up with R. The Background Materials lesson contains information about course mechanics and some videos on i... [MORE](#)

28 videos (Total 129 min), 9 readings, 8 quizzes [SEE ALL](#)

2  
SECTION

12 hours to complete

## Week 2: Programming with R

Welcome to Week 2 of R Programming. This week, we take the gloves off, and the lectures cover key topics like control structures and functions. We also intr... [MORE](#)

13 videos (Total 91 min), 3 readings, 5 quizzes [SEE ALL](#)

3  
SECTION

10 hours to complete

## Week 3: Loop Functions and Debugging

We have now entered the third week of R Programming, which also marks the halfway point. The lectures this week cover loop functions and the debuggi... [MORE](#)

8 videos (Total 61 min), 2 readings, 4 quizzes [SEE ALL](#)

4  
SECTION

11 hours to complete

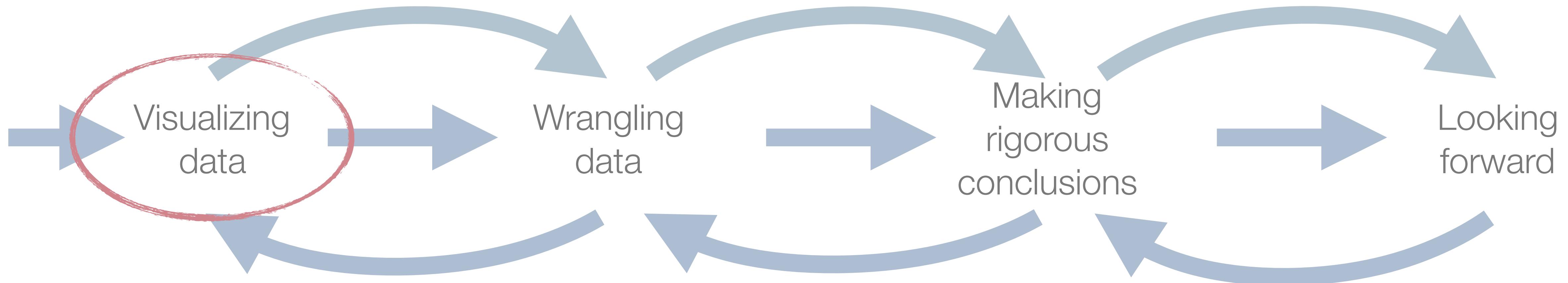
## Week 4: Simulation & Profiling

This week covers how to simulate data in R, which serves as the basis for doing simulation studies. We also cover the profiler in R which lets you collect det... [MORE](#)

6 videos (Total 42 min), 4 readings, 5 quizzes [SEE ALL](#)

# ex: Better Living with Data Science

Duke University



Fundamentals of  
data & data viz,  
confounding variables,  
Simpson's paradox  
(R + RStudio +  
R Markdown + git/GitHub)

Tidy data, data frames vs.  
summary tables,  
reencoding and transforming  
variables,  
web scraping and iteration

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

Data science ethics,  
interactive viz & reporting,  
text analysis,  
Bayesian inference,  
...

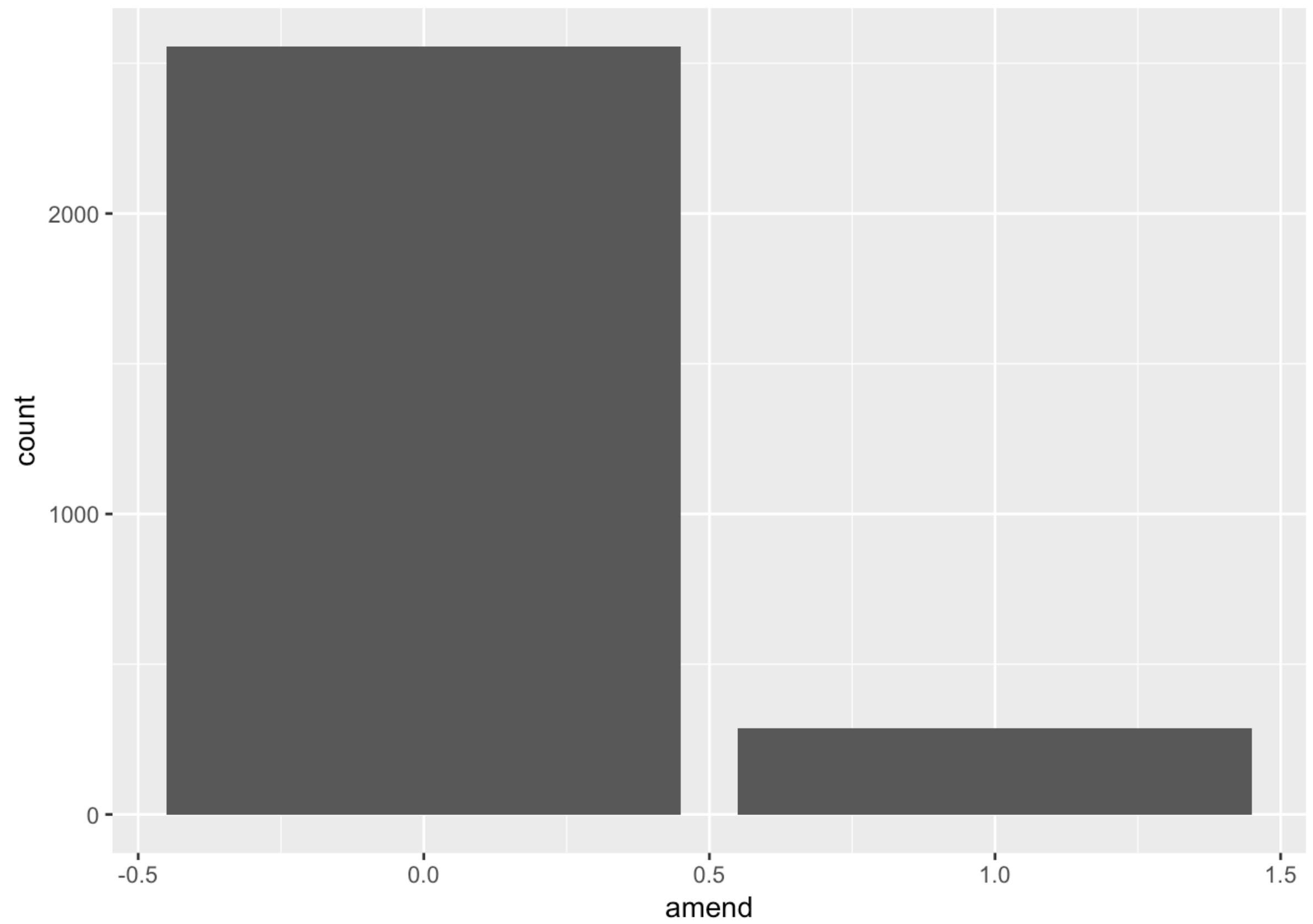
**skip  
baby  
steps**





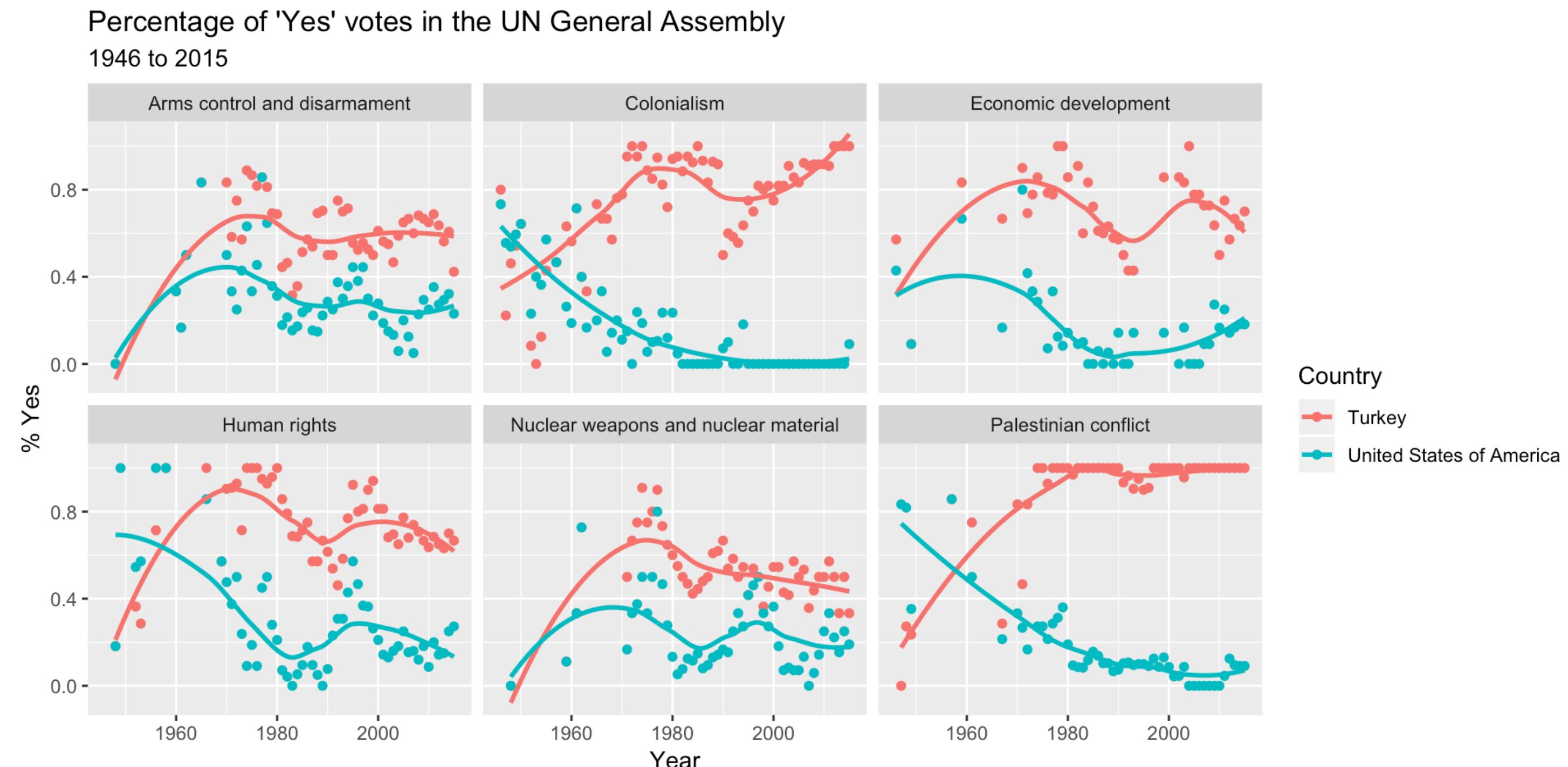
Which of the following two visualizations is more likely to **motivate** students to want to learn more?

**(a)** `ggplot(data = un_roll_calls, mapping = aes(x = amend)) +  
geom_bar()`

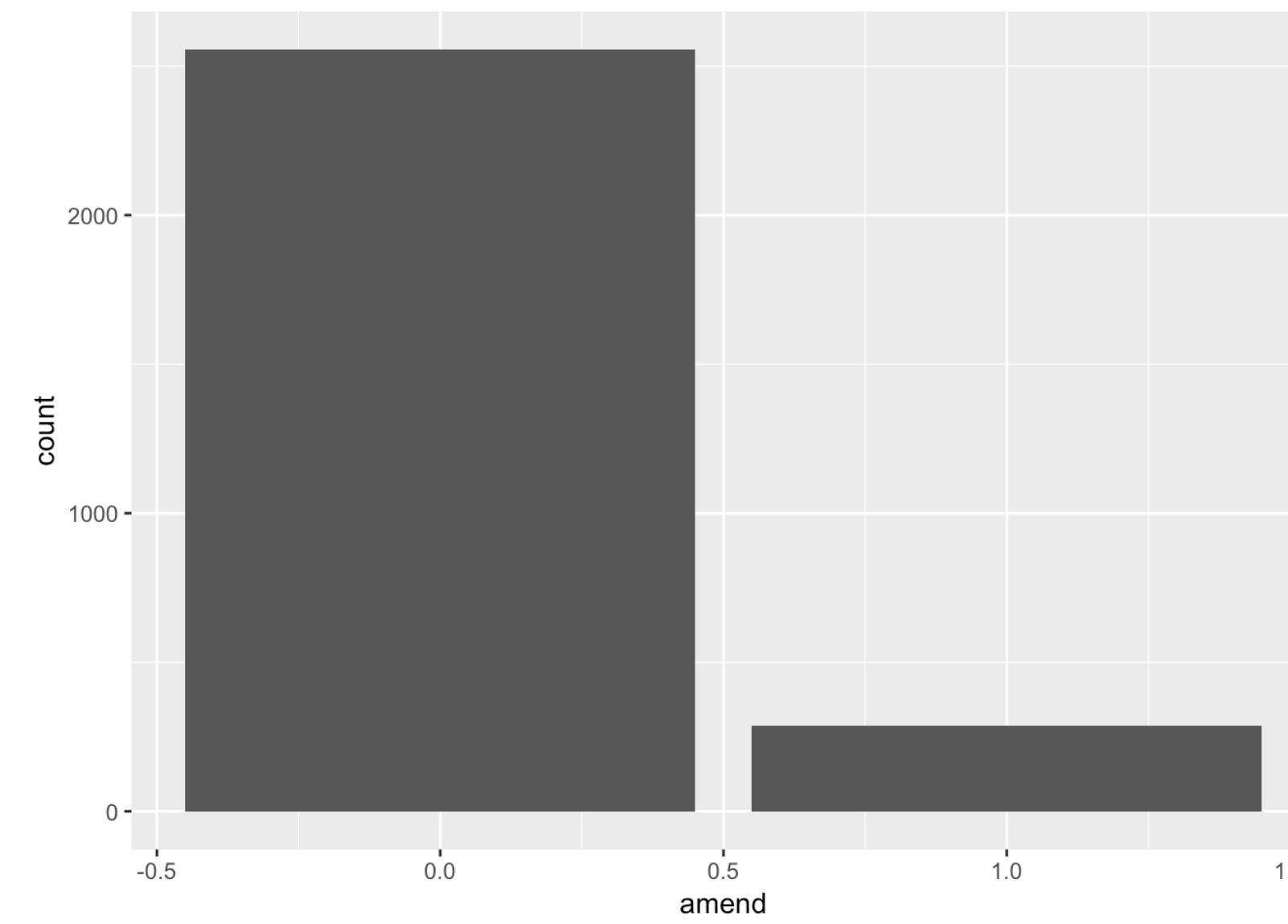


**(b)**

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

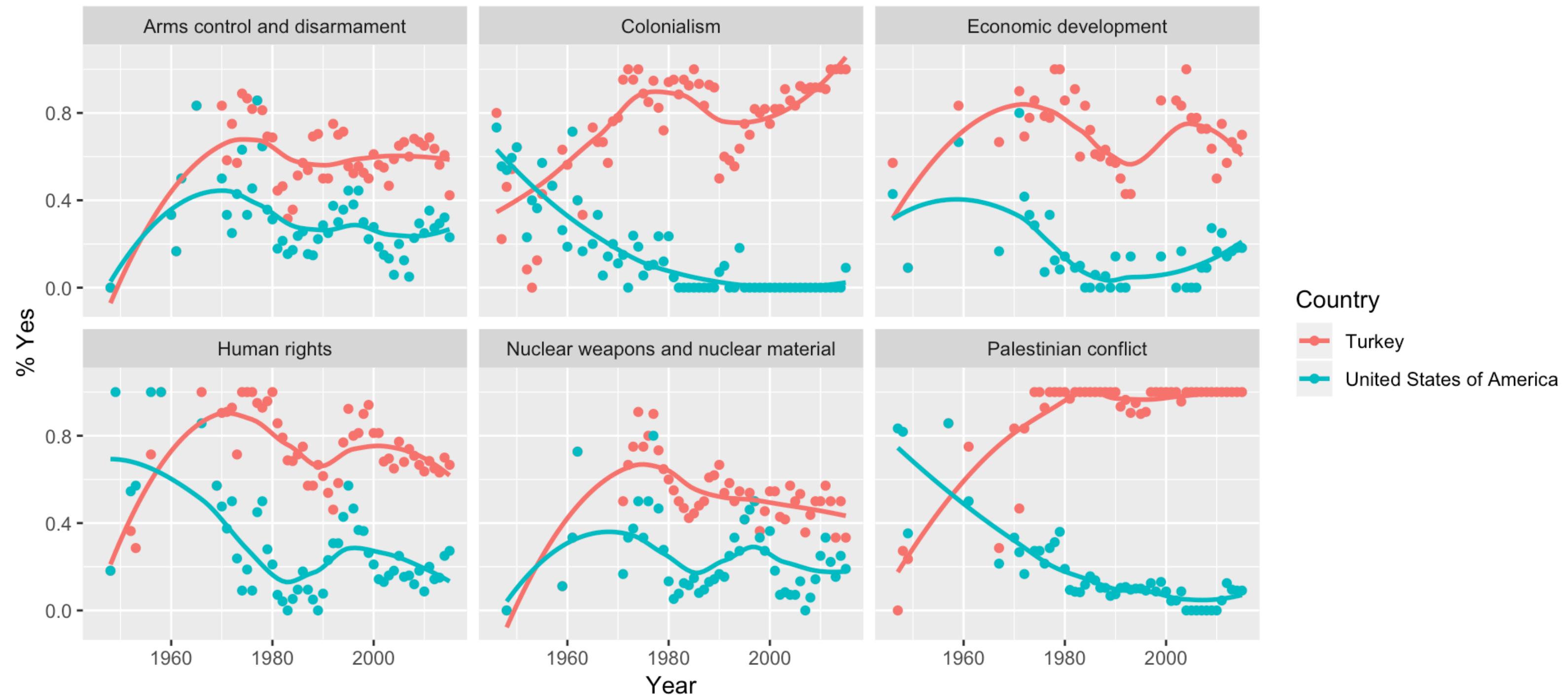


**(a)**



**(b)**

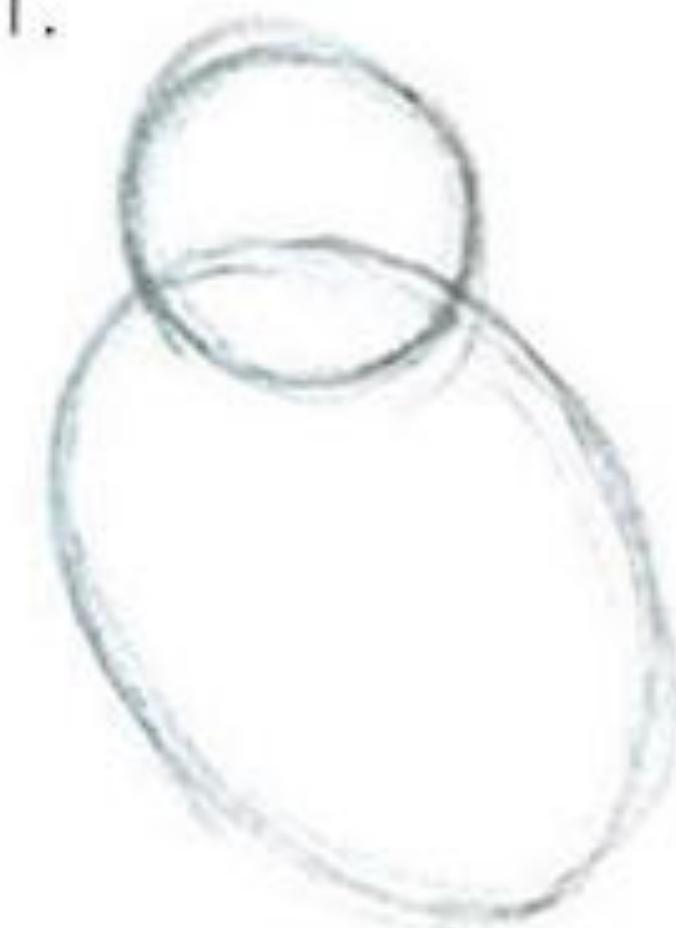
Percentage of 'Yes' votes in the UN General Assembly  
1946 to 2015



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

How to draw an owl

1.



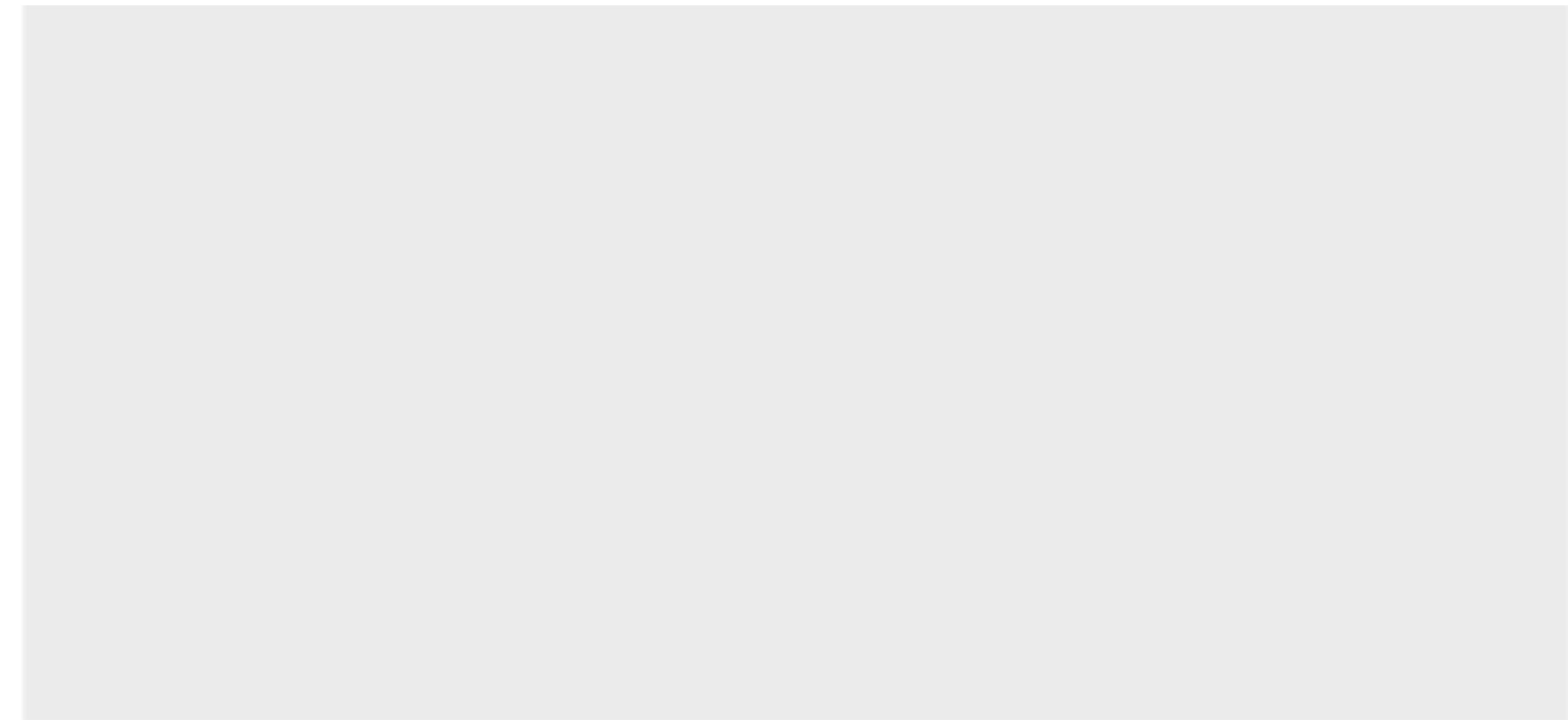
2.



1. Draw some circles

2. Draw the rest of the fucking 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))
```

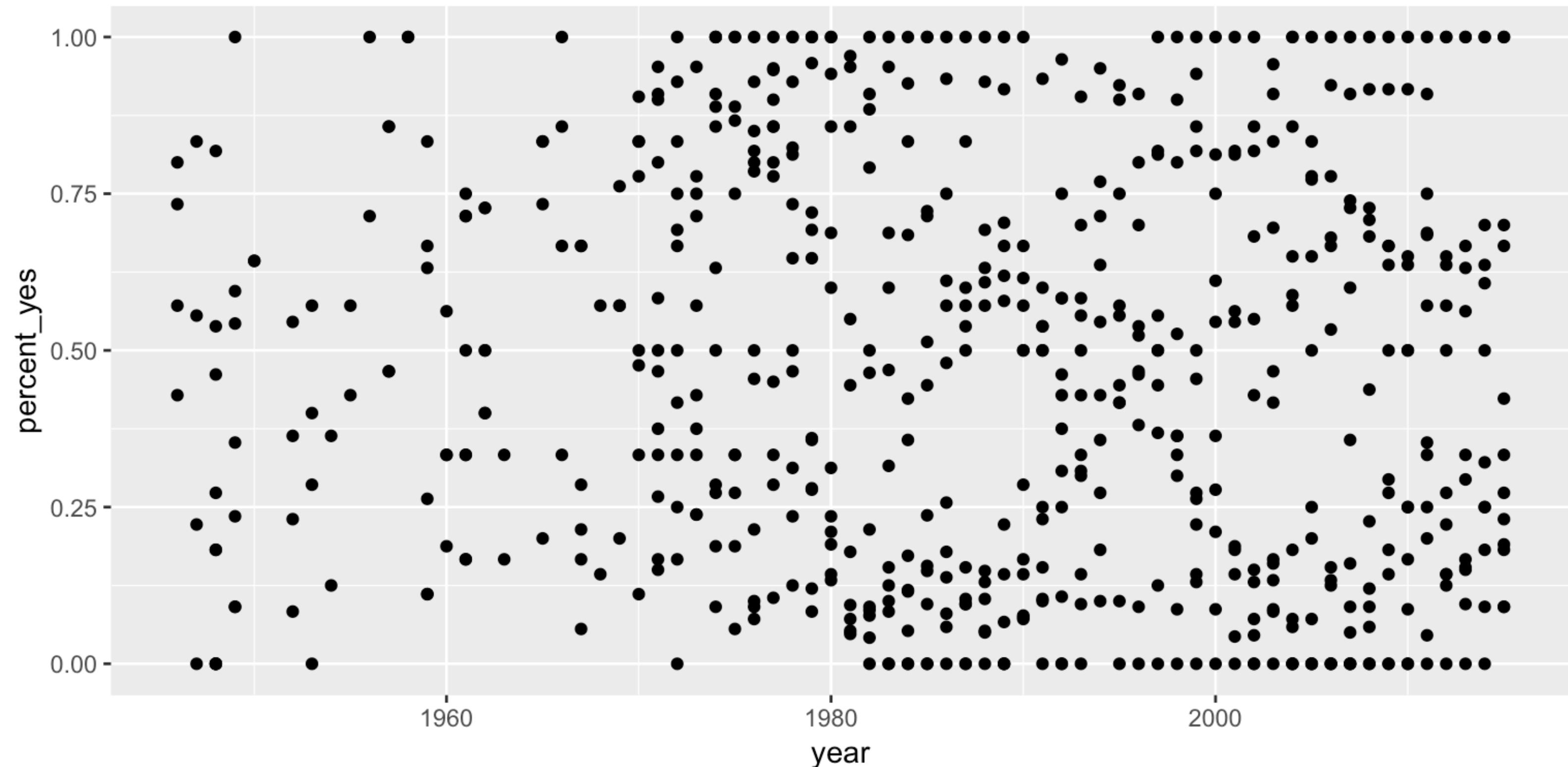
rows =  
observations

	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()
```



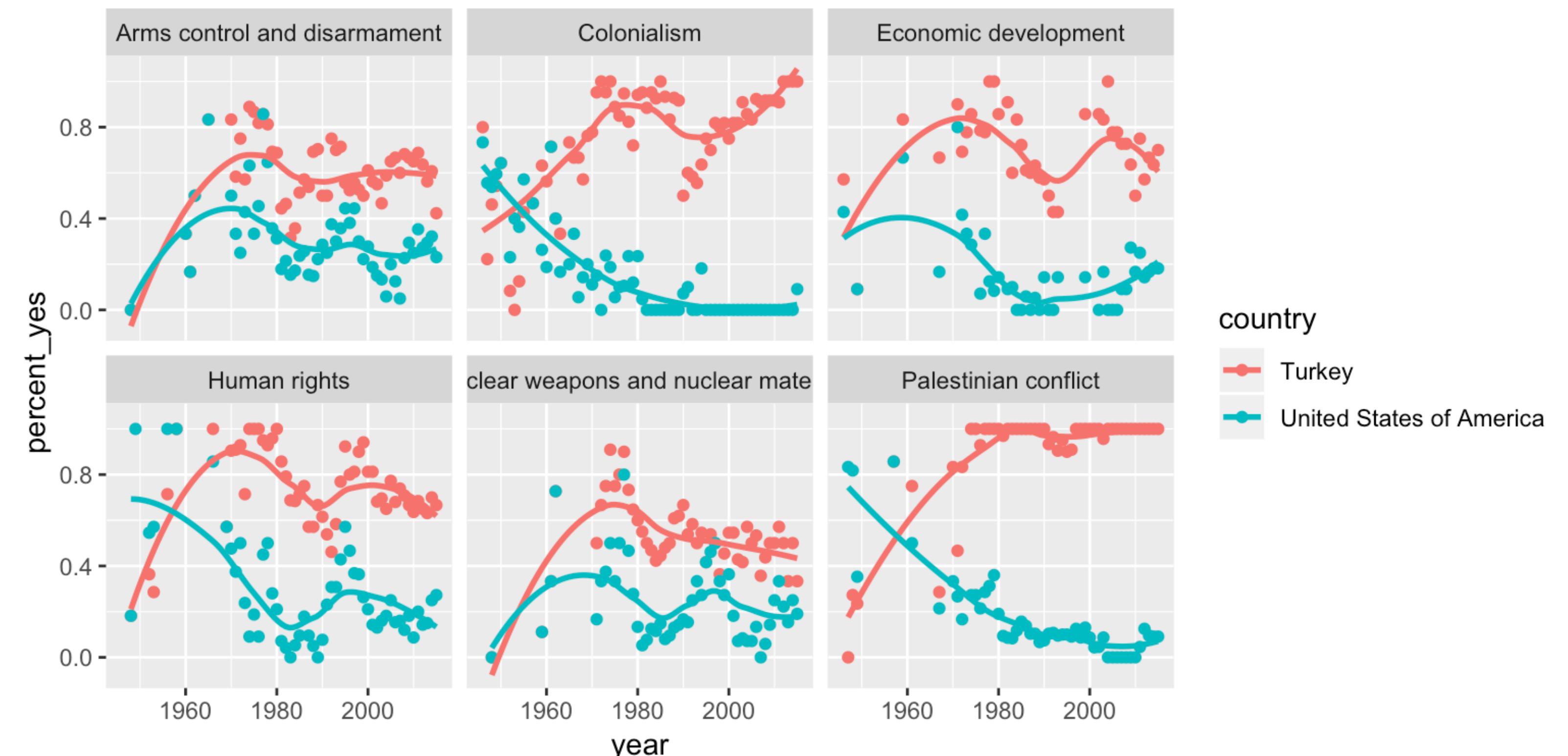
```
ggplot(data = un_votes_joined,  
       mapping = aes(x = year, y = percent_yes, color = country)) +  
geom_point()
```



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



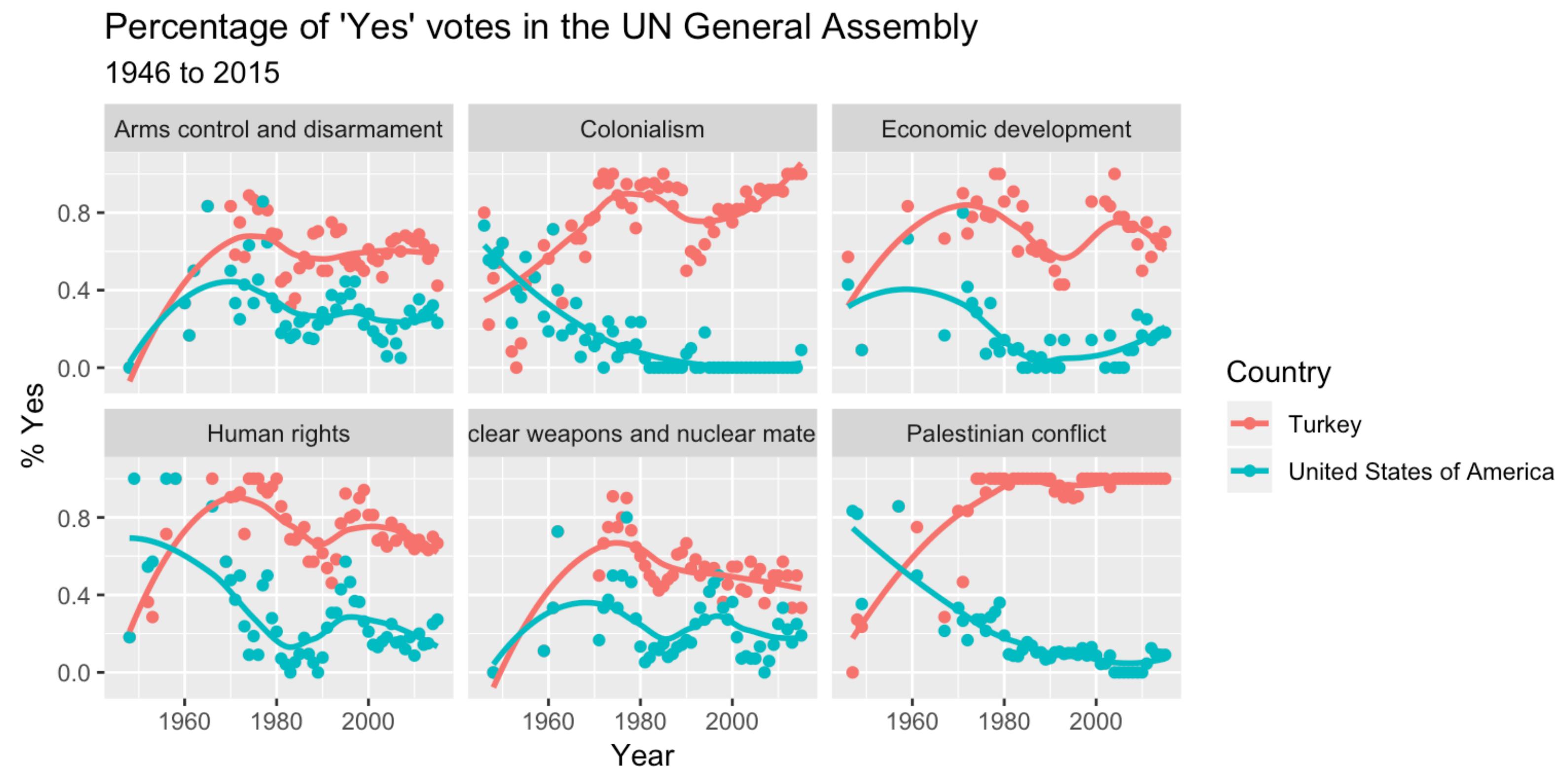
```
ggplot(data = un_votes_joined,
       mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ issue)
```



```

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

```



cherish  
day  
one

35



Which of the following two tasks is more likely to be **welcoming** for a wide range of students?

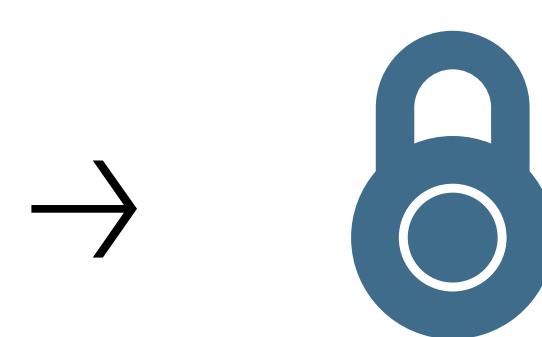
**(a)**

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

**(b)**

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

method of delivery,  
and medium of interaction matters



## 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

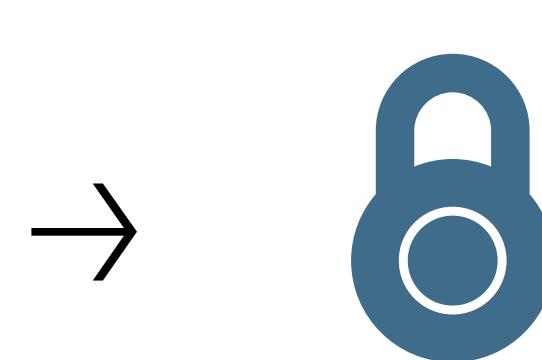
The `unvotes` package provides three datasets we can work with: `un_roll_calls`, `un_roll_call_issues`, and `un_votes`. Each of these datasets contains a variable called `roid`, the roll call id, which can be used as a unique identifier to join them with each other.

- 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> <dttm>    <dttm> <dttm> <dttm> <dttm> <dttm> <dttm>
## 1 3 1946-01-01 0 1946-01-01 1 0 AMEN. TO ADD.
## 2 4 1946-01-02 0 1946-01-02 0 0 SEC'D. TO ADD.
## 3 5 1946-01-04 0 1946-01-04 0 0 VOTED "TO ADD.
```



## 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> <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_
```

hide  
the  
veggies



ex2.

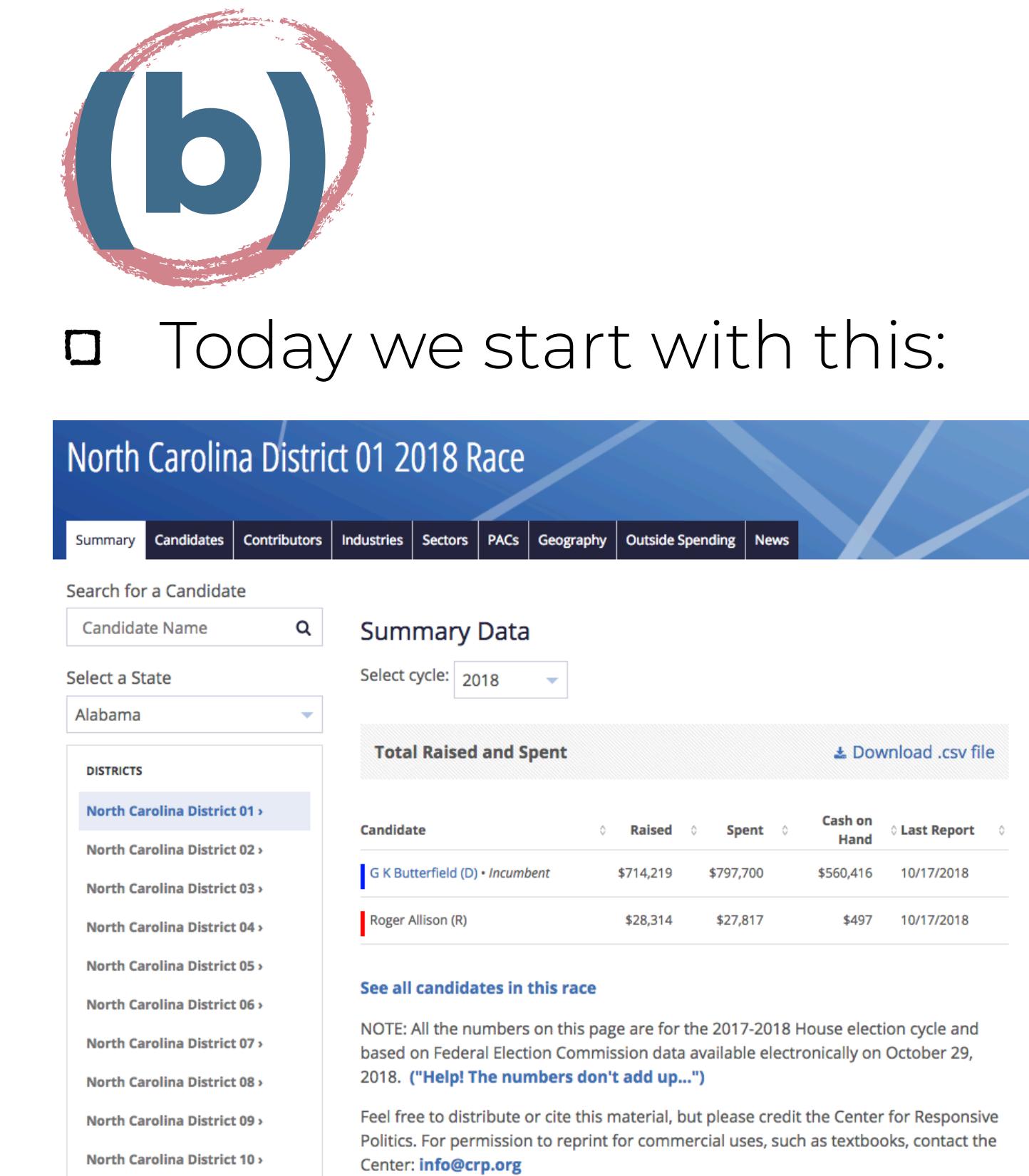
data acquisition



Which of the following two tasks is more likely to be **welcoming** for a wide range of students?

(a)

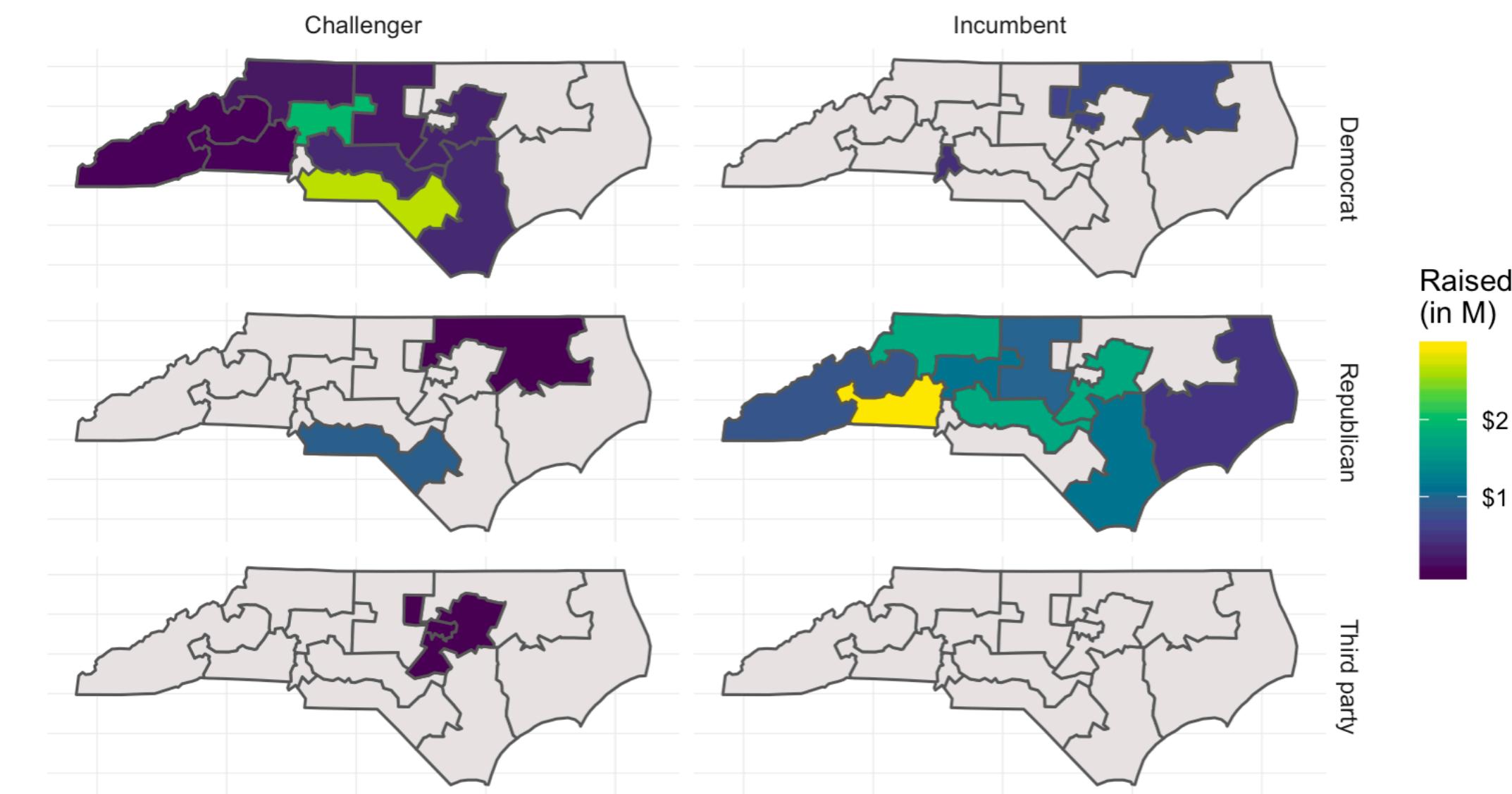
- Topic: Web scraping
- Tools: **rvest** and regular expressions



- 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.

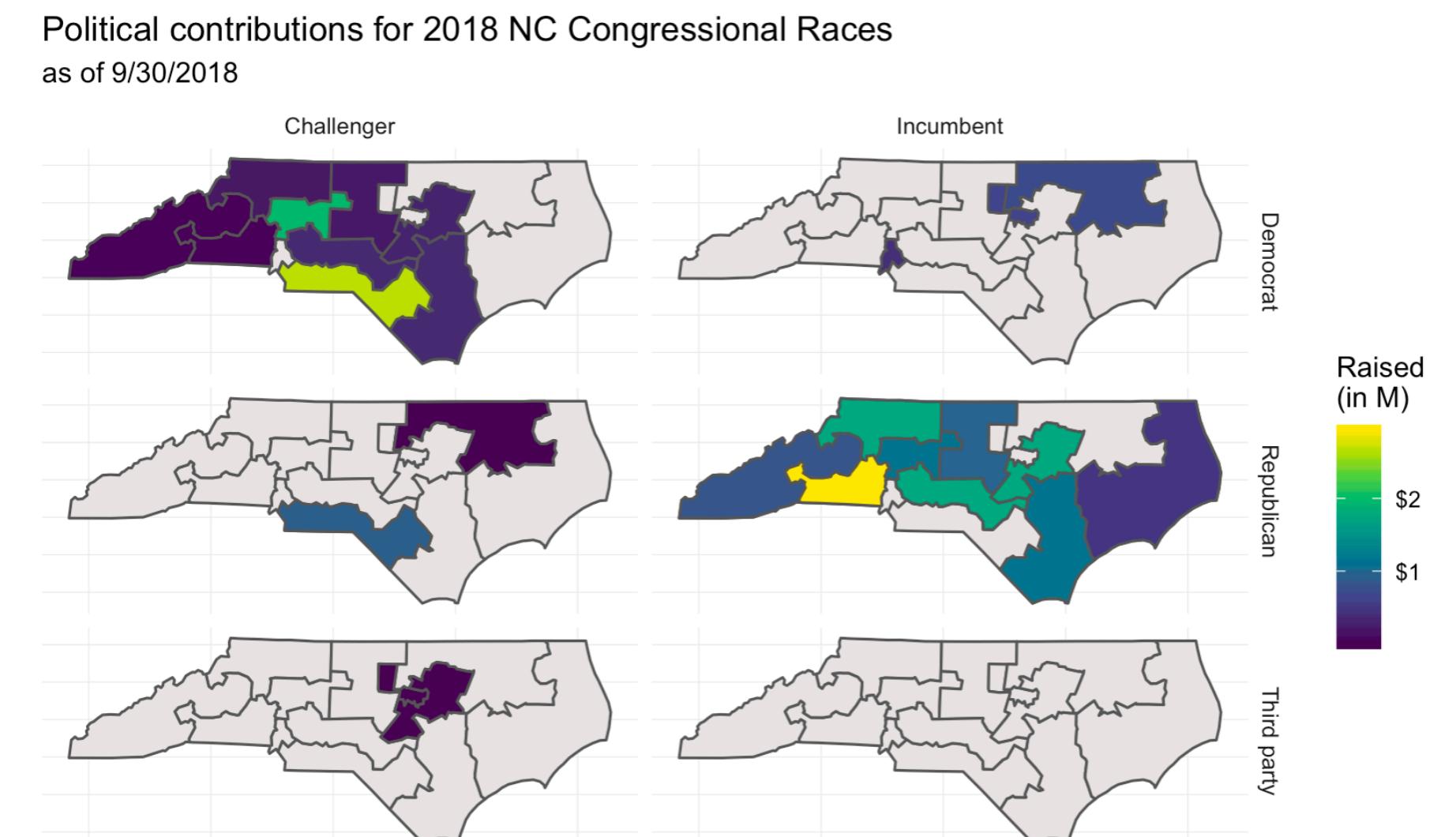
- **Ex 1:** Scrape the table off the web and save as a data frame.

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.

- **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

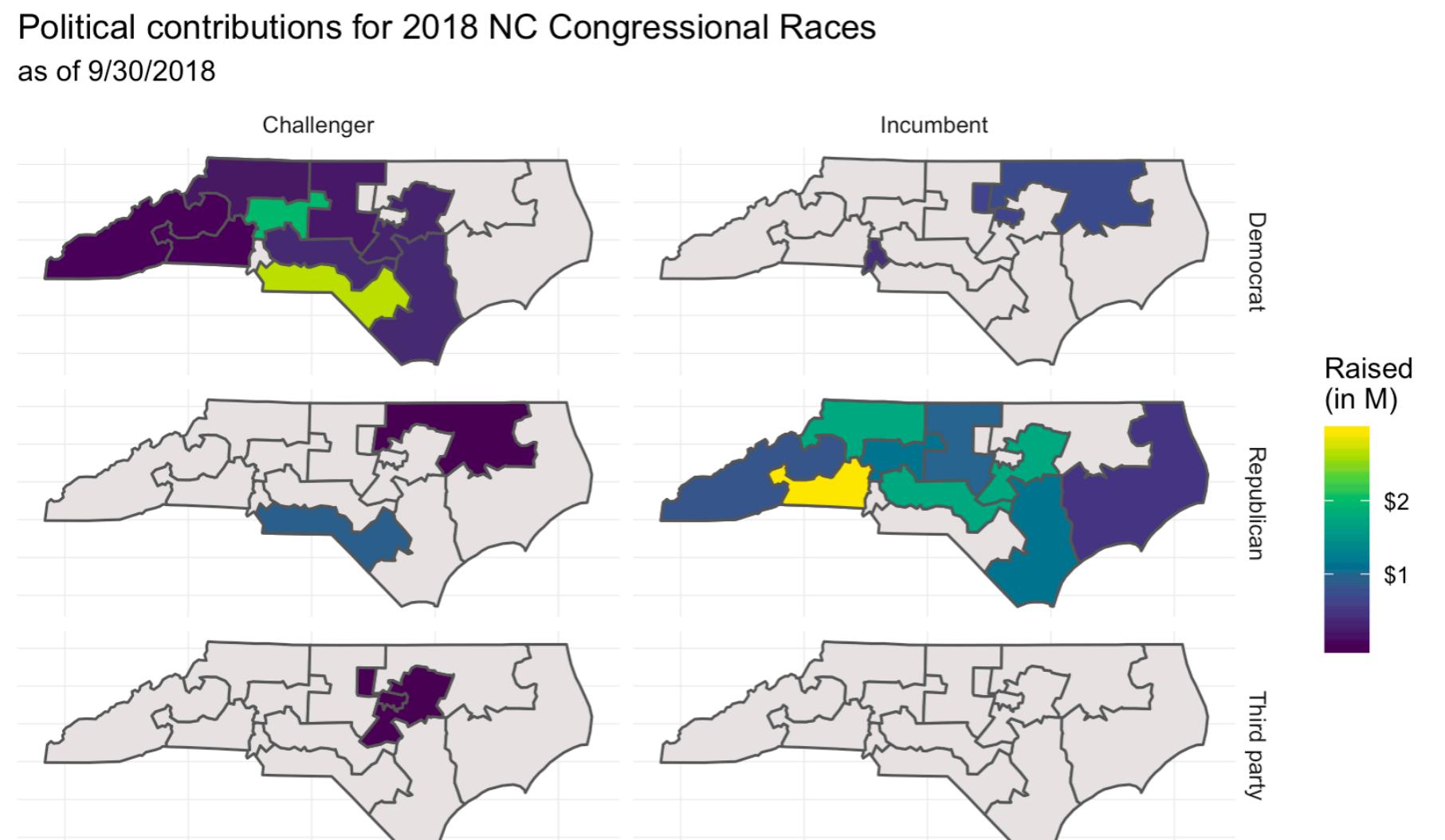
- **Lesson:** “Just enough” string parsing and regular expressions to go from

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

to

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

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



**focus  
on  
exposure**



ex<sup>3</sup>.

modeling

evaluation  
score  
(1-5)

beauty  
score  
(1-10)

	<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

```
library(broom)
```

```
lm(score ~ rank + ethnicity + gender + bty_avg, data = evals) %>%  
tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	3.78	0.114	33	4.84E-123
ranktenure track	-0.12	0.0741	-1.62	1.07E-01
ranktenured	-0.159	0.0625	-2.54	1.14E-02
Ethnicitynot minority	0.1	0.0723	1.39	1.66E-01
gendermale	0.182	0.052	3.5	5.10E-04
bty_avg	0.0728	0.0164	4.45	1.09E-05

- Write the linear model for male professors.
- Write the linear model for female professors.
- Interpret the slope of the beauty score for each.

```
library(broom)
```

```
lm(score ~ rank + ethnicity + gender*bty_avg, data = evals) %>%  
tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	3.93	0.144	27.2	1.58E-97
ranktenure track	-0.109	0.0742	-1.46	1.44E-01
ranktenured	-0.135	0.064	-2.1	3.6E-02
ethnicitynot minority	0.0764	0.0735	1.04	2.99E-01
gendermale	-0.0793	0.161	-0.493	6.23E-01
bty_avg	0.0416	0.0245	1.7	8.97E-02
gendermale:bty_avg	0.0579	0.0338	1.71	8.73E-02

- Write the linear model for male professors.
- Write the linear model for female professors.
- Interpret the slope of the beauty score for each. What changed?

twiday

- 1 start with cake
- 2 skip baby steps
- 3 cherish day one
- 4 hide the veggies
- 5 focus on exposure



Fine,  
I'm intrigued,  
but I need to see  
the big picture

# 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.

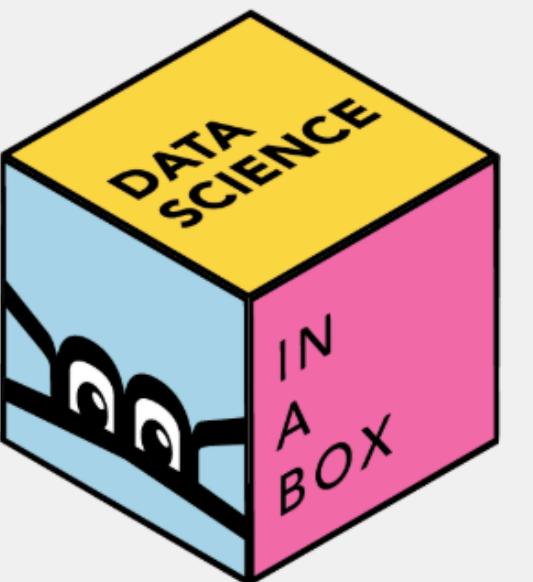
GAISE 2016, [http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege\\_Full.pdf](http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf).

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

 Search...

Hello #dsbox

Overview

Philosophy

Topics

Tech stack

Community

Course content

Infrastructure

Pedagogy



bit.ly/let-eat-cake

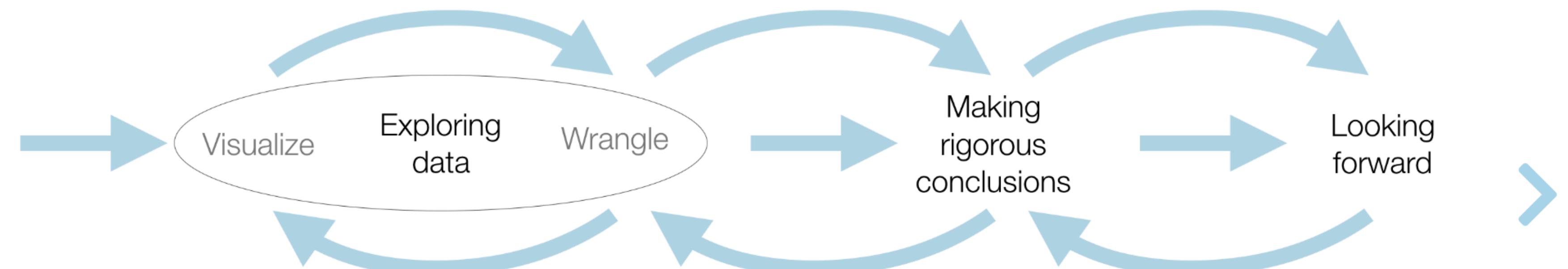


Data Science in a Box &gt; Hello #dsbox &gt; Topics



# 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.

Let them eat cake (first)!\*

↳ [bit.ly/let-eat-cake](http://bit.ly/let-eat-cake)

</> [bit.ly/repo-eat-cake](http://bit.ly/repo-eat-cake)

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



@minebocek



mine-cetinkaya-rundel



cetinkaya.mine@gmail.com

