Let them eat cake (first)!

@minebocek
mine-cetinkaya-rundel
mcetinka@ed.ac.uk

bit.ly/eat-cake-cetl-cerse

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?

bit.ly/eat-cake-cetl-cerse

Pineapple and Coconut Sandwich Cake

bit.ly/eat-cake-cetl-msor

Pineapple and Coconut Sandwich Cake

3 misconceptions

1 context

5 design principles

3

miscon-ceptions

bit.ly/eat-cake-cetl-cerse

“Students should learn a new programming language in the order in which **it’s been taught forever**.”

“Students should learn
a new programming language starting with
data structures and algorithms.”

“Students approach a new programming language
the same way a software engineer does.”

context

1

Introduction to Data Science

Fall 2019

University of Edinburgh

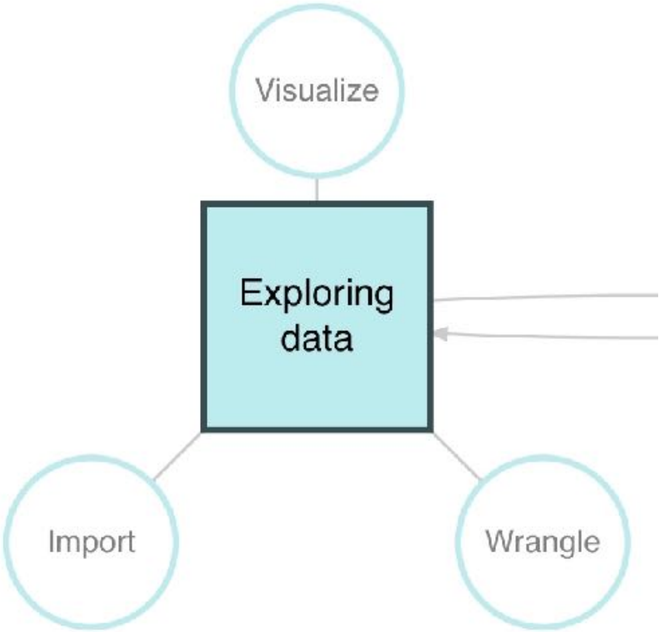
🔗 introds.org

bit.ly/eat-cake-cetl-cerse

Introduction to Data Science

Fall 2019

University of Edinburgh

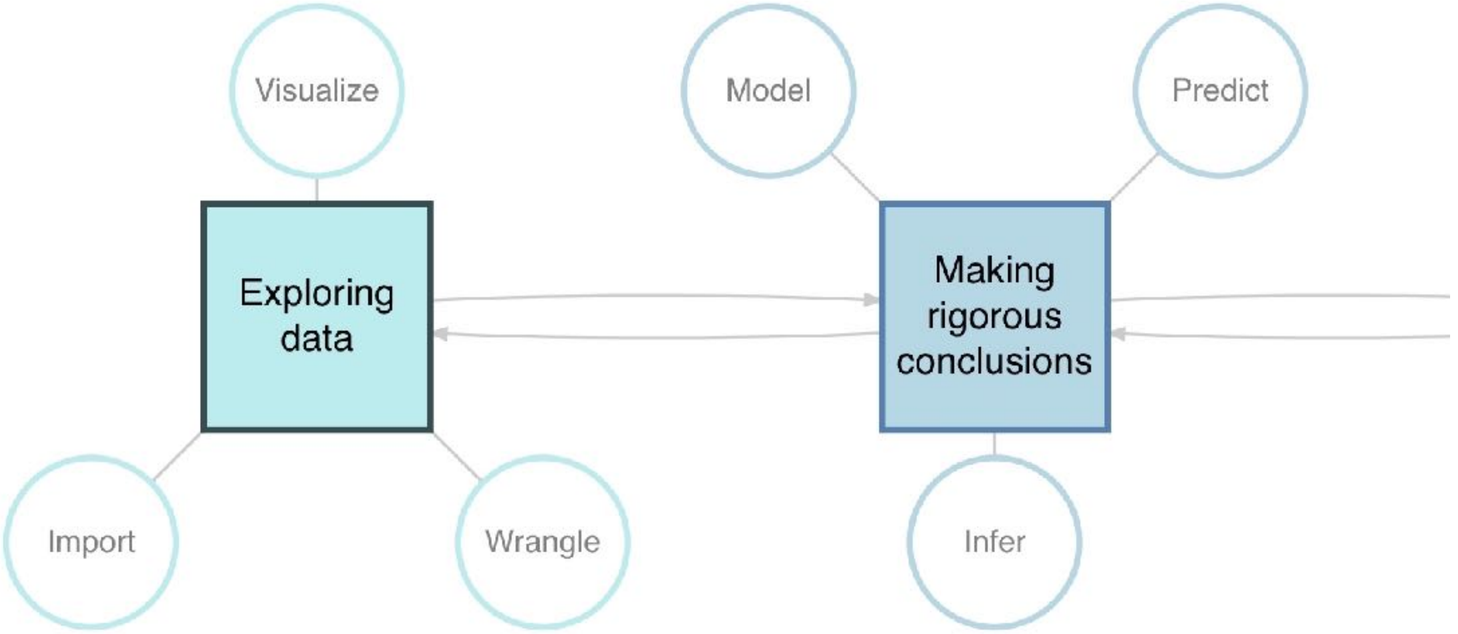


🔗 introds.org

Introduction to Data Science

Fall 2019

University of Edinburgh



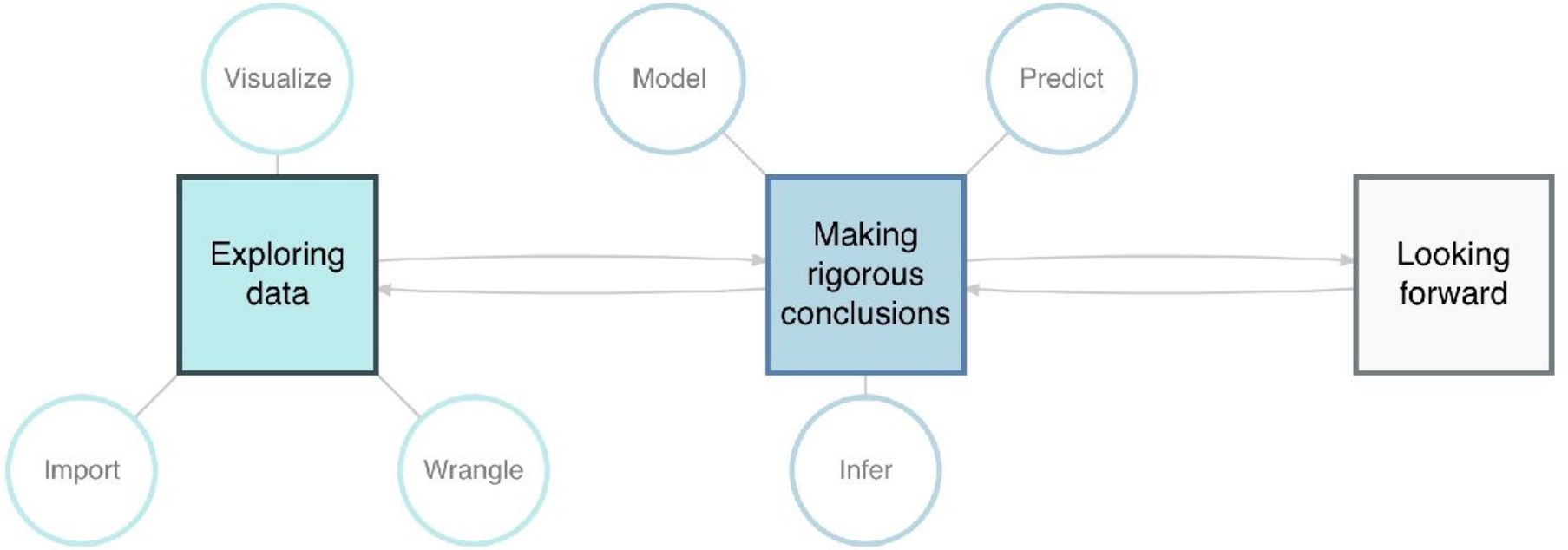
🔗 introds.org

Introduction to Data Science

Fall 2019

University of Edinburgh

🔗 introds.org



bit.ly/eat-cake-cetl-cerse

5 design principles

bit.ly/eat-cake-cetl-cerse

Q

Which kitchen would you rather bake a cake?

bit.ly/eat-cake-cetl-cerse

Q

Which kitchen would you rather bake a cake?

bit.ly/eat-cake-cetl-cerse

cherish

day

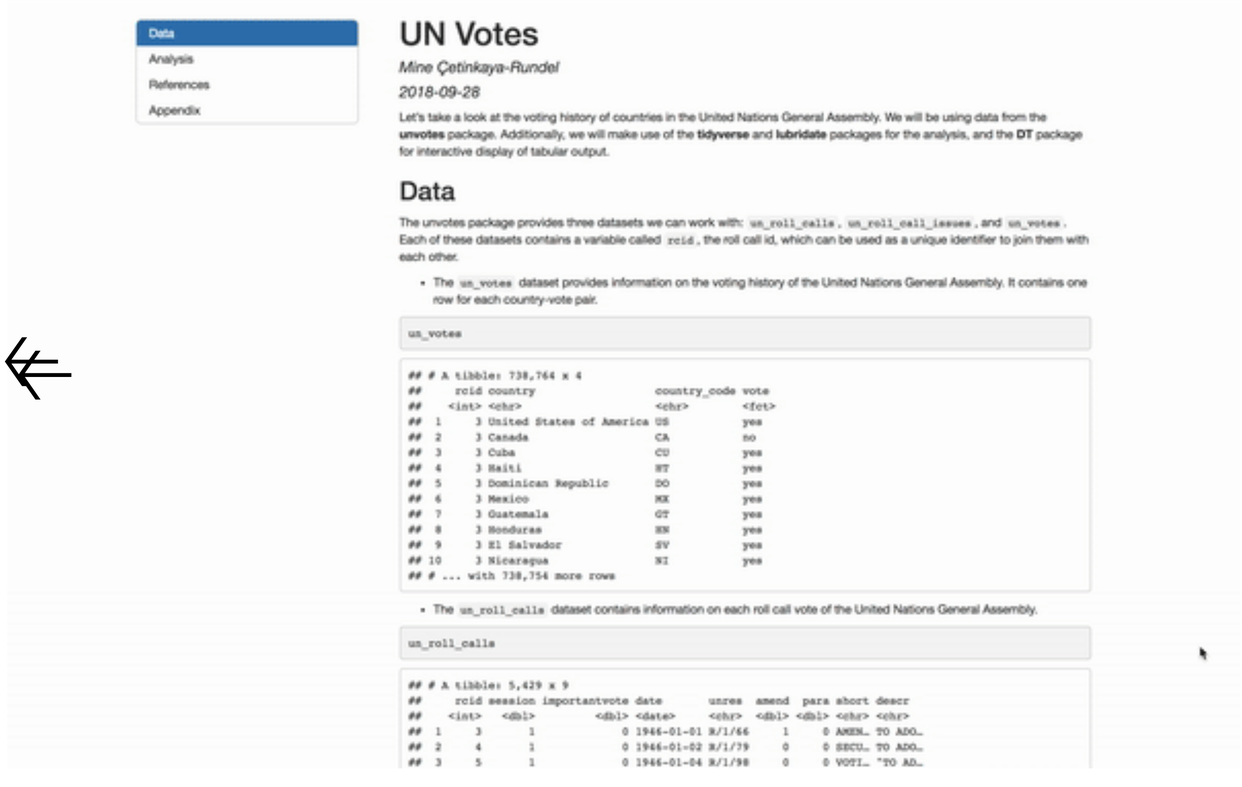
one

bit.ly/eat-cake-cetl-cerse

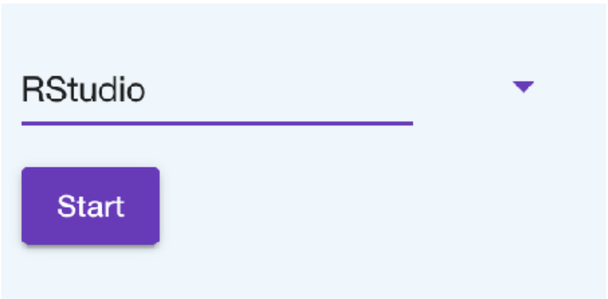
- Install R
- Install RStudio
- Install the following packages:
  - tidyverse
  - rmarkdown
  - ...
- Load these packages
- Install git

- Go to RStudio in the cloud
- Log in with your ID & pass

```
> hello R!
```



bit.ly/eat-cake-cetl-cerse


RStudio
Start

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 (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

bit.ly/eat-cake-cetl-cerse

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

bit.ly/eat-cake-cetl-cerse

start with cake

bit.ly/eat-cake-cetl-cerse

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

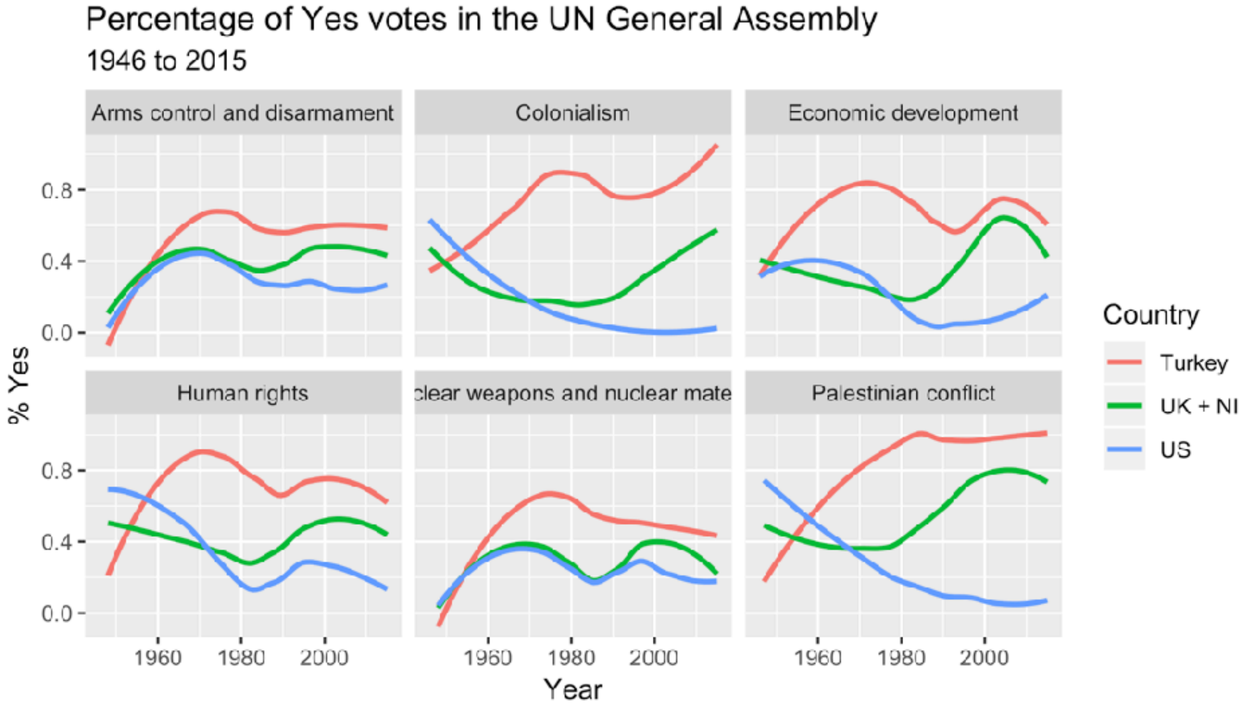
1. **For the Cake:** Adjust oven rack to lower-middle position and preheat to 350°F (180°C). Lightly grease three 8-inch anodized aluminum cake pans and line with parchment (explanation and tutorial here). If you don't have three pans, it's okay to bake the cakes in stages, the batter will keep at room 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

- 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"
class(z)
#> [1] "logical"
```

bit.ly/eat-cake-cetl-cerse

- 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...

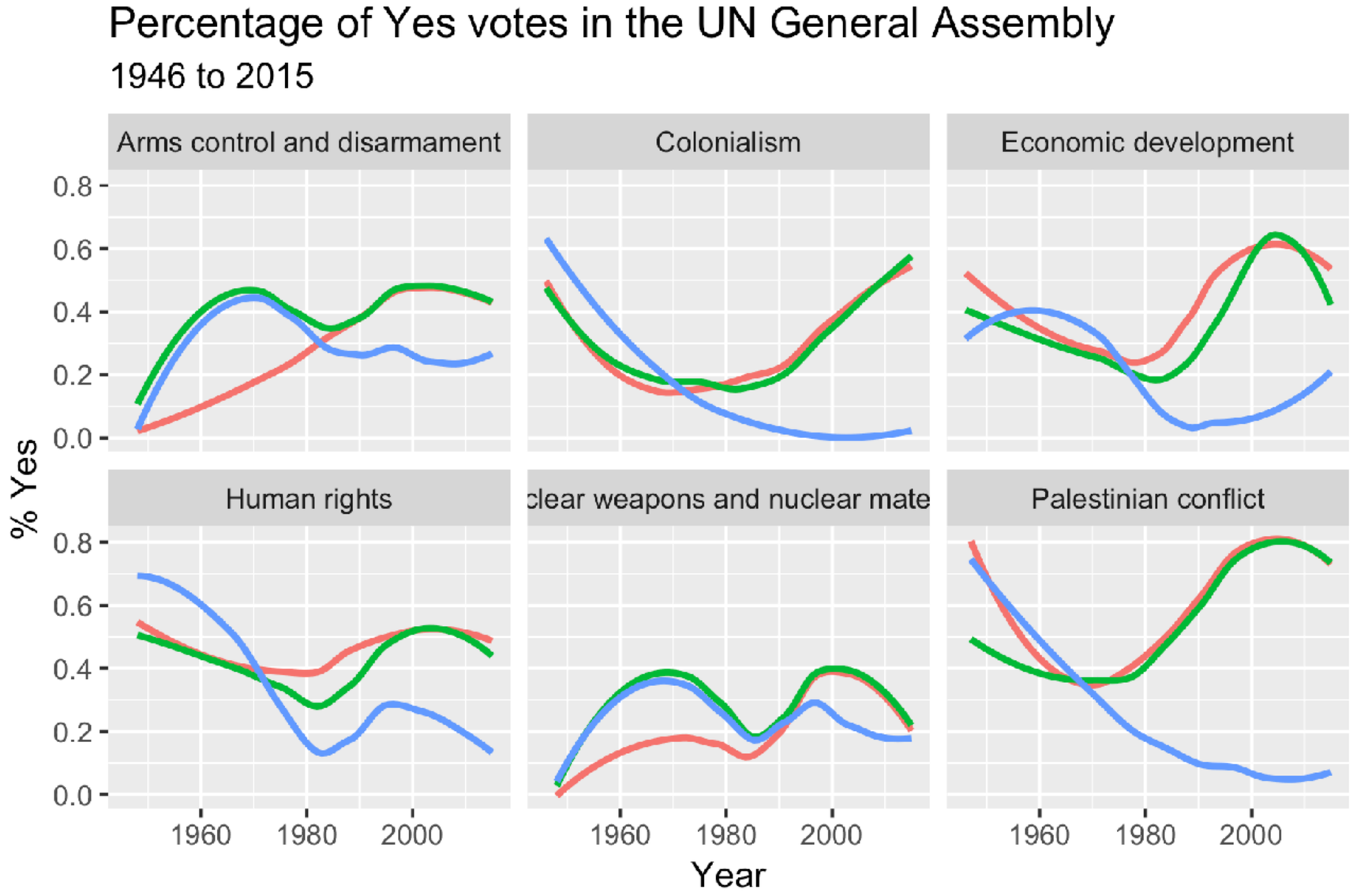
but let's focus on the task at hand...

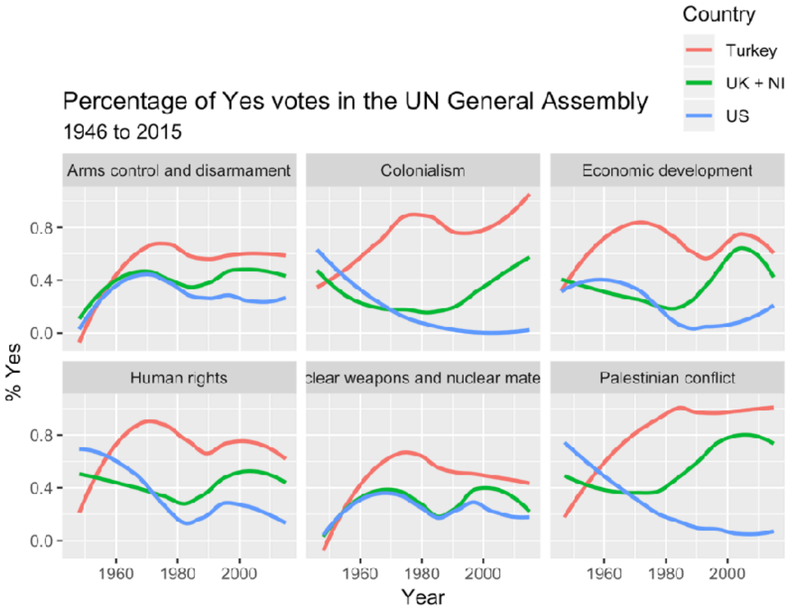
- Open today's demo project
- Knit the document and discuss the results with your neighbor
- Then, change `Turkey` to a different country, and plot again

```r
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"
  )
```

```r
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"
  )
```

```r
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"
  )
```





bit.ly/eat-cake-cetl-cerse

Q

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

bit.ly/eat-cake-cetl-cerse

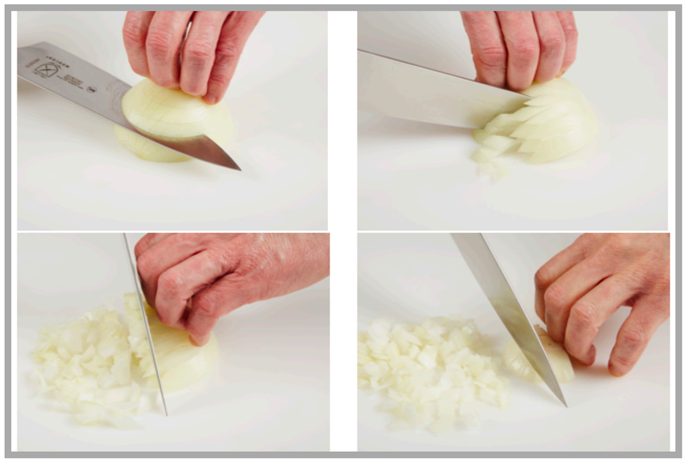
Q

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

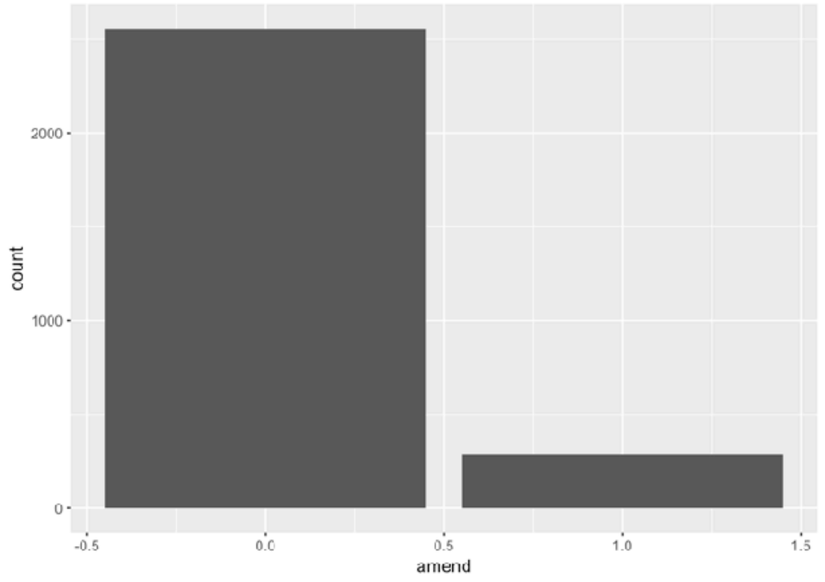
bit.ly/eat-cake-cetl-cerse

skip baby steps

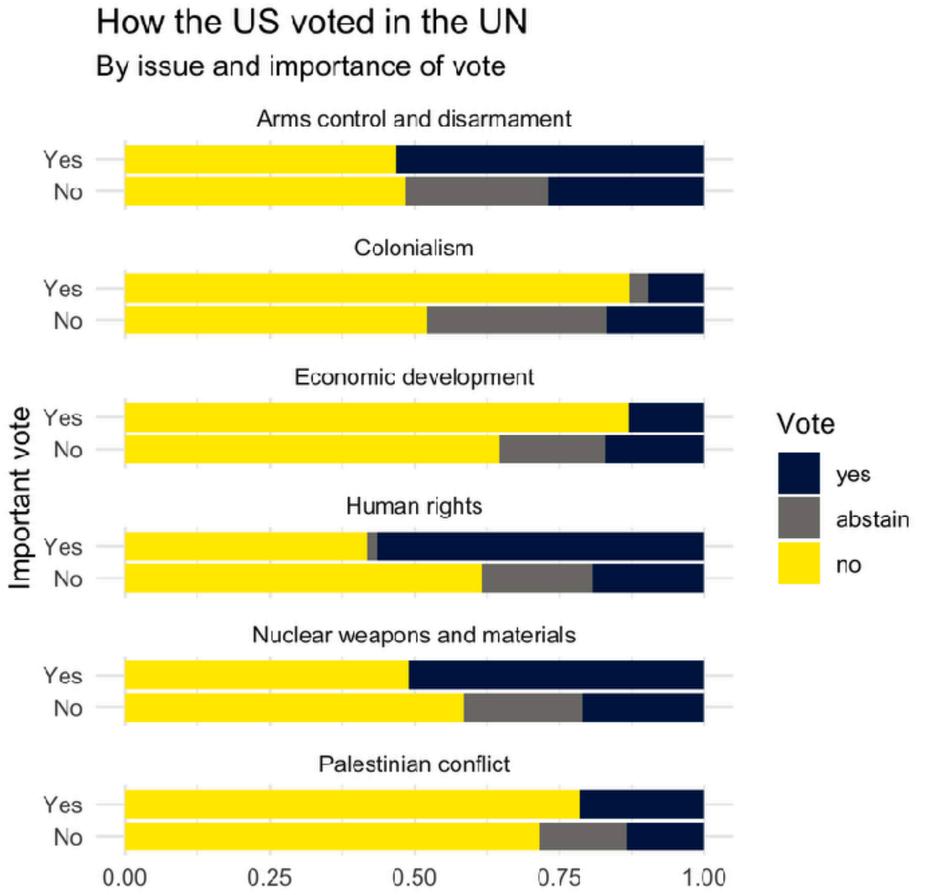
bit.ly/eat-cake-cetl-cerse



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

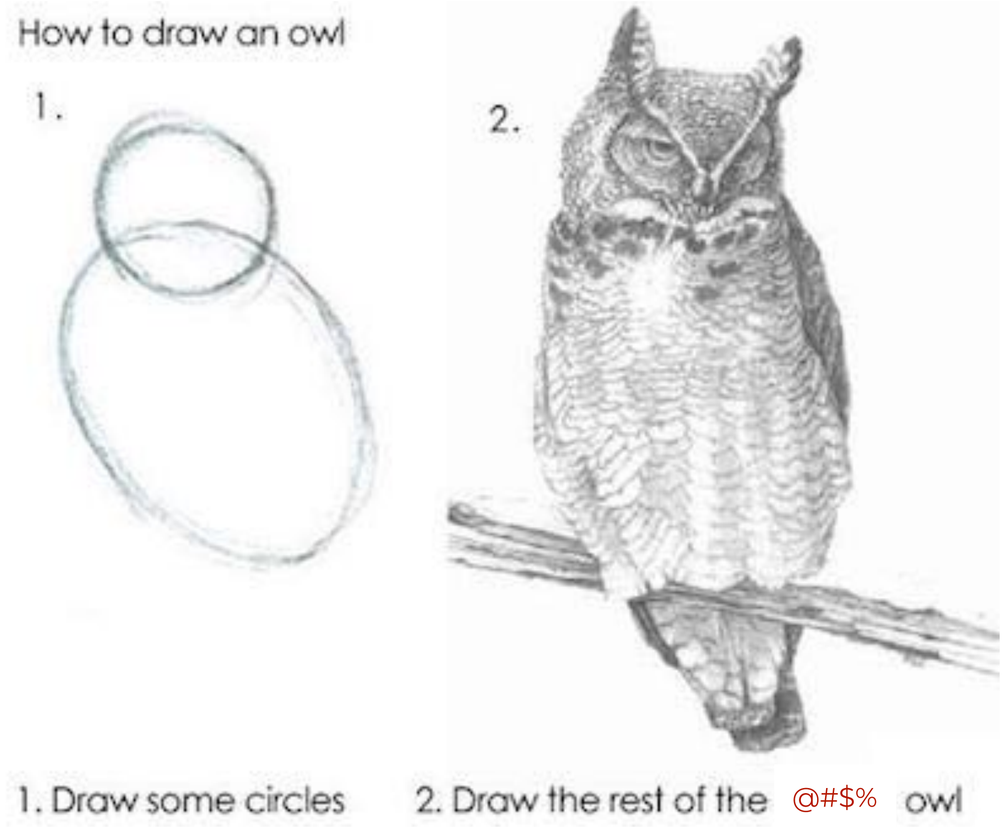


bit.ly/eat-cake-cetl-cerse



non-trivial examples can be motivating,

but need to avoid 👇!



bit.ly/eat-cake-cetl-cerse

How to draw an owl

1.

1. Draw some circles

2.

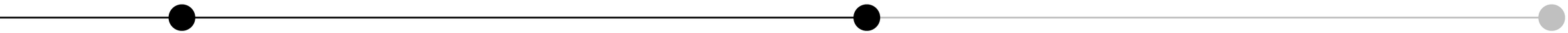
2. Draw the rest of the @#$% owl

scaffold + layer

bit.ly/eat-cake-cetl-cerse

Write it out to your
heart's desire
and polish it

bit.ly/eat-cake-cetl-cerse



Write it out to your heart's desire and polish it

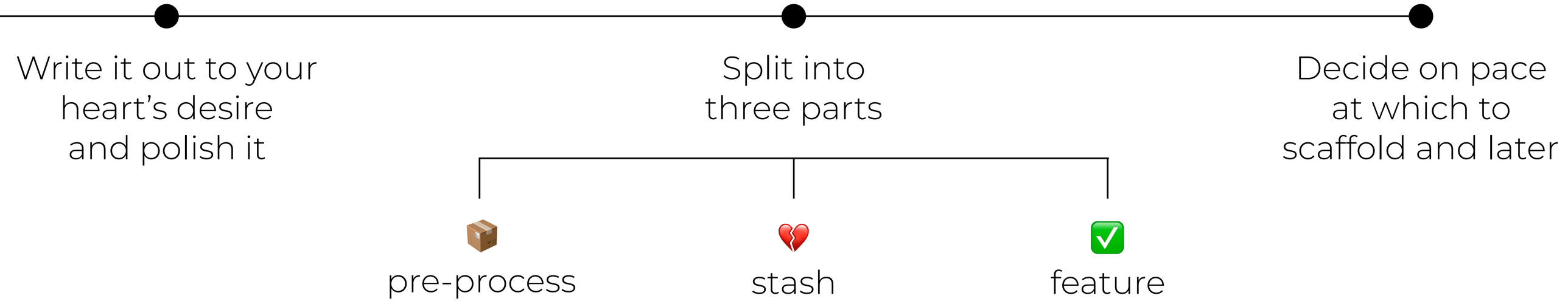
Split into three parts

📦 pre-process

💔 stash

✅ feature

bit.ly/eat-cake-cetl-cerse



bit.ly/eat-cake-cetl-cerse

```r
un_votes %>%
  filter(country %in% c("United States of America")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  mutate(
    importantvote = ifelse(importantvote == 0, "No", "Yes"),
    issue = ifelse(issue == "Nuclear weapons and nuclear material",
                   "Nuclear weapons and materials", issue)
  ) %>%
  ggplot(aes(y = importantvote, fill = vote)) +
  geom_bar(position = "fill") +
  facet_wrap(~ issue, ncol = 1) +
  labs(
    title = "How the US voted in the UN",
    subtitle = "By issue and importance of vote",
    x = "Important vote", y = "", fill = "Vote"
  ) +
 theme_minimal() +
  scale_fill_viridis_d(option = "E")
```

📦

pre-process

us_votes

```
un_votes %>%
  filter(country %in% c("United States of America")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  mutate(
    importantvote = ifelse(importantvote == 0, "No", "Yes"),
    issue = ifelse(issue == "Nuclear weapons and nuclear material",
                   "Nuclear weapons and materials", issue)
  ) %>%
  ggplot(aes(y = importantvote, fill = vote)) +
  geom_bar(position = "fill") +
  facet_wrap(~ issue, ncol = 1) +
  labs(
    title = "How the US voted in the UN",
    subtitle = "By issue and importance of vote",
    x = "Important vote", y = "", fill = "Vote"
  ) +
  theme_minimal() +
  scale_fill_viridis_d(option = "E")
```

✅

feature

💔

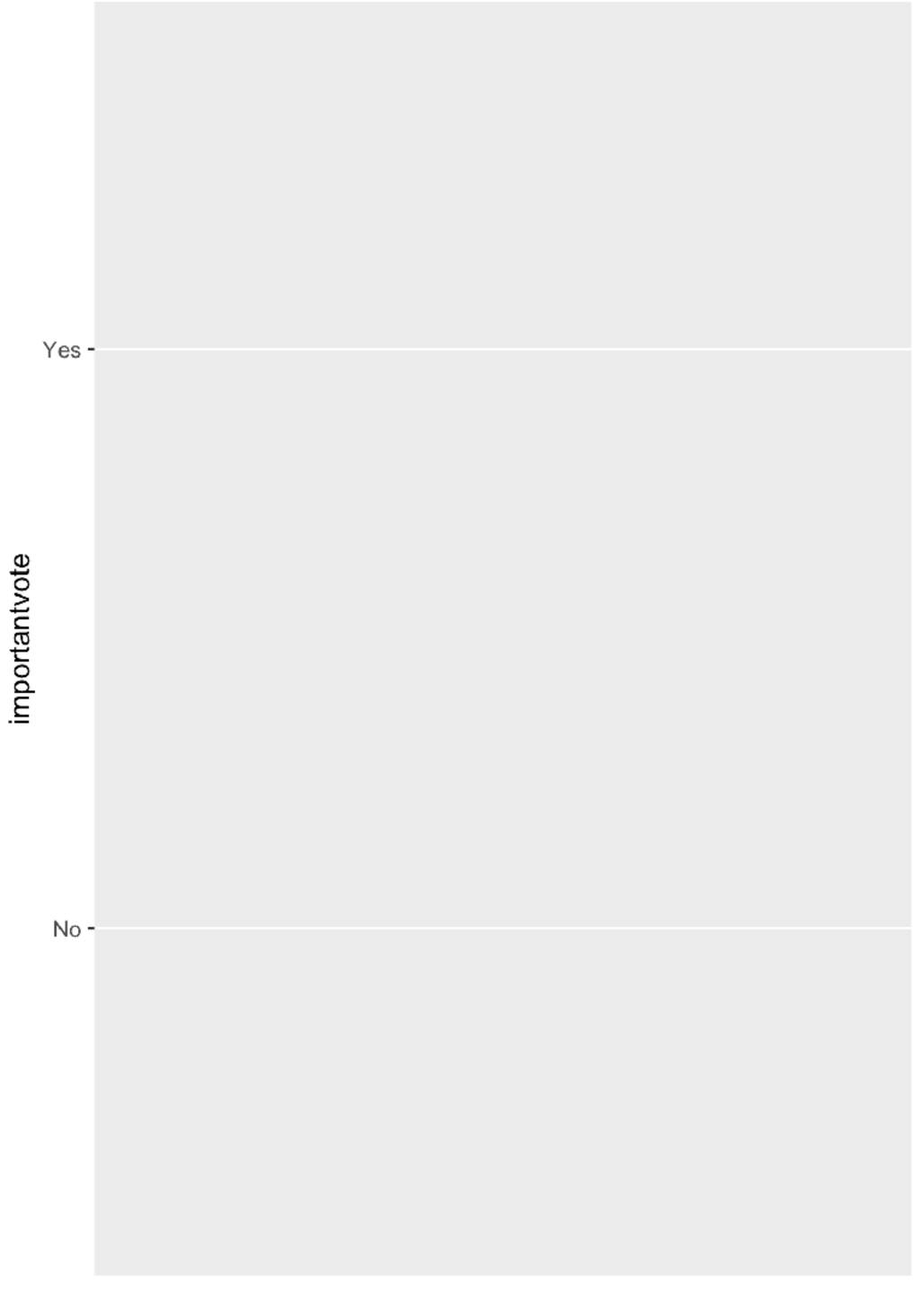
stash

bit.ly/eat-cake-cetl-cerse

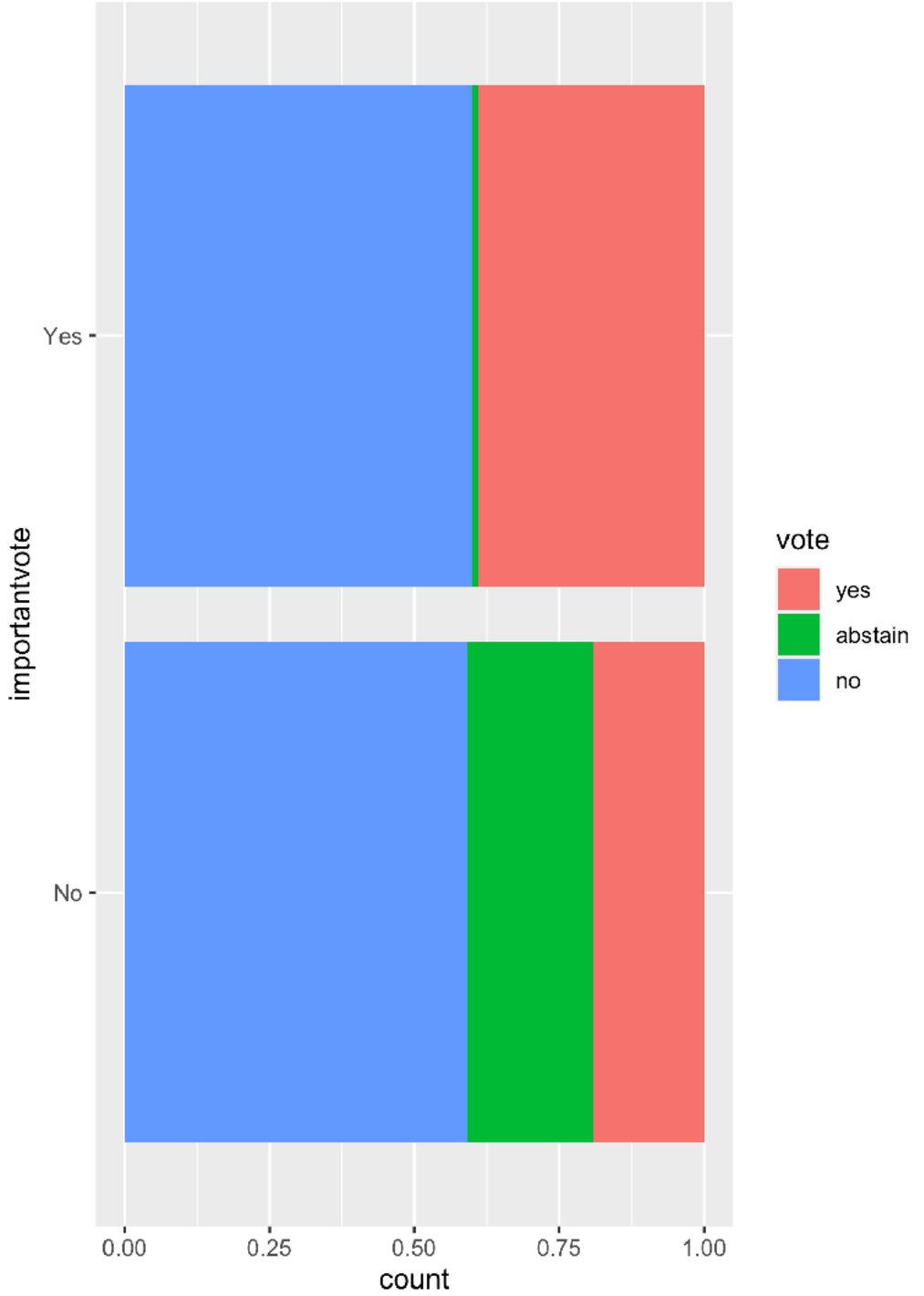
```
ggplot(us_votes)
```


bit.ly/eat-cake-cetl-cerse

```
ggplot(us_votes,
       aes(y = importantvote, fill = vote))
```

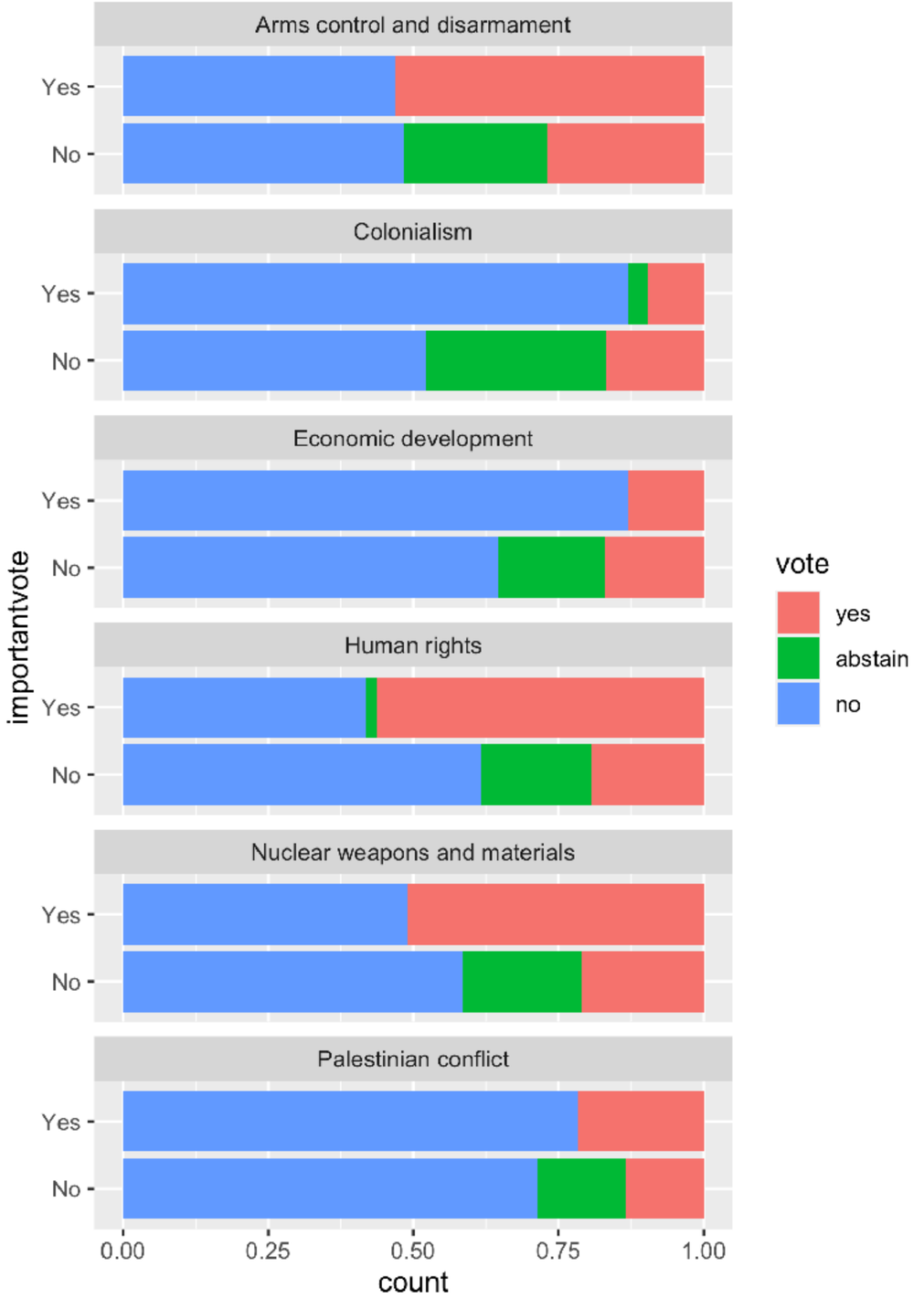


```
ggplot(us_votes,
       aes(y = importantvote, fill = vote)) +
  geom_bar(position = "fill")
```

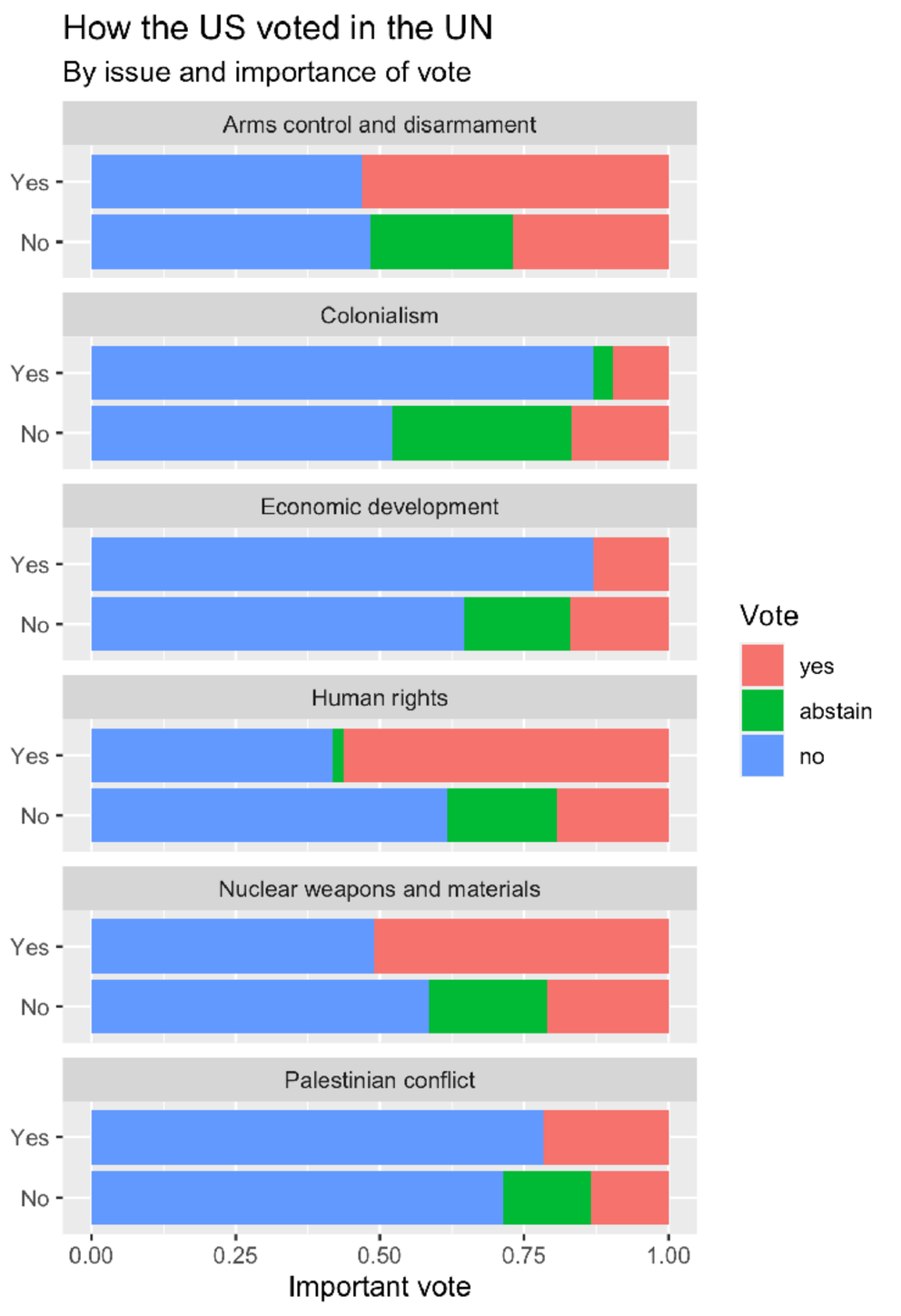


bit.ly/eat-cake-cetl-cerse

```
ggplot(us_votes,
       aes(y = importantvote, fill = vote)) +
  geom_bar(position = "fill") +
  facet_wrap(~ issue, ncol = 1)
```



```
ggplot(us_votes,
       aes(y = importantvote, fill = vote)) +
  geom_bar(position = "fill") +
  facet_wrap(~ issue, ncol = 1) +
  labs(
    title = "How the US voted in the UN",
    subtitle = "By issue and importance of vote",
    x = "Important vote", y = "", fill = "Vote"
  )
```



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

bit.ly/eat-cake-cetl-cerse

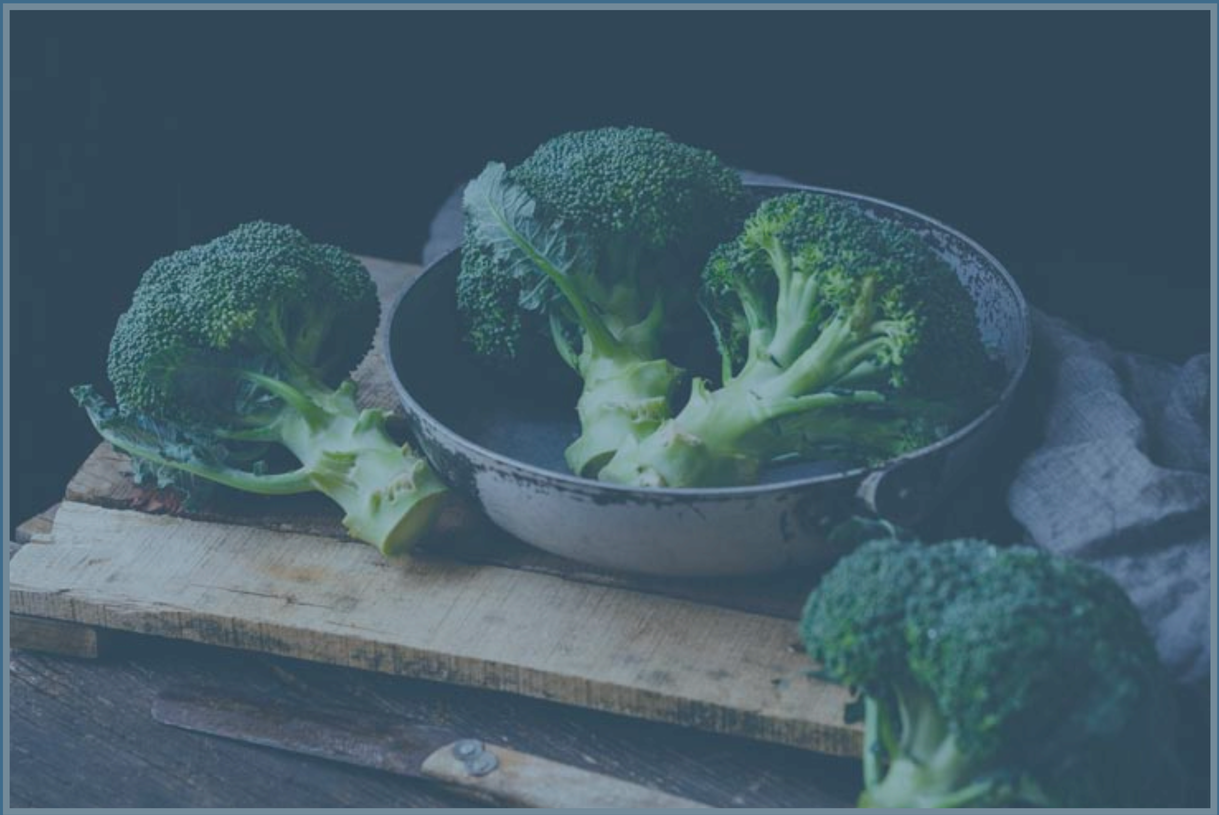
Q

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

bit.ly/eat-cake-cetl-cerse

Q

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



bit.ly/eat-cake-cetl-cerse

hide the veggies

bit.ly/eat-cake-cetl-cerse

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

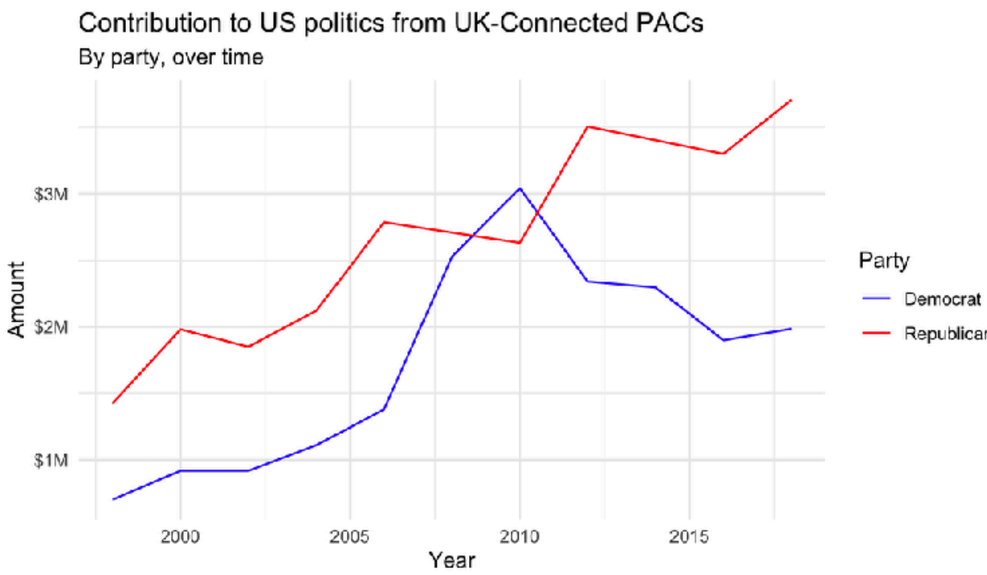
- Today we start with this:

SELECT A CYCLE

2020

PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$1,000	$1,000	$0
Accenture (Accenture)	Ireland/Accenture plc	$80,000	$48,000	$32,000
Air Liquide America	France/L'Air Liquide SA	$16,500	$10,000	$6,500
Airbus Group	Netherlands/Airbus Group	$108,500	$43,500	$65,000
Alkermes Inc	Ireland/Alkermes Plc	$62,750	$23,250	$39,500
Allergan PLC (Allergan PLC)	Ireland/Allergan PLC	$111,000	$6,000	$105,000
Allianz of America (Allianz)	Germany/Allianz AG Holding	$42,750	$18,100	$24,650
Anheuser-Busch (Anheuser-Busch InBev)	Belgium/Anheuser-Busch InBev	$239,000	$119,500	$119,500

- and end with this:



bit.ly/eat-cake-cetl-cerse

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.

PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$1,000	$1,000	$0

	PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
1	7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
2	ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$1,000	$1,000	$0

bit.ly/eat-cake-cetl-cerse

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

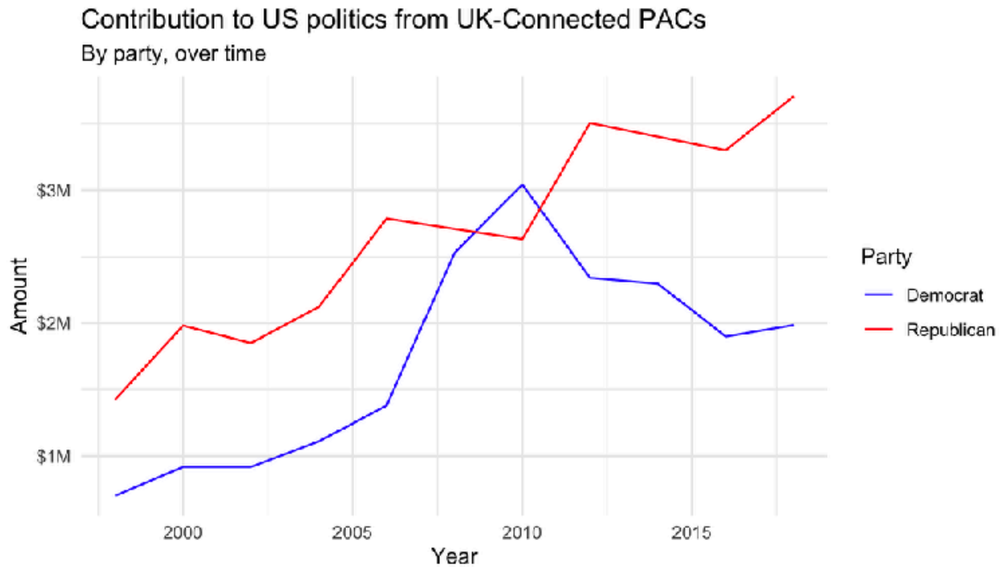
- **Ex 1:** Scrape the table off the web and save as a data frame.

PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$1,000	$1,000	$0

↓

	PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
1	7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
2	ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$1,000	$1,000	$0

- **Ex 2:** What other information do we need represented as variables to build the following visualisation?



bit.ly/eat-cake-cetl-cerse

- **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

	PAC Name (Affiliate)	Country of Origin/Parent Company
1	7-Eleven	Japan/Seven & I Holdings
2	ABB Group (ABB Group)	Switzerland/Asea Brown Boveri
3	Accenture (Accenture)	Ireland/Accenture plc
4	Air Liquide America	France/L'Air Liquide SA

to

	name	country	parent
1	ABB Daimler-Benz Transportation	Germany	Daimler-Benz AG
2	ABB Group	Switzerland	Asea Brown Boveri
3	AE Staley Manufacturing (Tate & Lyle)	UK	Tate & Lyle
4	AEGON USA (AEGON USA)	Netherlands	Aegon NV

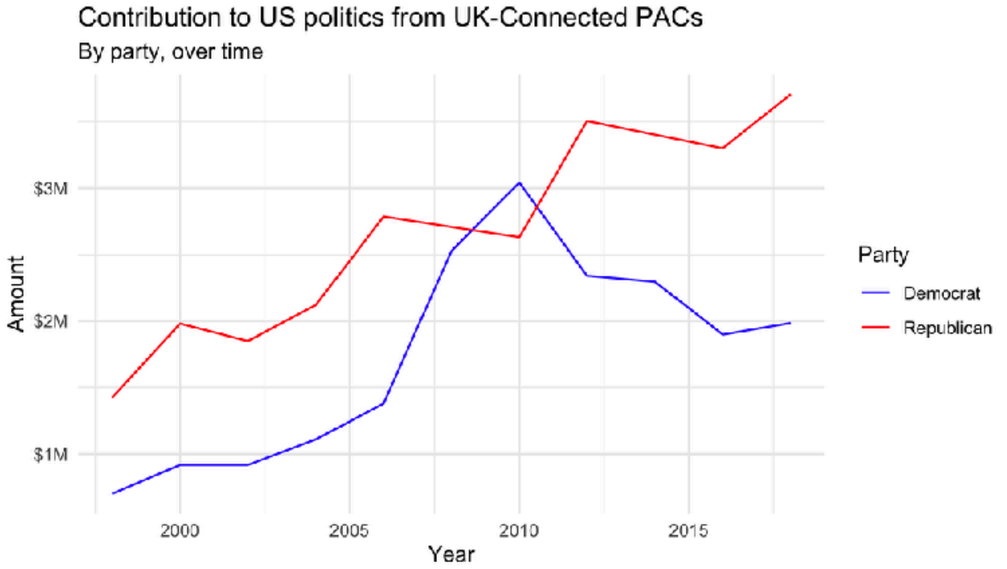
- **Ex 1:** Scrape the table off the web and save as a data frame.

PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$1,000	$1,000	$0

↓

	PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
1	7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
2	ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$1,000	$1,000	$0

- **Ex 2:** What other information do we need represented as variables to build the following visualisation?



bit.ly/eat-cake-cetl-cerse

Q

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

bit.ly/eat-cake-cetl-cerse

Q

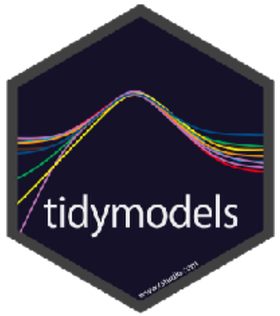
If you are already taking a baking class, which will be easier to venture on to?

bit.ly/eat-cake-cetl-cerse

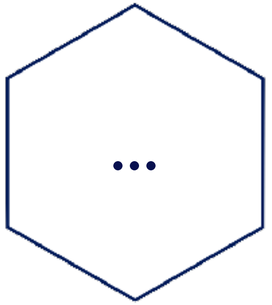
leverage the ecosystem

bit.ly/eat-cake-cetl-cerse

student + instructor



instructor



bit.ly/eat-cake-cetl-cerse

Let them eat cake (first)!*

- bit.ly/eat-cake-cetl-cerse
- </> bit.ly/repo-eat-cake

* You can tell them all about the ingredients later!

@minebocek
mine-cetinkaya-rundel
mcetinka@ed.ac.uk