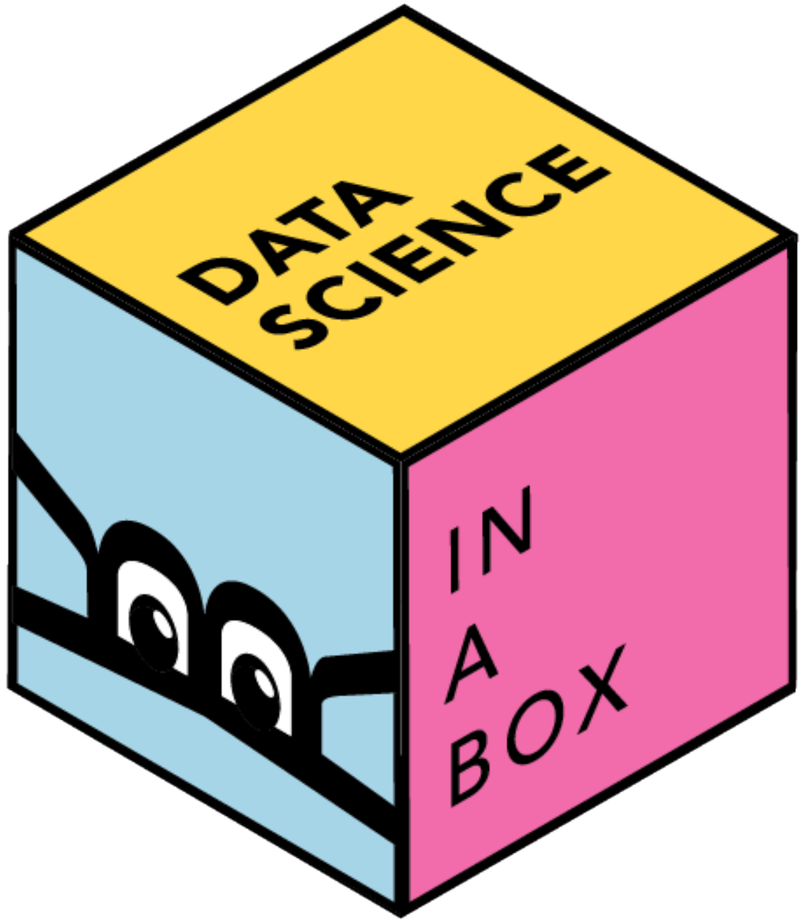


MINE ÇETINKAYA-RUNDEL

DUKE UNIVERSITY + RSTUDIO

@minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com

bit.ly/dsbox-dscwav

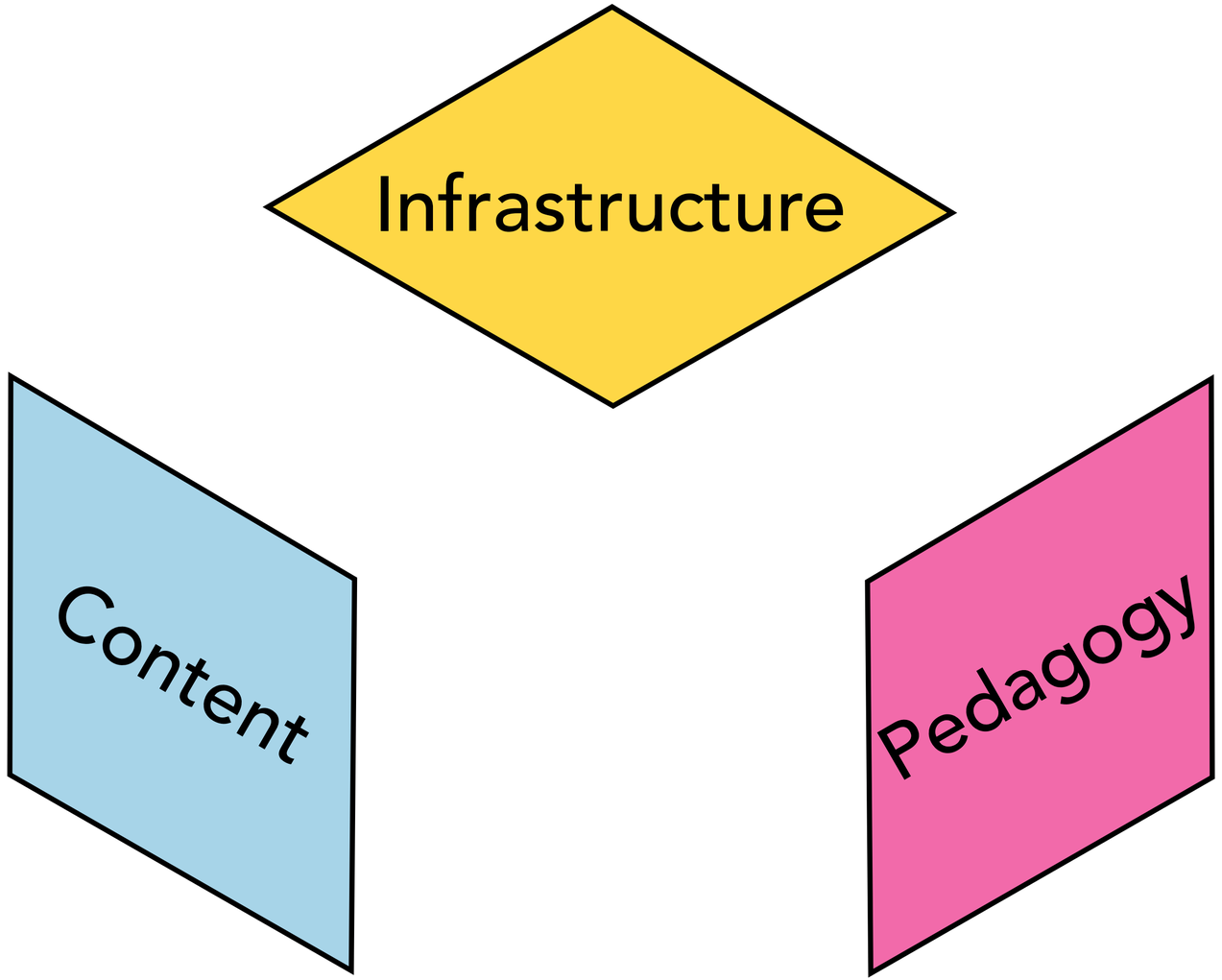
Three questions that keep me up at night…

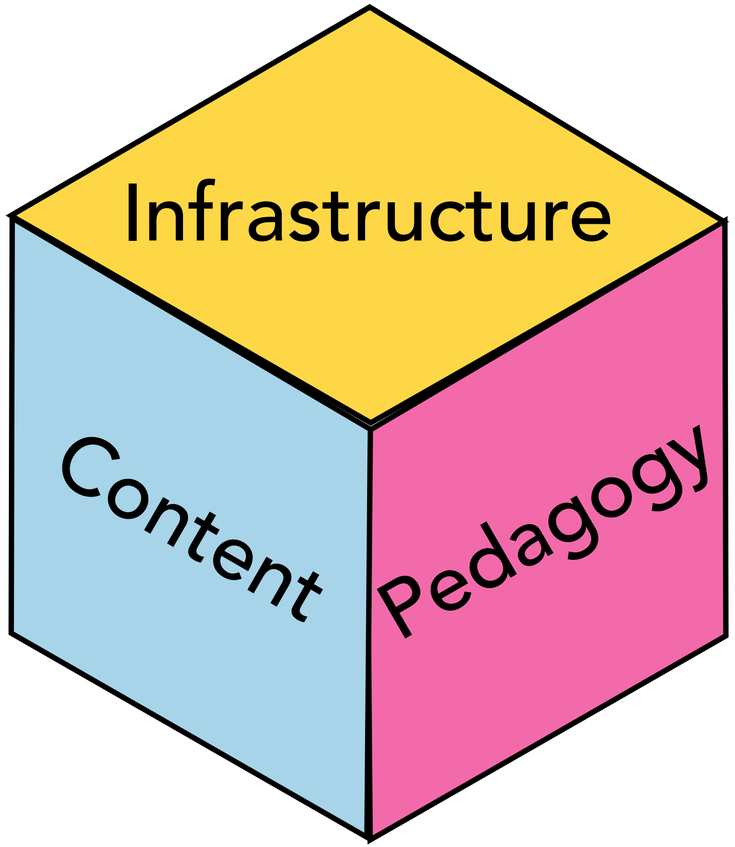
1. What should my students learn?
2. How will my students learn best?
3. What tools will enhance my students' learning?

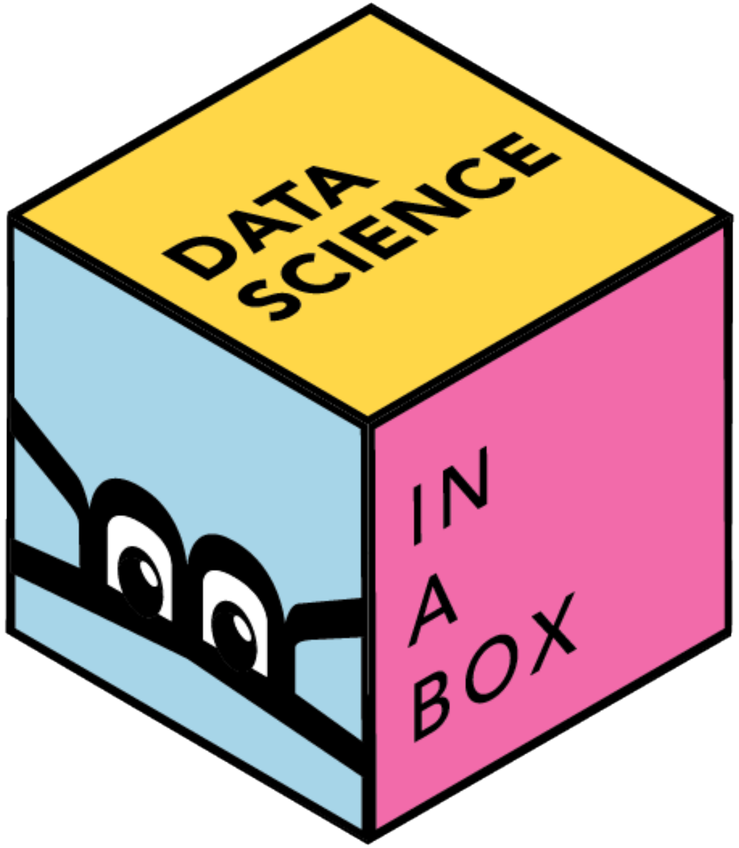
Three questions that keep me up at night…

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

🔗 bit.ly/dsbox-dscwav


Infrastructure
Content
Pedagogy


Infrastructure
Content
Pedagogy



🔗 bit.ly/dsbox-dscwav

Data Science Course in a Box

rstudio-education.github.io/dsbox/

dsbox 0.1.1 Reference Tutorials Data Science in a Box Code of Conduct

dsbox

The goal of dsbox is to supplement the Data Science Course in a Box project. The package contains the datasets that are used in the materials in Data Science Course in a Box as well as the learnr tutorials.

Installation

dsbox is not yet on CRAN. For now, you can install it from GitHub with

```
# install.packages("devtools")
devtools::install_github("rstudio-education/dsbox")
```

Questions, bugs

You can file an issue

When

a minimal **reproducible example** using the reprex package. If you haven't

seriously, reprex will make all of your R-question-asking endeavors easier (which is

s it'll take you to learn what it's all about). For additional reprex pointers, check out the Get

issue, be sure to search issues and pull requests to make sure the bug hasn't been reported and/or already fixed

pment version. By default, the search will be pre-populated with `is:issue is:open`. You can edit the qualifiers

g. `is:pr`, `is:closed`) as needed. For example, you'd simply remove `is:open` to search *all* issues in the repo, open or closed.

Code of Conduct

Please note that the dsbox project is released with a Contributor Code of Conduct. By contributing to this project, you agree to abide by its terms.

Links

Browse source code

Report a b

LICENSE

Community

Code of conduct

Citation

Citing dsbox

Developers

Mine Çetinkaya-Rundel
Author, maintainer

RStudio
Copyright holder, funder

More about authors...

Dev status

R-CMD-check passing

rstudio-education/datascience-box

Chapter 3 Topics | Data Scienc

datasciencebox.org/topics.html

Data Science in a Box

Search

Table of contents

Welcome

Hello #dsbox!

1 Overview

2 Design principles

3 **Topics**

4 Tech stack

5 Community

Course content

6 Hello world

7 Exploring data

8 Data science ethics

9 Making rigorous conclusions

10 Looking further

11 Interactive tutorials

12 Project

13 Exam

17 Sharing

3 Topics

The course content is organized in three units:

Unit 1 - Hello world: This unit is an introduction to the conte

of the course.

Unit 2 - Exploring data: This unit f

ngling.

Specifically we cover fun

onfounding

variables, and P

tidy data, data import,

data

with web scraping and introduce the

next unit. Also in this unit students are

RStudio, R Markdown, Git, and GitHub.

a science ethics: In this unit we discuss misrepresentation of findings,

rticularly in data visualisations, breaches of data privacy, and algorithmic bias.

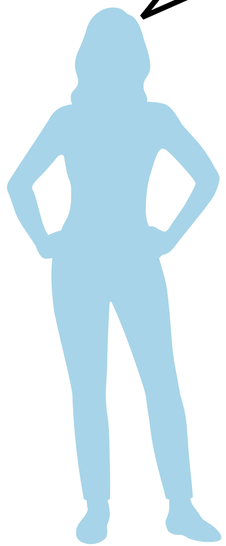
Unit 4 - Making rigorous conclusions: In this unit we introduce modelling and statistical inference for making data-based conclusions. We discuss building, interpreting, and selecting models, visualizing interaction effects, and prediction and model validation. Statistical inference is introduced from a simulation based

datasciencebox.org

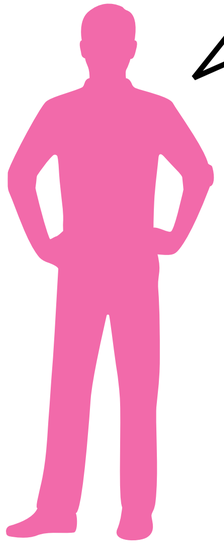
bit.ly/dsbox-dscwav

AUDIENCE

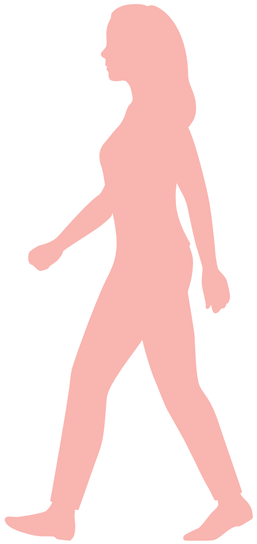
I have been teaching with R for a while, but I want to update my teaching materials



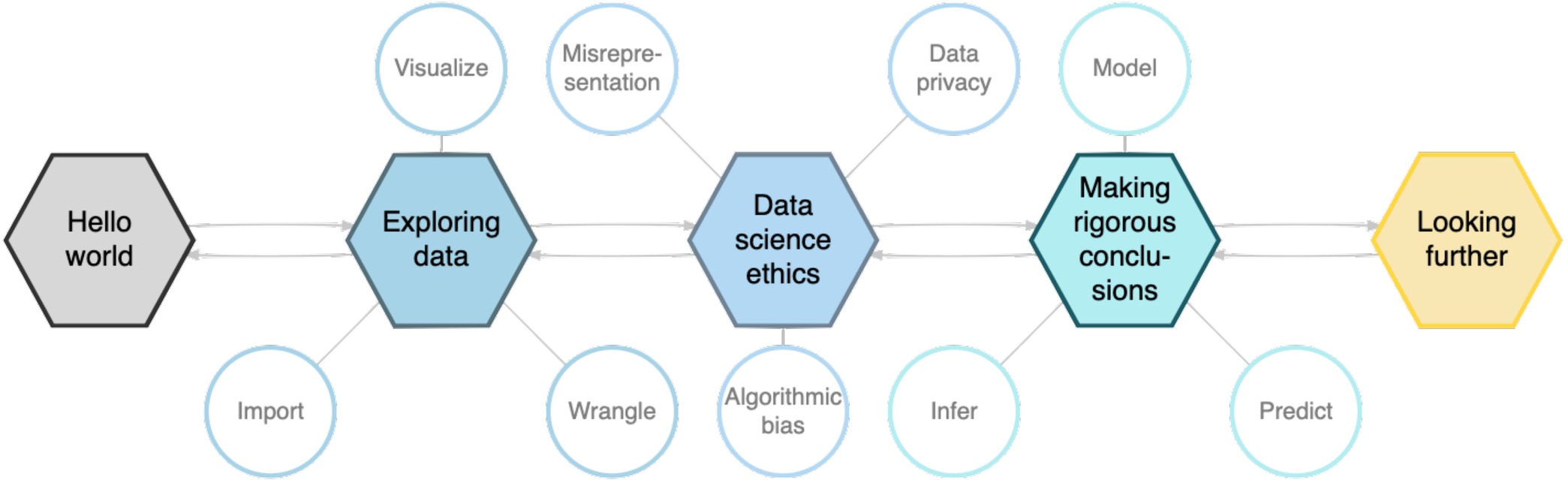
I'm new to teaching with R and need to build up my course materials



This teaching slide deck I came across on Twitter is pretty cool, but I have no idea what type of course it belongs in



TOPICS



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

Interactive viz & reporting, text analysis, Bayesian inference

+

communication & dissemination

🔗 bit.ly/dsbox-dscwav

CONTENTS

website

datasciencebox.org

48
slide
decks

48
videos

10
application
exercises

14
computing
labs

repository

10
homework
assignments

2
take-home
exams

1
open-ended
project

package

dsbox

9
interactive
tutorials

9
interactive
tutorials

bit.ly/dsbox-dscwav

DESIGN PRINCIPLES

cherish day one

start with cake

skip baby steps

hide the veggies

leverage the ecosystem

🔗 bit.ly/dsbox-dscwav

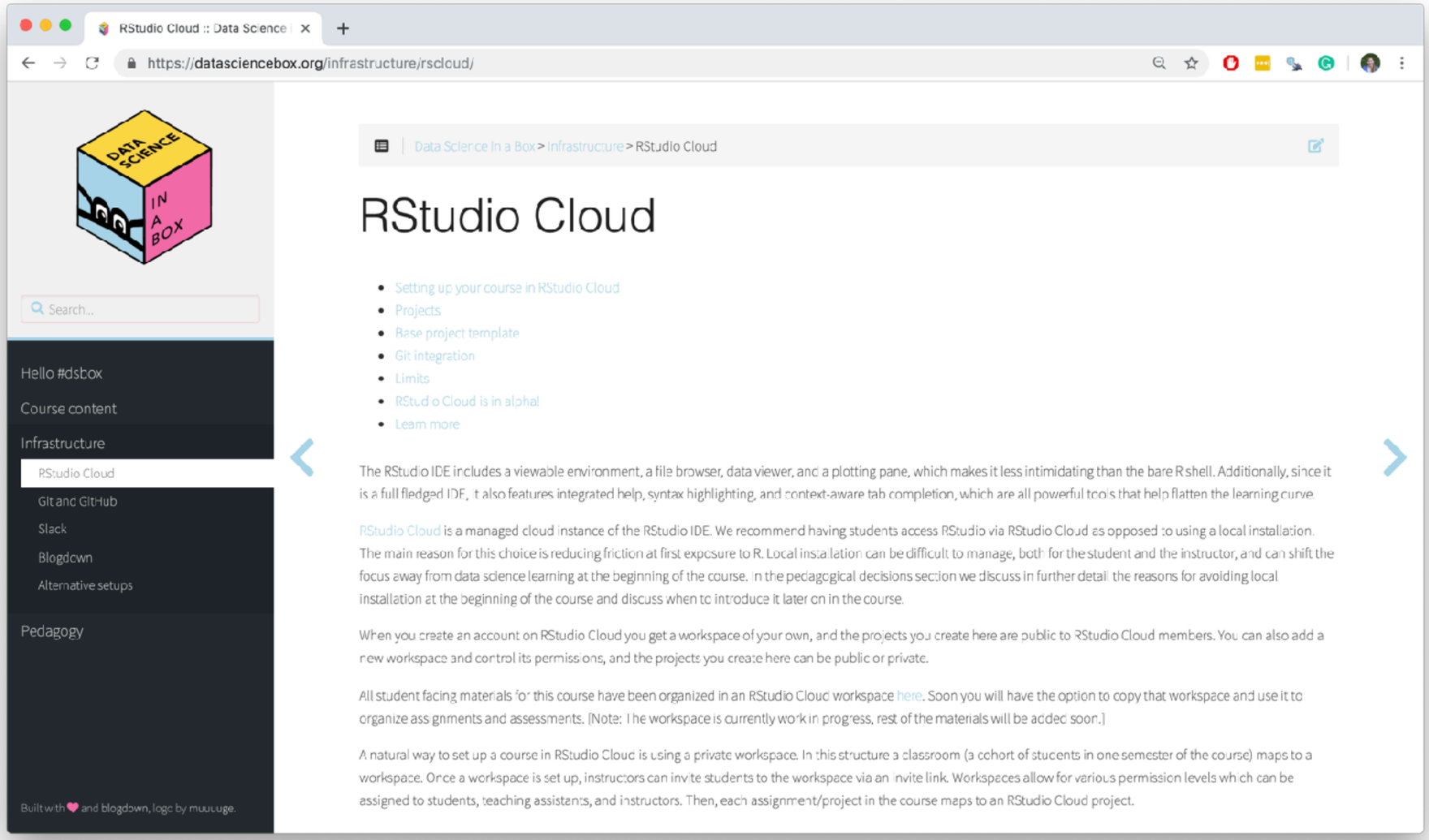
DESIGN PRINCIPLES

Which kitchen would you rather bake a cake?

DESIGN PRINCIPLES

Which kitchen would you rather bake a cake?

Cherish day one



DESIGN PRINCIPLES

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?

DESIGN PRINCIPLES

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

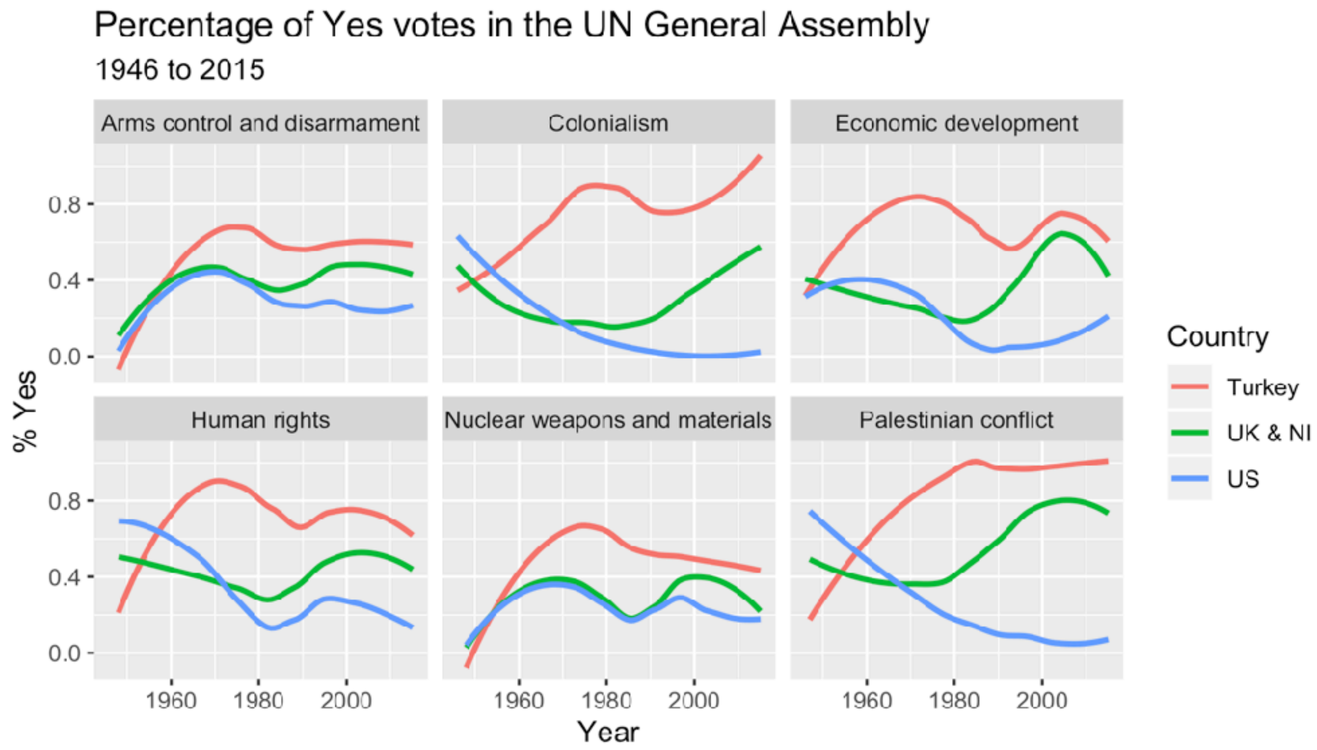
5. Fold batter once or twice from the bottom up with a flexible spatula, then divide evenly between prepared cake pans (about 20 ounces or 565g if you have a scale). Stagger pans together on the oven rack, and bake until puffed, firm, and pale gold, about 32 minutes. If your oven has very uneven heat, 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

- 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

DESIGN PRINCIPLES

Start with cake

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

Start with cake

```
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"
  )
```


Start with cake

```
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"
  )
```

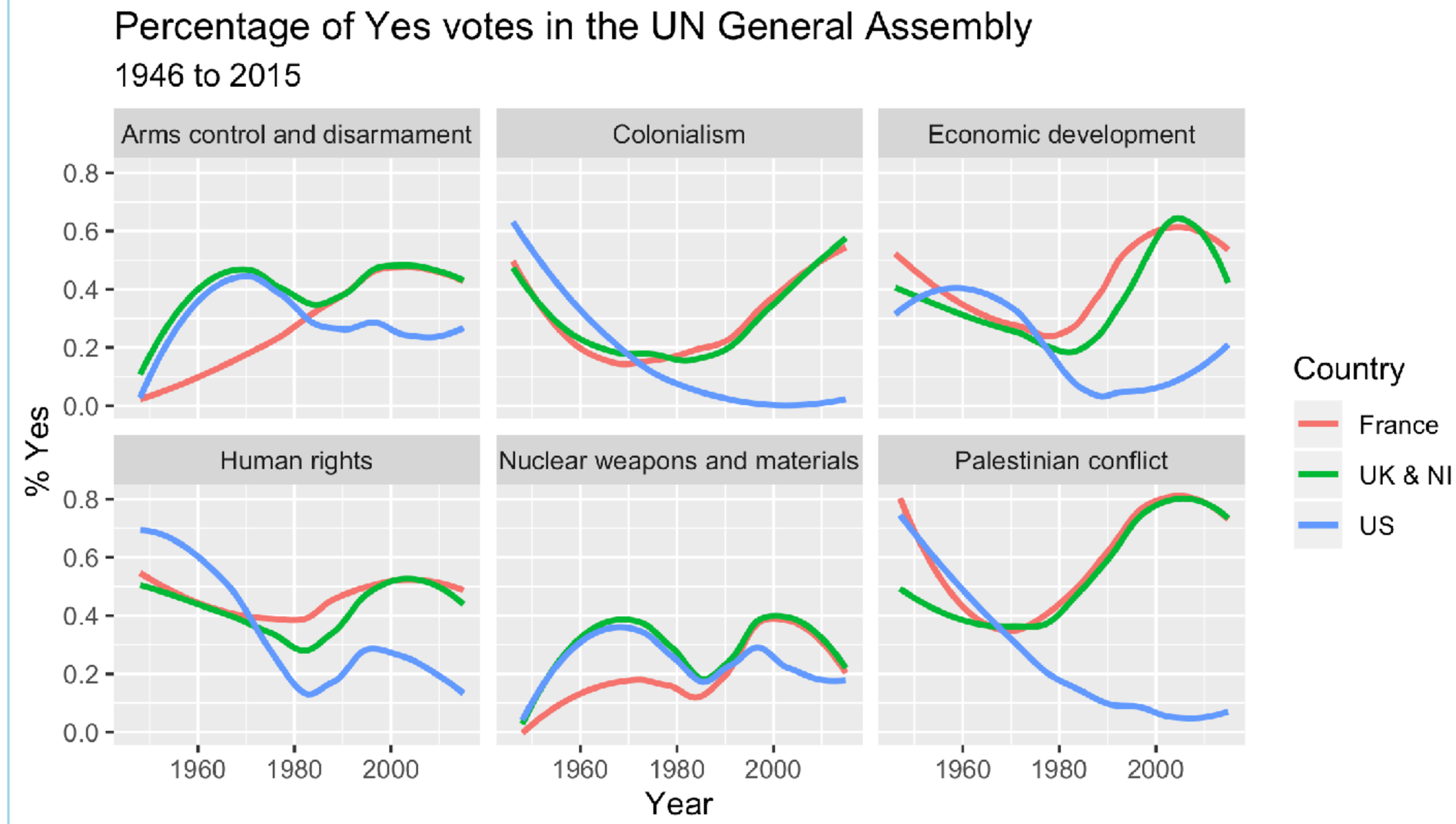

Start with cake

```
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"
  )
```


Start with cake

```
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"
  )
```


Start with cake



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?

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

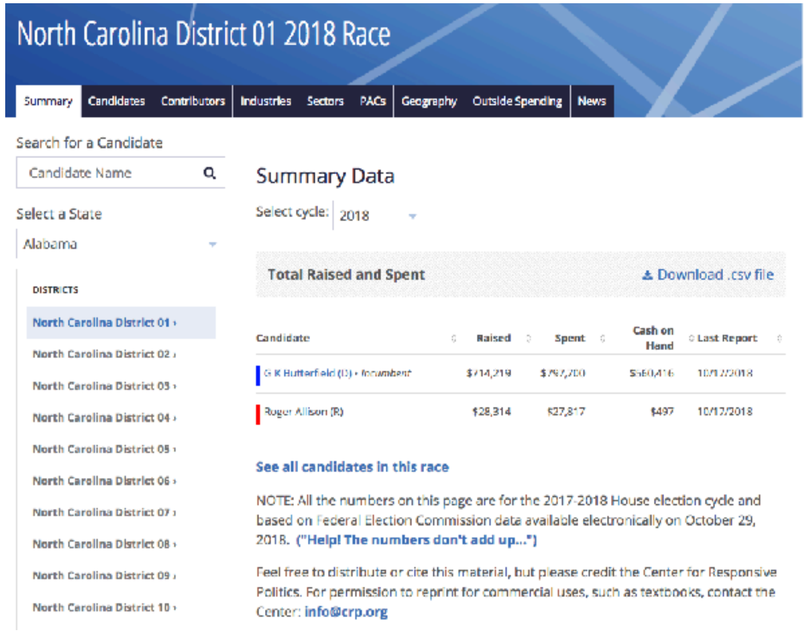
Submit Answer

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

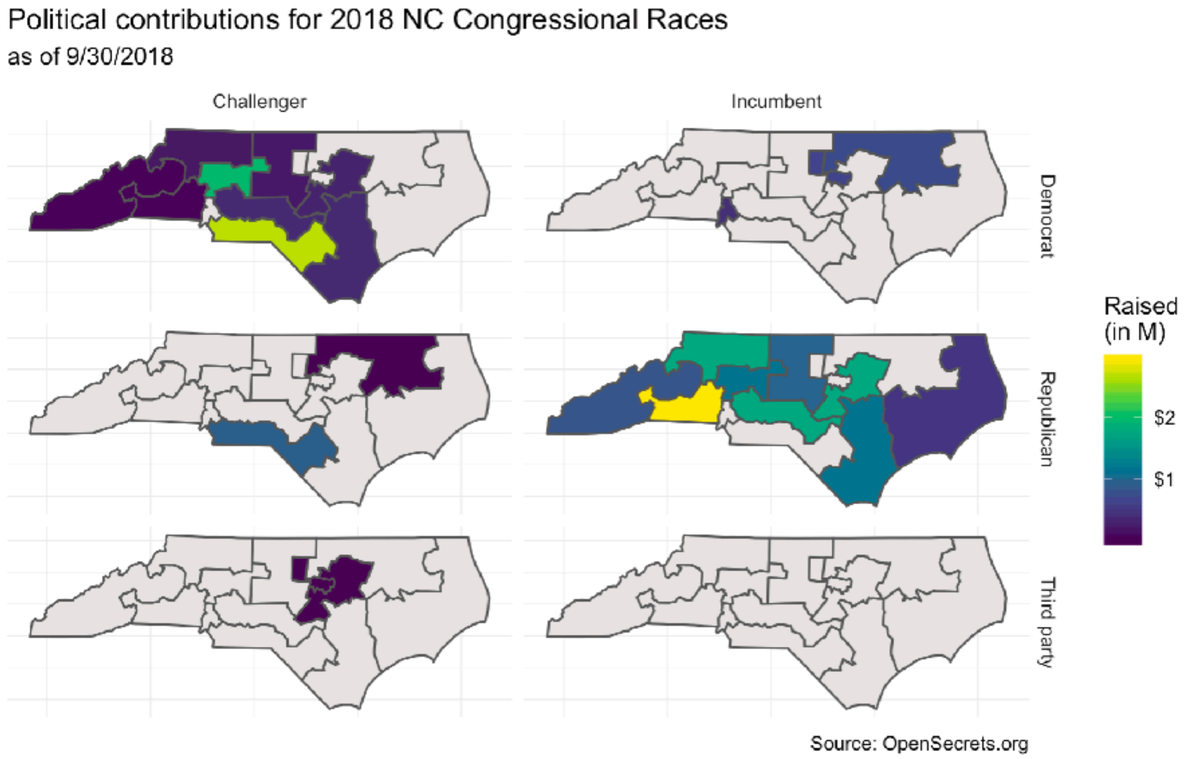
Which is more likely to appeal to someone who has never tried broccoli?

Hide the veggies

- Today we go from this to that



→



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

Hide the veggies

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

Hide the veggies

DESIGN PRINCIPLES

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

Hide the veggies

DESIGN PRINCIPLES

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

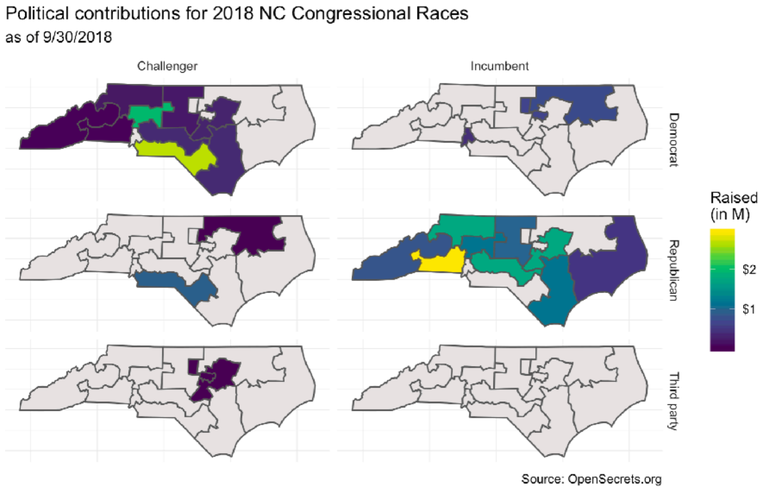
Ex 1: Scrape the table off the web and save as a data frame.

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?



Hide the veggies

DESIGN PRINCIPLES

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

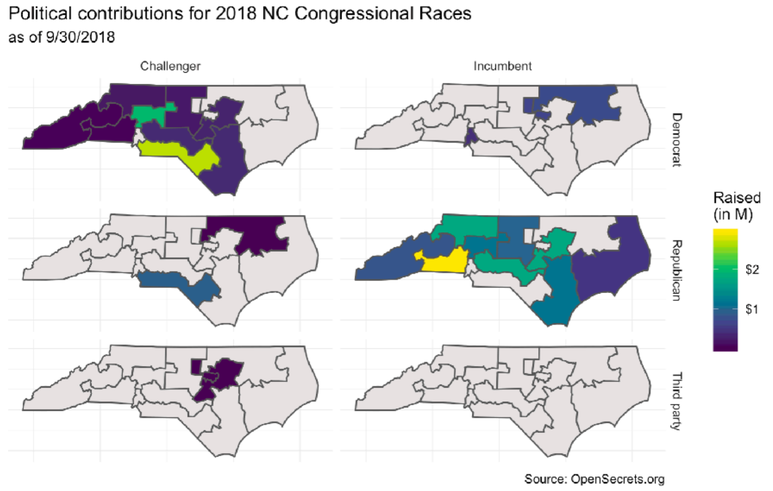
	candidate_info	raised	spent	cash_on_hand	last_report	race
1	G K Butterfield (D) • Incumbent	714219	797700	560416	2018-10-17	North Carolina District 01
2	Roger Allison (R)	28314	27817	497	2018-10-17	North Carolina District 01

Lesson: "Just enough" regex

	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?



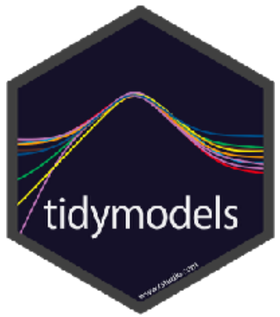
If you are already taking a baking class, which will be easier to venture on to?

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

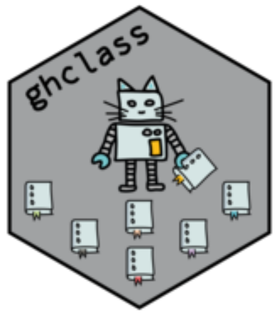
Leverage the ecosystem

DESIGN PRINCIPLES

student + instructor

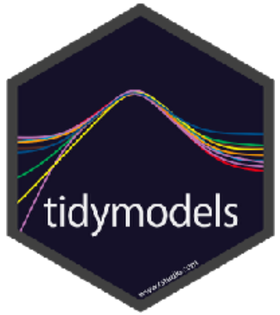


instructor

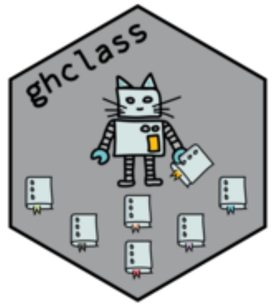


Leverage the ecosystem

student + instructor



instructor



USAGE

in full
to jumpstart / overhaul your teaching

in bits & pieces
to supplement your teaching

LICENSE

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.

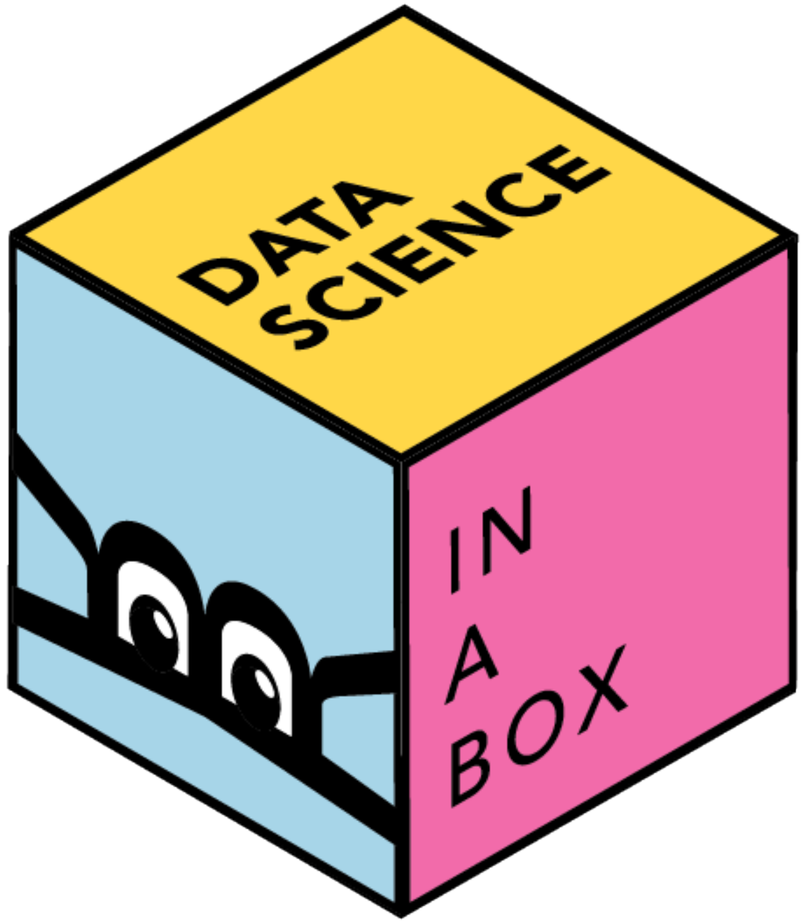
Under the following terms:

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.

bit.ly/dsbox-dscwav



📦 datasciencebox.org

📦 rstudio-education.github.io/dsbox

🖥️ bit.ly/dsbox-dscwav

MINE ÇETINKAYA-RUNDEL

DUKE UNIVERSITY + RSTUDIO

@minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com