



the art and science  
of teaching data science  
mine çetinkaya-rundel

**[bit.ly/art-sci-ucl](http://bit.ly/art-sci-ucl)**



@minebocek



mine-cetinkaya-rundel



[cetinkaya.mine@gmail.com](mailto:cetinkaya.mine@gmail.com)

Image credit: Thomas Pedersen, [data-imaginist.com/art](http://data-imaginist.com/art)

# 2016 GAISE

## 1. Teach statistical thinking.

- ▶ **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.
- ▶ **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 purpose.

## 4. Foster active learning.

## 5. Use technology to explore concepts and analyse data.

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

## 1. Teach statistical thinking.

- ▶ **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.
- ▶ **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 purpose.

## 4. Foster active learning.

## 5. Use technology to explore concepts and analyse 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

# 2016 GAISE

## 1. Teach statistical thinking.

- ▶ **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.
- ▶ **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 purpose.

## 4. Foster active learning.

## 5. Use technology to explore concepts and analyse data.

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

② Multivariate analysis requires the use of computing

# 2016 GAISE

## 1. Teach statistical thinking.

- ▶ **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.
- ▶ **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 purpose.

## 4. Foster active learning.

## 5. Use technology to explore concepts and analyse data.

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

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

## 1. Teach statistical thinking.

- ▶ **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.
- ▶ **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.

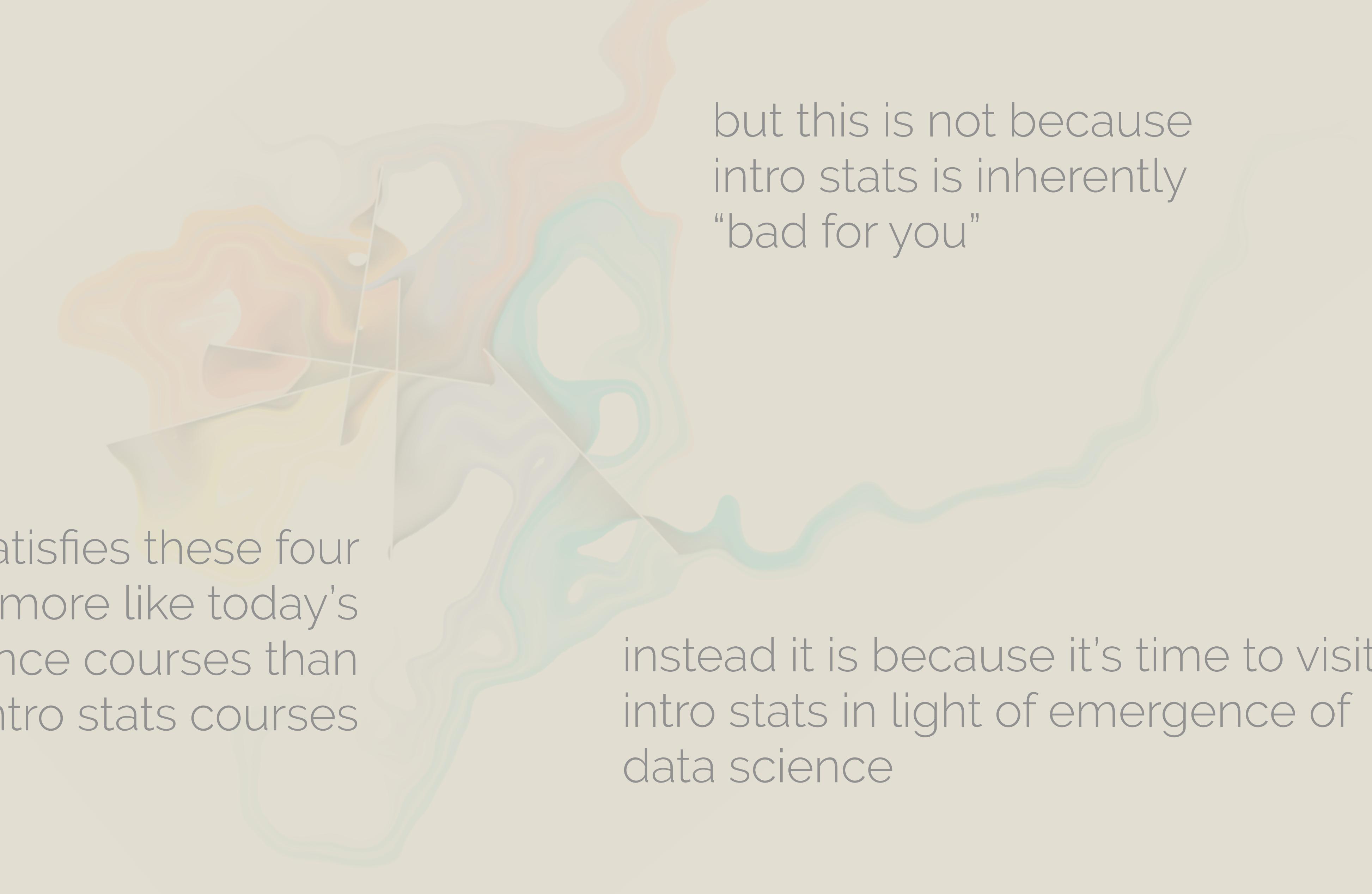
## 3. Integrate real data with a context and purpose.

## 4. Foster active learning.

## 5. Use technology to explore concepts and analyse data.

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

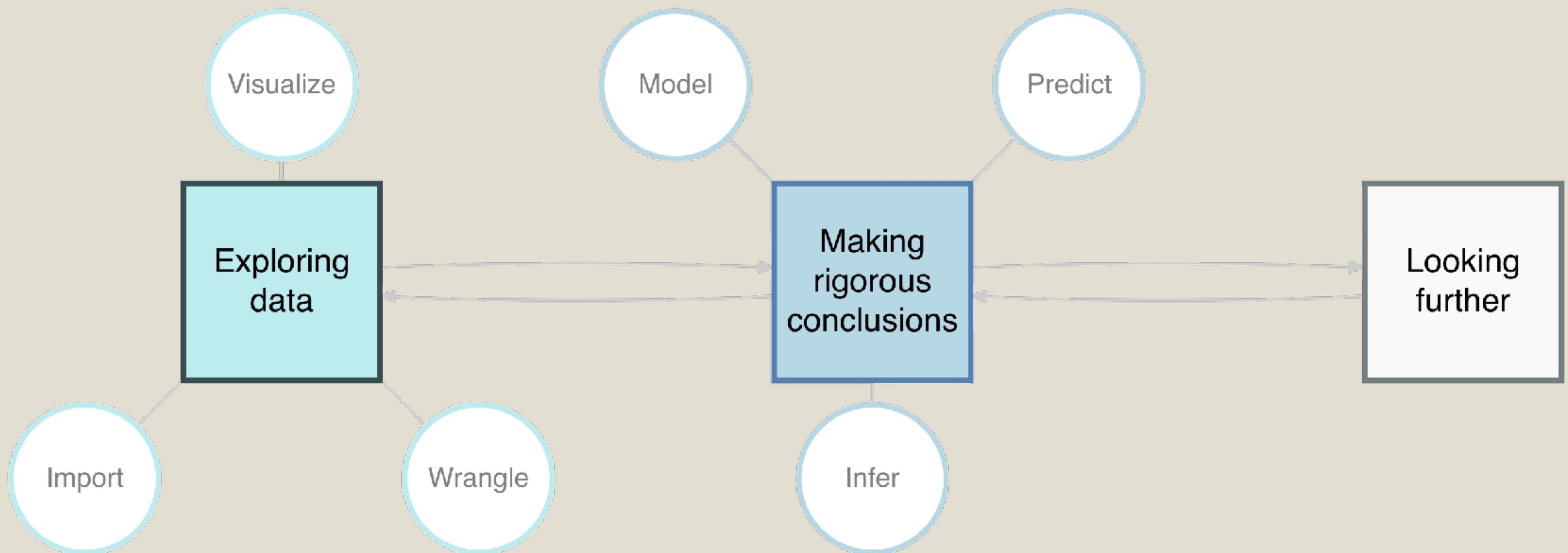
④ Data analysis isn't just inference and modelling, it's also data importing, cleaning, preparation, exploration, and visualisation

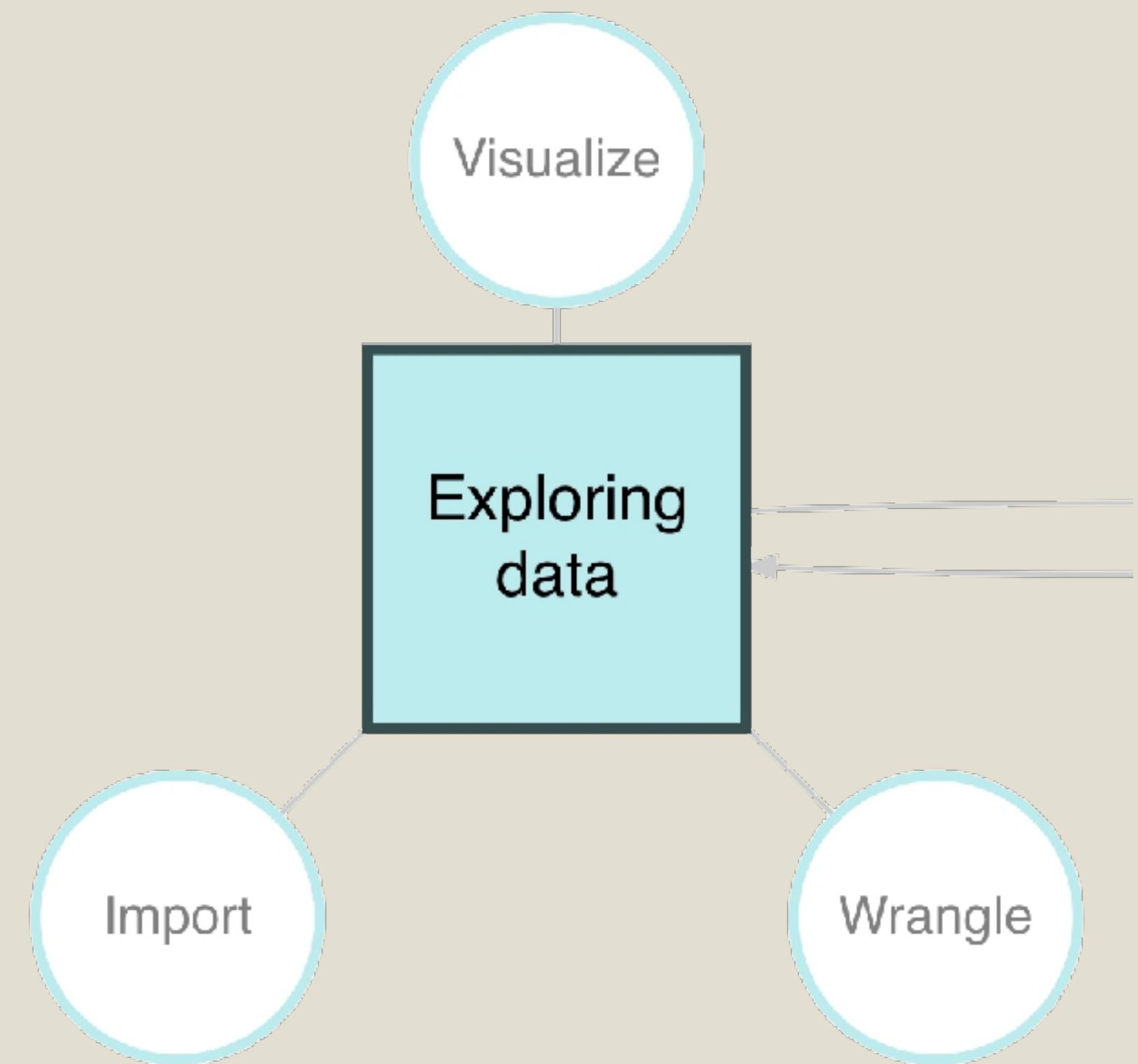


a course that satisfies these four points is looking more like today's intro data science courses than (most) intro stats courses

but this is not because intro stats is inherently "bad for you"

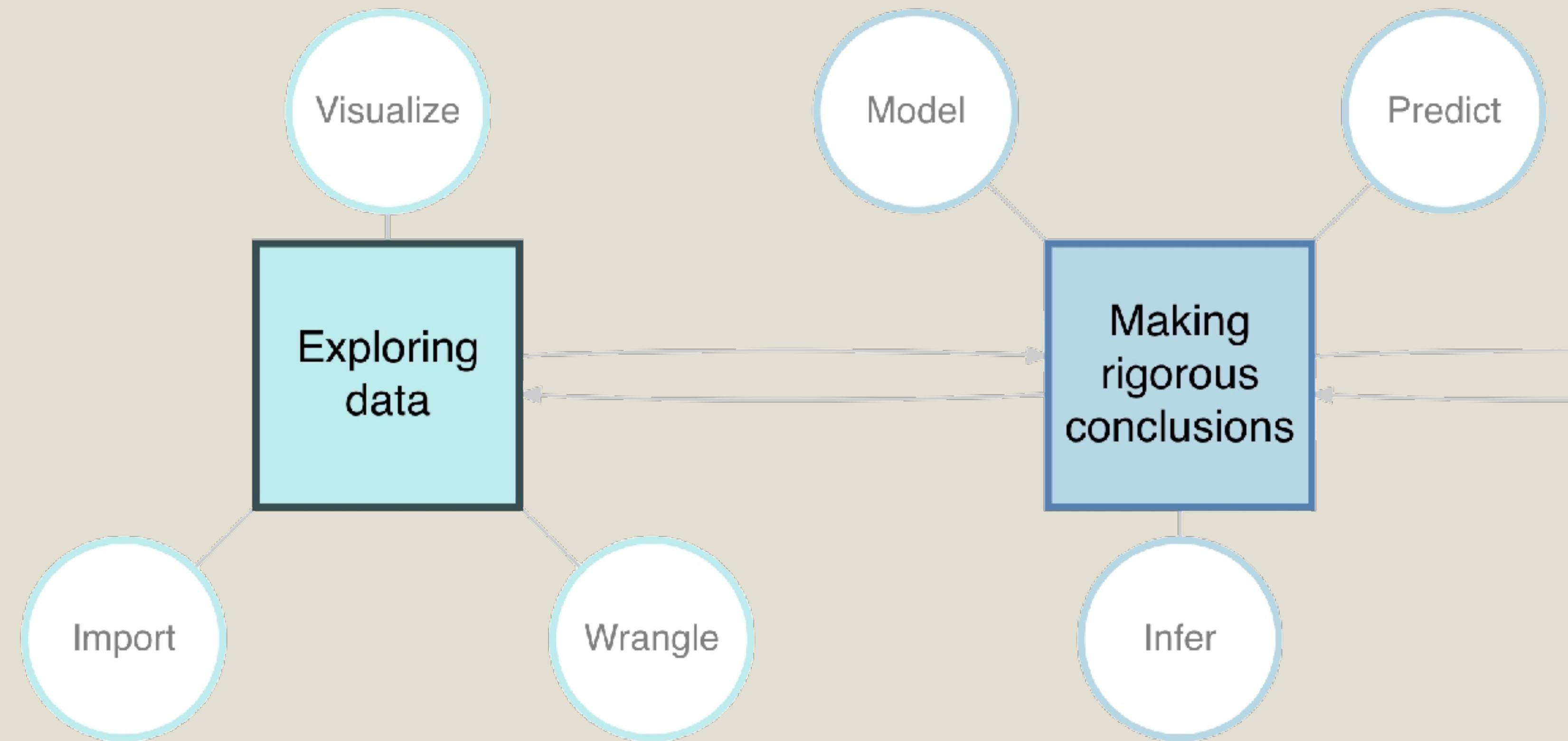
instead it is because it's time to visit intro stats in light of emergence of data science





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

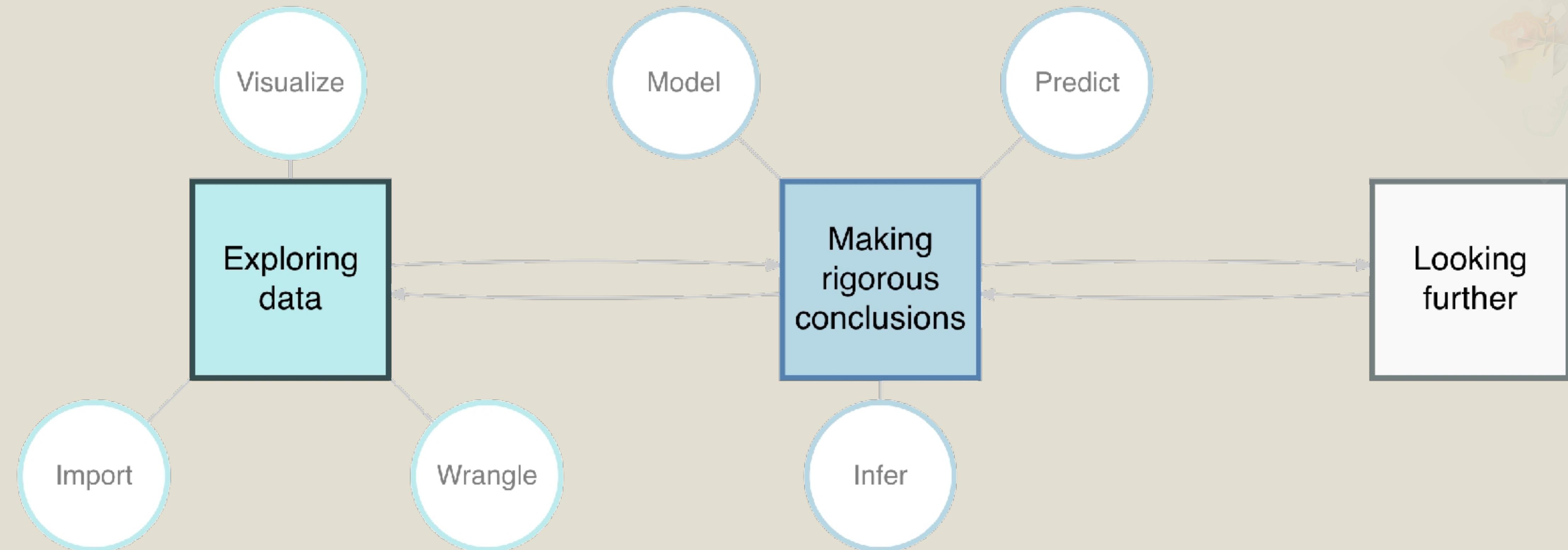


fundamentals of  
data & data viz,  
confounding variables,  
Simpson's paradox  
+  
R / RStudio,  
R Markdown, simple Git

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

building & selecting  
models,  
visualising interactions,  
prediction & validation,  
inference via simulation





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,  
visualising interactions,  
prediction & validation,  
inference via simulation

data science ethics,  
text analysis,  
Bayesian inference  
+  
communication &  
dissemination





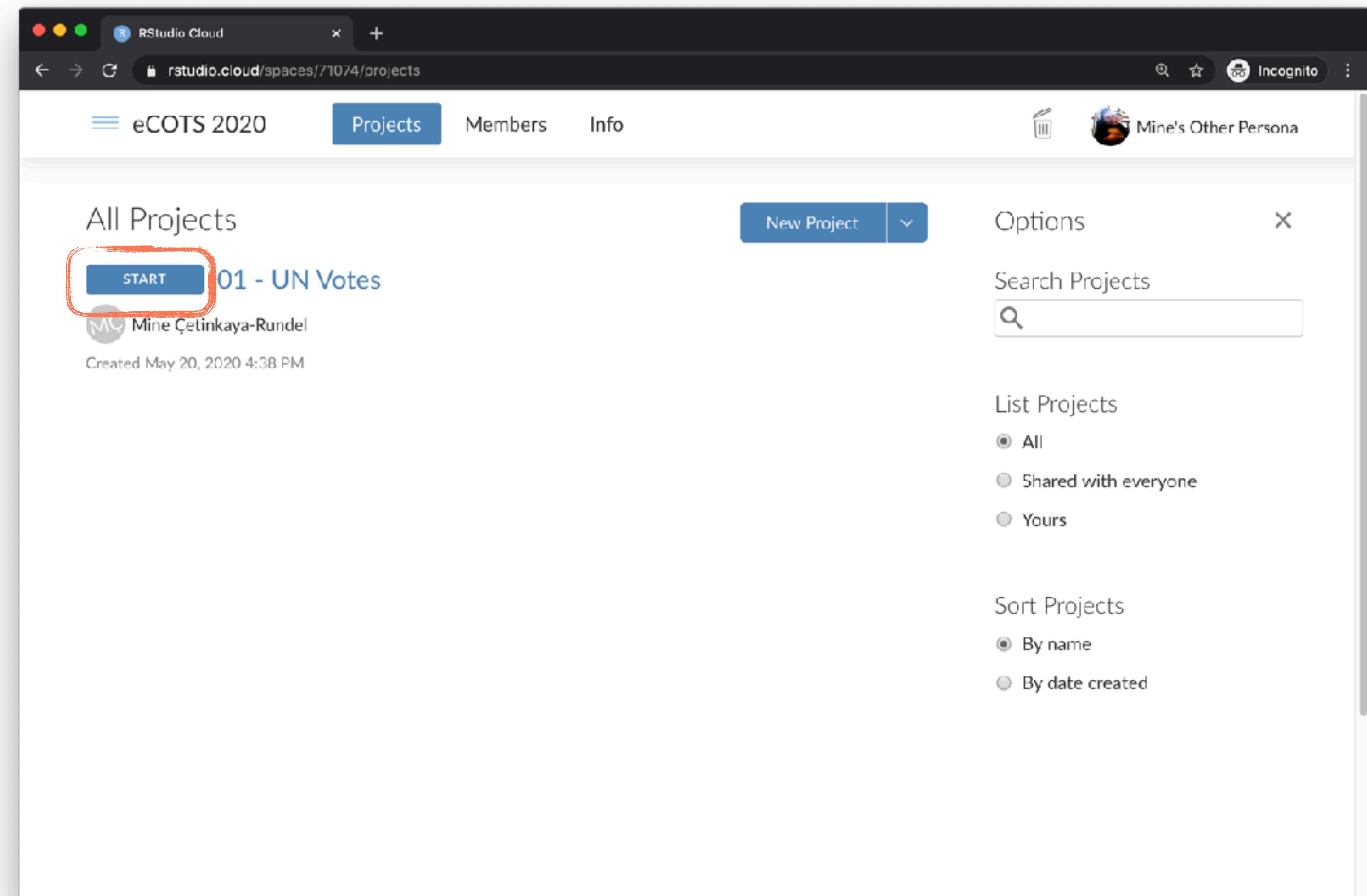
fundamentals of  
data & data viz,  
confounding variables,  
Simpson's paradox  
+  
R / RStudio,  
R Markdown, simple Git

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

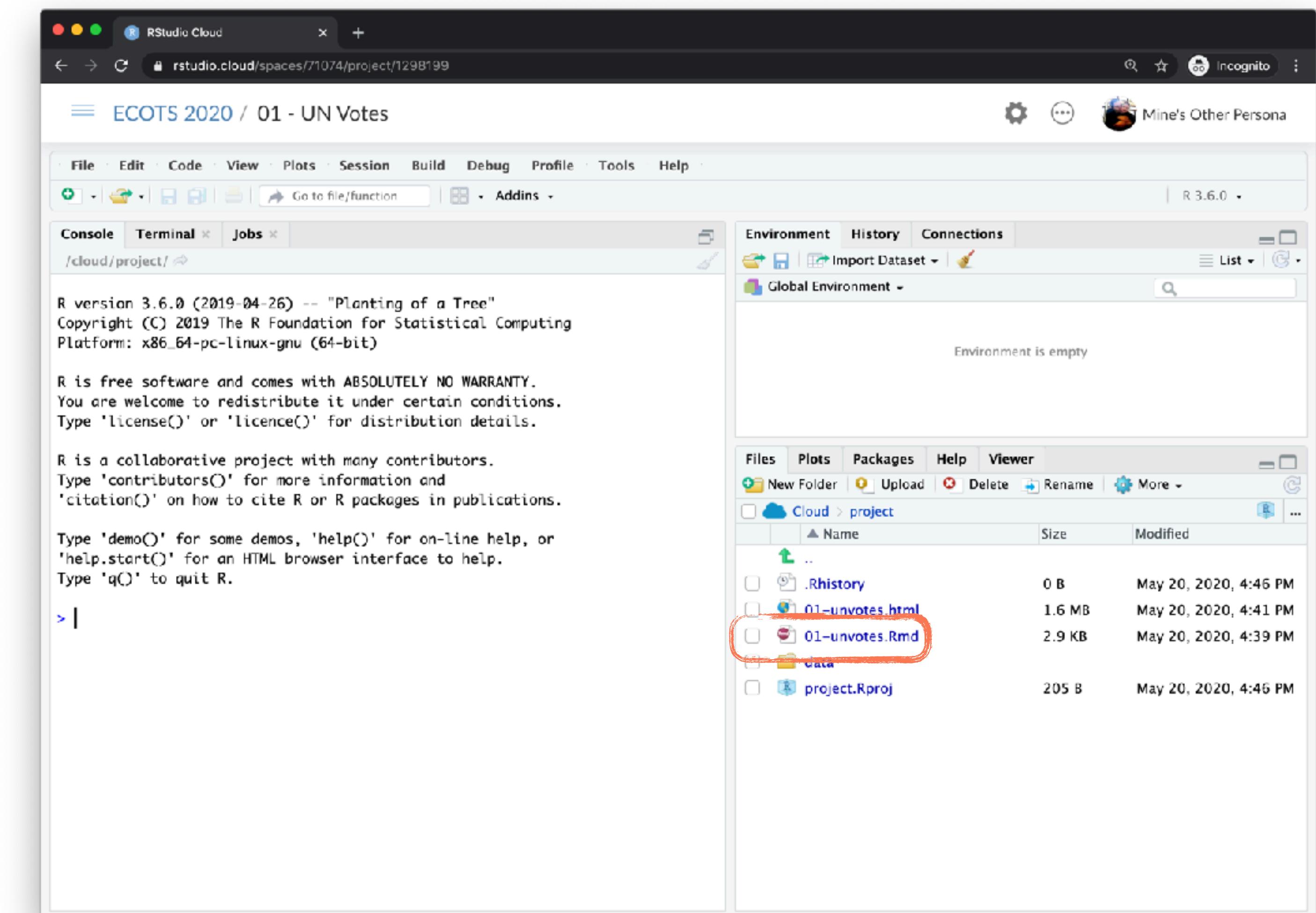
building & selecting  
models,  
visualising interactions,  
prediction & validation,  
inference via simulation

data science ethics,  
text analysis,  
Bayesian inference  
+  
communication &  
dissemination

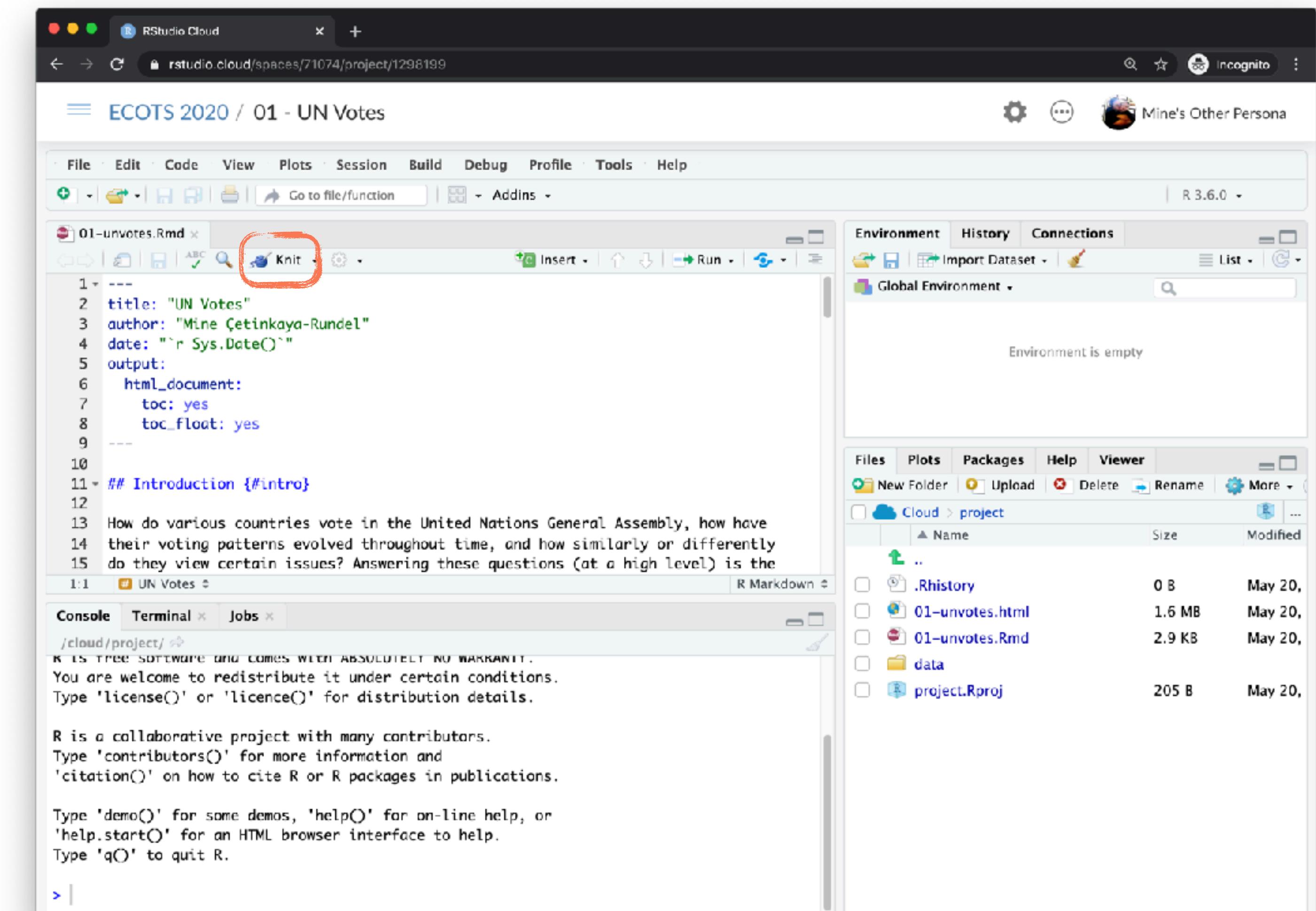
- ▶ Go to [bit.ly/rscloud-ecots2020](https://bit.ly/rscloud-ecots2020)
- ▶ Start the project titled UN Votes



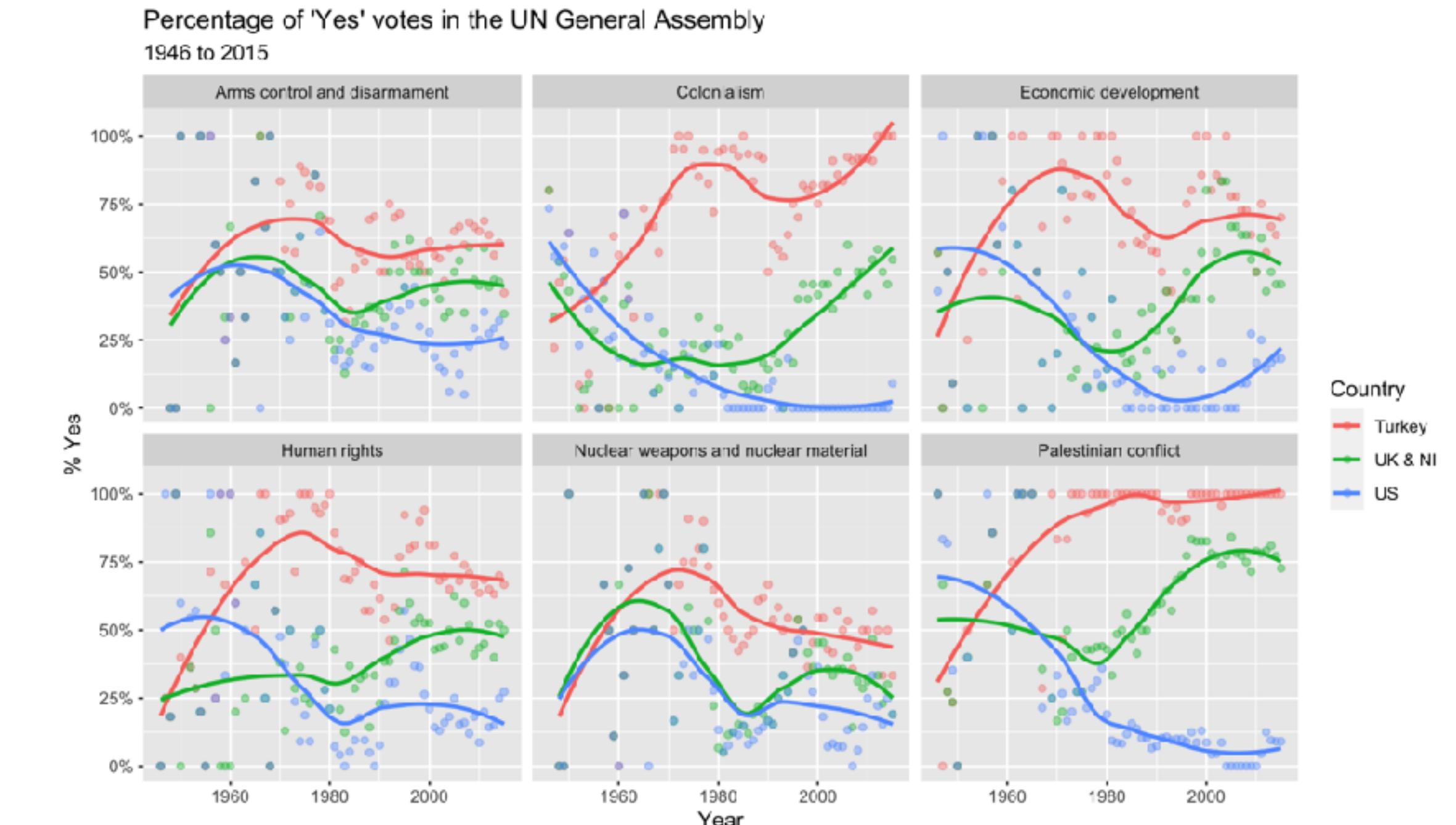
- ▶ Go to [bit.ly/rscloud-ecots2020](https://bit.ly/rscloud-ecots2020)
- ▶ Start the project titled UN Votes
- ▶ Open the R Markdown document called `unvotes.Rmd`



- ▶ Go to [bit.ly/rscloud-ecots2020](https://bit.ly/rscloud-ecots2020)
- ▶ Start the project titled UN Votes
- ▶ Open the R Markdown document called `unvotes.Rmd`
- ▶ Knit the document and review the data visualisation you just produced



- ▶ Go to [bit.ly/rscloud-ecots2020](http://bit.ly/rscloud-ecots2020)
- ▶ Start the project titled UN Votes
- ▶ Open the R Markdown document called `unvotes.Rmd`
- ▶ Knit the document and review the data visualisation you just produced
- ▶ Then, look for the character string “Turkey” in the code and replace it with another country of your choice
- ▶ Knit again, and review how the voting patterns of the country you picked compares to the United States and United Kingdom & Northern Ireland



# three questions that keep me up at night...

- 
- 1 what should students learn?
  - 2 how will students learn best?
  - 3 what tools will enhance student learning?

# three questions that keep me up at night...

**content**

1

what should students learn?

**pedagogy**

2

how will students learn best?

**infrastructure**

3

what tools will enhance student learning?





**content**

ex. 1

## money in politics



[About](#) | [Resources](#) | [Sign Up](#)Search  MENU

DONATE

## SECTIONS

[Top PACs](#) ›[Top Recipients of PAC Money](#) ›[Industry Breakdown](#) ›[Leadership PACs](#) ›[Super PACs](#) ›[Search](#) ›[Foreign-connected PACs](#) ›[What is a PAC?](#) ›

Ad closed by Google

[Report this ad](#)[Why this ad? ▶](#)

# Foreign-Connected PACs

## SELECT A CYCLE

2020 

| PAC Name (Affiliate)            | Country of Origin/Parent Company | Total    | Dems     | Repubs   |
|---------------------------------|----------------------------------|----------|----------|----------|
| ABB Group (ABB Group)           | Switzerland/Asea Brown Boveri    | \$1,000  | \$1,000  | \$0      |
| Accenture (Accenture)           | Ireland/Accenture plc            | \$73,500 | \$45,000 | \$28,500 |
| Advance America Cash            |                                  |          |          |          |
| Advance Centers (Grupo Salinas) | Mexico/Grupo Salinas             | \$3,000  | \$1,000  | \$2,000  |

## SELECT A CYCLE

2020

| PAC Name (Affiliate)                                       | Country of Origin/Parent Company | Total     | Dems     | Repubs    |
|--|----------------------------------|-----------|----------|-----------|
| ABB Group (ABB Group)                                      | Switzerland/Asea Brown Boveri    | \$1,000   | \$1,000  | \$0       |
| Accenture (Accenture)                                      | Ireland/Accenture plc            | \$73,500  | \$45,000 | \$28,500  |
| Advance America Cash<br>Advance Centers<br>(Grupo Salinas) | Mexico/Grupo Salinas             | \$3,000   | \$1,000  | \$2,000   |
| Air Liquide America  | France/L'Air Liquide SA          | \$11,500  | \$5,000  | \$6,500   |
| Airbus Group   | Netherlands/Airbus Group         | \$81,500  | \$26,500 | \$55,000  |
| Alkermes Inc   | Ireland/Alkermes Plc             | \$40,250  | \$11,250 | \$29,000  |
| Allergan PLC (Allergan<br>PLC)                             | Ireland/Allergan PLC             | \$111,000 | \$6,000  | \$105,000 |
| Allianz of America<br>(Allianz)                            | Germany/Allianz AG Holding       | \$35,500  | \$17,100 | \$18,400  |

- \* web scraping
- \* text parsing
- \* data types
- \* regular expressions

## SELECT A CYCLE

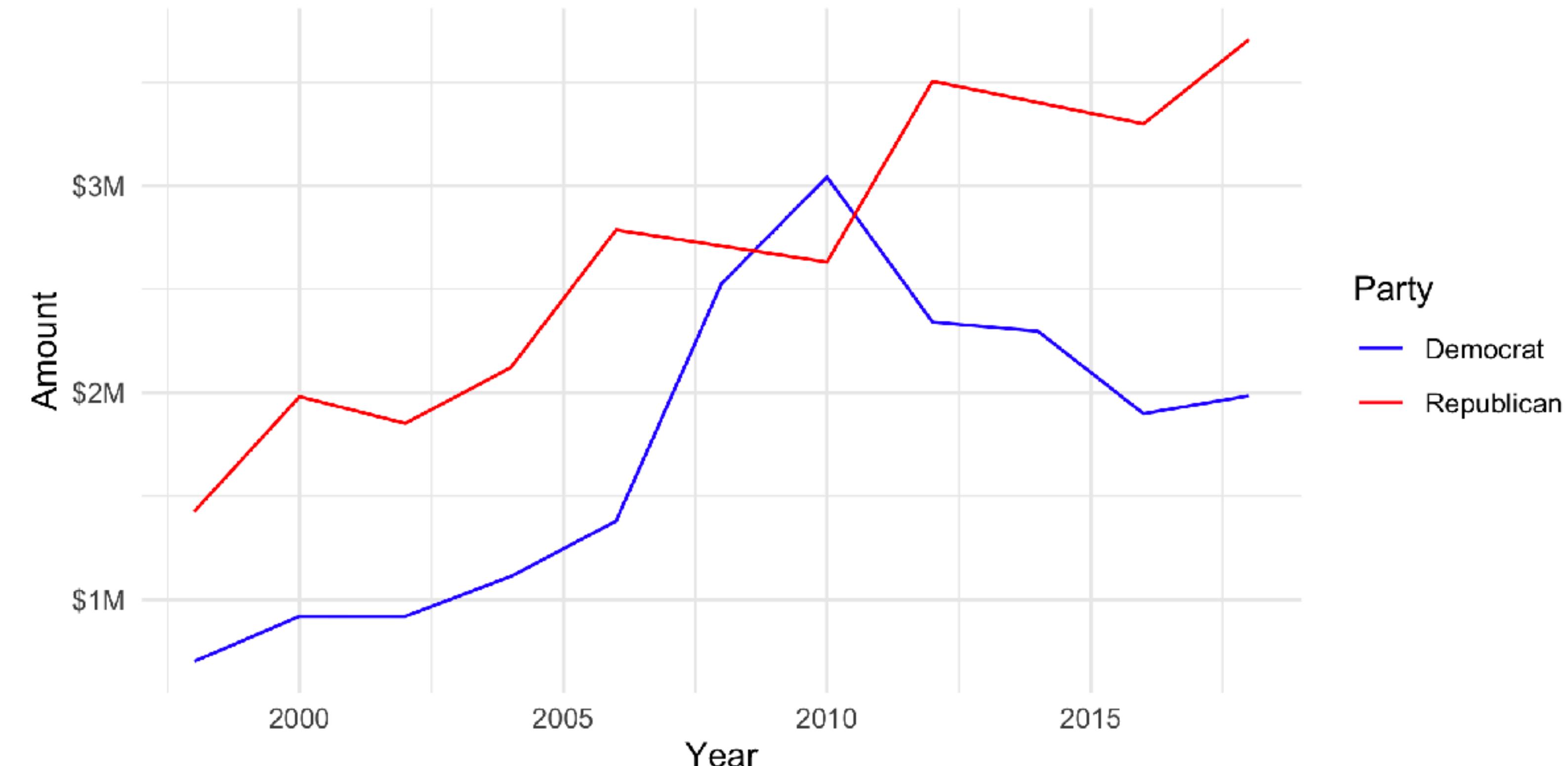
- 2020
- 2018
- 2016
- 2014
- 2012
- 2010
- 2008
- 2006
- 2004
- 2002
- 2000
- 1998

Advance America Case

|                                    |                            | Repubs    |          |           |
|------------------------------------|----------------------------|-----------|----------|-----------|
| Advance Centers<br>(Grupo Salinas) | Mexico/Grupo Salinas       | \$3,000   | \$1,000  | \$2,000   |
| Air Liquide America                | France/L'Air Liquide SA    | \$11,500  | \$5,000  | \$6,500   |
| Airbus Group                       | Netherlands/Airbus Group   | \$81,500  | \$26,500 | \$55,000  |
| Alkermes Inc                       | Ireland/Alkermes Plc       | \$40,250  | \$11,250 | \$29,000  |
| Allergan PLC (Allergan<br>PLC)     | Ireland/Allergan PLC       | \$111,000 | \$6,000  | \$105,000 |
| Allianz of America<br>(Allianz)    | Germany/Allianz AG Holding | \$35,500  | \$17,100 | \$18,400  |

- ★ web scraping
- ★ text parsing
- ★ data types
- ★ regular expressions
- ★ iteration

## Contribution to US politics from UK-Connected PACs By party, over time



- ★ web scraping
- ★ text parsing
- ★ data types
- ★ regular expressions
- ★ iteration
- ★ data visualisation
- ★ interpretation

```
paths_allowed("https://www.opensecrets.org")
```

```
## [1] TRUE
```

- ★ web scraping
- ★ text parsing
- ★ data types
- ★ regular expressions
- ★ iteration
- ★ data visualisation
- ★ interpretation
- ★ data science ethics

# Project: The North South Divide: University Edition

**Question:** Does the geographical location of a UK university affect its university score?

**Team:** Fried Egg Jelly Fish

## University League Tables 2020

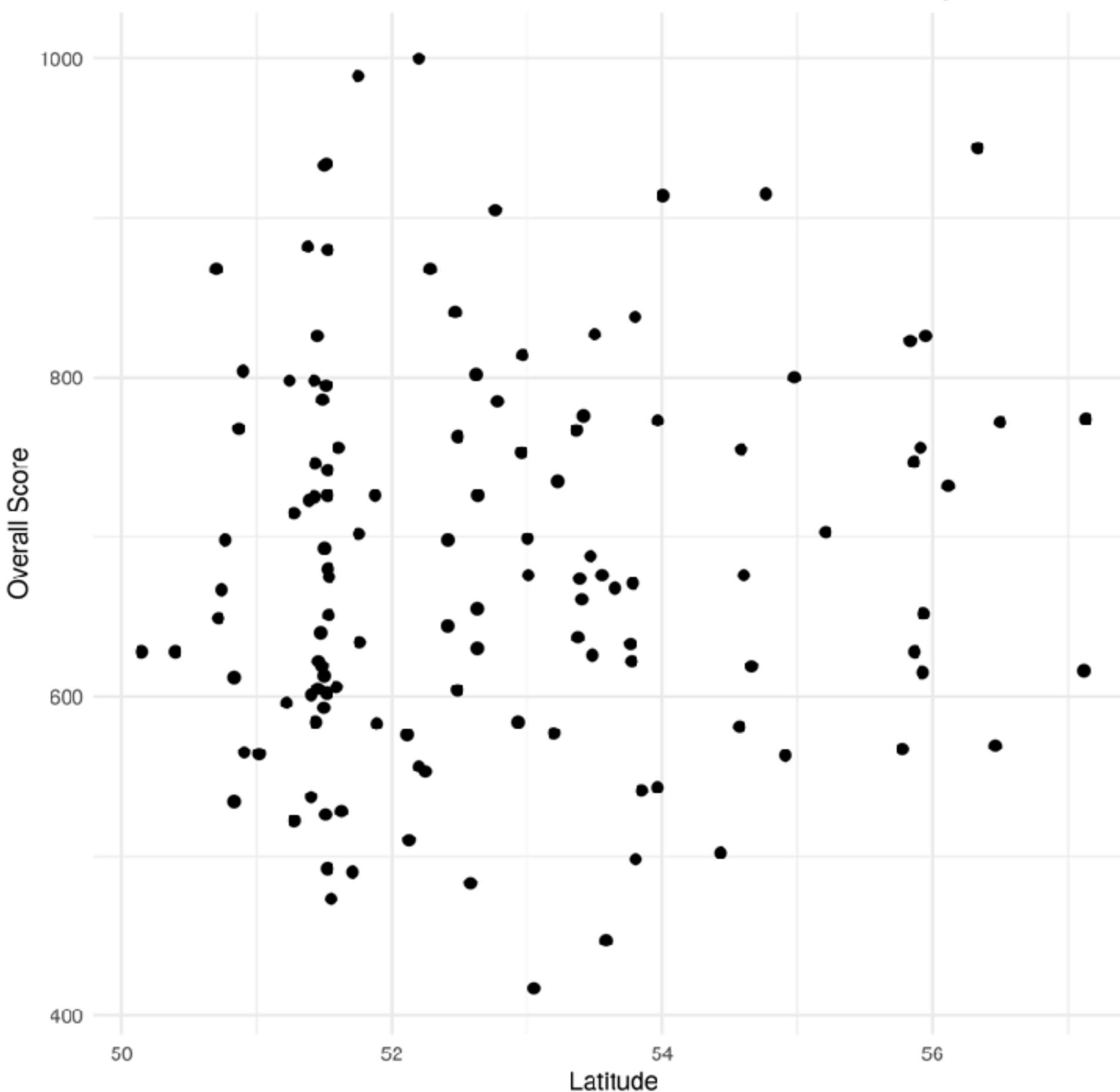
How to use | Methodology | FullTable | Print

Our League Tables rank UK universities both nationally and in 70 subjects. You can sort each table on the measure important to you and compare universities. Read university profiles for more information about the institutions or search for courses. We also rank 12 specialist colleges and conservatoires separately in the Arts, Drama & Music League Table.

Filter by Subjects Year Region Group Order by Overall Score

|   | Rank | University Name                | Entry Standards | Student Satisfaction | Research Quality | Graduate Prospects | Overall Score | Next Steps      |
|---|------|--------------------------------|-----------------|----------------------|------------------|--------------------|---------------|-----------------|
| ▼ | 1st  | ► 0 Cambridge                  | 224             | 4.09                 | 3.55             | 86.7               | 1000          | PROFILE COURSES |
| ▼ | 2nd  | ► 0 Oxford                     | 215             | 4.10                 | 3.34             | 83.4               | 989           | PROFILE COURSES |
| ▼ | 3rd  | ▲ 2 St Andrews                 | 207             | 4.26                 | 3.13             | 79.7               | 941           | PROFILE COURSES |
| ▼ | 4th  | ▼ 1 London School of Economics | 189             | 3.67                 | 3.35             | 86.1               | 934           | PROFILE COURSES |
| ▼ | 5th  | ▼ 1 Imperial College London    | 205             | 4.02                 | 3.36             | 90.4               | 933           | PROFILE COURSES |
| ▼ | 6th  | ► 0 Durham                     | 191             | 4.01                 | 3.14             | 84.8               | 915           | PROFILE COURSES |

Association between overall score and the latitude of a university



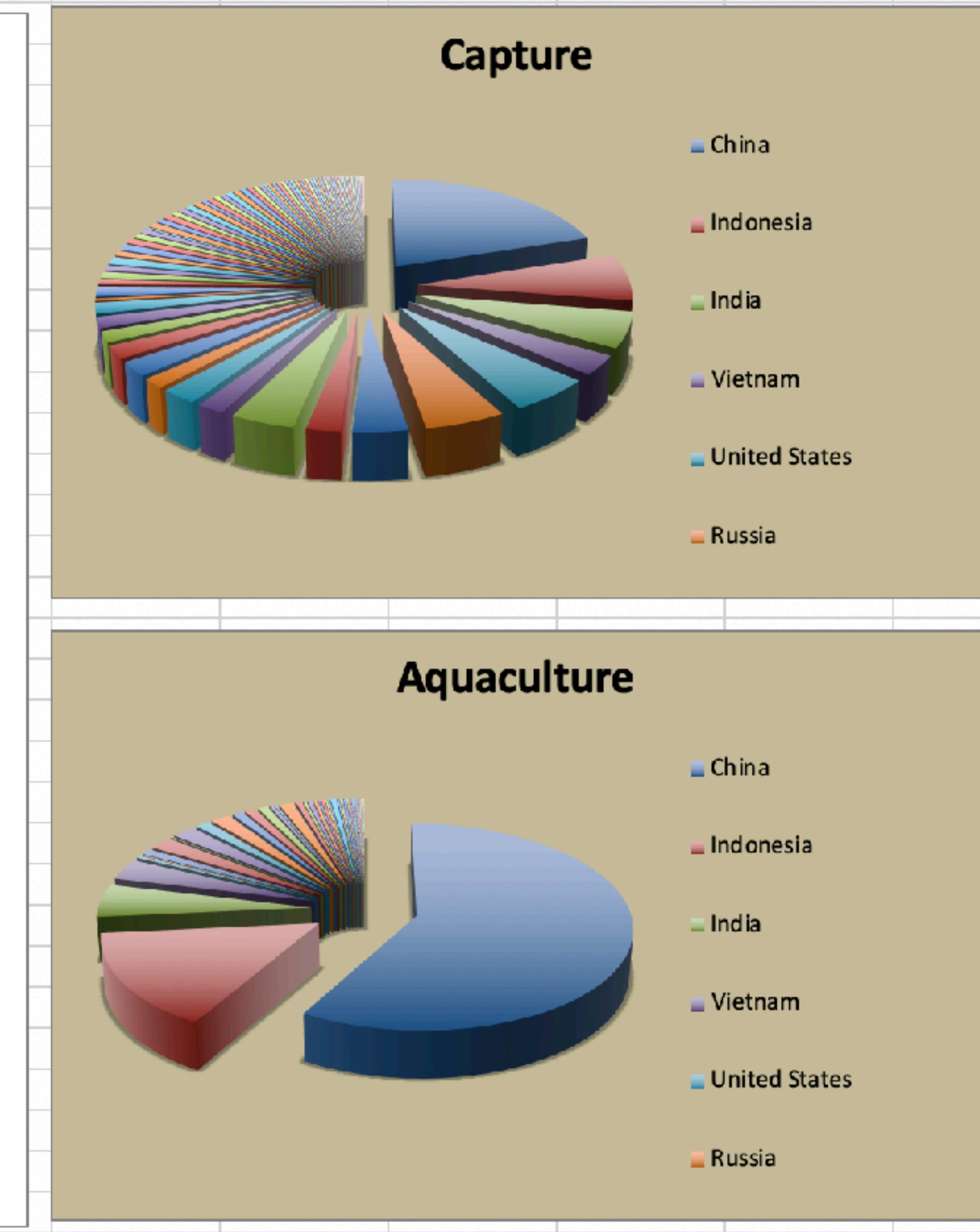
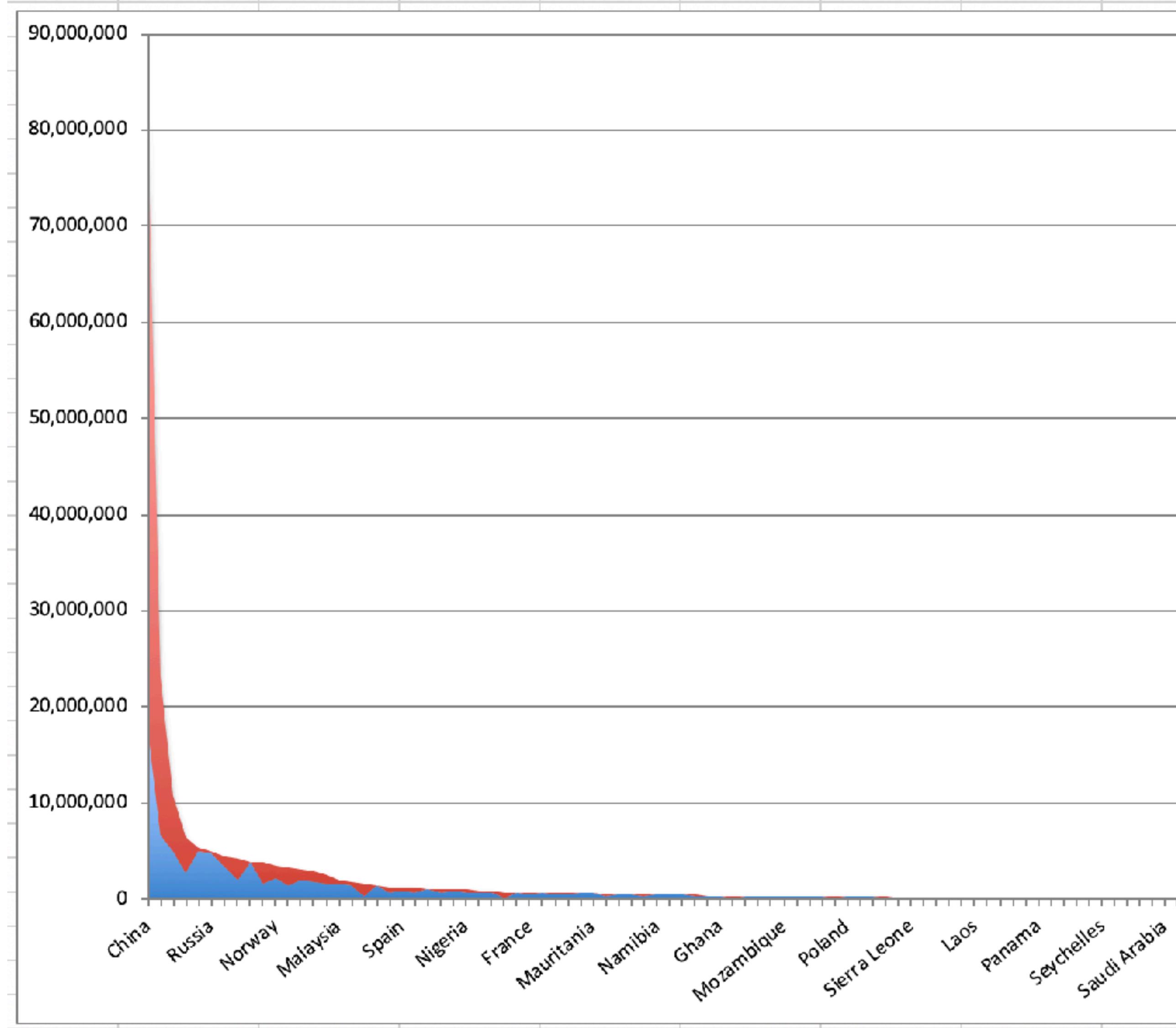
# Resources

- ▶ **Sample assignment:** [introds.org/hw/hw-06/hw-06-money-in-politics.html](https://introds.org/hw/hw-06/hw-06-money-in-politics.html)
- ▶ **Code:** Go to [bit.ly/rscloud-ecots2020](https://bit.ly/rscloud-ecots2020), start the project titled **02 - Money in politics**
- ▶ **Paper:** Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities (Dogucu & Çetinkaya-Rundel, 2020)  
[github.com/mdogucu/web-scrape](https://github.com/mdogucu/web-scrape) (conditionally accepted to JSE)

ex. 2

# fisheries of the world





```
fisheries %>% select(country)
```

```
## # A tibble: 82 x 1
##   country
##   <chr>
## 1 Angola
## 2 Argentina
## 3 Australia
## 4 Bangladesh
## 5 Brazil
## 6 Cambodia
## 7 Cameroon
## 8 Canada
## 9 Chad
## 10 Chile
```

```
continents
```

```
## # A tibble: 245 x 2
##   country      continent
##   <chr>        <chr>
## 1 Afghanistan Asia
## 2 Åland Islands Europe
## 3 Albania      Europe
## 4 Algeria      Africa
## 5 American Samoa Oceania
## 6 Andorra      Europe
## 7 Angola       Africa
## 8 Anguilla     Americas
## 9 Antigua & Barbuda Americas
## 10 Argentina   Americas
```

\* data joins

```
fisheries <- left_join(fisheries, continents)
```

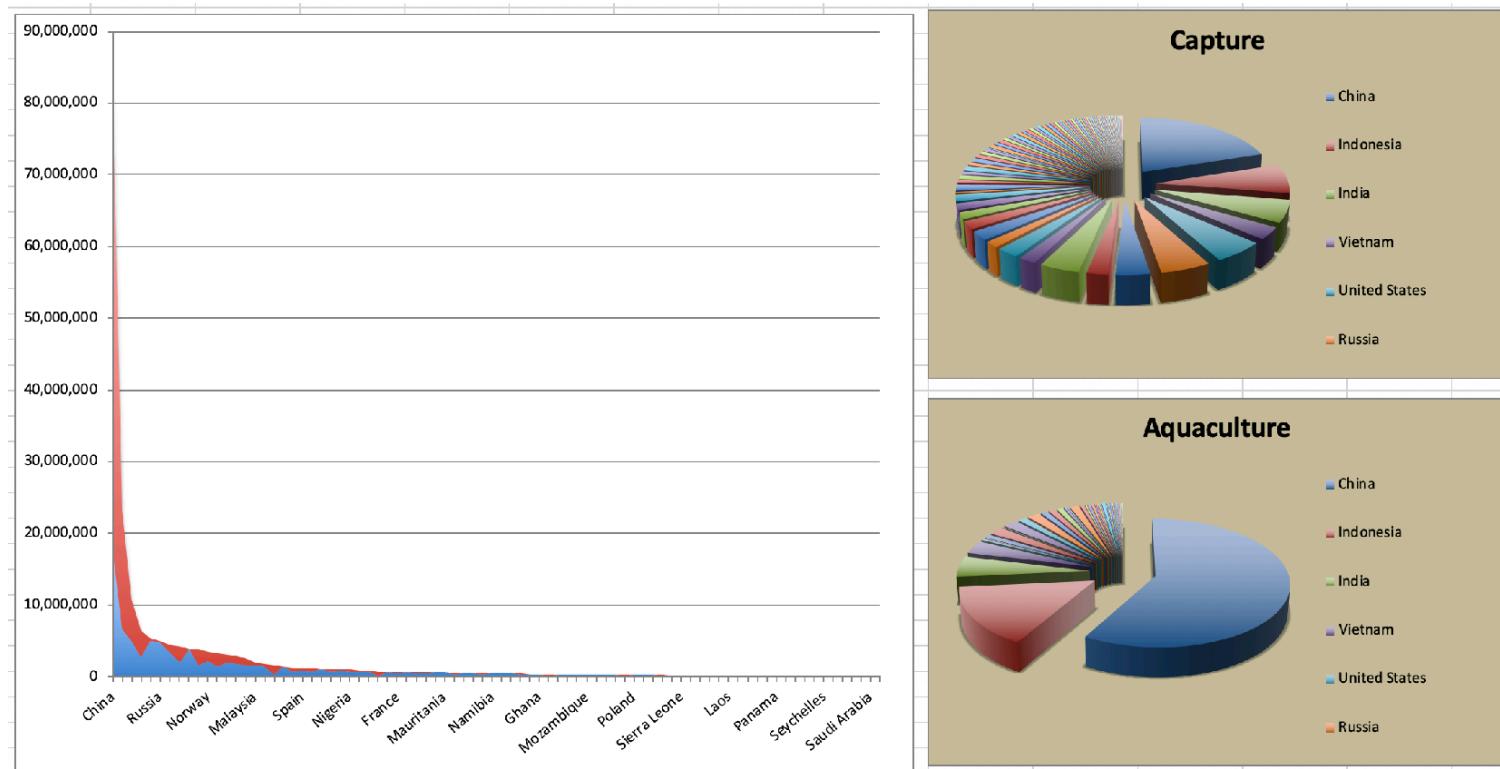
```
## Joining, by = "country"
```

- ★ data joins
- ★ data science ethics

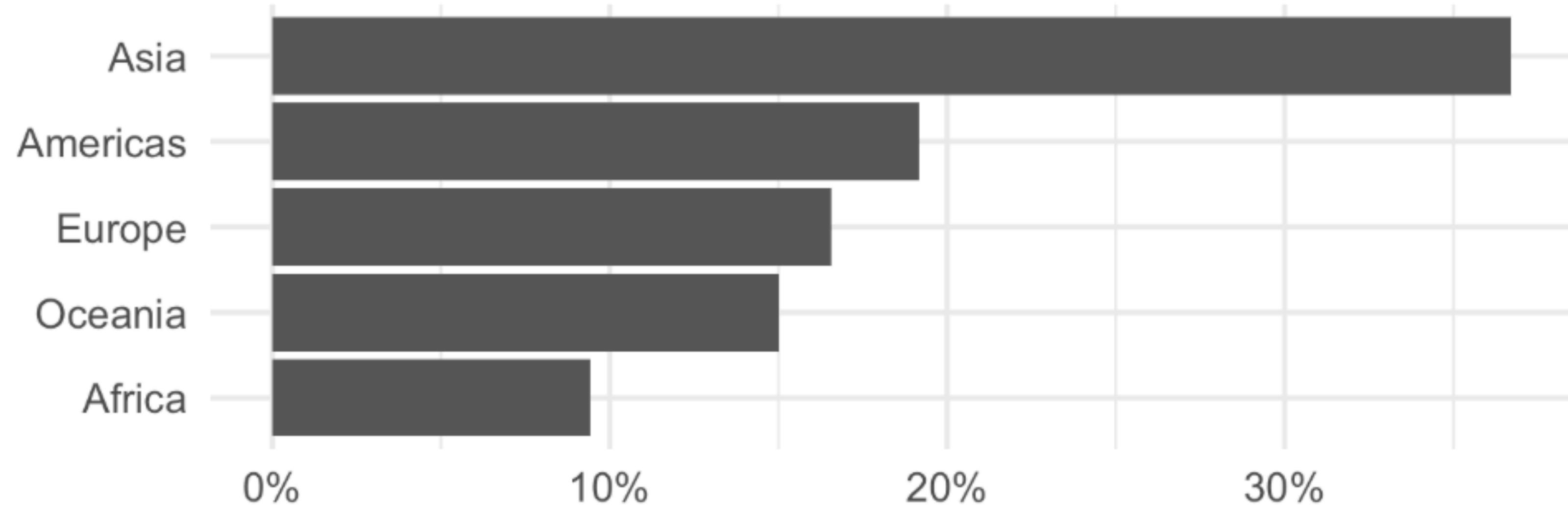
```
fisheries %>%
  filter(is.na(continent))
```

```
## # A tibble: 3 x 5
##   country           capture aquaculture    total continent
##   <chr>            <dbl>      <dbl>     <dbl> <chr>
## 1 Democratic Republic of the Congo 237372       3161  240533 <NA>
## 2 Hong Kong          142775       4258  147033 <NA>
## 3 Myanmar            2072390      1017644 3090034 <NA>
```

```
fisheries <- fisheries %>%
  mutate(continent = case_when(
    country == "Democratic Republic of the Congo" ~ "Africa",
    country == "Hong Kong" ~ "Asia",
    country == "Myanmar" ~ "Asia",
    TRUE ~ continent
  ))
```

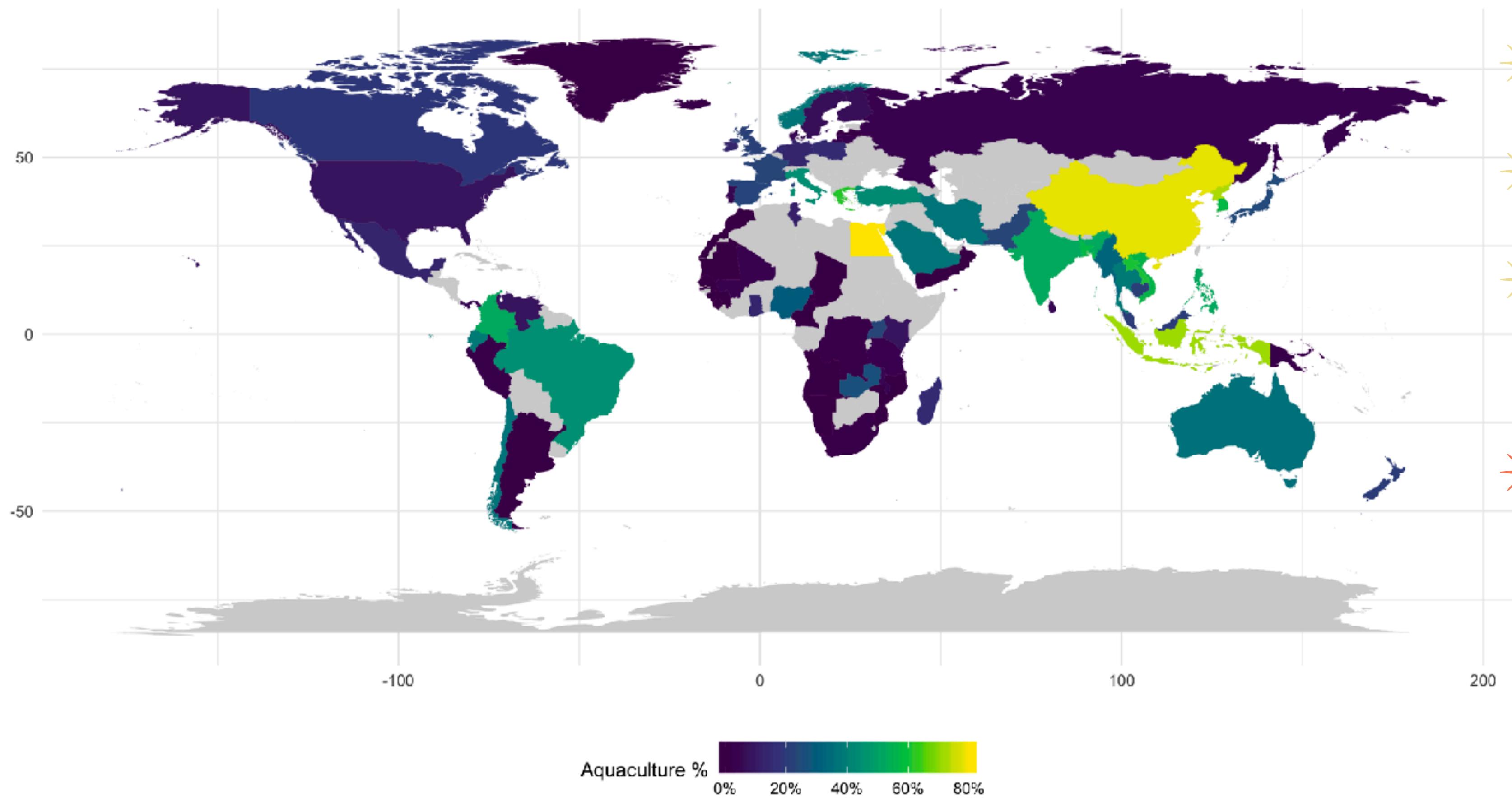


## Average share of aquaculture by continent out of total fisheries harvest, 2016



- ★ data joins
- ★ data science ethics
- ★ critique
- ★ improving data visualisations

Average share of aquaculture by country  
out of total fisheries harvest, 2016



- ★ data joins
- ★ data science ethics
- ★ critique
- ★ improving data visualisations
- ★ mapping

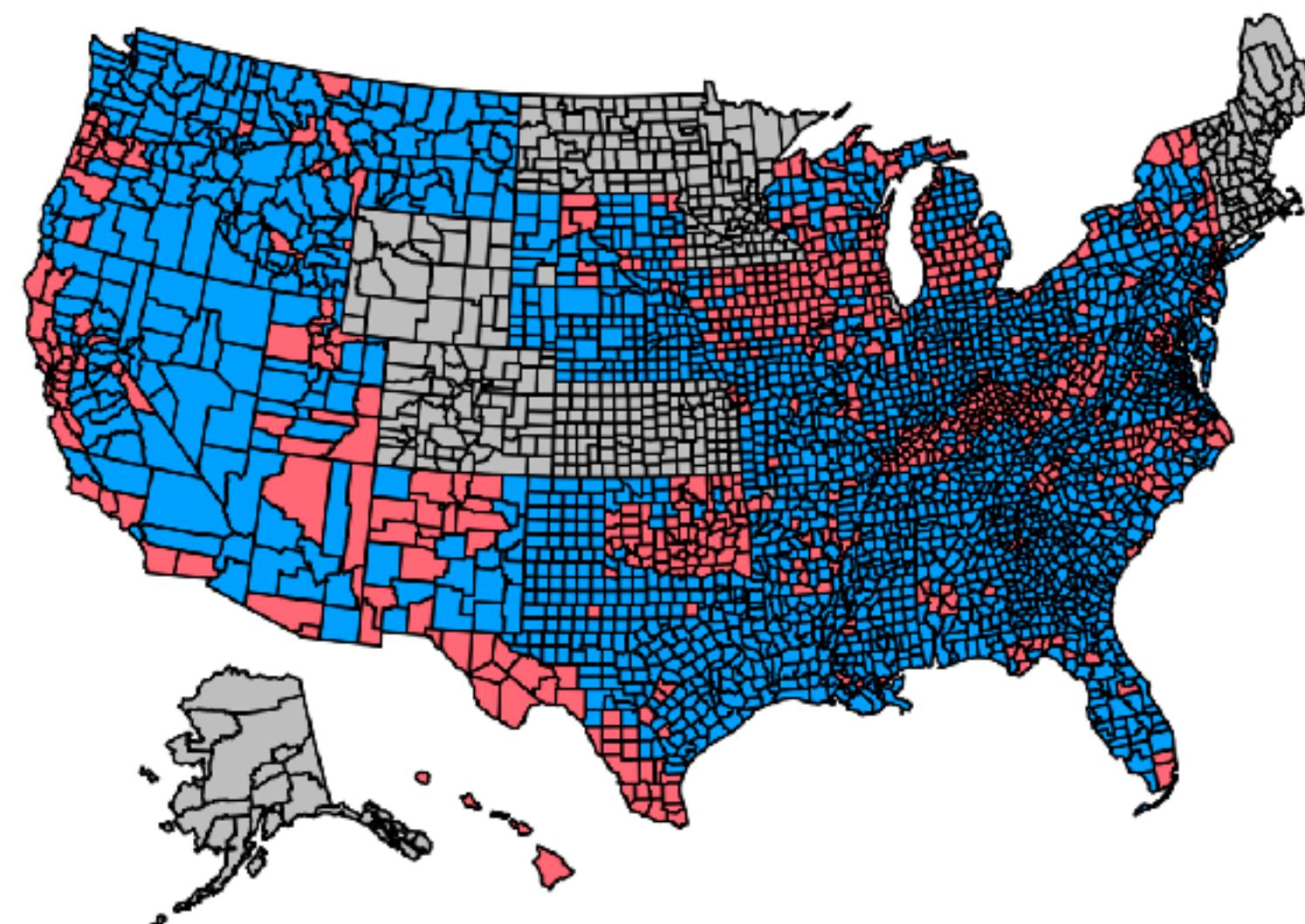
Source: [bit.ly/2VrawTt](https://bit.ly/2VrawTt)

# Project: 2016 US Election Redux

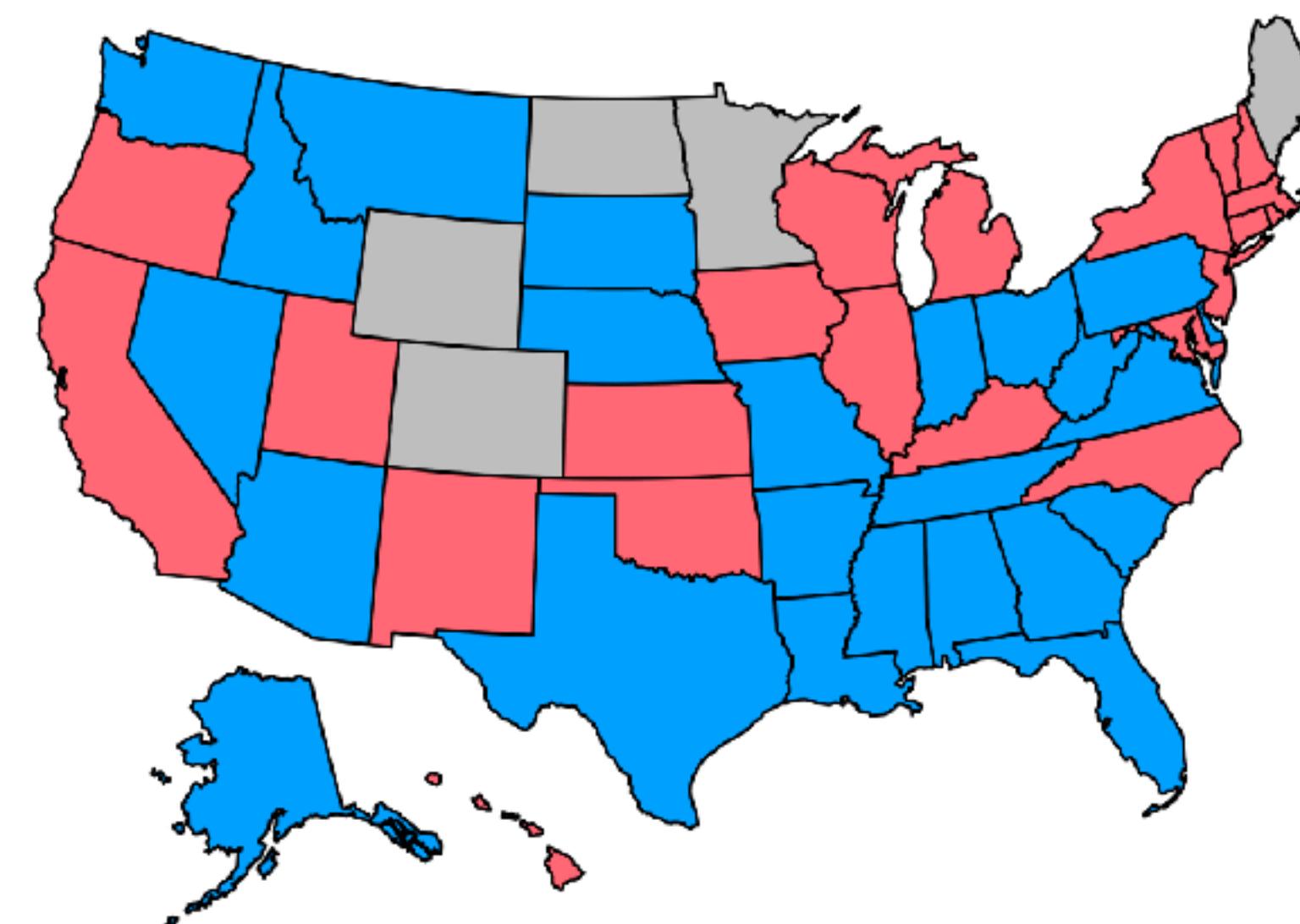
**Question:** Would the outcome of the 2016 US Presidential Elections been different had Bernie Sanders been the Democrat candidate?

**Team:** 4 Squared

Predicted Results  
By County



Predicted Results  
By State



| Candidate      |
|----------------|
| Bernie Sanders |
| Donald Trump   |
| NA             |

# Resources

- ▶ Sample lab: [introds.org/labs/lab-04/lab-04-ugly-charts.html](https://introds.org/labs/lab-04/lab-04-ugly-charts.html)
- ▶ Code: Go to [bit.ly/rscloud-ecots2020](https://bit.ly/rscloud-ecots2020), start the project titled **03 - Fisheries of the world**
- ▶ Sample lecture: [introds.org/slides/w4\\_d1-effective-dataviz/w4\\_d1-effective-dataviz.html](https://introds.org/slides/w4_d1-effective-dataviz/w4_d1-effective-dataviz.html)

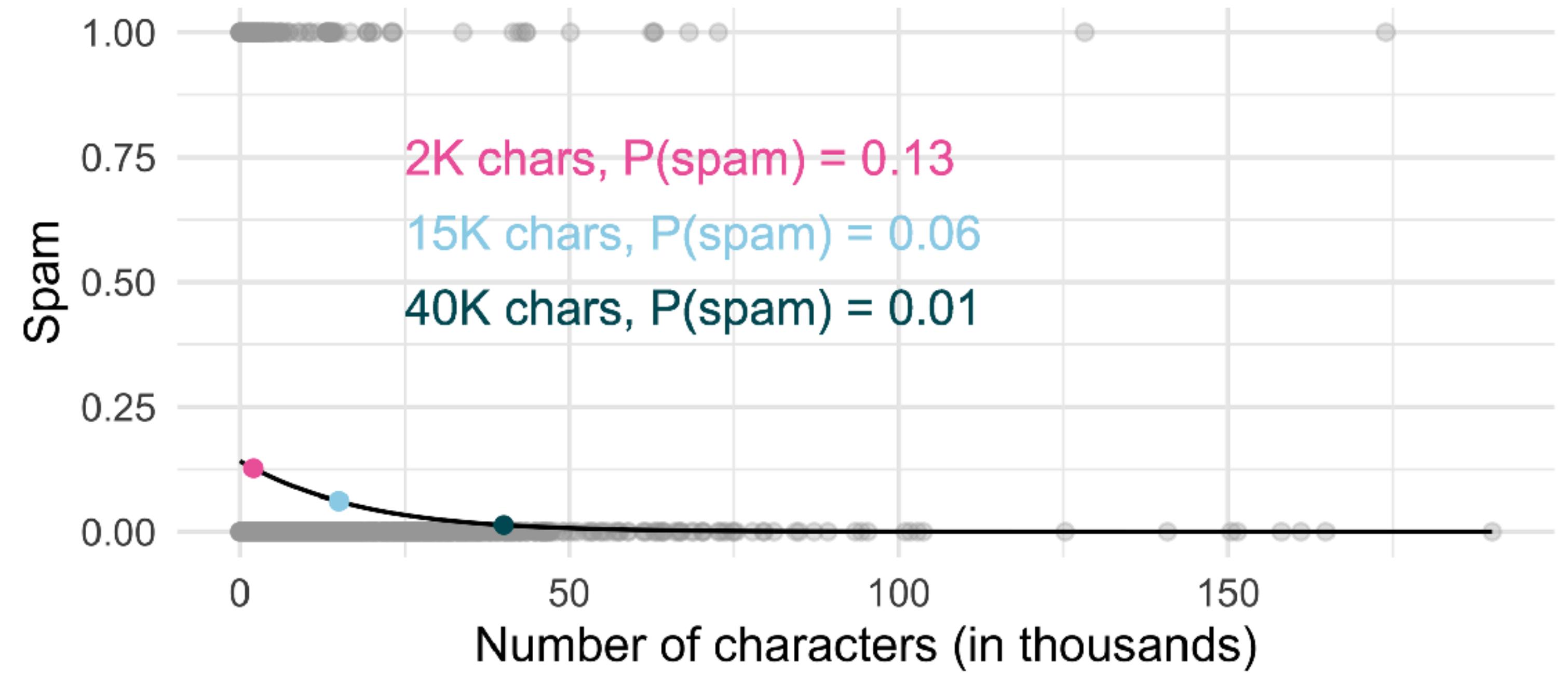
ex. 3

spam filters



## Spam vs. number of characters

- \* logistic regression
- \* prediction



- ★ logistic regression
- ★ prediction
- ★ decision errors
- ★ sensitivity / specificity
- ★ intuition around loss functions

|                         | <b>Email is spam</b>          | <b>Email is not spam</b>      |
|-------------------------|-------------------------------|-------------------------------|
| Email labelled spam     | True positive                 | False positive (Type 1 error) |
| Email labelled not spam | False negative (Type 2 error) | True negative                 |

# Project: Spotify Top 100 Tracks of 2017/18

**Question:** Is it possible to predict the year a song made the Top Tracks playlist based on its metadata?

**Team:** weR20

```
year ~ danceability + energy + key + loudness + mode + speechiness +
      acousticness + instrumentalness + liveness + valence + tempo +
      duration_s
```

## 2017

|                   | name         | artists          |
|-------------------|--------------|------------------|
| I'm the One       |              | DJ Khaled        |
| Redbone           |              | Childish Gambino |
| Sign of the Times | Harry Styles |                  |

## 2018

|  | name                               | artists      |
|--|------------------------------------|--------------|
|  | Everybody Dies In Their Nightmares | XXXTENTACION |
|  | Jocelyn Flores                     | XXXTENTACION |
|  | Plug Walk                          | Rich The Kid |
|  | Moonlight                          | XXXTENTACION |
|  | Nevermind                          | Dennis Lloyd |
|  | In My Mind                         | Dynoro       |
|  | changes                            | XXXTENTACION |

# Resources

- ▶ **Sample lecture:** [introds.org/slides/w10\\_d1-logistic-regression/w10\\_d1-logistic-regression.html](https://introds.org/slides/w10_d1-logistic-regression/w10_d1-logistic-regression.html)
- ▶ **Code:** Go to [bit.ly/rscloud-ecots2020](https://bit.ly/rscloud-ecots2020), start the project titled **04 - Spam filter**
- ▶ **Book chapter:** **OpenIntro Statistics**, 4th Edition (Diez, Çetinkaya-Rundel, and Barr, 2019), Chapter 9.5 with randomised controlled trial data on discrimination on job application evaluation [openintro.org/book/os](https://openintro.org/book/os)



pedagogy



**teams:** weekly labs in teams +  
periodic team evaluations +  
term project in teams

**peer feedback:** used  
minimally so far, but  
positive experience

**“minute paper”:** weekly online  
quizzes ending with a brief  
reflection of the week’s material



# Week 07 - Simple linear regression

Teacher salaries

Single numerical predictor

Prediction

Assessing model fit

Working backwards

Single categorical predictor

Finish up

Start Over



## Finish up

To finish up the quiz go to the form linked below and answer a few simple questions.

Hi Mine, when you submit this form, the owner will be able to see your name and email address.

\* Required

1. What is your name? \*

Enter your answer

2. What is your student ID? \*

This is the number that starts with s.

Enter your answer

3. Write about one or two questions you didn't get right initially but were able to solve after a few tries. What was difficult about them? What did you ultimately learn?

OR

If you got every single question correct on the first try, write one question you would still like clarified on the topics covered in this quiz. \*

Your answers can be brief / in bullet point form. The goal isn't to make you write too much, but instead to make you quickly reflect on your learning.

Enter your answer

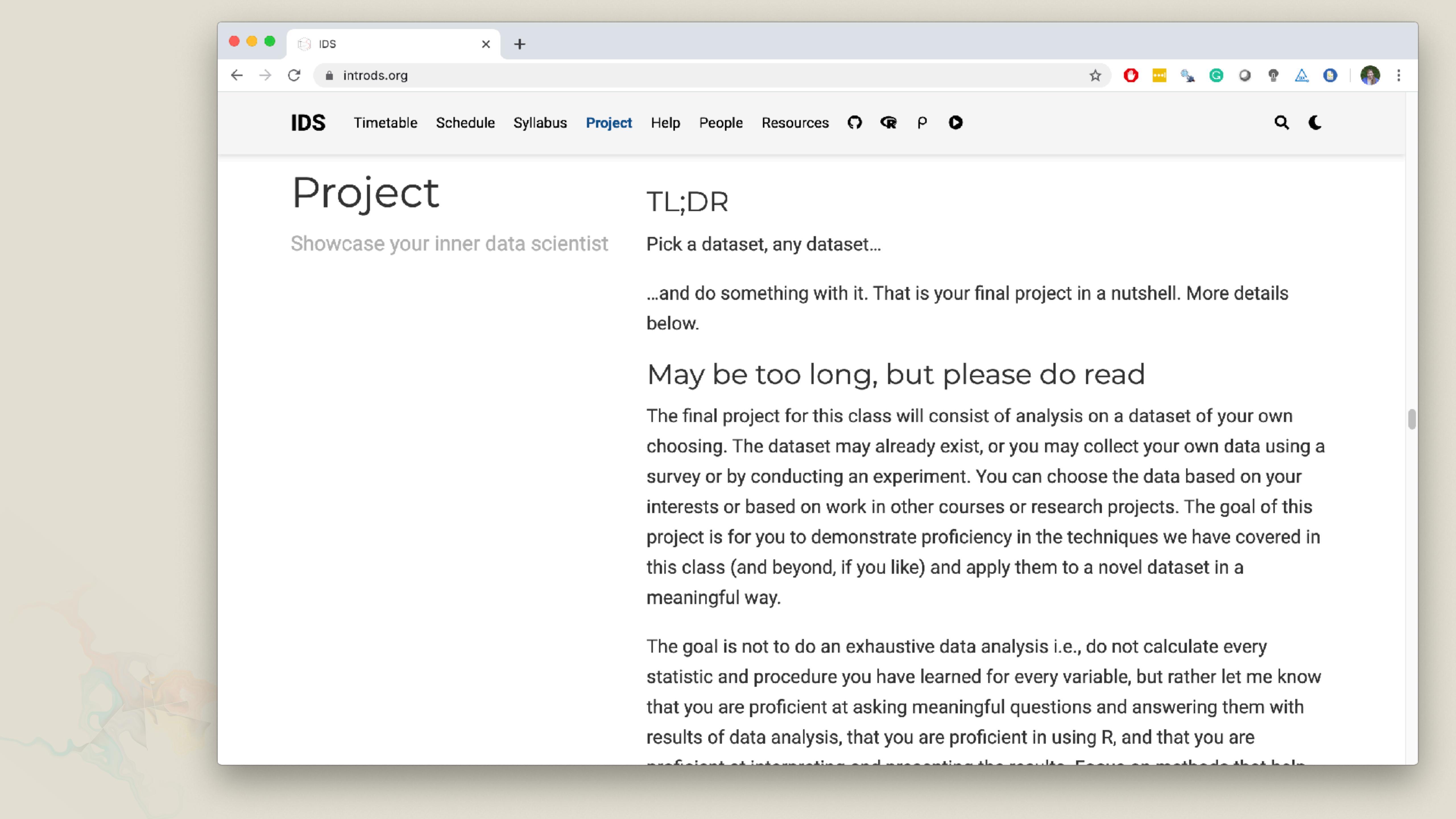


**teams:** weekly labs in teams +  
periodic team evaluations +  
term project in teams

**peer feedback:** used  
minimally so far, but  
positive experience

**“minute paper”:** weekly online  
quizzes ending with a brief  
reflection of the week's material

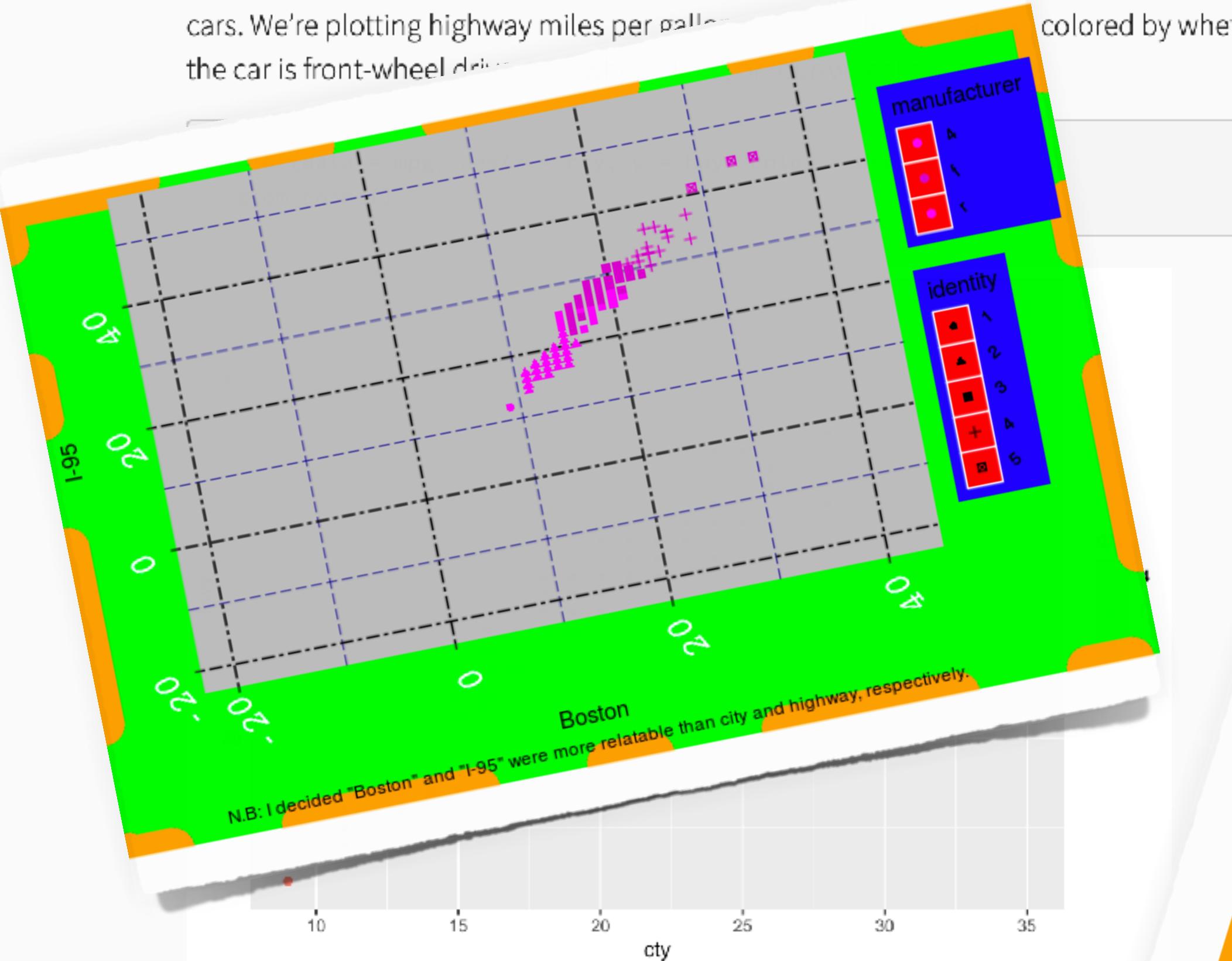
**creativity:** assignments that  
make room for creativity





## Part 3 - Mirror, mirror on the wall, who's the ugliest of them all?

Here is a simple plot using the `mpg` dataset, which contains info on fuel economy of cars. We're plotting highway miles per gallon (`hwy`) on the y-axis against city miles per gallon (`cty`) on the x-axis. The points are colored by whether the car is front-wheel drive (`drv`), with four categories: 4 (red), f (green), t (blue), and r (orange).

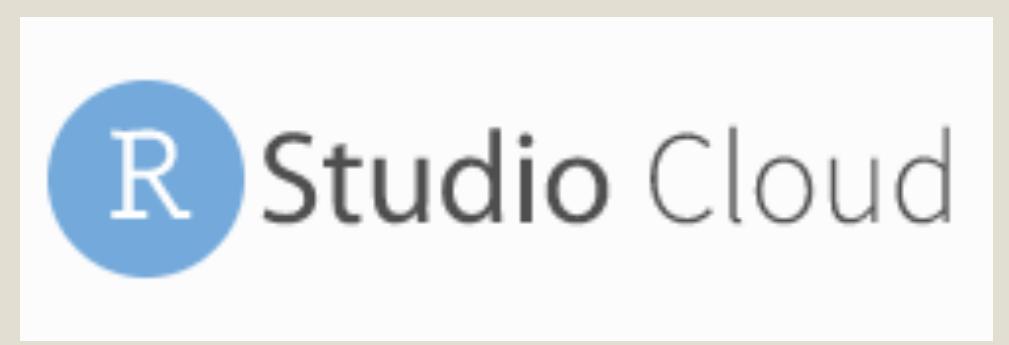


**Exercise 11.** Make this plot as ugly as possible by changing colors, background color, fonts, or anything else you can think of. You will probably want to play around with [theme options](#), but you can do more. You can also search online for other themes, fonts, etc. that you want to tweak.

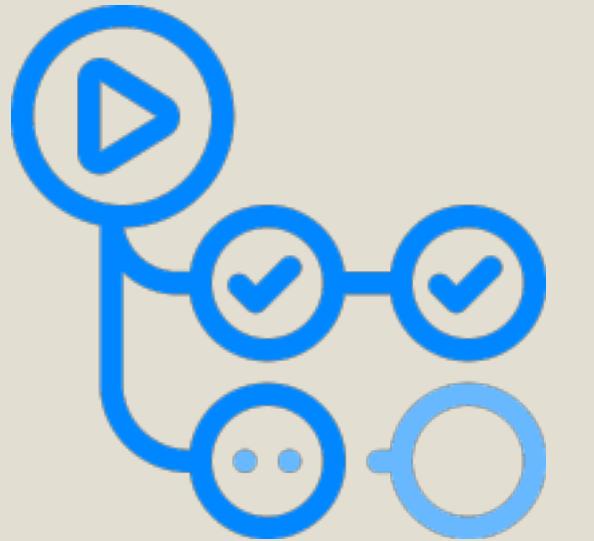




**infrastructure**



+

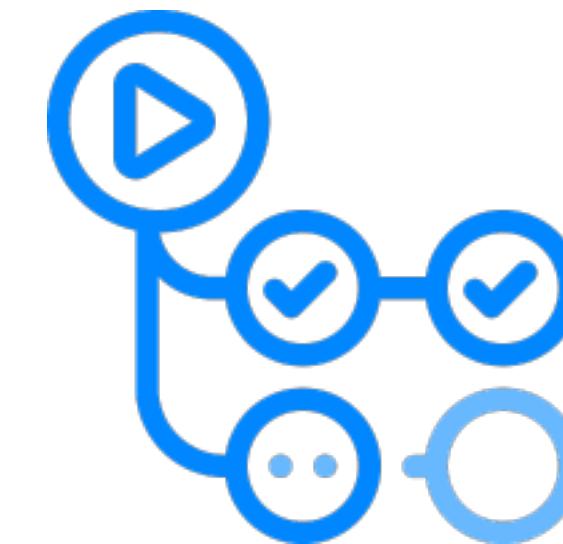
**ghclass**



+



ghclass



ghclass · GitHub

Code Issues Pull requests Projects Wiki Security Insights Settings

## styler revisions #1

Draft mine-cetinkaya... wants to merge 1 commit into master from styler

Conversation 0 Commits 1 Checks 0 Files changed 1

+16 -18

Changes from all commits ▾ File filter... ▾ Jump to... ▾

Review changes ▾

### Results of running styler:

Styling 1 files:  
hw-03-ncbikecrash.Rmd

| Status | Count | Legend                  |
|--------|-------|-------------------------|
| ✓      | 0     | File unchanged.         |
| i      | 1     | File changed.           |
| *      | 0     | Styling threw an error. |

Please review the changes carefully!

styler (#1)

mine-cetinkaya-rundel committed 16 minutes ago

commit 721c39cf8b6c0e6b3bc50f2918747415910b980e

34 hw-03-ncbikecrash.Rmd

@@ -7,7 +7,7 @@ output: github\_document

```
7
8  ```{r load-packages, message=FALSE}
9  # load packages
10 -library(tidyverse)
```

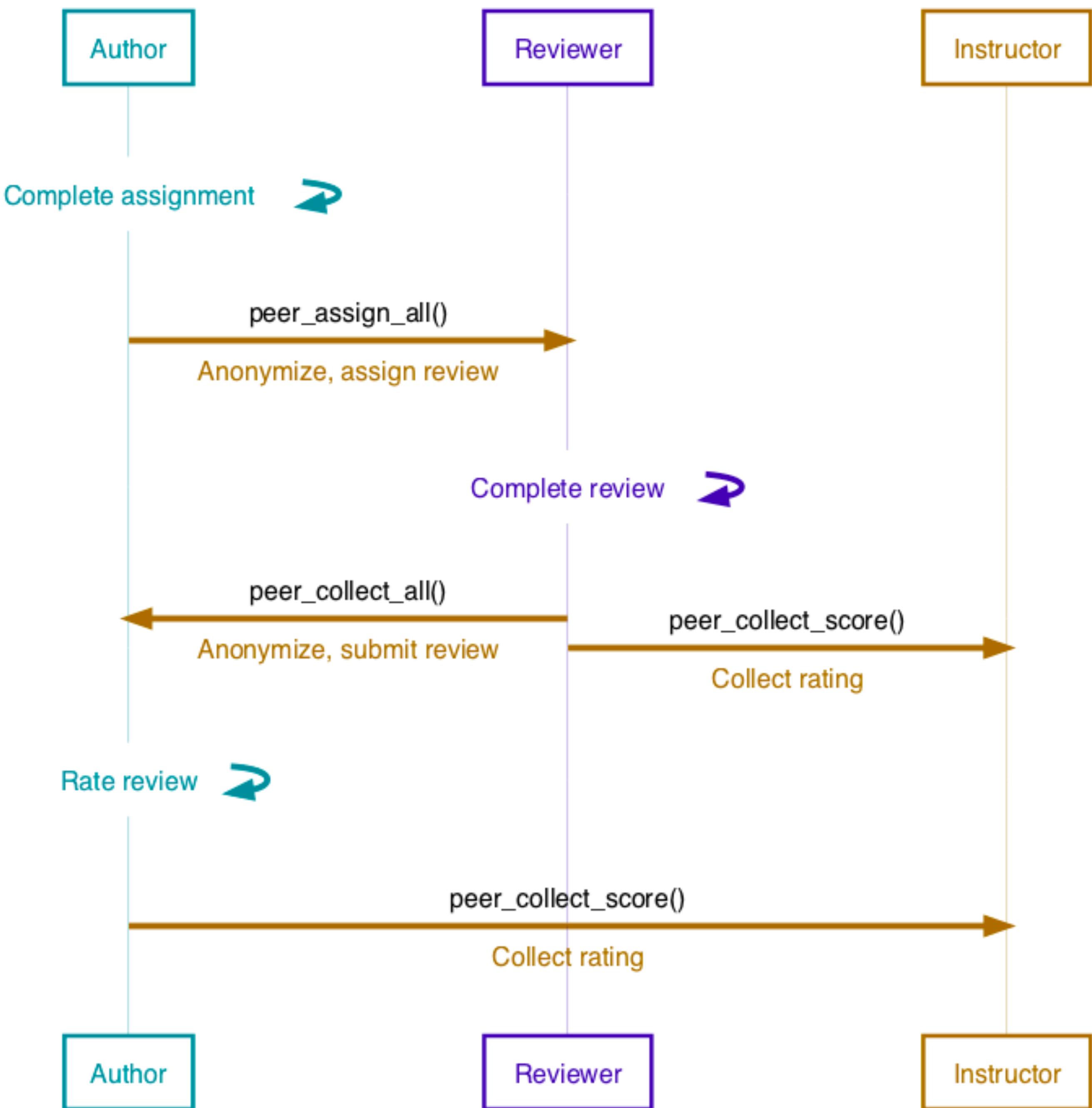
```
7
8  ```{r load-packages, message=FALSE}
9  # load packages
10 +library(tidyverse)
```

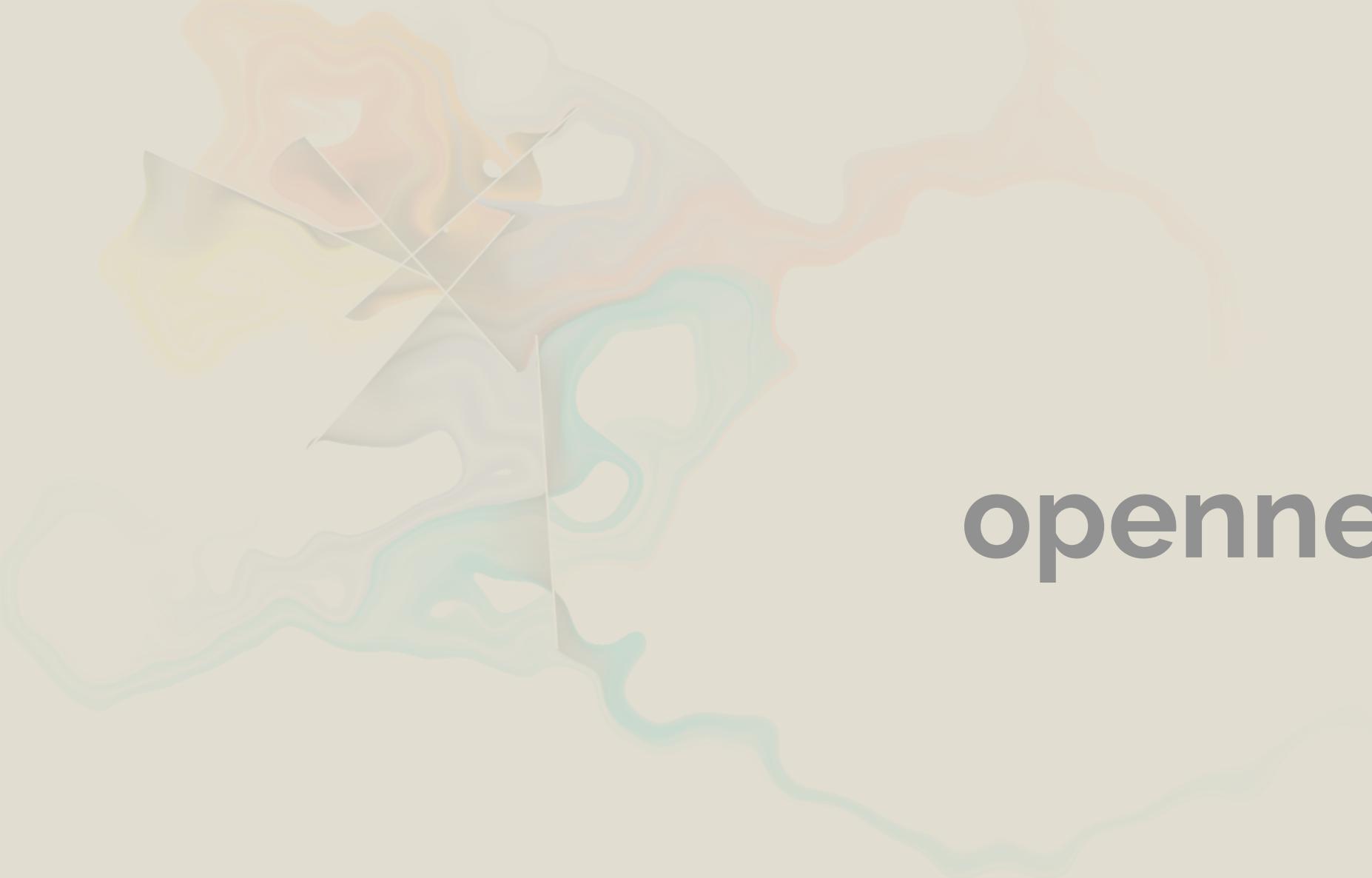


+



ghclass





**openness**

IDS introds.org Incognito

IDS Timetable Schedule Syllabus Project Help People Resources Q R P ▶

Introduction to Data Science

Learn to explore, visualize, and analyze data to understand natural phenomena, investigate patterns, model outcomes, and make predictions, and do so in a reproducible and shareable manner. Gain experience in data collection, wrangling, and visualization, exploratory data analysis, predictive modeling, and effective communication of results while working on problems and case studies inspired by and based on real-world questions. The course will focus on the R statistical computing language. No statistical or computing background is necessary.

Introduction to Data Science

Fall 2019

University of Edinburgh

Q R P ▶

The screenshot shows a web browser window for 'Data Science in a Box :: Data S X' at [datasciencebox.org](https://datasciencebox.org). The page features a large yellow 3D cube logo with 'DATA SCIENCE' on top, 'IN A BOX' on the side, and a small 'ds' icon on the front. To the right of the logo is a search bar with a magnifying glass icon and the placeholder 'Search...'. Below the search bar is a sidebar with links: 'Hello #dsbox', 'Course content', 'Infrastructure', and 'Pedagogy'. To the right of the sidebar is a large section titled 'Data Science in a Box' with a descriptive paragraph. Further down is another section with text and a blue right-pointing arrow. At the bottom left is a footer note.

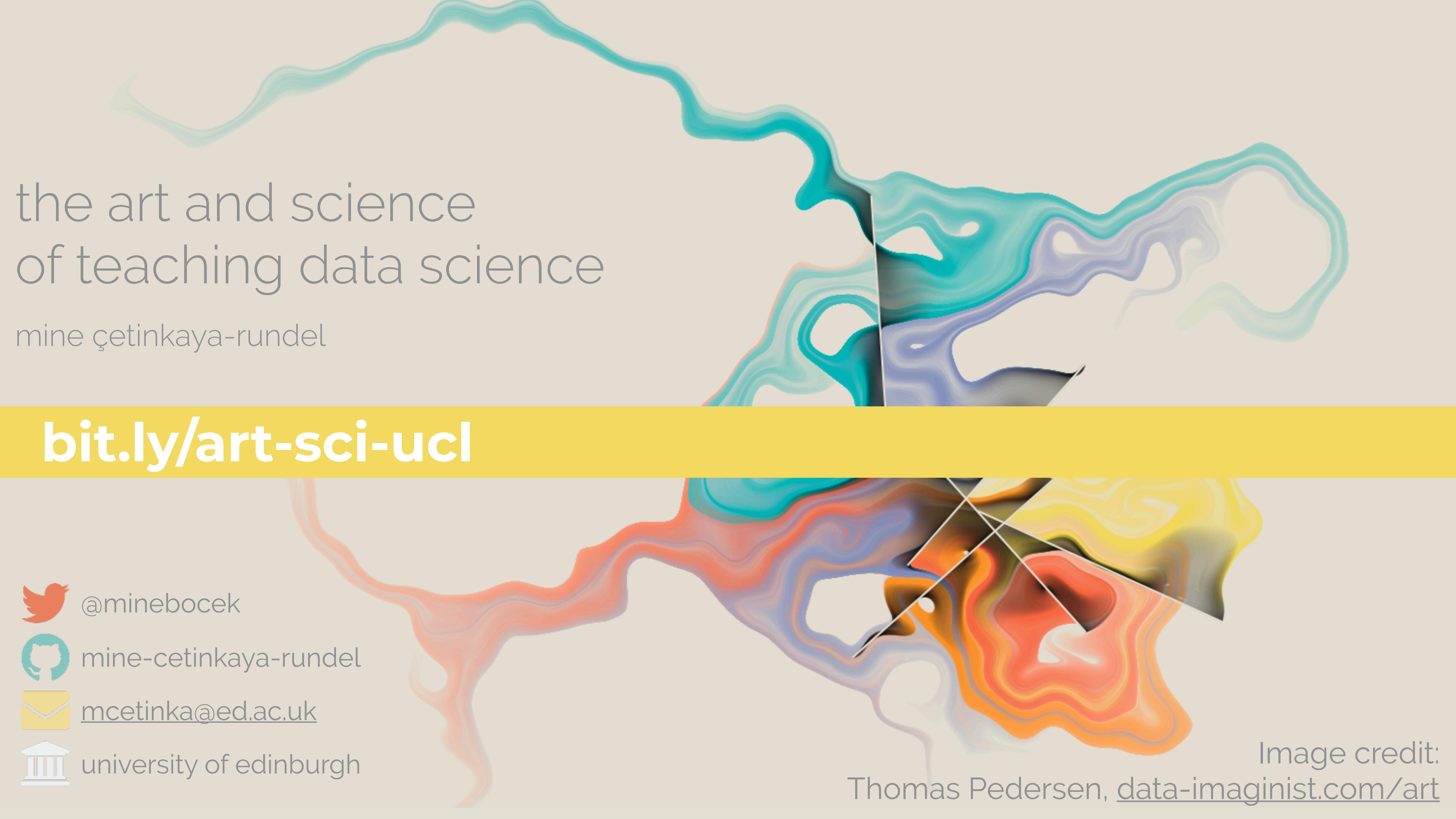
# Data Science in a Box

How can we effectively and efficiently teach data science to students with little to no background in computing and statistical thinking? How can we equip them with the skills and tools for reasoning with various types of data and leave them wanting to learn more? This introductory data science course is our (working) answer to this question.

The core content of the course focuses on data acquisition and wrangling, exploratory data analysis, data visualization, inference, modeling, and effective communication of results. Time permitting, the course also introduces additional concepts and tools like interactive visualization and reporting, text analysis, and Bayesian inference. A heavy emphasis is placed on a consistent syntax (with tools from the [tidyverse](#)), reproducibility (with [R Markdown](#)), and version control and collaboration (with Git and GitHub). In addition, out-of-class learning is supplemented with interactive [tutorials](#). The goal of the course is to bring students from zero to being able to work in a team on a fully reproducible data science project analyzing a dataset of their choice and answering questions they care about.

Data Science in a Box contains the materials required to teach (or learn from) the course described above, all of which are [freely-available and open-source](#). They include course materials such as slide decks, homework assignments, guided labs, sample exams, a final project assignment, as well as materials for instructors such as pedagogical tips, information on computing infrastructure, technology stack, and course logistics.

Built with ❤️ and [blogdown](#), logo by [muuuuge](#).



# the art and science of teaching data science

mine çetinkaya-rundel

[bit.ly/art-sci-ucl](https://bit.ly/art-sci-ucl)

 @minebocek

 mine-cetinkaya-rundel

 [mcetinka@ed.ac.uk](mailto:mcetinka@ed.ac.uk)

 university of edinburgh

Image credit:  
Thomas Pedersen, [data-imaginist.com/art](http://data-imaginist.com/art)

*four*

~~three~~ questions that keep me up at night...

**content**

1 what should students learn?

**pedagogy**

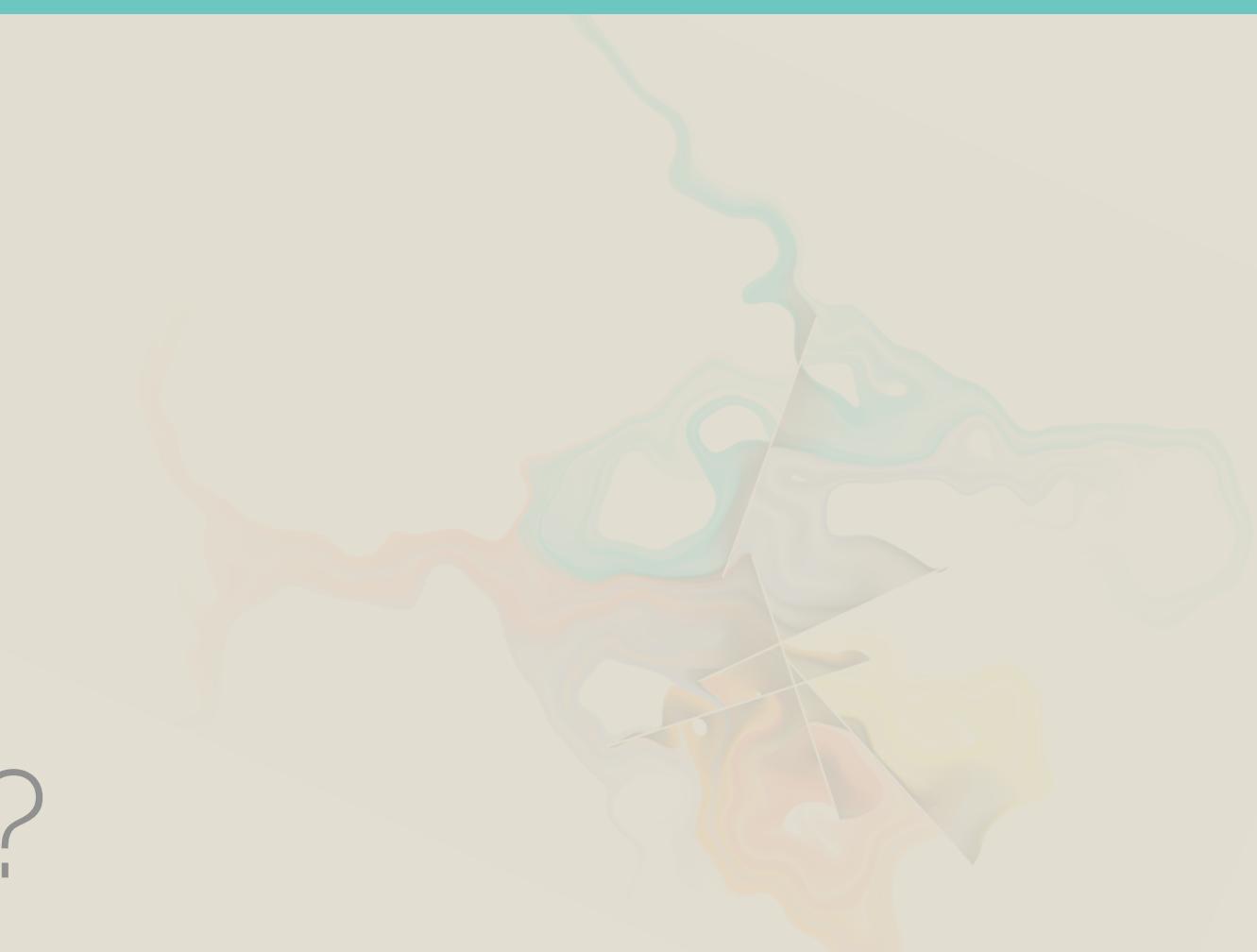
2 how will students learn best?

**infrastructure**

3 what tools will enhance student learning?

**assessment**

4 how can we assess any of this?



**in progress:** retrospective study

**data:** 205 open-ended student projects  
over 4 years

**group 1:**

learned R & intro  
statistics using  
base R

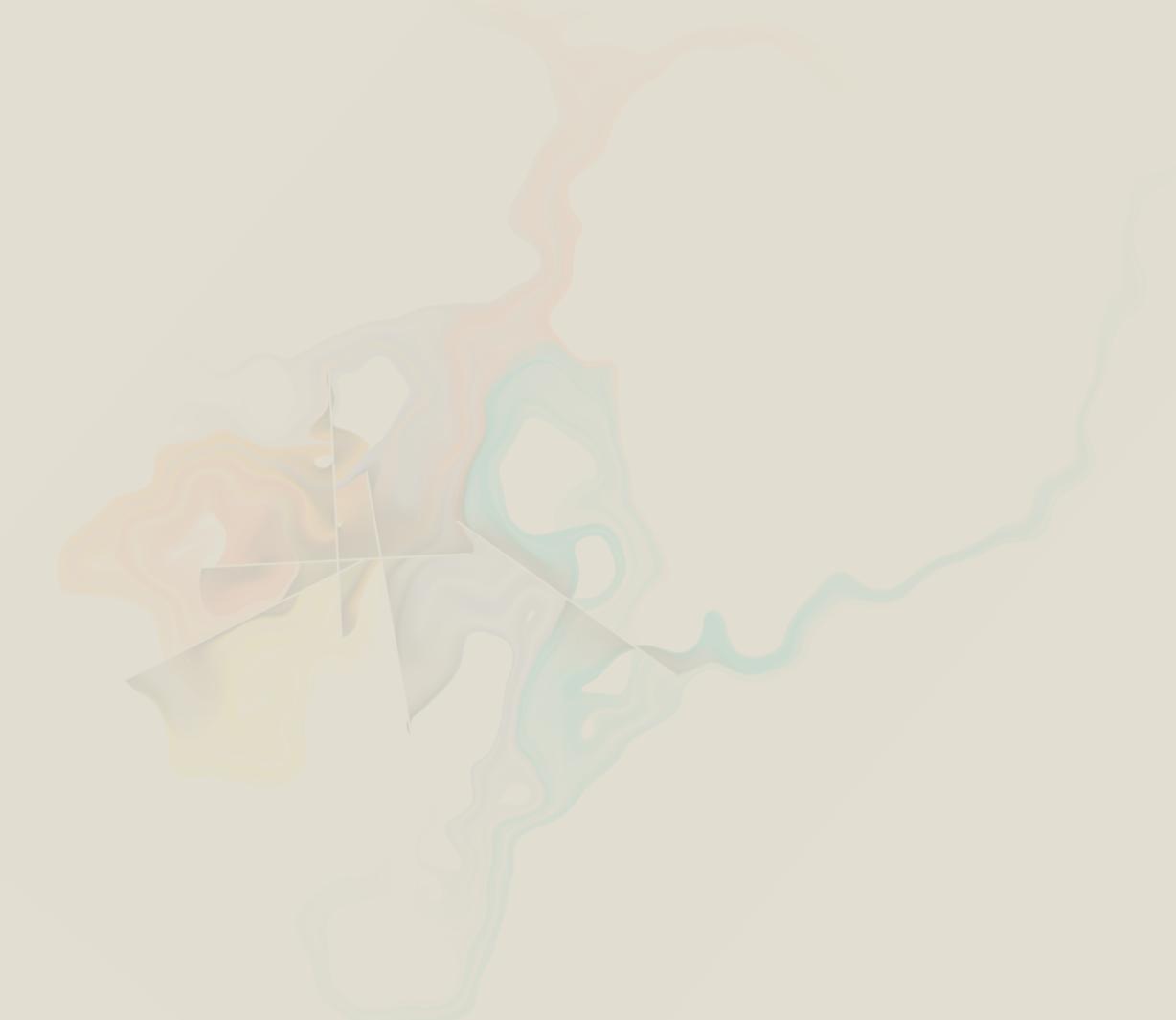
**group 2:**

learned R & intro  
statistics using  
tidyverse\*

\* starting before the term  
tidyverse was coined.

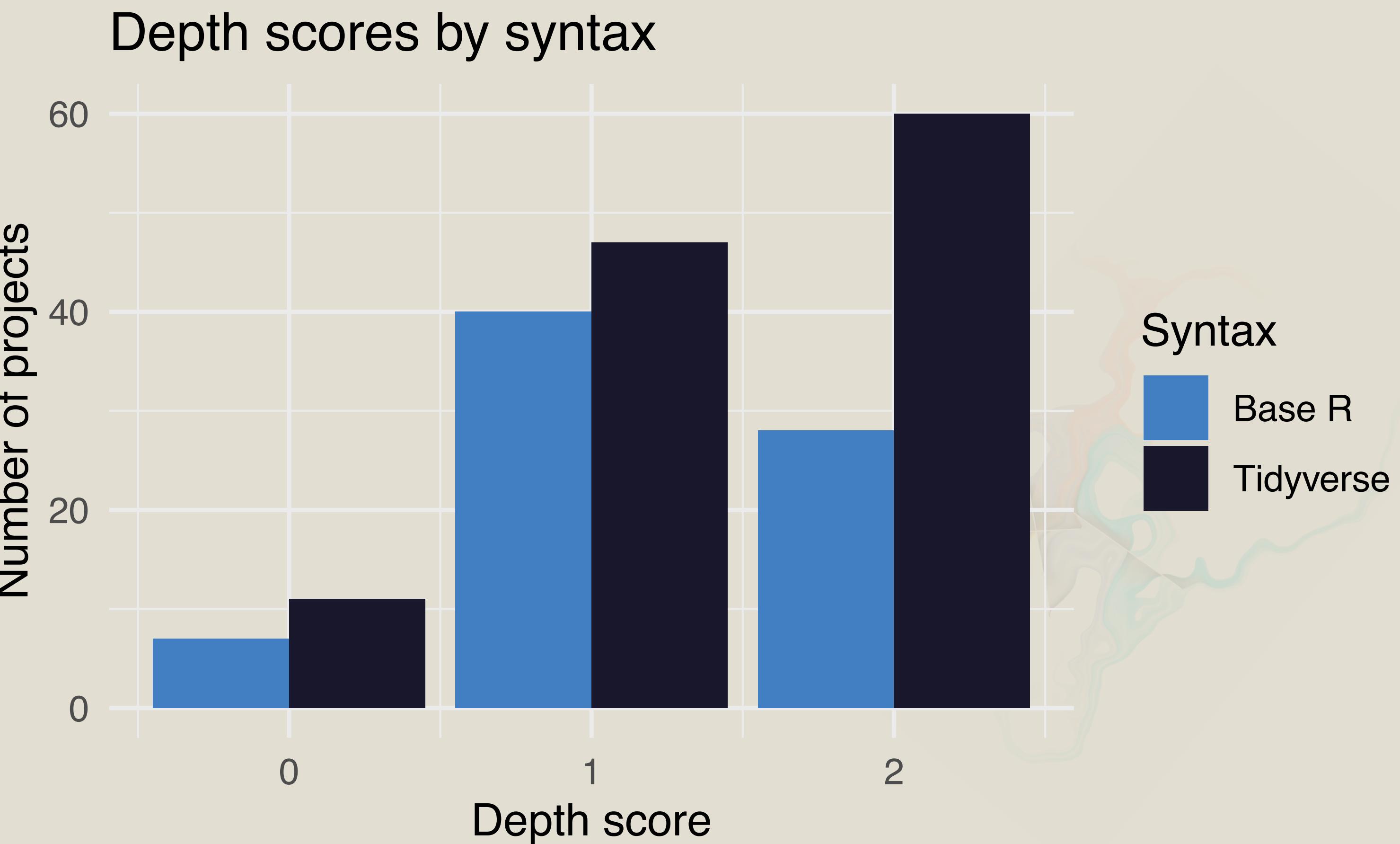
same assignment, same(ish) dataset

**measures:** *creativity*, *depth* and the  
complexity of ***multivariate visualisations***



## depth

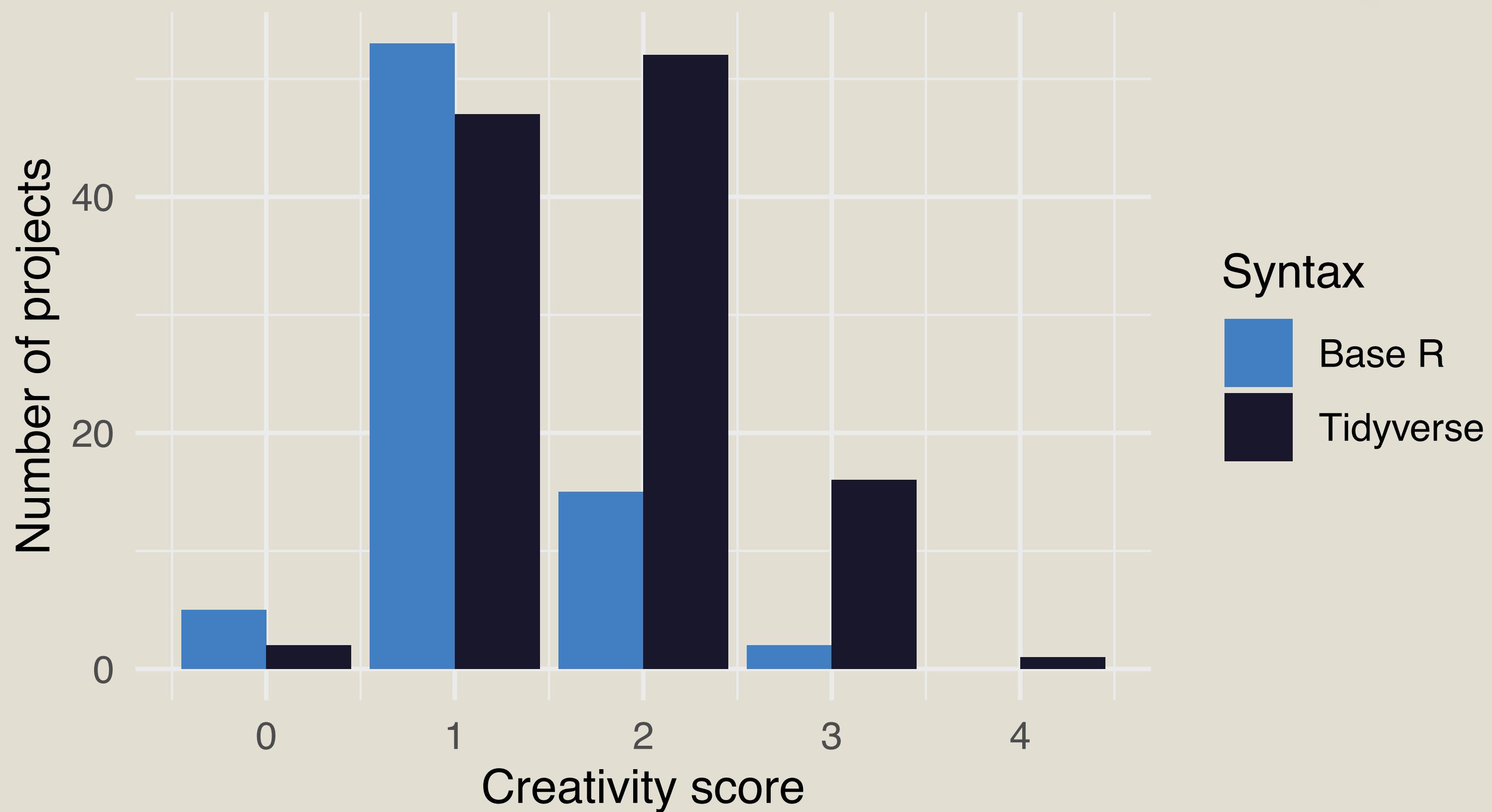
- consistent theme throughout the project
- relevant data for each analysis



# creativity

- creation of new variables
- transformation of existing variables
- subgroup analysis
- use of a subset of data for the entire project

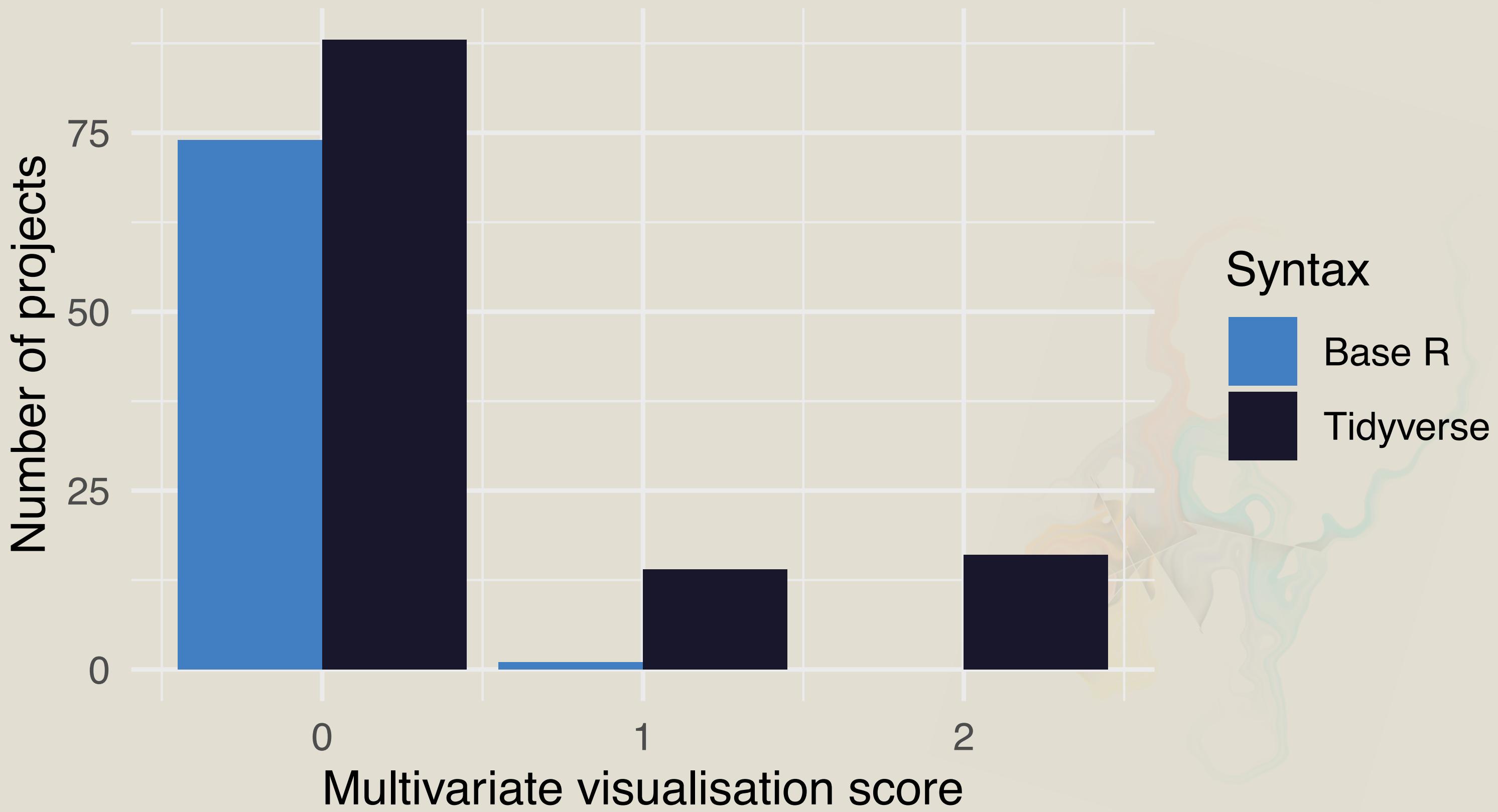
Creativity scores by syntax



# multivariate visualisation

- visualisation with 3+ variables
- effective interpretations of visualisations

Multivariate visualisation by syntax



# summary

|   | Estimate | Std. error | Statistic | P-value | 95% CI          |
|---|----------|------------|-----------|---------|-----------------|
| <b>Creativity (0 - 4)</b>                 |          |            |           |         |                 |
| Intercept                                 | 1.187    | 0.080      | 14.876    | 0.000   | (1.029, 1.344)  |
| Tidyverse                                 | 0.534    | 0.102      | 5.231     | 0.000   | (0.332, 0.735)  |
| <b>Depth (0 - 2)</b>                      |          |            |           |         |                 |
| Intercept                                 | 1.280    | 0.075      | 17.153    | 0.000   | (1.133, 1.427)  |
| Tidyverse                                 | 0.135    | 0.095      | 1.417     | 0.158   | (-0.053, 0.324) |
| <b>Multivariate visualisation (0 - 2)</b> |          |            |           |         |                 |
| Intercept                                 | 0.013    | 0.065      | 0.204     | 0.838   | (-0.115, 0.142) |
| Tidyverse                                 | 0.376    | 0.083      | 4.509     | 0.000   | (0.212, 0.541)  |



**planned:** longitudinal study

**motivation:** higher conversion rate to stat 2

**explorations:**

retention, especially of  
students from under-  
represented  
backgrounds

preparation and  
confidence for applied  
and collaborative  
projects