teaching introductory data science

mine çetinkaya-rundel

minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com

🔗 bit.ly/ptt21-intro-ds

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?

goals

demonstrate concrete course examples

share a few tips

provide open-source teaching resources

focus on

data visualisation
data wrangling, tidying, acquisition
exploratory data analysis
predictive modeling + uncertainty quantification
effective communication of results

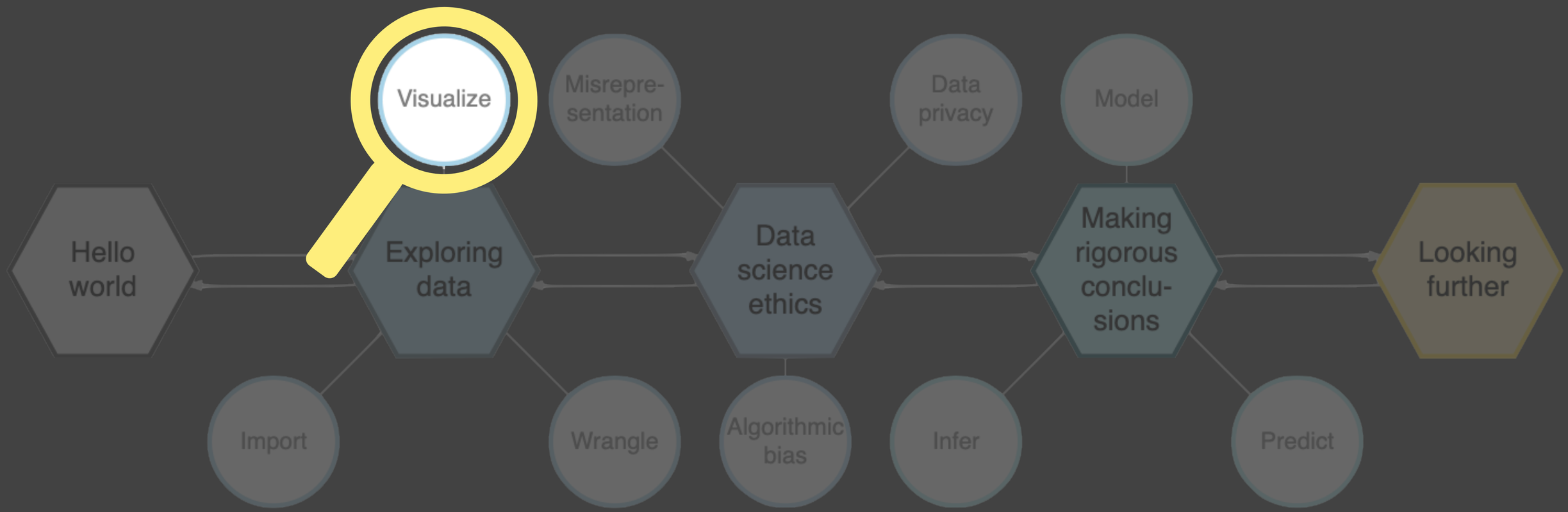
foray into

interactive visualizations
text analysis
machine learning
Bayesian inference
...

emphasise

consistent syntax | tidyverse
reproducibility | R Markdown
version control and collaboration | Git + GitHub

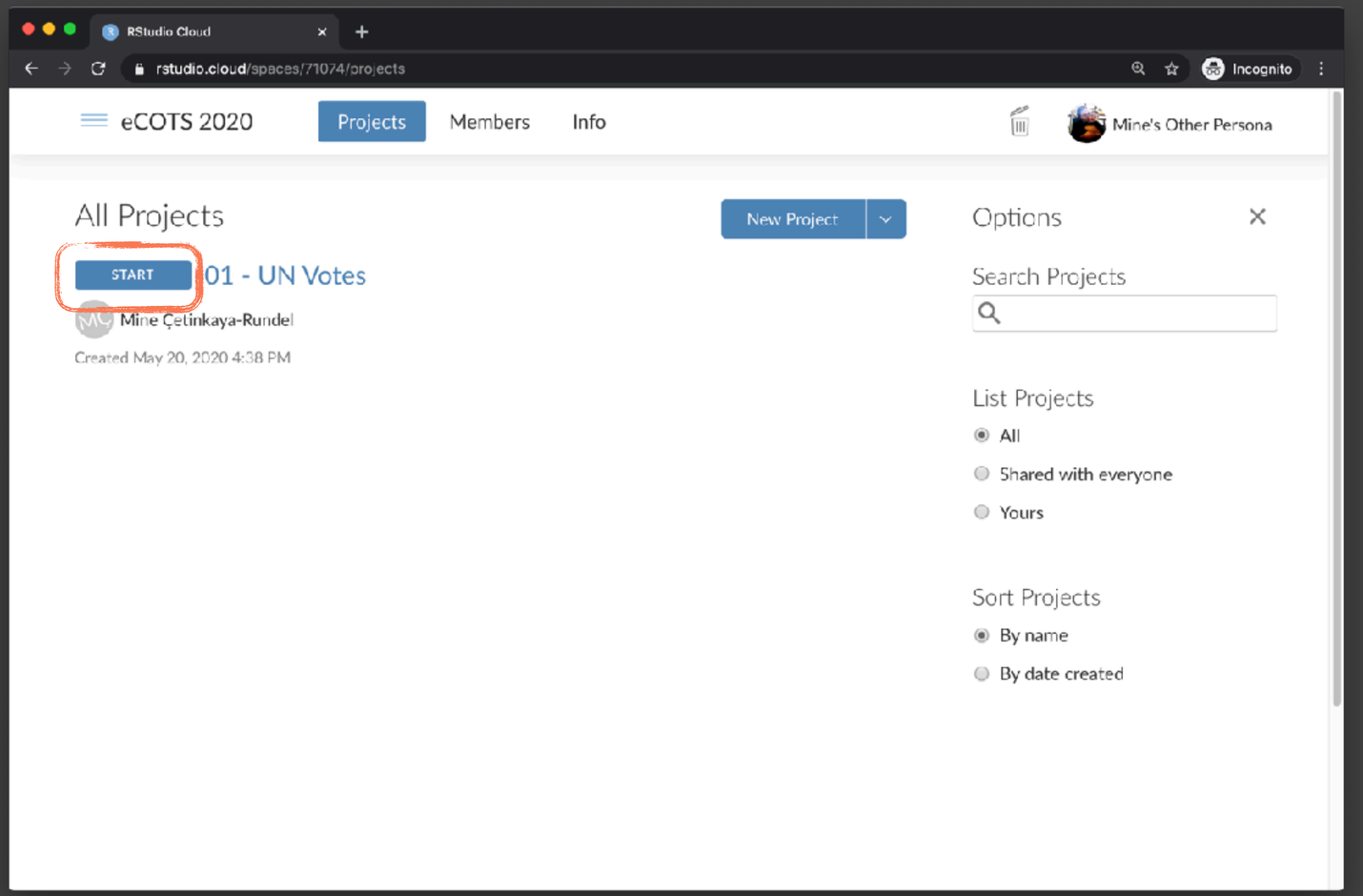
topics


Visualize
Misrepre-
sentation
Data
privacy
Model
Hello
world
Exploring
data
Data
science
ethics
Making
rigorous
conclu-
sions
Looking
further
Import
Wrangle
Algorithmic
bias
Infer
Predict

ex. 1
united nations

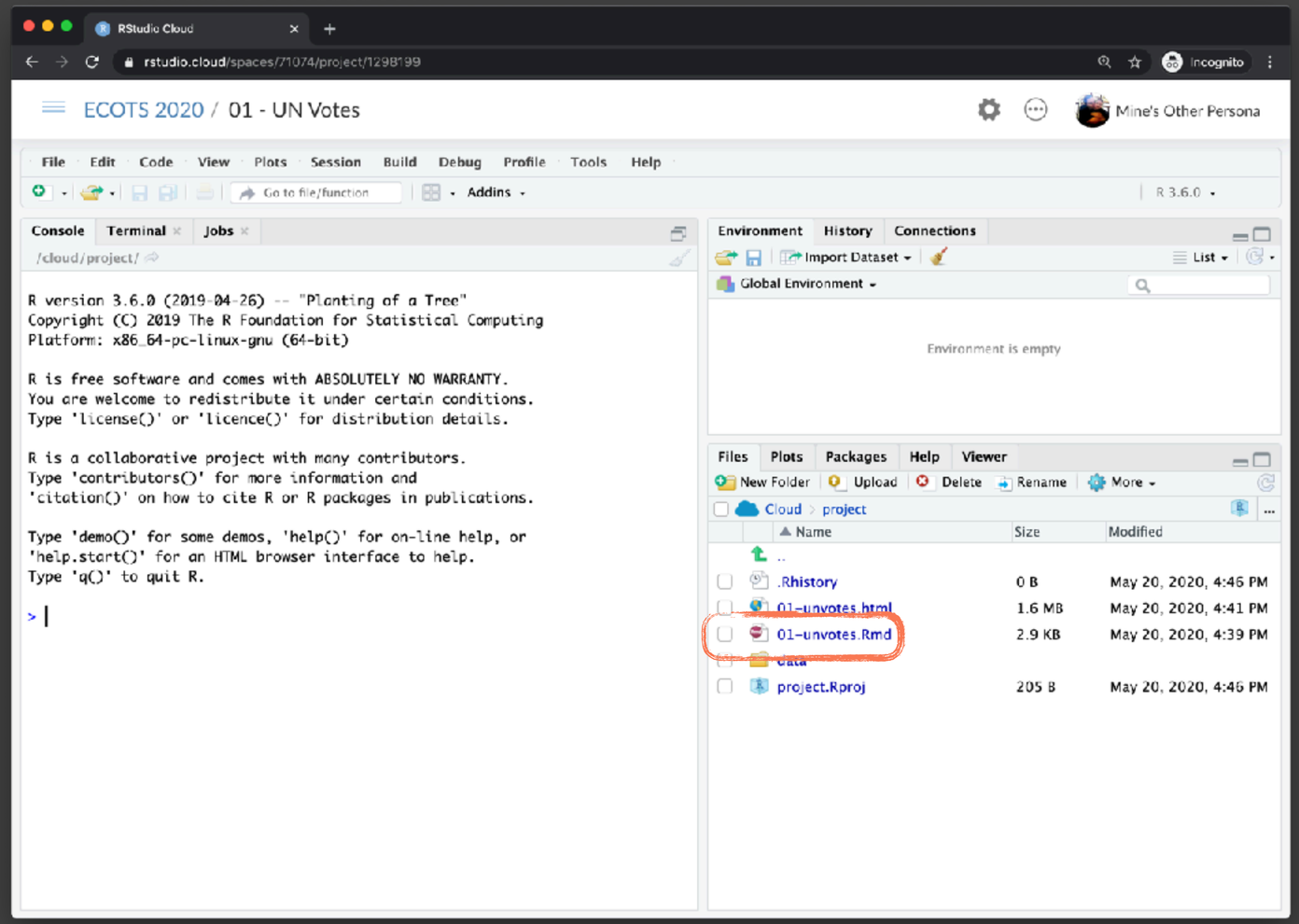
- Go to **RStudio Cloud**
- Start the project titled UN Votes

🔗 **rstd.io/dsbox-cloud**



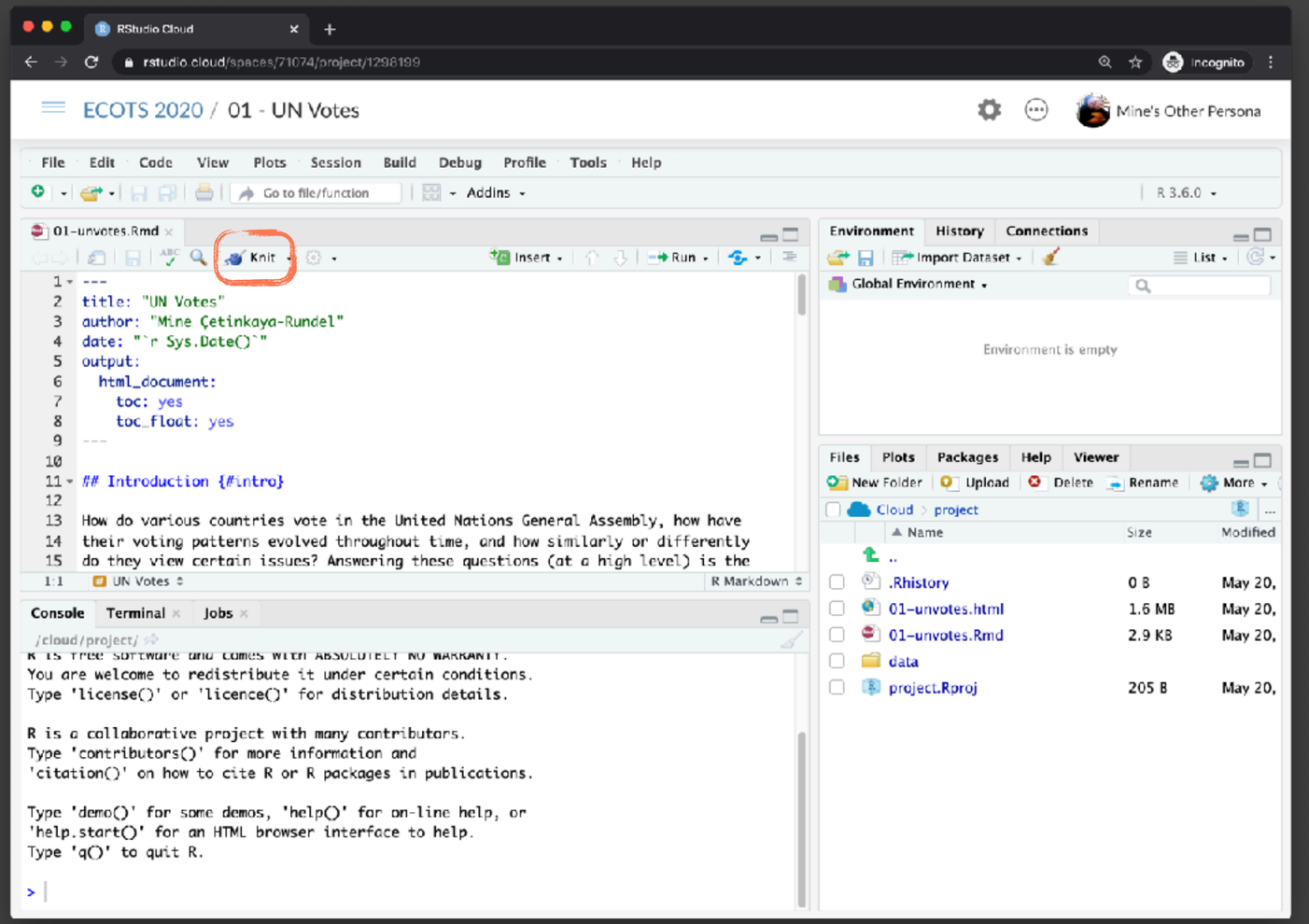
- Go to **RStudio Cloud**
- Start the project titled UN Votes
- Open the R Markdown document called `unvotes.Rmd`

🔗 **rstd.io/dsbox-cloud**



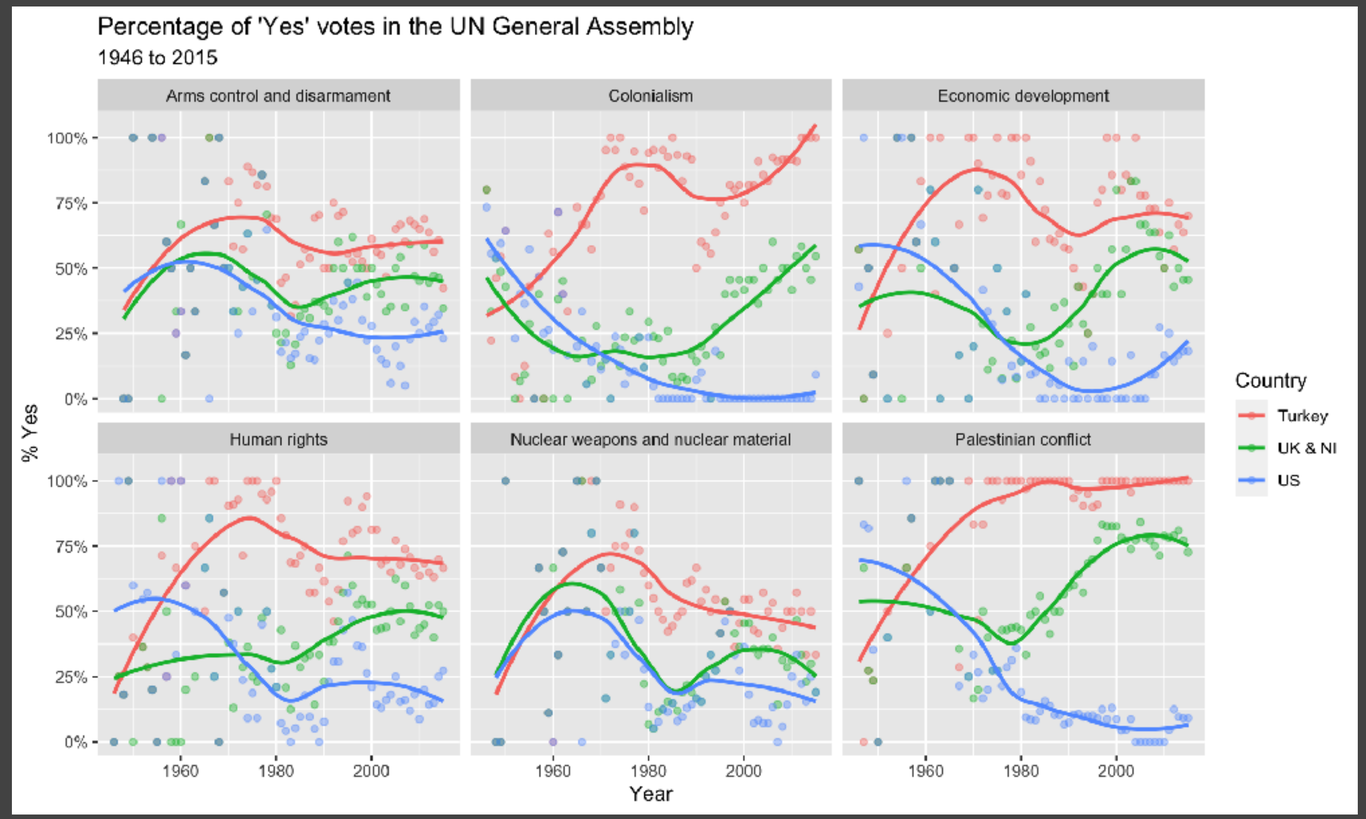
- Go to **RStudio Cloud**
- Start the project titled UN Votes
- Open the R Markdown document called `unvotes.Rmd`
- Knit the document and review the data visualisation you just produced

🔗 **rstd.io/dsbox-cloud**



- Go to **RStudio Cloud**
- 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

🔗 **rstd.io/dsbox-cloud**



takeaways

seamless onboarding to computing
(with RStudio Cloud or other server-based access)

data visualisation on day one

hands on computing on day one

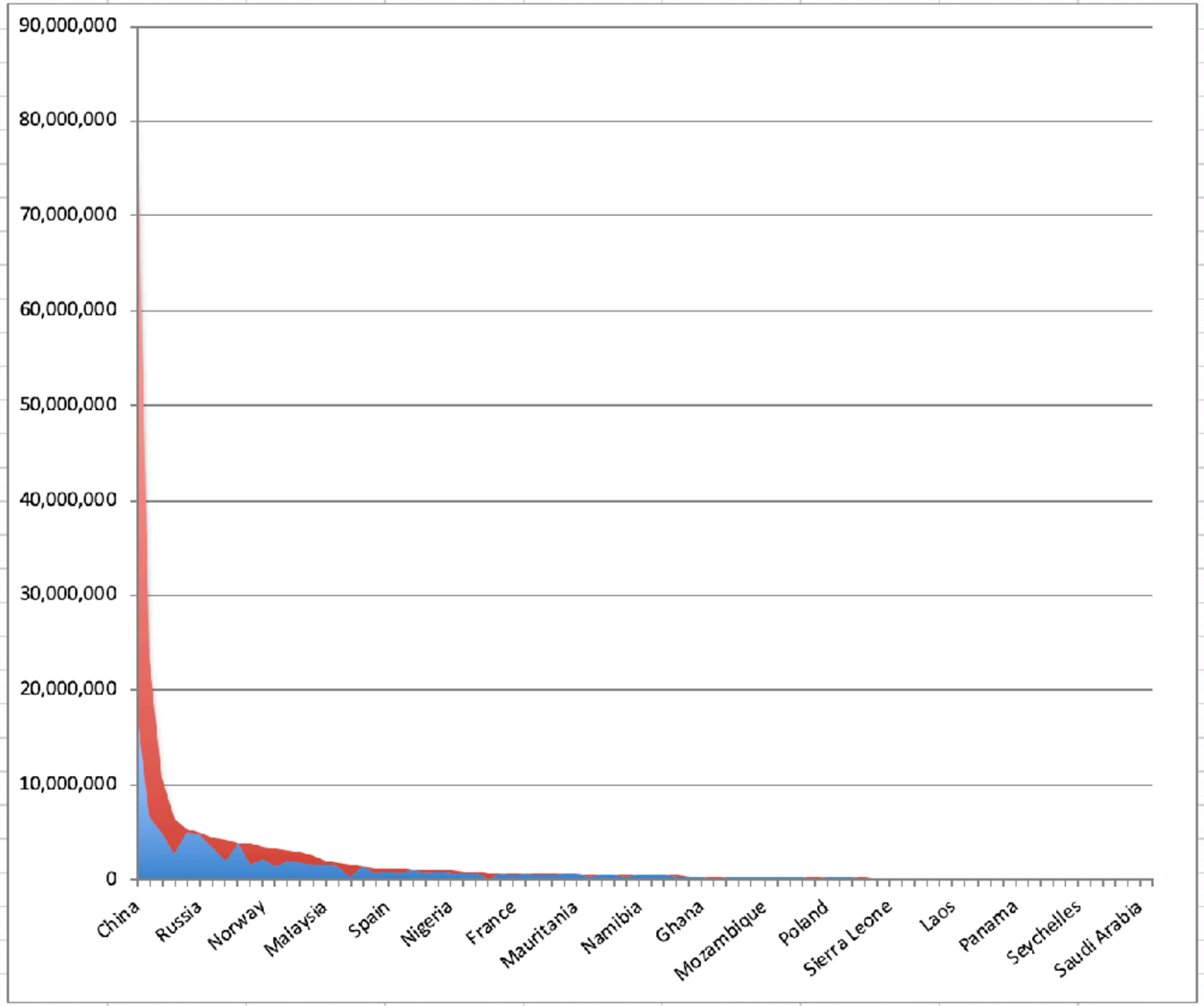
resources

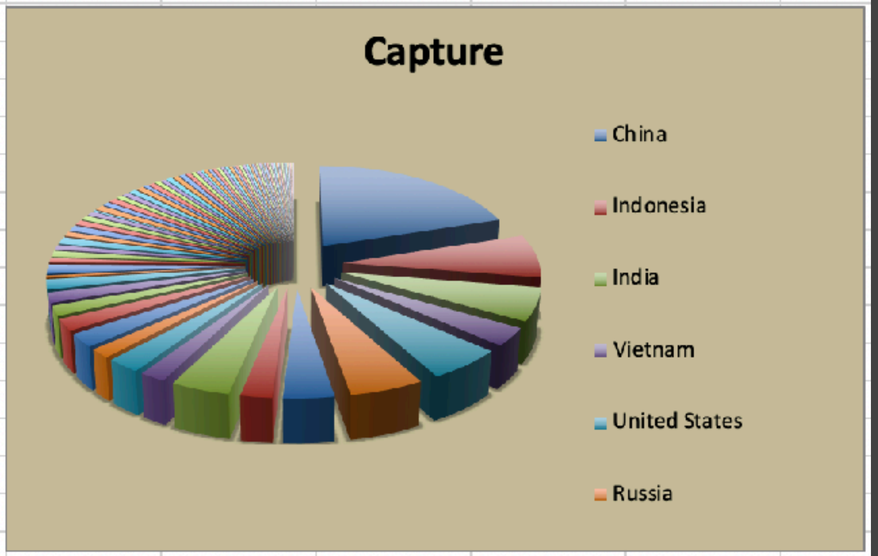
Data Science in a Box - Hello World
datasciencebox.org/hello-world.html

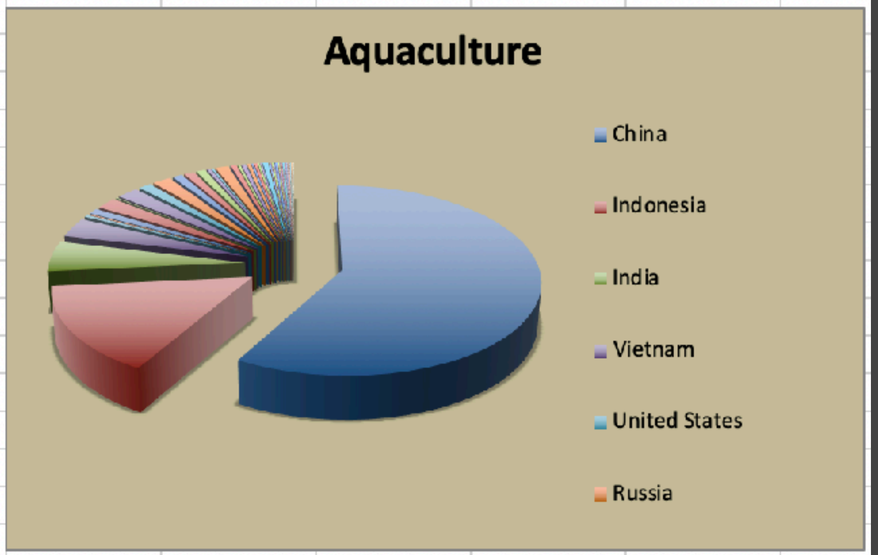

Hello world
Exploring data
Visualize
Import
Wrangle
Data science ethics
Misrepre-sentation
Algorithmic bias
Data privacy
Making rigorous conclu-sions
Model
Infer
Predict
Looking further

ex. 2

fisheries of the world


90,000,000
80,000,000
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
China
Russia
Norway
Malaysia
Spain
Nigeria
France
Mauritania
Namibia
Ghana
Mozambique
Poland
Sierra Leone
Laos
Panama
Seychelles
Saudi Arabia


Capture
China
Indonesia
India
Vietnam
United States
Russia


Aquaculture
China
Indonesia
India
Vietnam
United States
Russia

```
fisheries %>% select(country)
#> # A tibble: 75 x 1
#>    country
#>    <chr>
#>  1 Algeria
#>  2 Angola
#>  3 Argentina
#>  4 Australia
#>  5 Bangladesh
#>  6 Brazil
#>  7 Cambodia
#>  8 Canada
#>  9 Chile
#> 10 Colombia
#> # … with 65 more rows
```

✓ data joins

continents

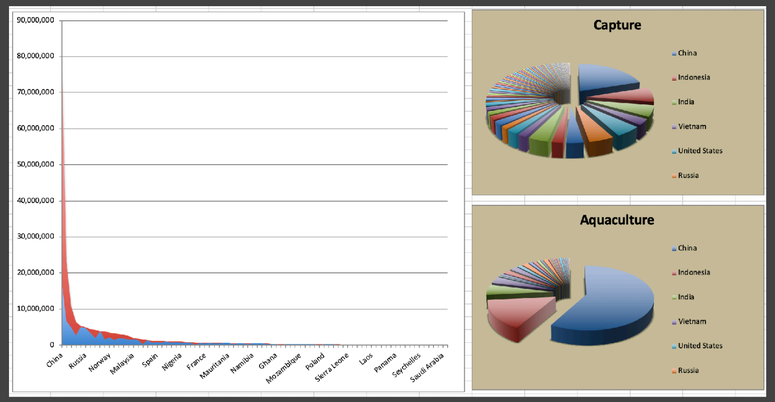
```
#> # A tibble: 245 x 2
#>    country            continent
#>    <chr>              <chr>
#>  1 Afghanistan        Asia
#>  2 Åland Islands      Europe
#>  3 Albania            Europe
#>  4 Algeria            Africa
#>  5 American Samoa     Oceania
#>  6 Andorra            Europe
#>  7 Angola             Africa
#>  8 Anguilla           Americas
#>  9 Antigua & Barbuda  Americas
#> 10 Argentina          Americas
#> # … with 235 more rows
```

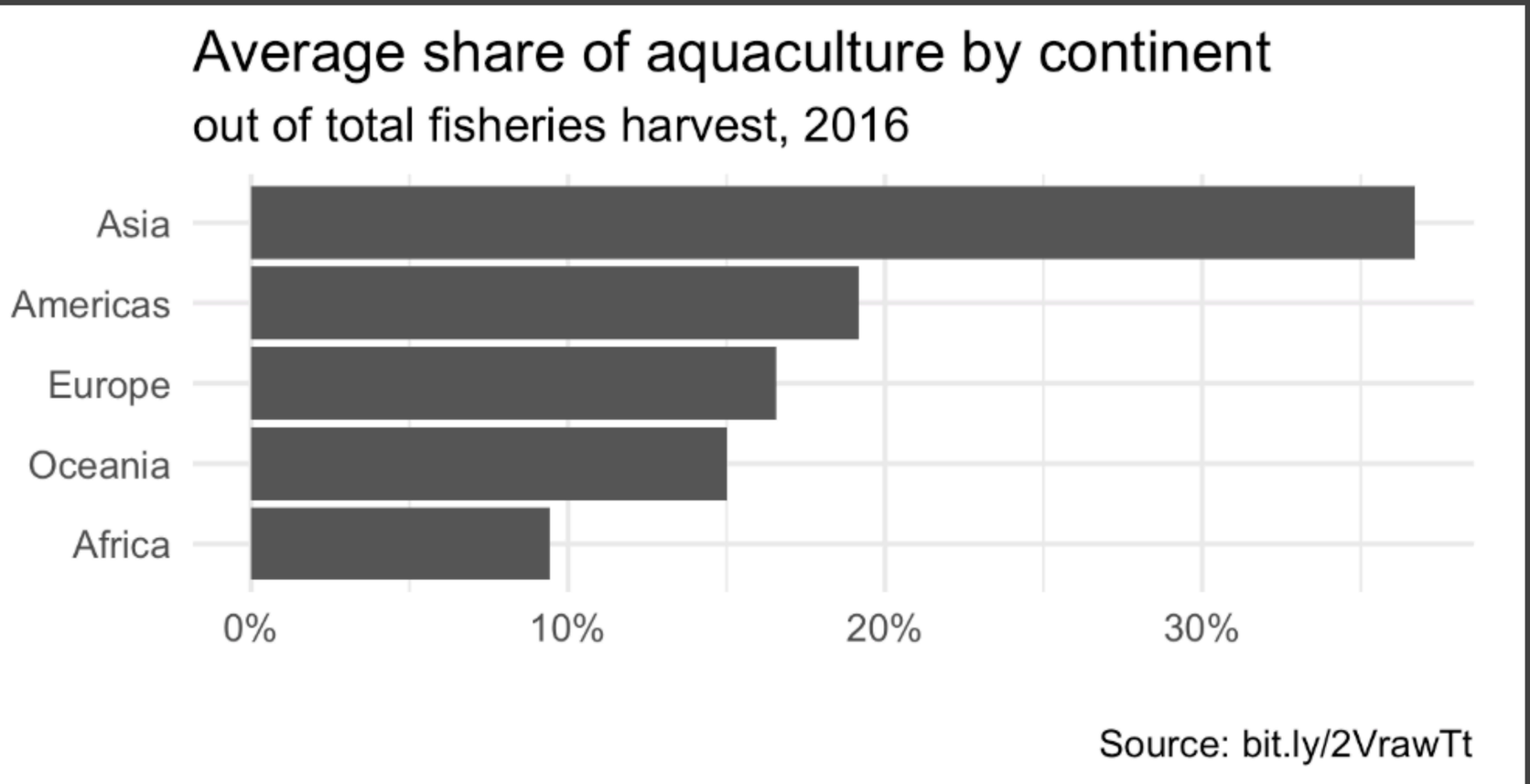
```
fisheries <- left_join(fisheries, continents)
Joining, by = “country"
```

✓ data joins

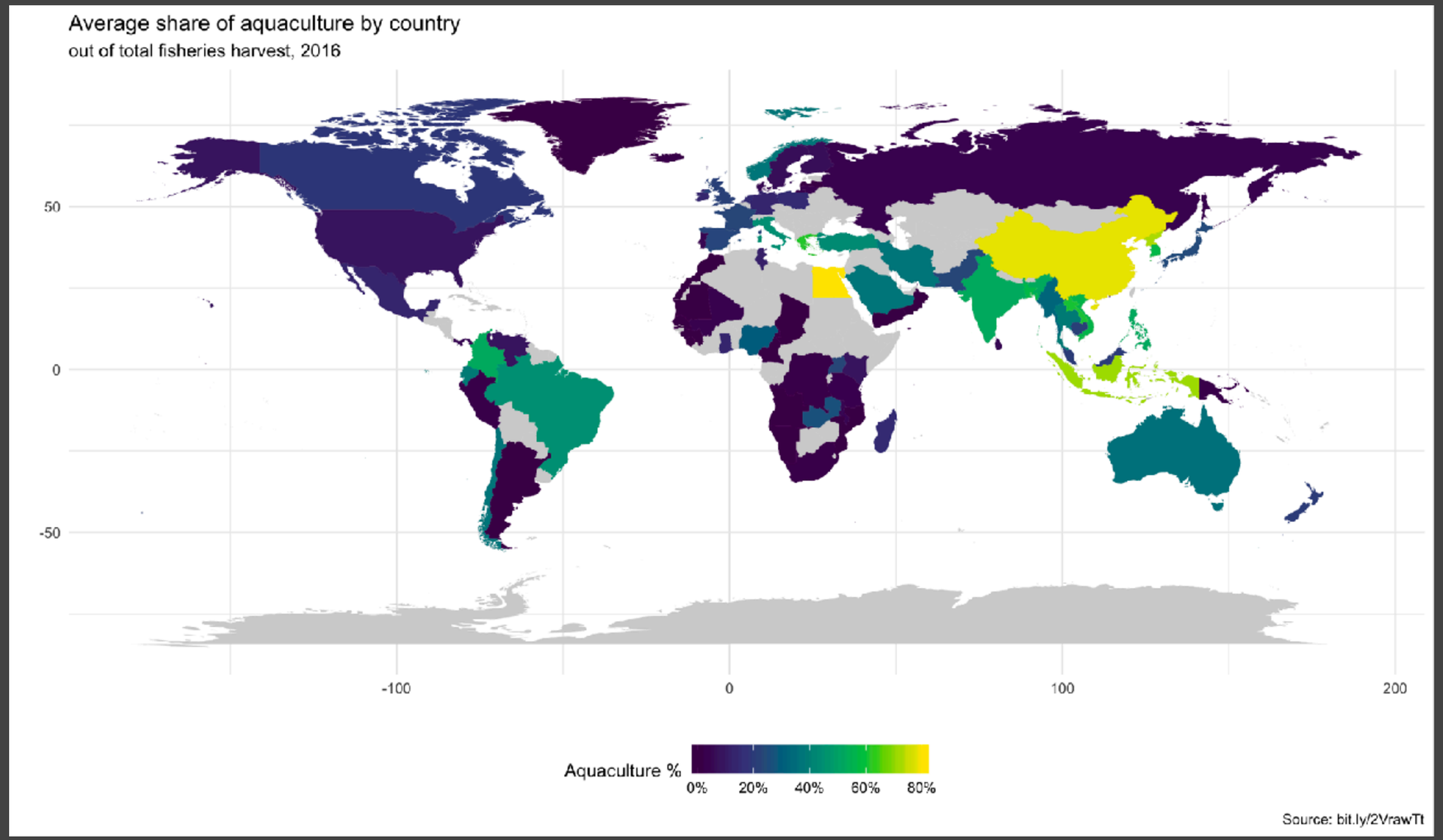
✓ ethics

```
fisheries %>%
  filter(is.na(continent))#> # A tibble: 75 x 1
#> # A tibble: 5 x 4
#>   country                                capture aquaculture continent
#>   <chr>                                    <dbl>       <dbl> <chr>
#> 1 Congo, Democratic Republic of the  220000        2965 NA
#> 2 Hong Kong                           161964        4130 NA
#> 3 Myanmar                            1742956      474510 NA
#> 4 Other                              9685851      786993 NA
#> 5 Taiwan (Republic of China)         1017243      304756 NA
```





- ✓ data joins
- ✓ ethics
- ✓ critique
- ✓ improving visualisations



- ✓ data joins
- ✓ ethics
- ✓ critique
- ✓ improving
- ✓ visualisations
- ✓ mapping

takeaways

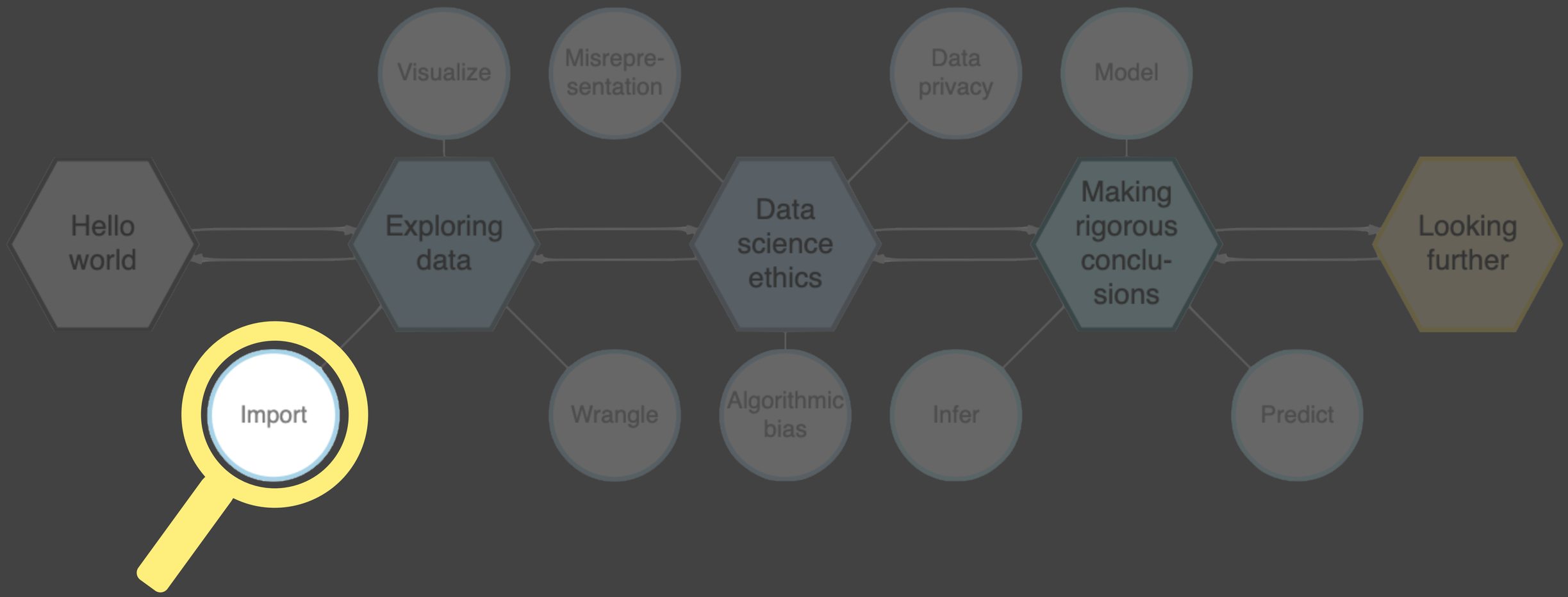
critque as a motivator for improvement

resources

Lab: Take a sad plot and make it better
rstudio-education.github.io/datascience-box/course-materials/lab-instructions/lab-06/lab-06-sad-plots.html

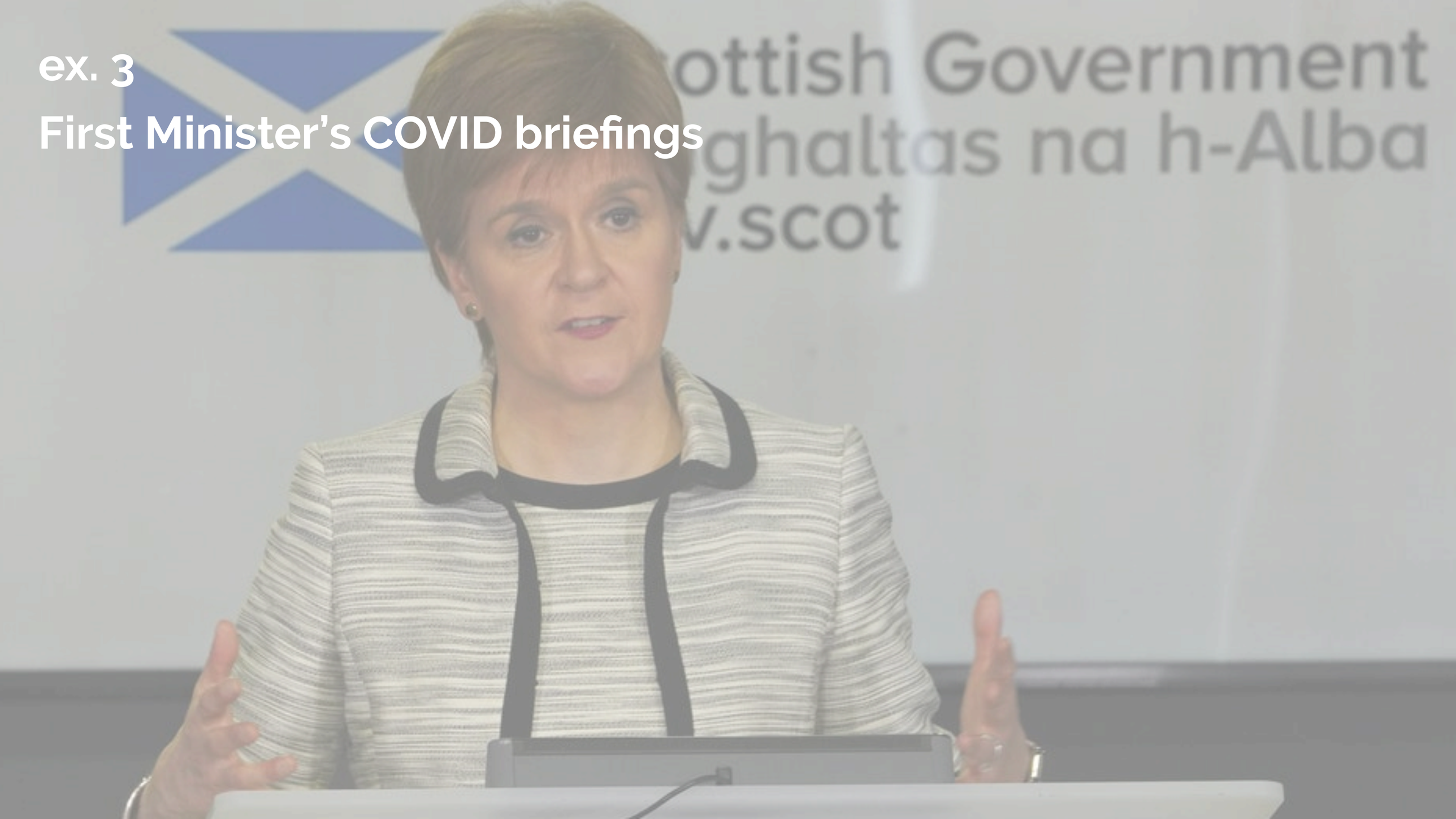
CHANCE column: **From drab to fab** (Mine Çetinkaya-Rundel & Maria Tackett)

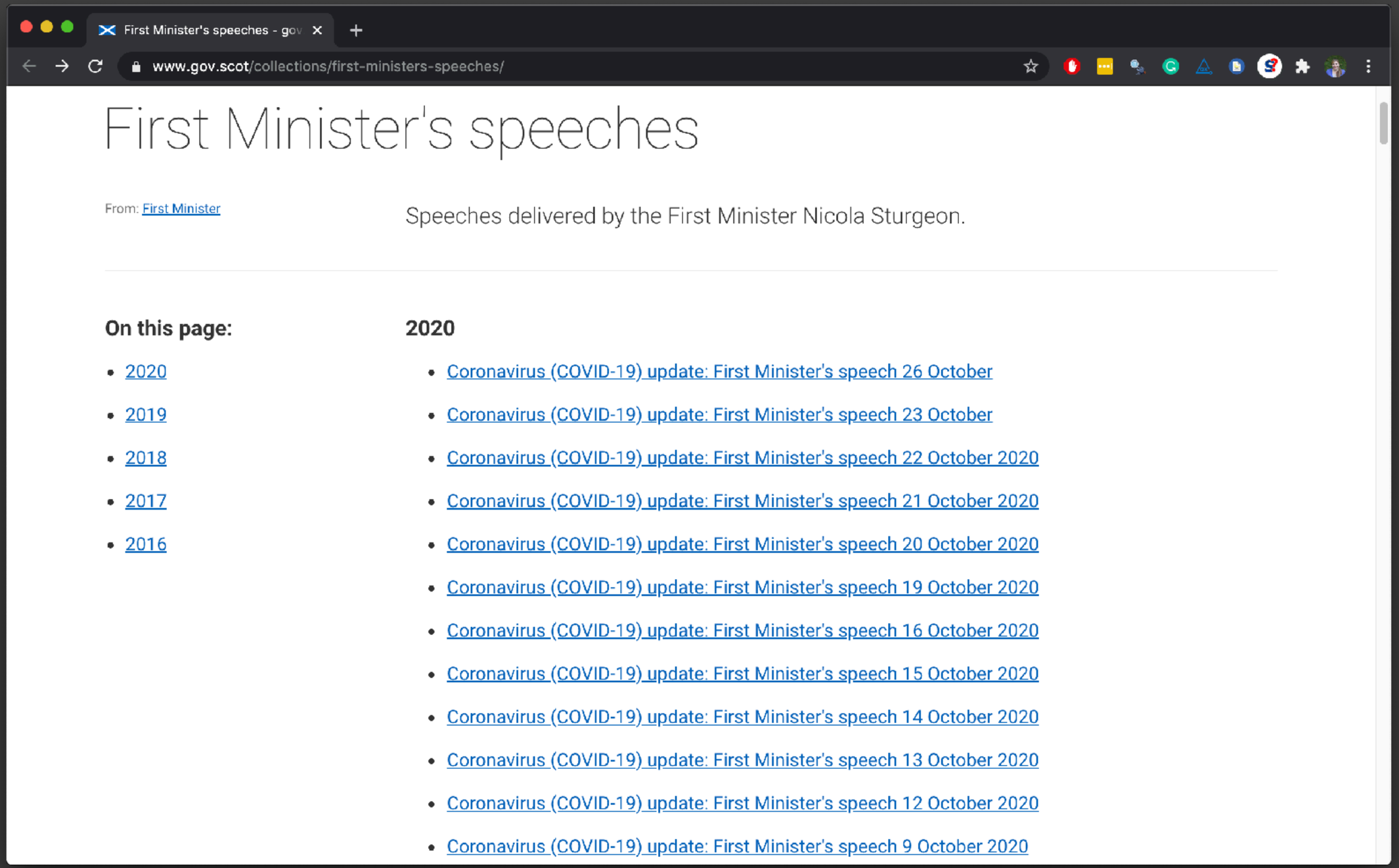
Talks:
- **Take a Sad Plot and Make it Better** (Alison Hill)
- **Tidy up your data science workflow with the tidyverse** (Mine Çetinkaya-Rundel)


Hello world
Exploring data
Visualize
Import
Wrangle
Data science ethics
Misrepre-sentation
Algorithmic bias
Data privacy
Making rigorous conclu-sions
Model
Infer
Predict
Looking further

ex. 3

First Minister’s COVID briefings





First Minister's speeches

From: First Minister

Speeches delivered by the First Minister Nicola Sturgeon.

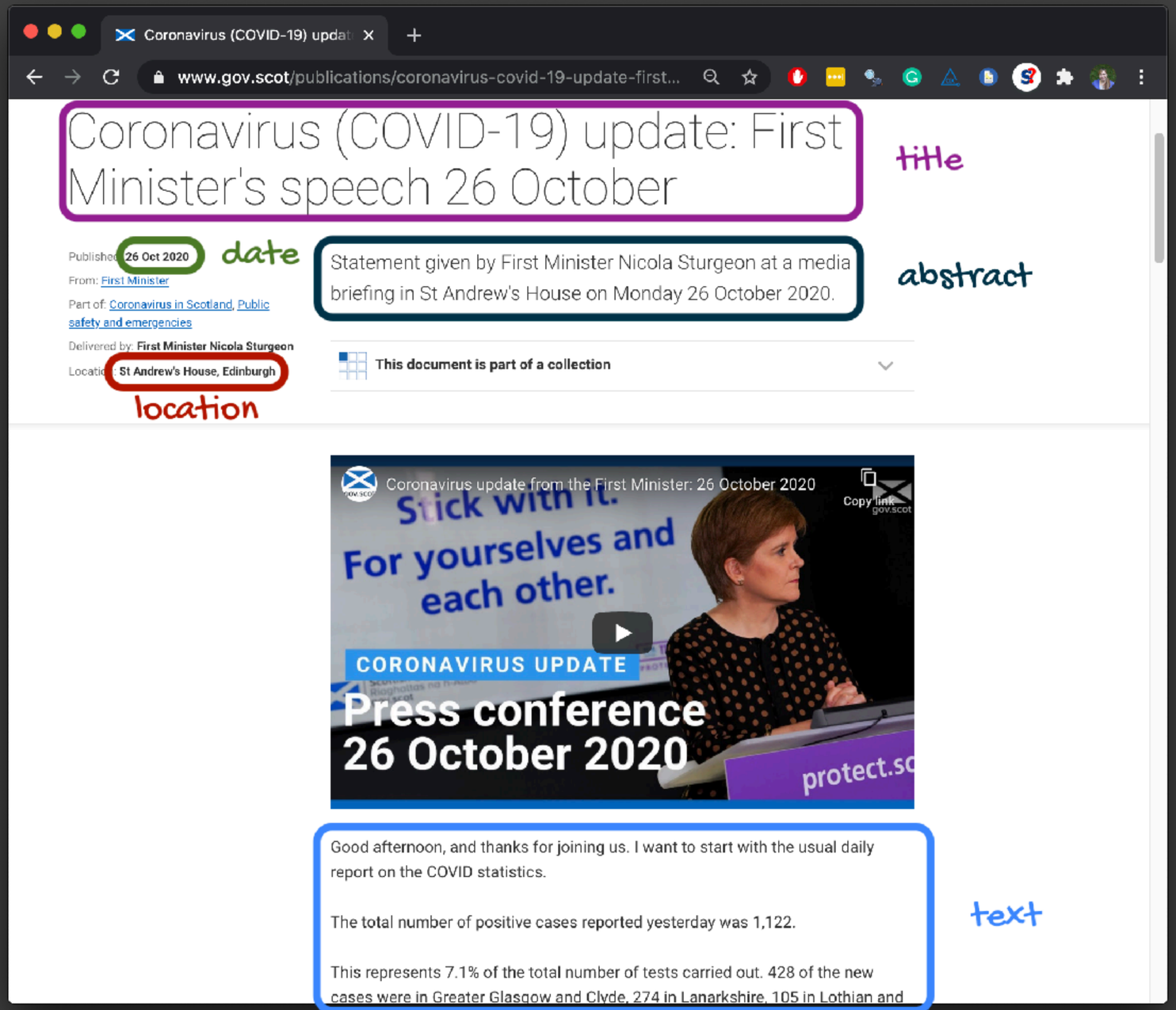
On this page:

2020

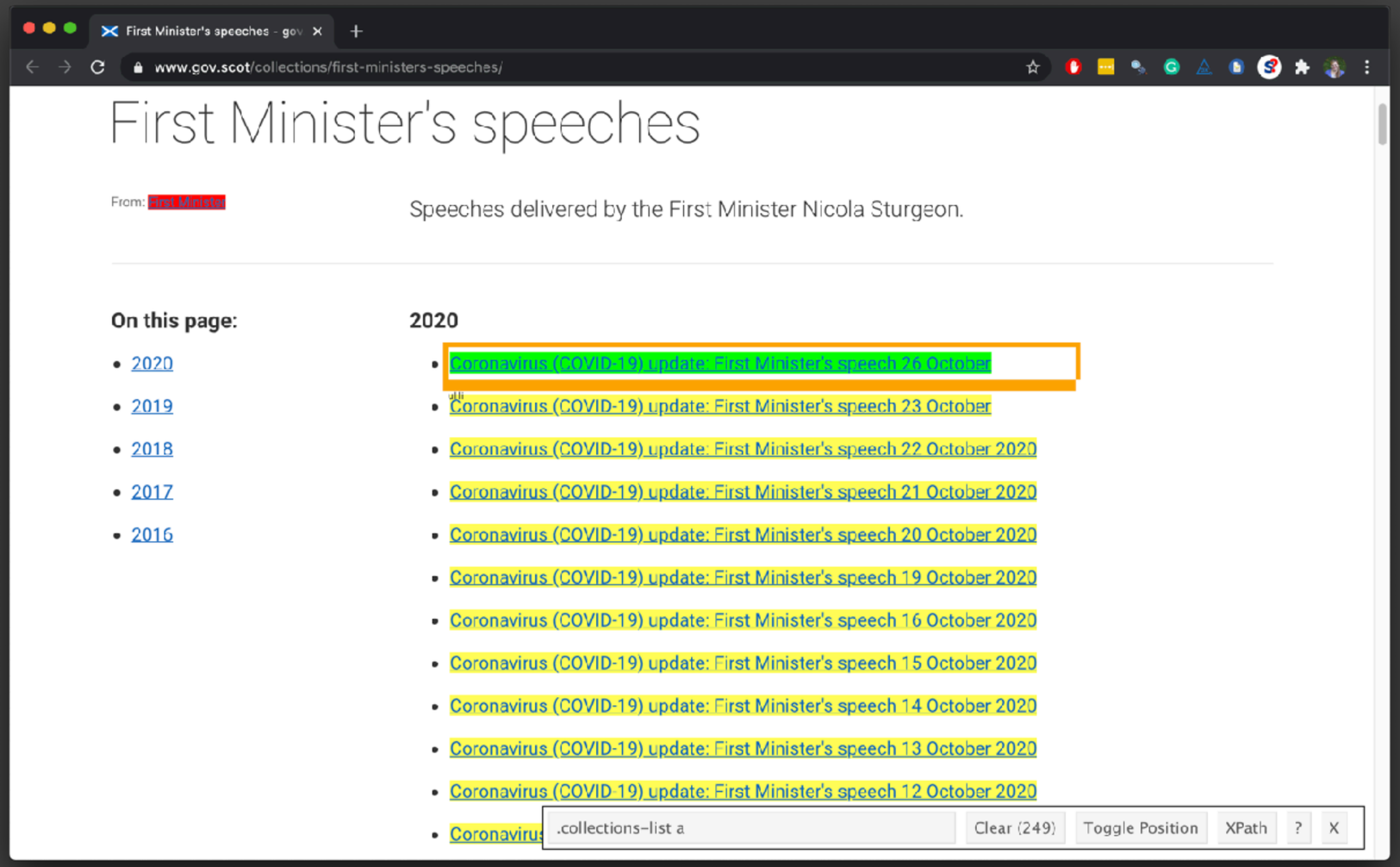
- Coronavirus (COVID-19) update: First Minister's speech 26 October
- Coronavirus (COVID-19) update: First Minister's speech 23 October
- Coronavirus (COVID-19) update: First Minister's speech 22 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 21 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 20 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 19 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 16 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 15 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 14 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 13 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 12 October 2020
- Coronavirus (COVID-19) update: First Minister's speech 9 October 2020

✓ ethics

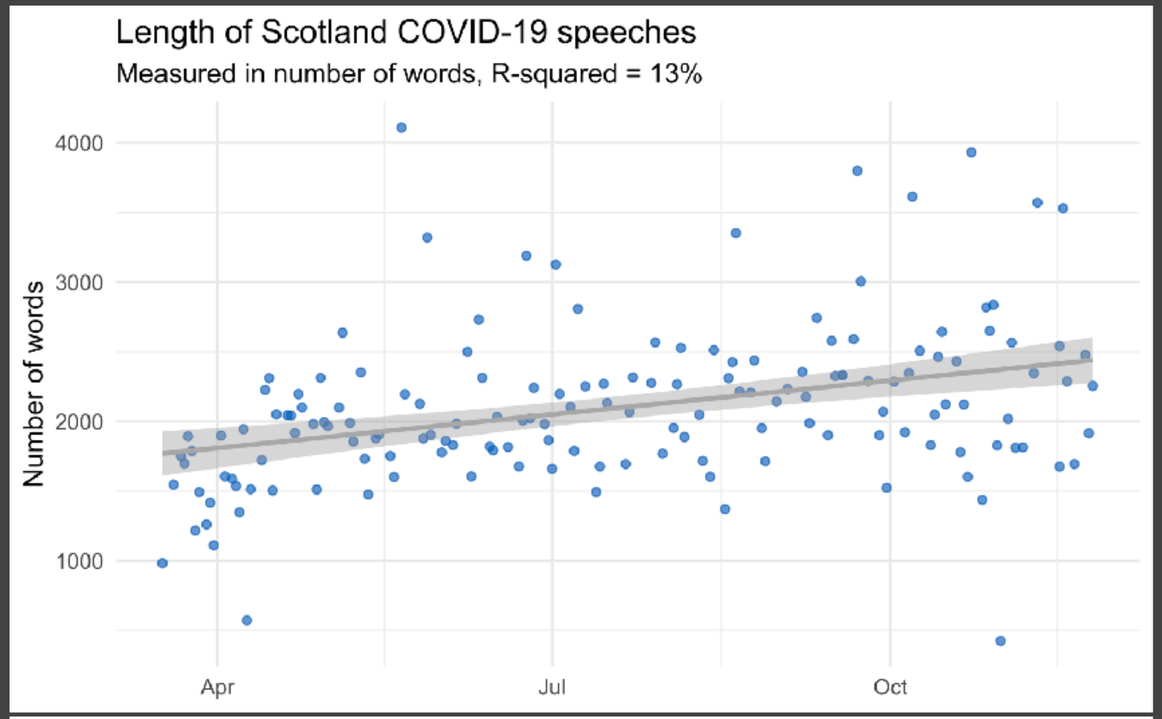
```
robotstxt::paths_allowed("https://www.gov.scot/")
 www.gov.scot

[1] TRUE
```

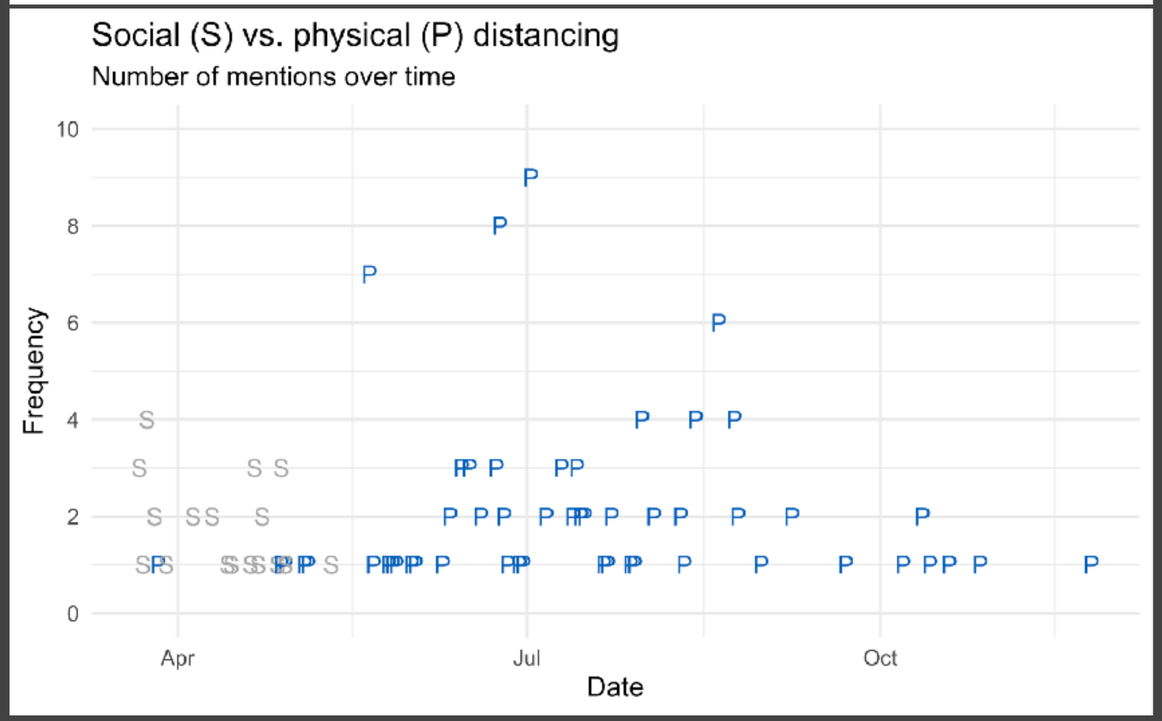


- ✓ ethics
- ✓ web scraping
- ✓ text parsing
- ✓ data types
- ✓ regular expressions

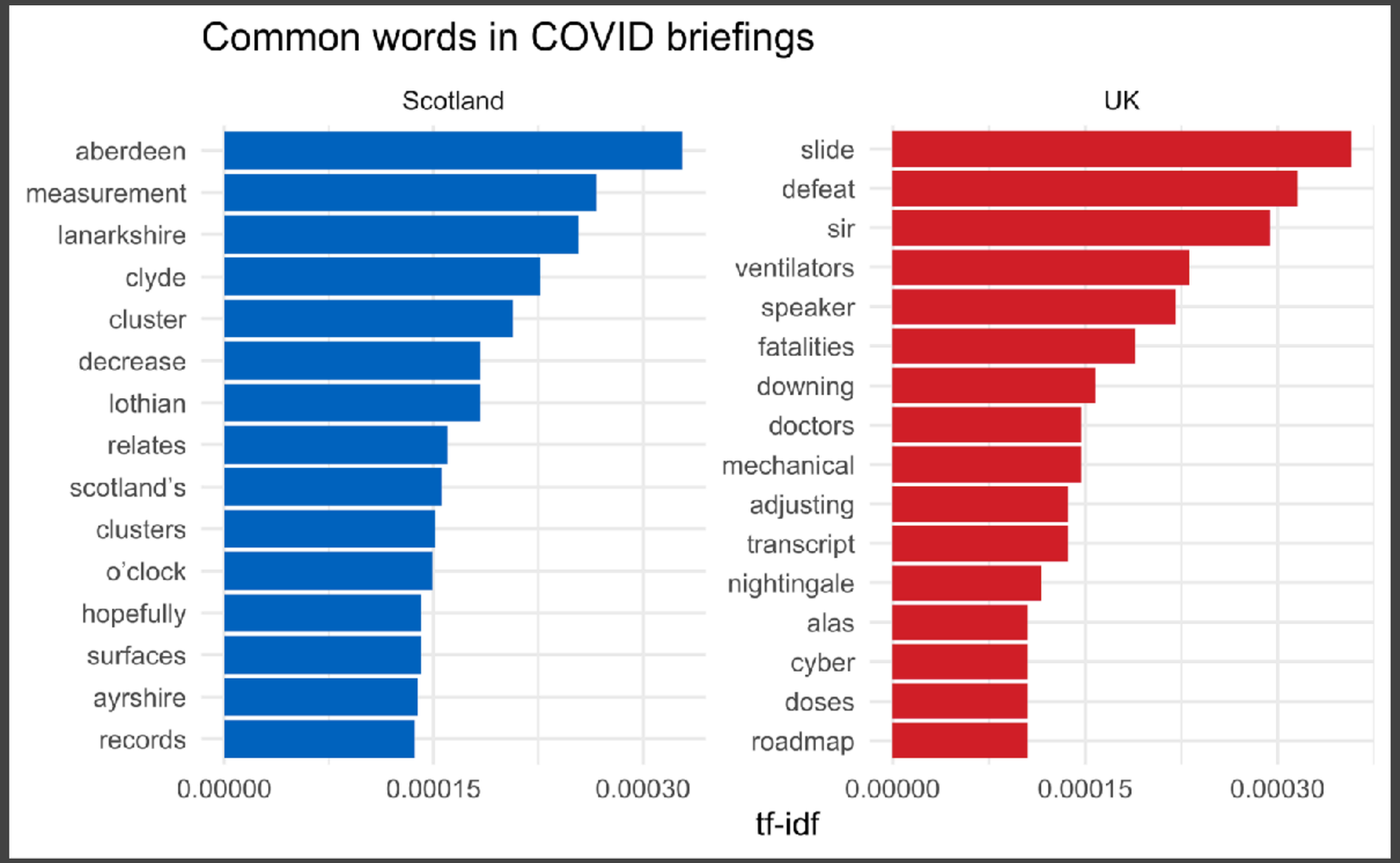


- ✓ ethics
- ✓ web scraping
- ✓ text parsing
- ✓ data types
- ✓ regular expressions
- ✓ functions
- ✓ iteration





- ✓ ethics
- ✓ web scraping
- ✓ text parsing
- ✓ data types
- ✓ regular expressions
- ✓ functions
- ✓ iteration
- ✓ visualisation
- ✓ interpretation



- ✓ ethics
- ✓ web scraping
- ✓ text parsing
- ✓ data types
- ✓ regular expressions
- ✓ functions
- ✓ iteration
- ✓ visualisation
- ✓ interpretation
- ✓ text analysis

takeaways

web scraping to motivate DRY
(do not repeat yourself)

resources

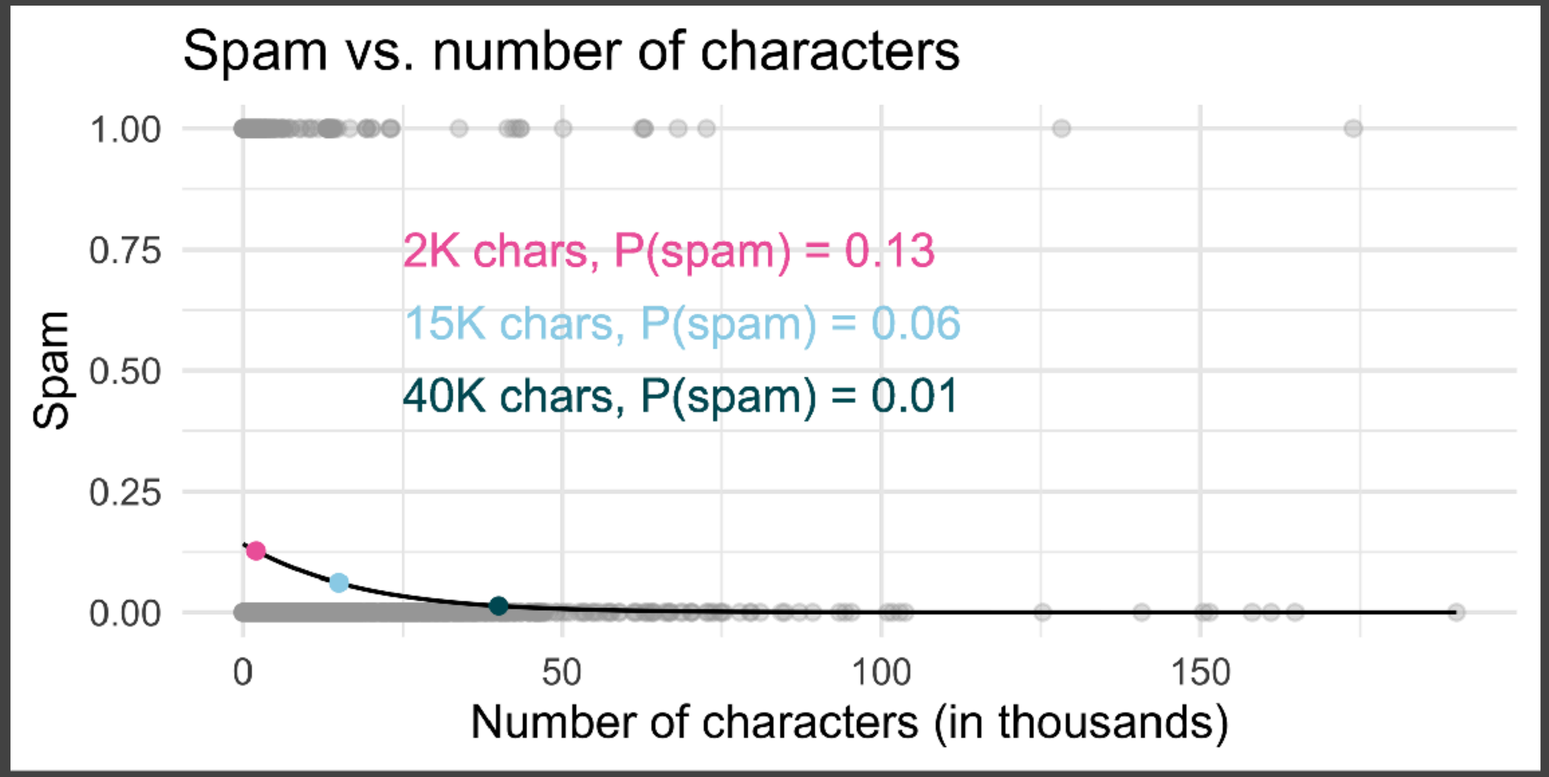
Lessons: Web scraping and programming
datasciencebox.org/exploring-data.html#web-scraping-and-programming

Lab: University of Edinburgh Art Collection
rstudio-education.github.io/datascience-box/course-materials/lab-instructions/lab-08/lab-08-uoe-art.html

Paper: Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities (Dogucu & Çetinkaya-Rundel, 2021)
doi.org/10.1080/10691898.2020.1787116


Hello world
Exploring data
Visualize
Import
Wrangle
Data science ethics
Misrepre-sentation
Algorithmic bias
Data privacy
Making rigorous conclu-sions
Model
Infer
Predict
Looking further

ex. 3
spam filters



- ✓ logistic regression
- ✓ prediction

	Email is spam	Email is not spam
Email labelled spam	True positive	False positive (Type 1 error)
Email labelled not spam	False negative (Type 2 error)	True negative

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

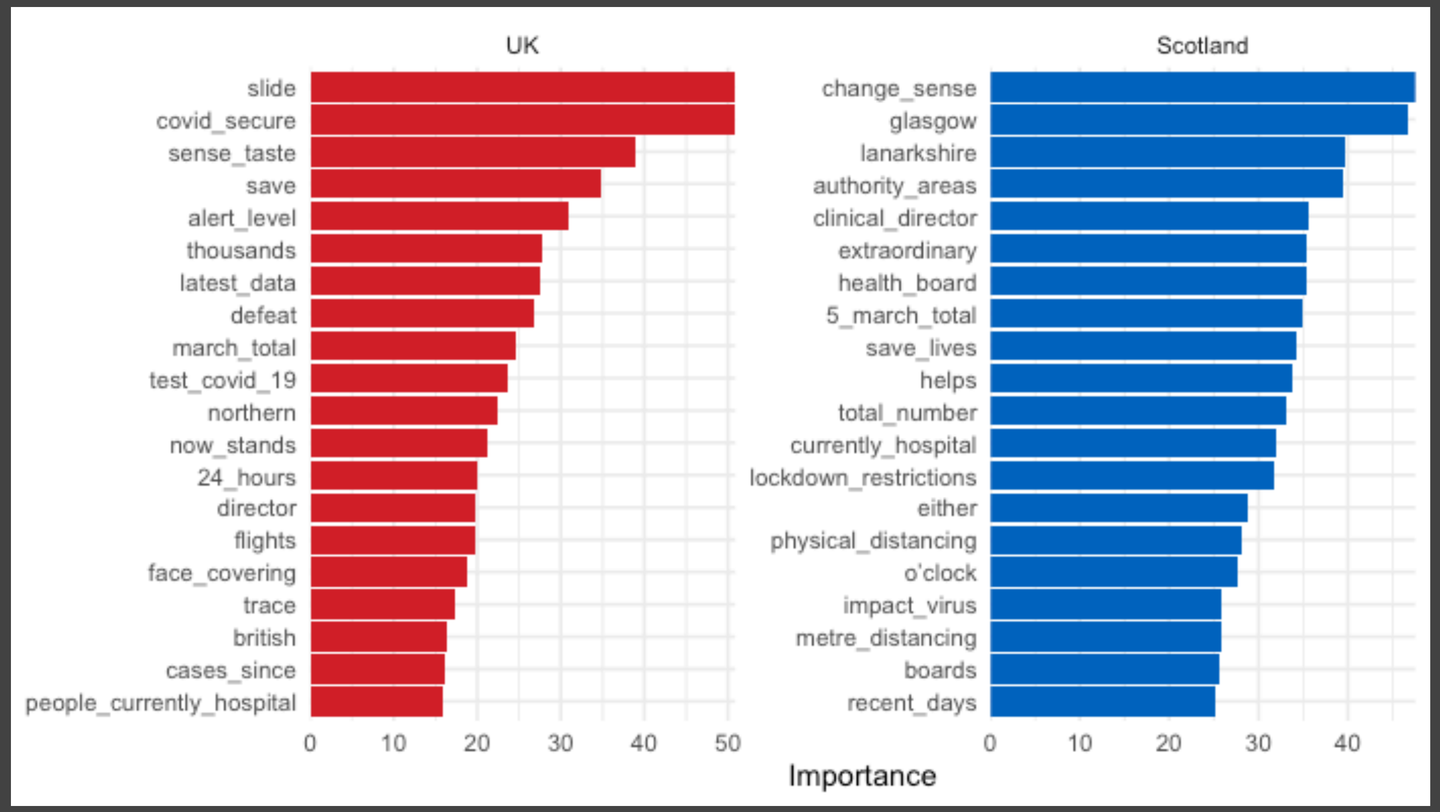
takeaways

decision tables to motivate intuition around loss functions

resources

Lessons: Classification and model building
datasciencebox.org/making-rigorous-conclusions.html#classification-and-model-building


Hello world
Exploring data
Visualize
Import
Wrangle
Data science ethics
Misrepre-sentation
Algorithmic bias
Data privacy
Making rigorous conclu-sions
Model
Infer
Predict
Looking further



✓ machine learning for text data

✓ repetition

Road Traffic Accidents

minecr.shinyapps.io/dsbox-02-accidents/#section-accident-severity

Road Traffic Accidents

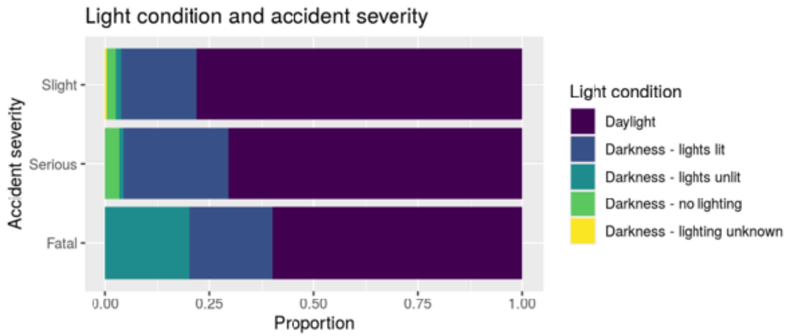
- Introduction
- Data
- Multi-vehicle accidents
- Speed limits
- Accident severity
- Wrap up

Start Over

Accident severity

Visualizing

Recreate the following plot. To match the colors, you can use `scale_fill_viridis_d()`.



R code | Start Over | Hints | Run Code | Submit Answer

```
1 ggplot(data = ___, aes(x = ___, ___ = ___)) +
2   geom____(___) +
3   ___() +
4   ___(y = ___, x = ___,
5        ___ = ___,
6        title = ___)
```

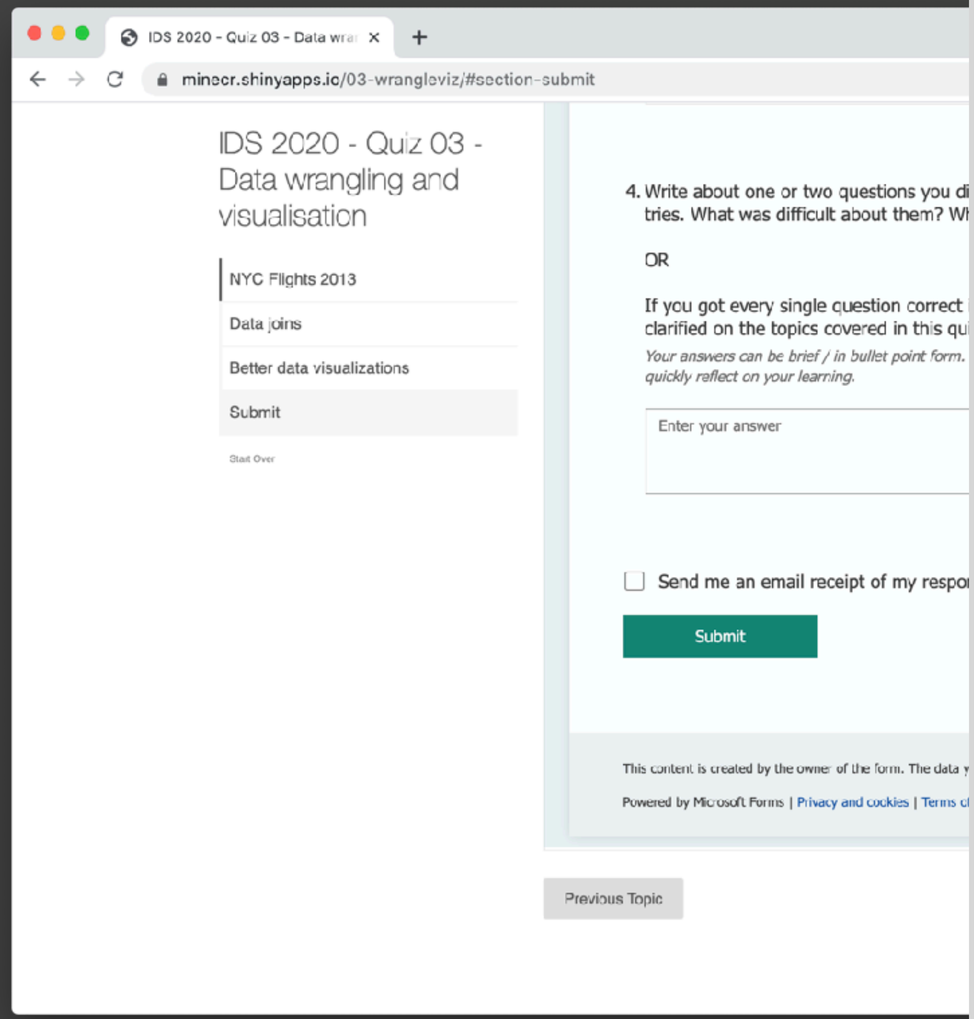
Which of the following are true? Check all that apply.

- [ ] Most accidents occur in daylight
- [ ] Roughly 20 percent of serious accidents occurred in the darkness without lighting
- [ ] Crashes in the darkness tend to be more severe
- [ ] Fatal crashes have the highest proportion of crashes in the darkness where the lights are lit
- [ ] Most slight accidents in the darkness happen without lighting.

Submit Answer

Continue

tips




```
# A tibble: 19 x 2
   bigram                 n
   <chr>              <int>
 1 question 7            19
 2 question 8            16
 3 questions 7           12
 4 join function          9
 5 question 2             9
 6 choice questions       7
 7 first question         7
 8 multiple choice        7
 9 correct answer         6
10 necessarily improve    6
11 join functions         5
12 question 1             5
13 7 8                    4
14 airline names          4
15 data frames            4
16 feel like              4
17 many options           4
18 right answer           4
19 x axis                 4
```

✓ repetition

✓ reflection

tips

✓ repetition

✓ reflection

✓ creativity

IDS

introds.org

IDS Timetable Schedule Syllabus

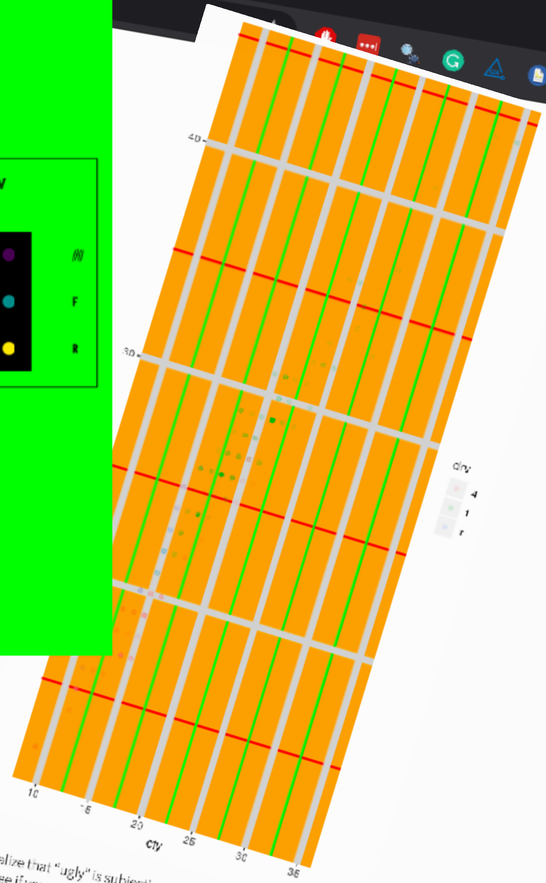
Project

Showcase your

HW 04 - Potpourri

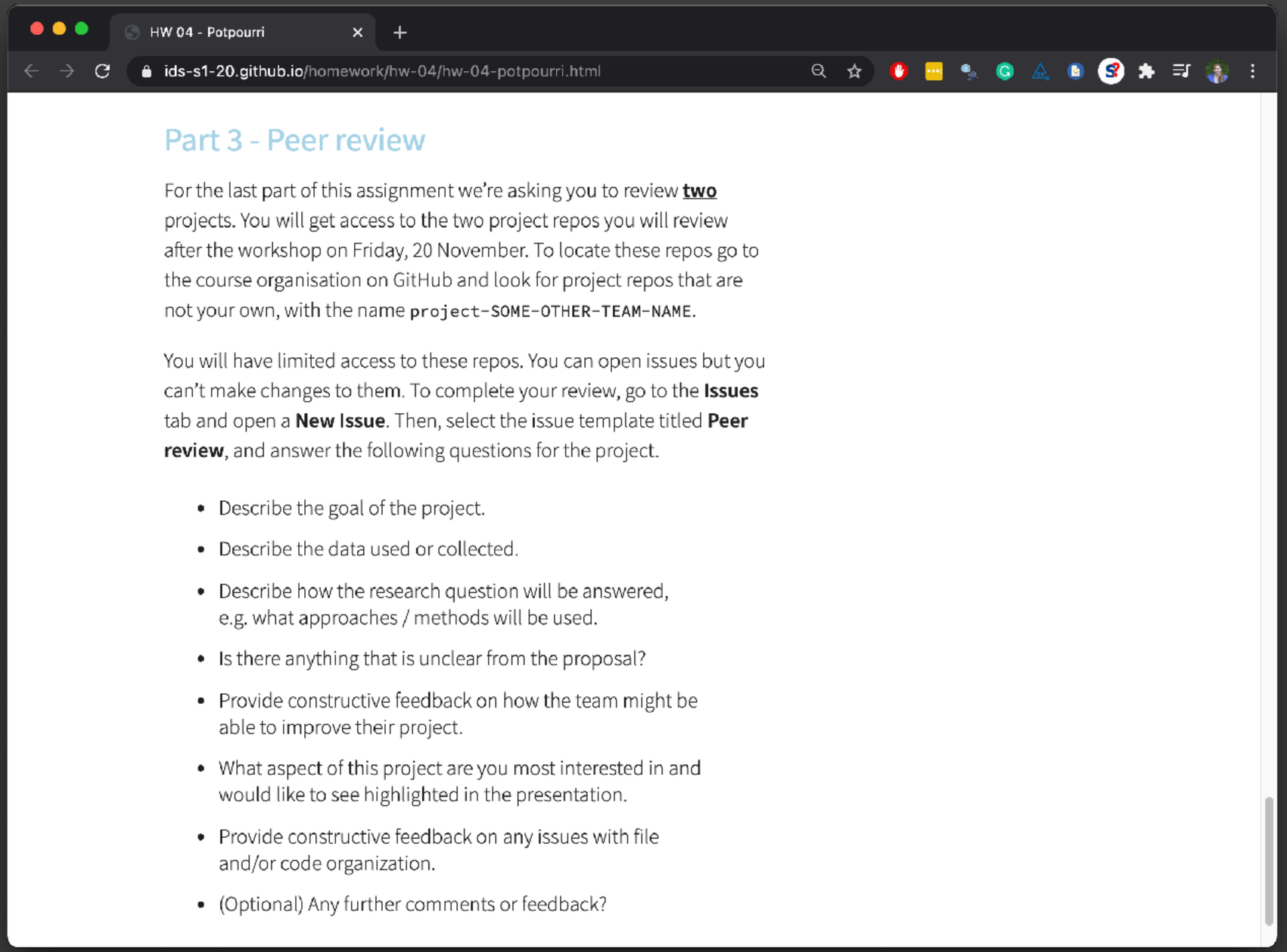
ids-s1-20.github.io/homework/hw-04/hw-04-potpourri

Exercise 1. Make this plot as ugly as possible by changing colours, 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. Try to make it as ugly as possible, the sky is the limit!



I realize that "ugly" is subjective, so we're mostly looking to see if you can figure out how to change the look of a plot using help files of functions you haven't learned before.

tips

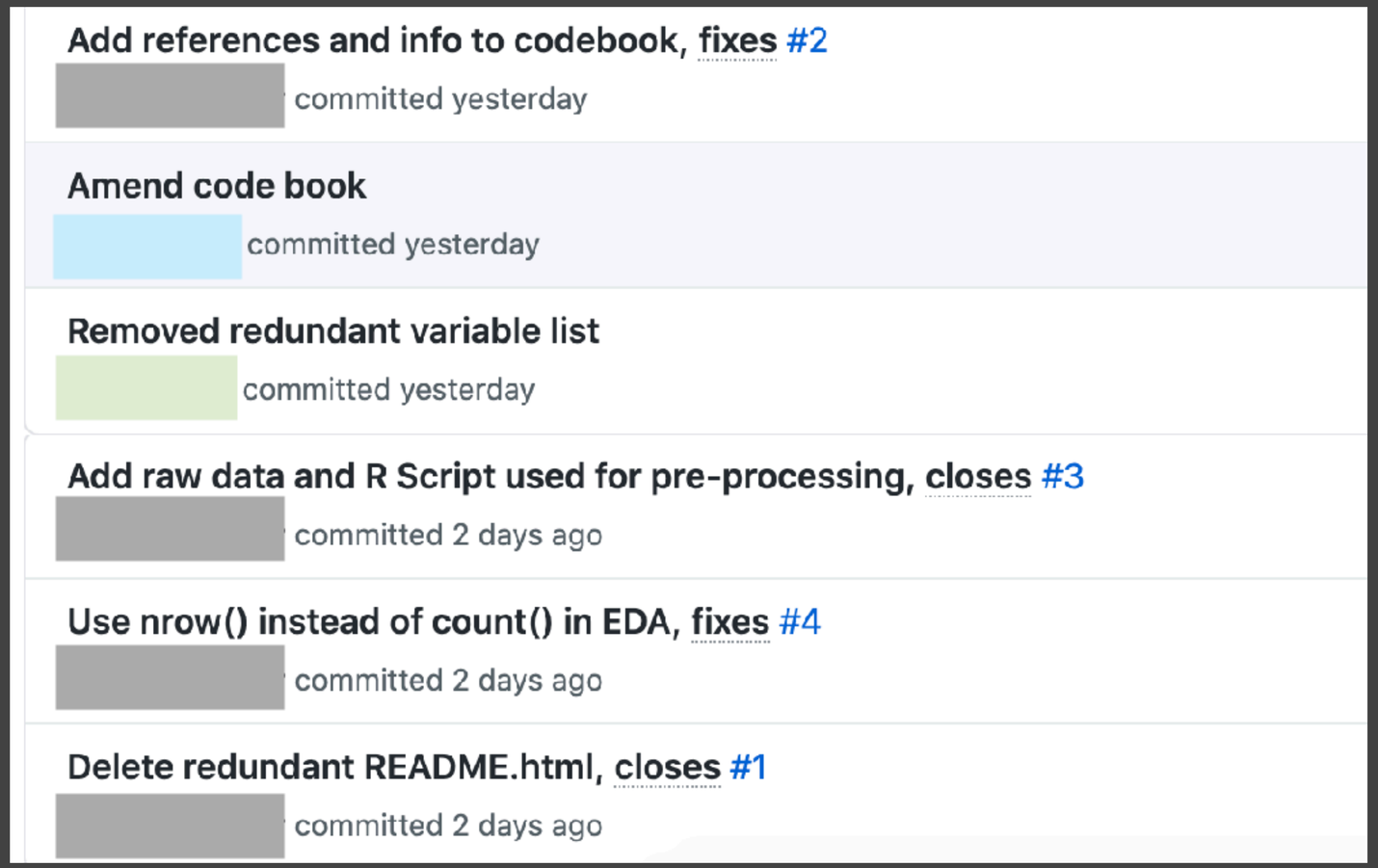


✓ reflection

✓ creativity

✓ peer review

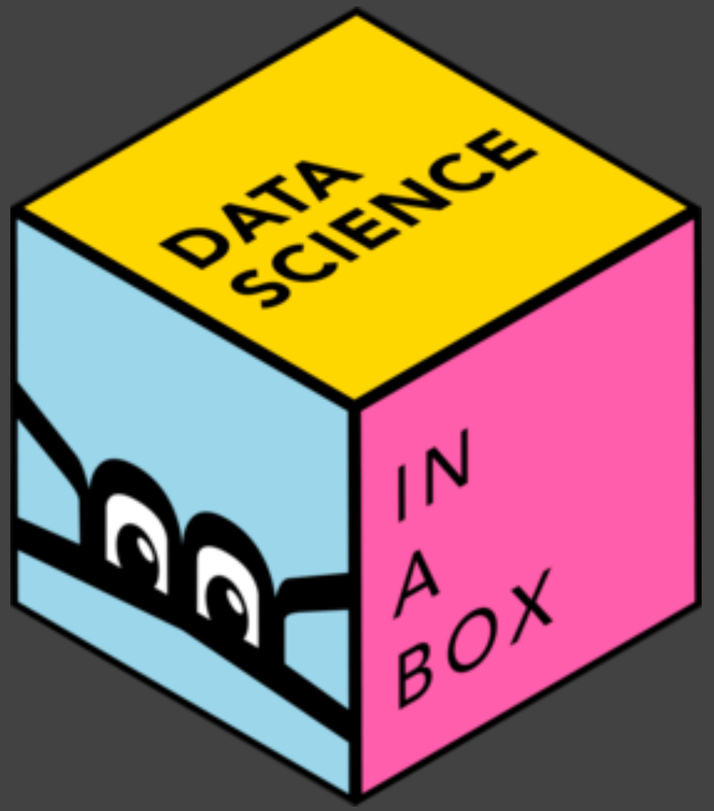
tips



- ✓ repetition
- ✓ reflection
- ✓ creativity
- ✓ peer review
- ✓ real workflows

tips

student
toolbox



instructor
toolbox

Chapter 7 Exploring data | Data

datasciencebox.org/exploring-data.html#slides-videos-and-application-exercises-1

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 Exams

Infrastructure

14 Accessing R
15 Version control
16 Discussion
17 Sharing
18 Alternative setups

Pedagogy

19 Pedagogy

7.1 Slides, videos, and application exercises

7.1.1 Visualising data

Unit 2 - Deck 1: Data and visualisation

Slides
Source
Video

Unit 2 - Deck 2: Visualising data with ggplot2

Slides
Source
Video

Reading:
R4DS :: Chp 3 - Data visualization

Unit 2 - Deck 3: Visualising numerical data

Slides

On this page

View source
Edit this page

datasciencebox.org

JOURNAL OF STATISTICS EDUCATION
2020, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10691898.2020.1804497

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

A Fresh Look at Introductory Data Science

Mine Çetinkaya-Rundel[a,b,c] and Victoria Ellison[b]

[a]School of Mathematics, University of Edinburgh, Edinburgh, UK; [b]Department of Statistical Science, Duke University, Durham, NC; [c]RStudio, Boston, MA

ABSTRACT

The proliferation of vast quantities of available datasets that are large and complex in nature has challenged universities to keep up with the demand for graduates trained in both the statistical and the computational set of skills required to effectively plan, acquire, manage, analyze, and communicate the findings of such data. To keep up with this demand, attracting students early on to data science as well as providing them a solid foray into the field becomes increasingly important. We present a case study of an introductory undergraduate course in data science that is designed to address these needs. Offered at Duke University, this course has no prerequisites and serves a wide audience of aspiring statistics and data science majors as well as humanities, social sciences, and natural sciences students. We discuss the unique set of challenges posed by offering such a course, and in light of these challenges, we present a detailed discussion into the pedagogical design elements, content, structure, computational infrastructure, and the assessment methodology of the course. We also offer a repository containing all teaching materials that are open-source, along with supplementary materials and the R code for reproducing the figures found in the article.

KEYWORDS
Data science curriculum; Data visualization; Exploratory data analysis; Modeling; Reproducibility; R

1. Introduction

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 article describes an introductory data science course that is our (working) answer to these questions.

At its core, 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 provides very brief forays into additional tools and concepts such as interactive visualizations, 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). The course design builds on the three key recommendations from Nolan and Temple Lang (2010): (1) broaden statistical computing to include emerging areas, (2) deepen computational reasoning skills, and (3) combine computational topics with data analysis. The goal of the course is to bring students from zero experience to being able to complete a fully reproducible data science project on a dataset of their choice and answer questions that they care about within the span of a semester.

In Section 2 of this article, we start with a review of the most recent curriculum guidelines for undergraduate programs in data science, statistics, and computer science. In this section, we also present a synopsis of the course content and structure of introductory data science courses at four other institutions with the goal of providing a snapshot of the current state of affairs in undergraduate introductory data science curricula. In Section 3, we outline the overall design goals of the Duke University introductory data science course that is the focus of this article and discuss how this course addresses current undergraduate curriculum guidelines in statistics and data science. In Section 4, we expand on the course content, flow, and pacing, and present examples of case studies from the course. In Section 5, we detail the pedagogical methods employed by this course, specifically addressing how these methods can support a large class with students with a diverse range of previous experiences in statistics and programming. Section 6 presents the computing infrastructure of the course, Section 7 presents the methods of assessment, and finally in Section 8, we provide a synthesis of where this course sits in the landscape of introductory data science curriculum guidelines, future design plans for the course, and opportunities and challenges for faculty wanting to adopt this course.

2. Background and Related Work

An exact characterization of what the field of data science is meant to encompass is still debated. However, in this article,

CONTACT Mine Çetinkaya-Rundel cetinkaya.mine@gmail.com School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, UK.

Mine Çetinkaya-Rundel &
Victoria Ellison (2020)

A Fresh Look at Introductory Data Science

Journal of Statistics Education

DOI: 10.1080/10691898.2020.1804497

IDS

introds.org

IDS Timetable Schedule Syllabus Help Extra credit Project Resources People

Course Schedule

Overview

This is a tentative course schedule. The flow of topics might change slightly depending on how quickly / slowly it feels right to ...

Introduction to Data Science
Last updated on 20 Oct 2020

Week 1 - Welcome to IDS

Get acquainted with the course, the technology, the workflow, and the skills you will acquire throughout the semester.

Introduction to Data Science
Last updated on 5 Oct 2020

Week 2 - Visualizing data

Data visualization and interpretation of graphical information.

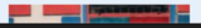
Introduction to Data Science
Last updated on 5 Oct 2020

Week 3 - Wrangling and tidying data

Data wrangling, joining, and tidying.

Introduction to Data Science
Last updated on 15 Oct 2020

Week 4 - Importing and recoding data



🔗 introds.org

🔗 bit.ly/ptt21-intro-ds

🔗 datasciencebox.org

minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com