introductory data science

a fresh look

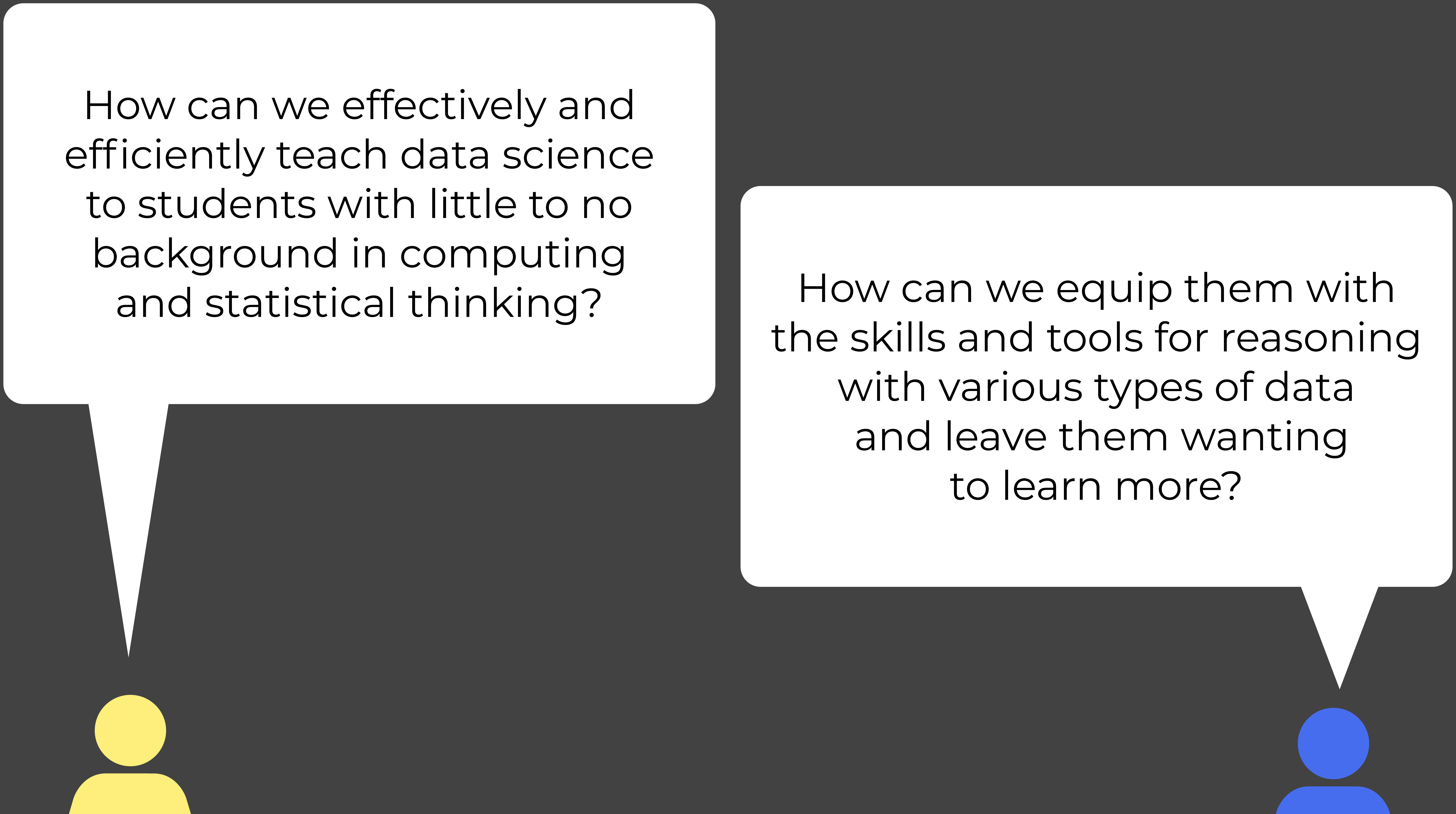
mine çetinkaya-rundel

🔗 bit.ly/fresh-ds-jmm

minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com ✉



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

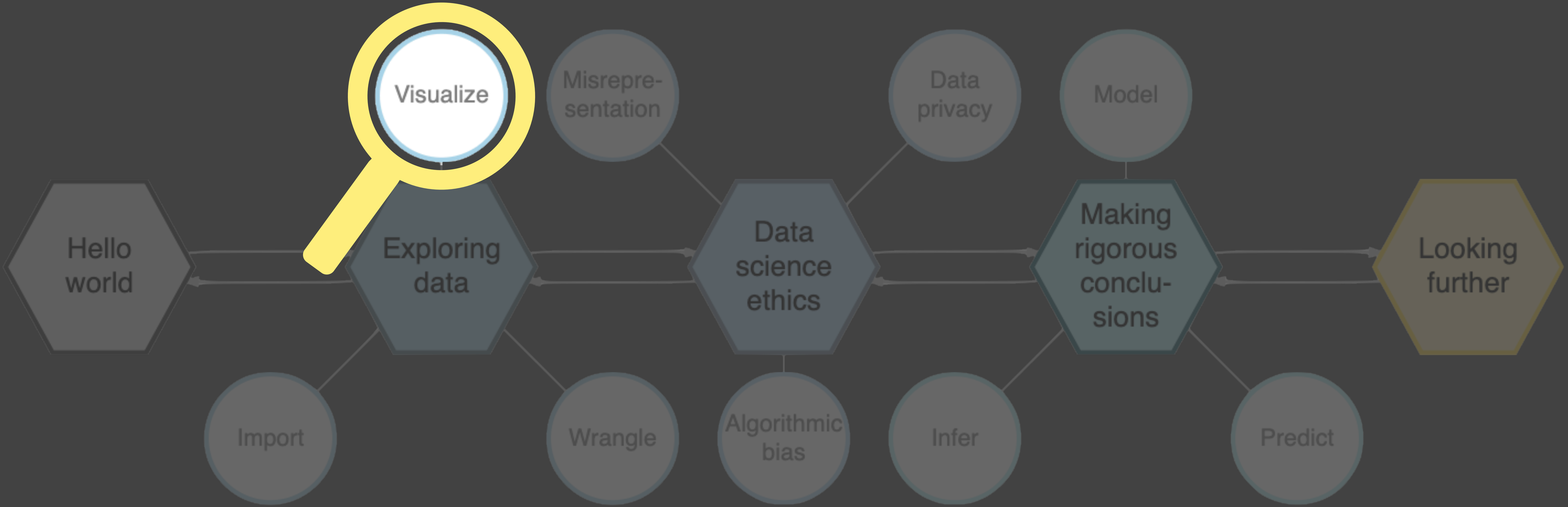
data visualisation
data wrangling, tidying, acquisition
exploratory data analysis
predictive modeling + uncertainty quantification
effective communication of results

foray into

interactive visualizations
text analysis
machine learning
Bayesian inference
...

emphasise

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

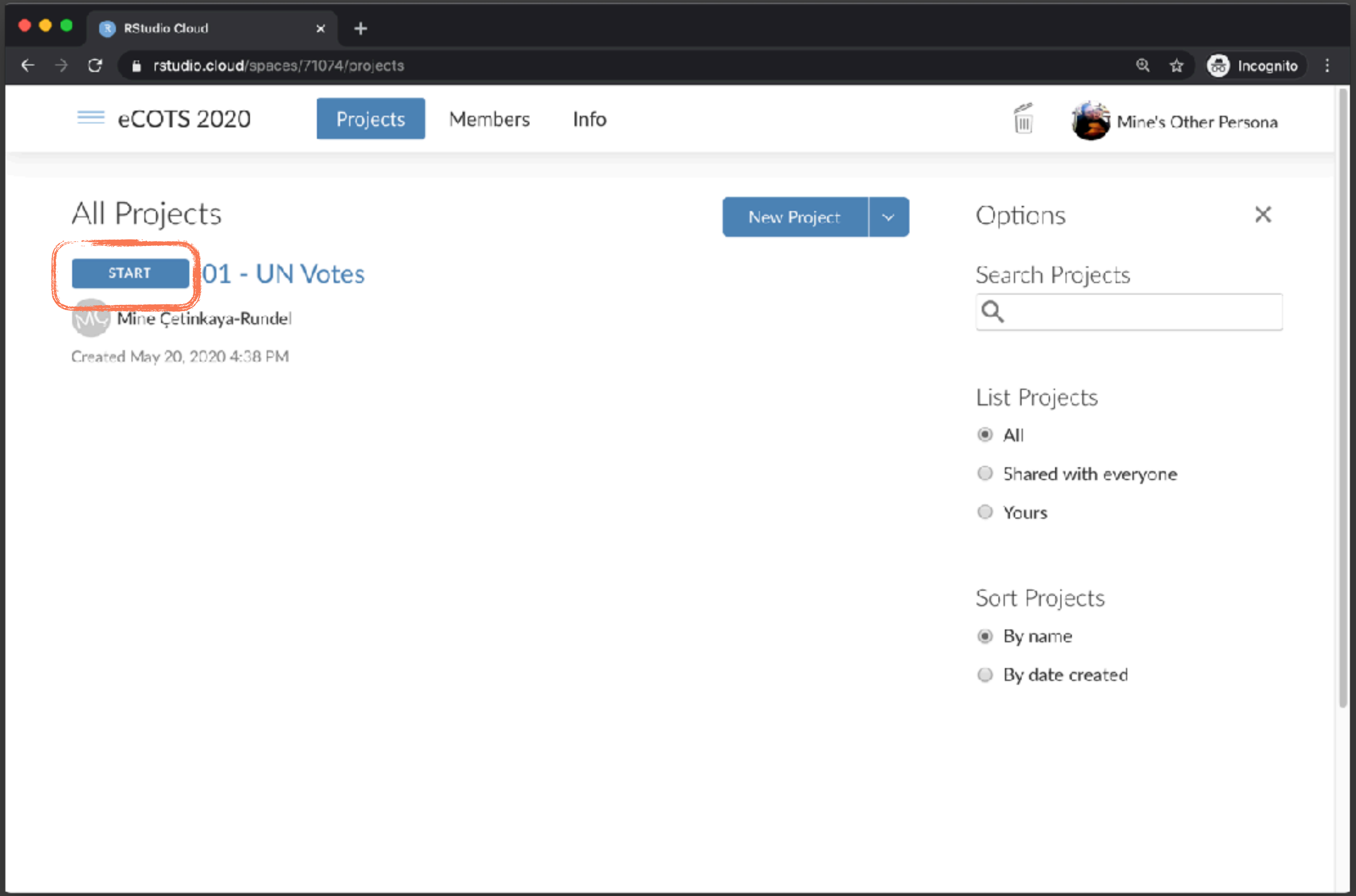

Visualize
Misrepre-sentation
Data privacy
Model
Hello world
Exploring data
Data science ethics
Making rigorous conclu-sions
Looking further
Import
Wrangle
Algorithmic bias
Infer
Predict

ex. 1

united nations

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

🔗 **rstd.io/dsbox-cloud**



All Projects

New Project

START 01 - UN Votes

Mine Çetinkaya-Rundel

Created May 20, 2020 4:38 PM

Options

Search Projects

List Projects

● All

○ Shared with everyone

○ Yours

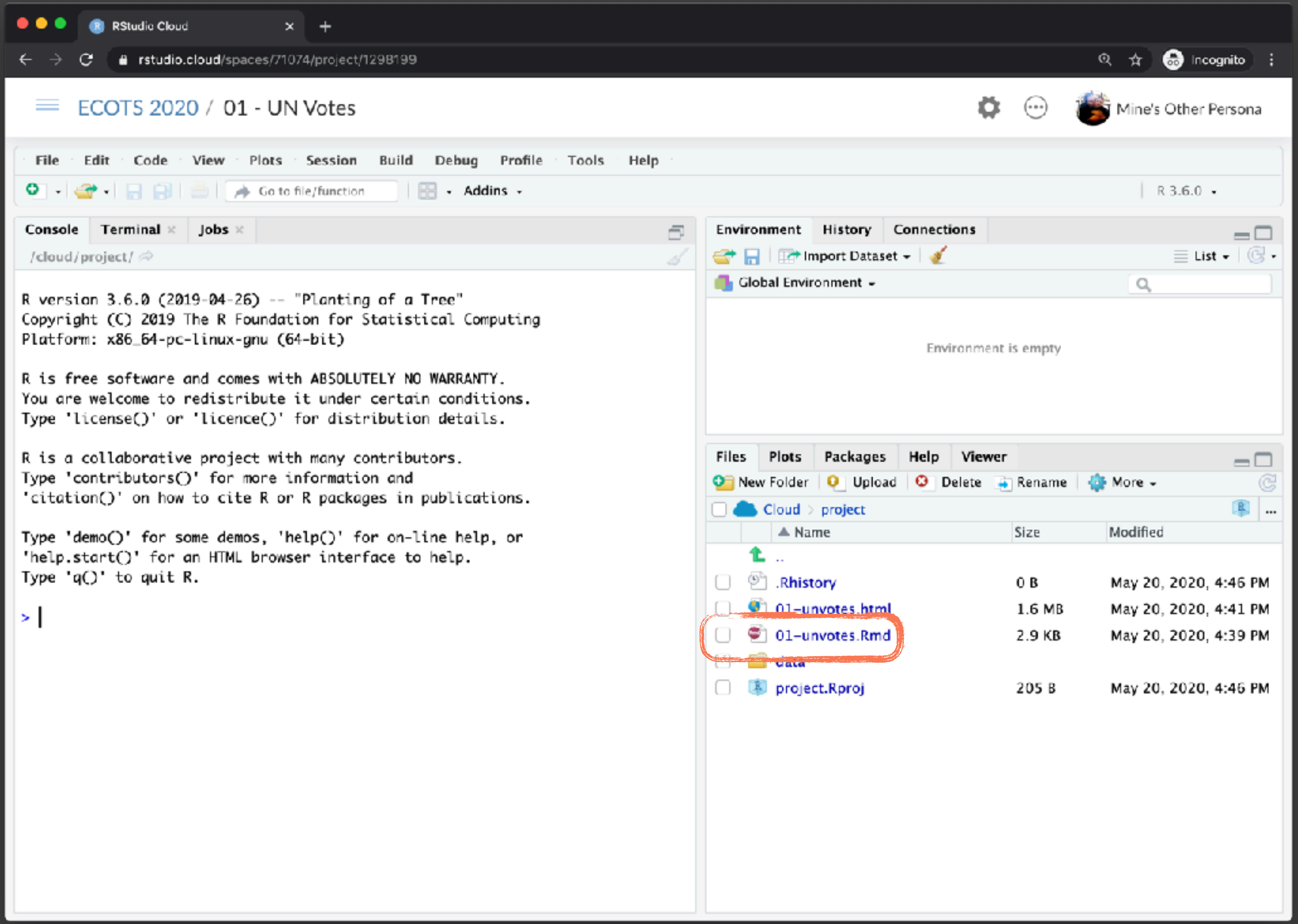
Sort Projects

● By name

○ By date created

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

🔗 **rstd.io/dsbox-cloud**

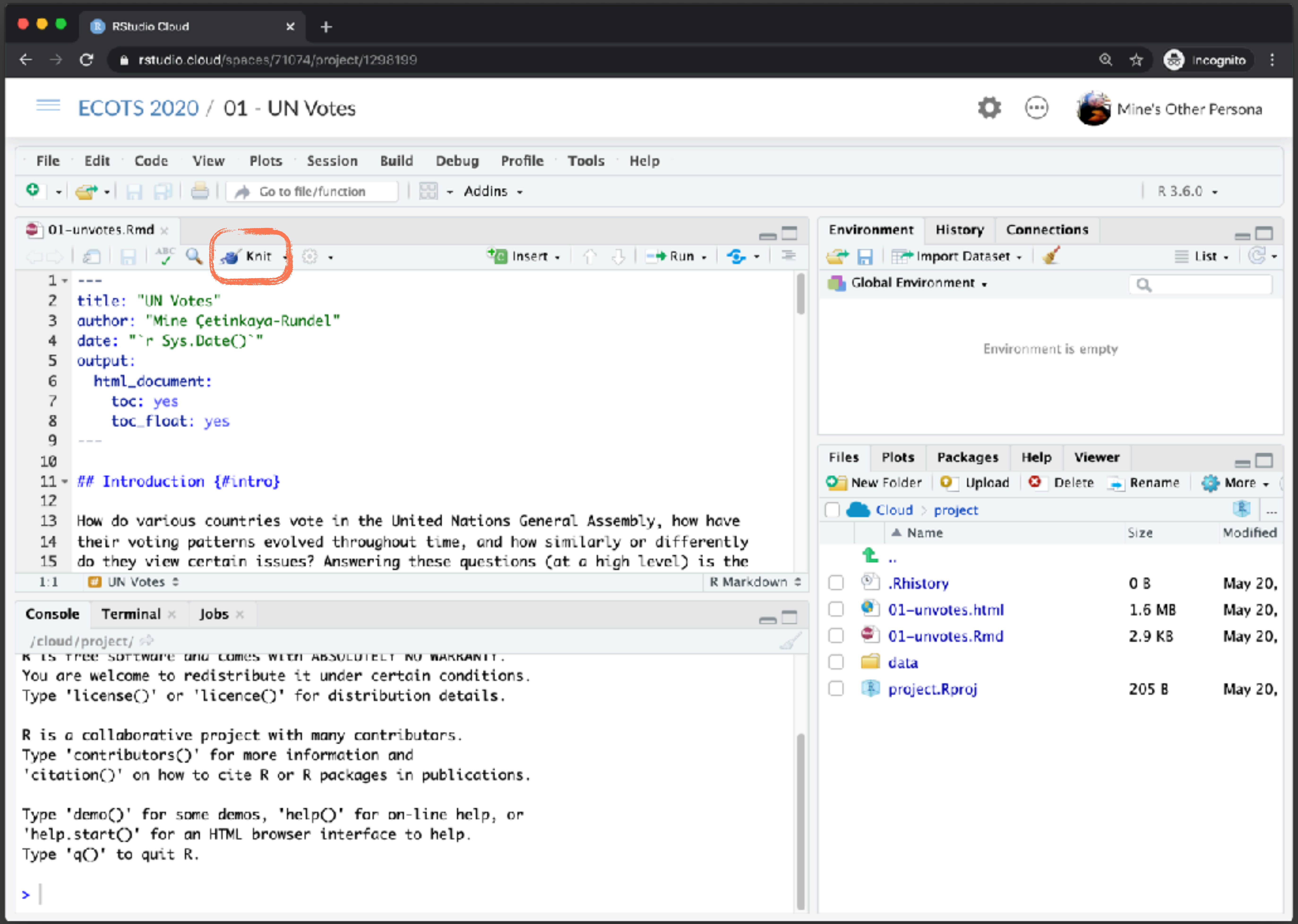


```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Name	Size	Modified
..		
.Rhistory	0 B	May 20, 2020, 4:46 PM
01-unvotes.html	1.6 MB	May 20, 2020, 4:41 PM
01-unvotes.Rmd	2.9 KB	May 20, 2020, 4:39 PM
data		
project.Rproj	205 B	May 20, 2020, 4:46 PM

- 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**



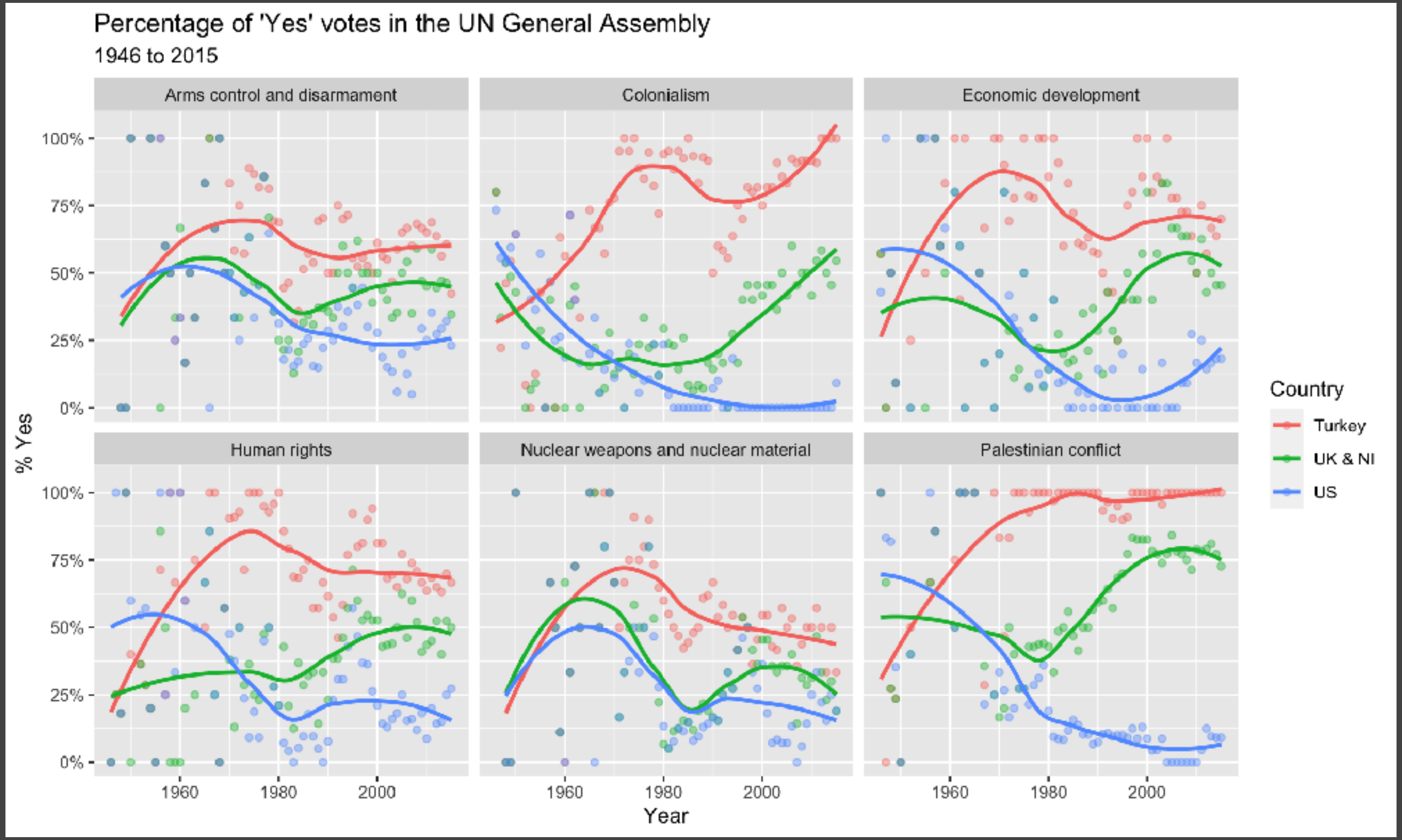
```
---
title: "UN Votes"
author: "Mine Çetinkaya-Rundel"
date: "`r Sys.Date()`"
output:
  html_document:
    toc: yes
    toc_float: yes
---

## Introduction {#intro}

How do various countries vote in the United Nations General Assembly, how have
their voting patterns evolved throughout time, and how similarly or differently
do they view certain issues? Answering these questions (at a high level) is the
```

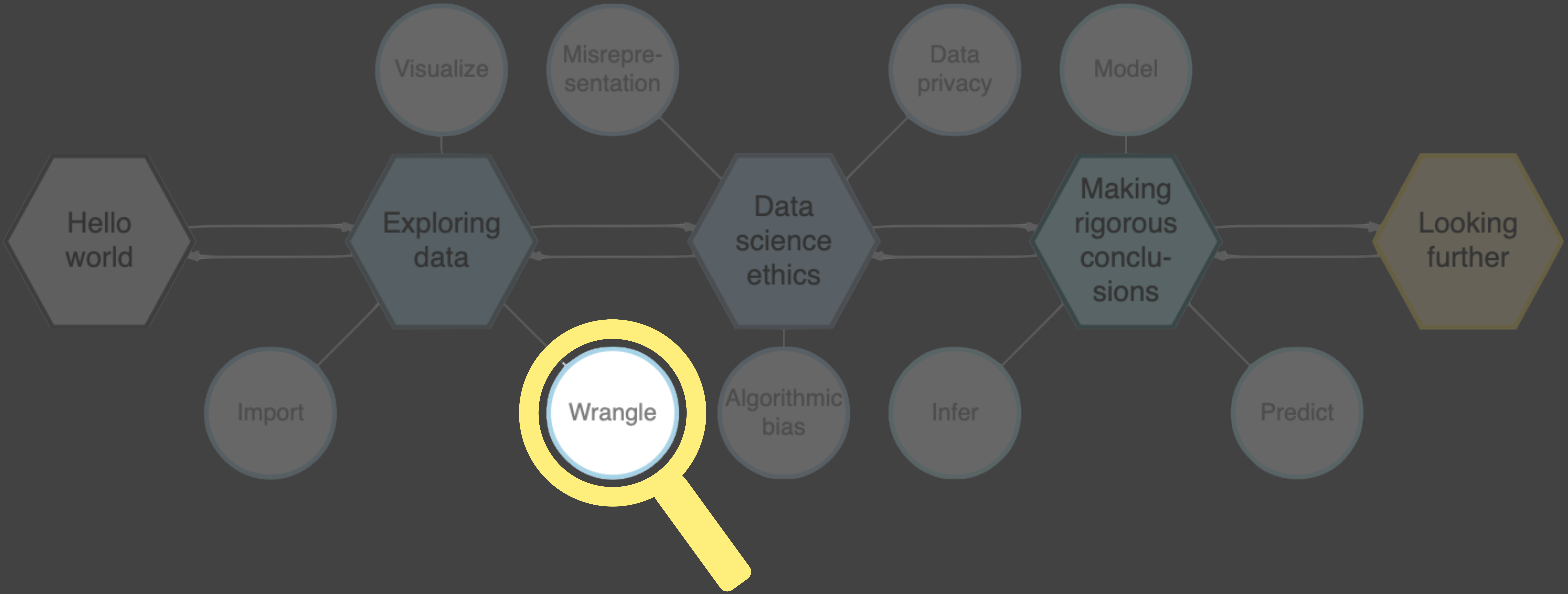
```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Name	Size	Modified
..		
.Rhistory	0 B	May 20,
01-unvotes.html	1.6 MB	May 20,
01-unvotes.Rmd	2.9 KB	May 20,
data		
project.Rproj	205 B	May 20,

- 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

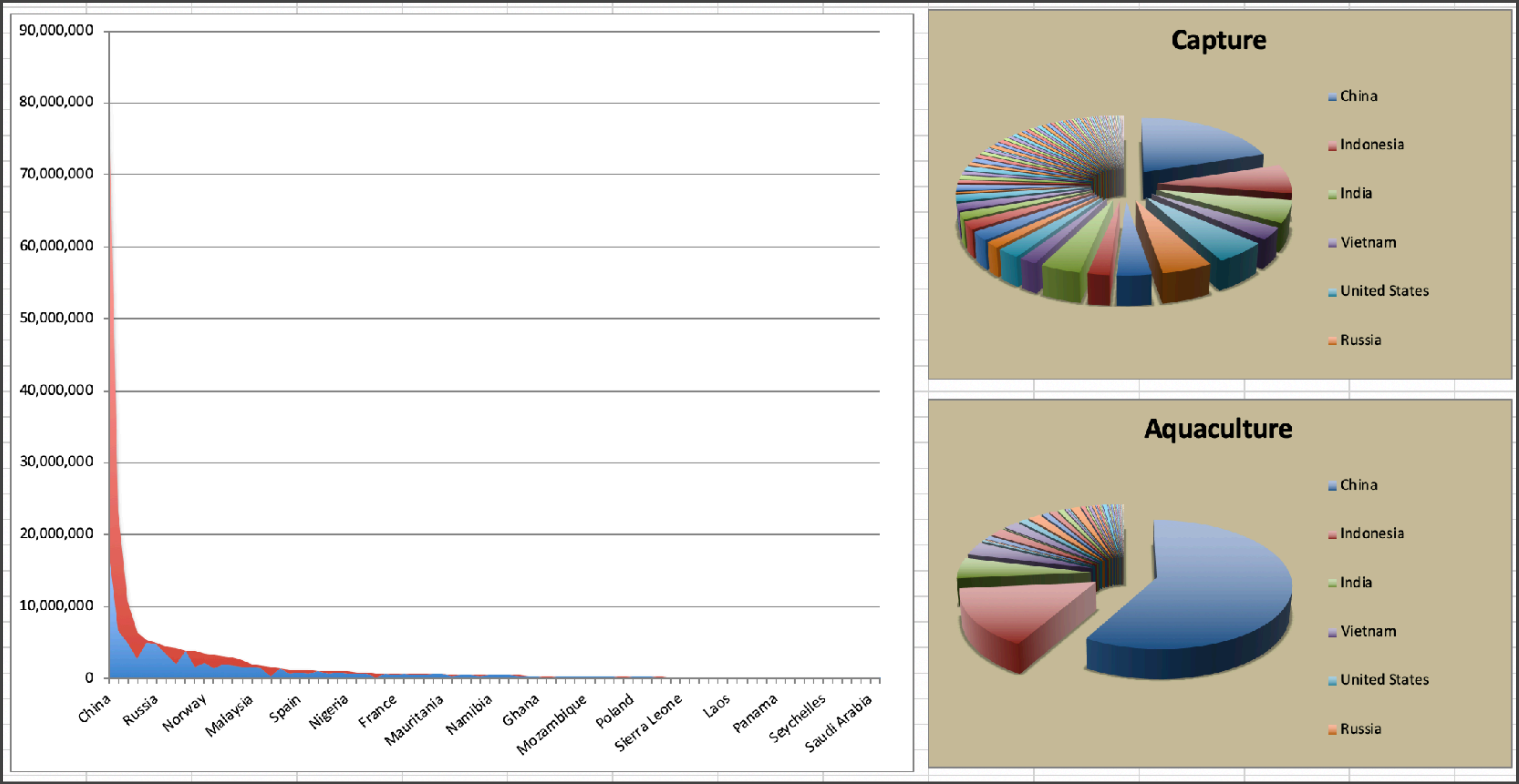
🔗 **rstd.io/dsbox-cloud**




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. 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

✓ data joins

```
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
```

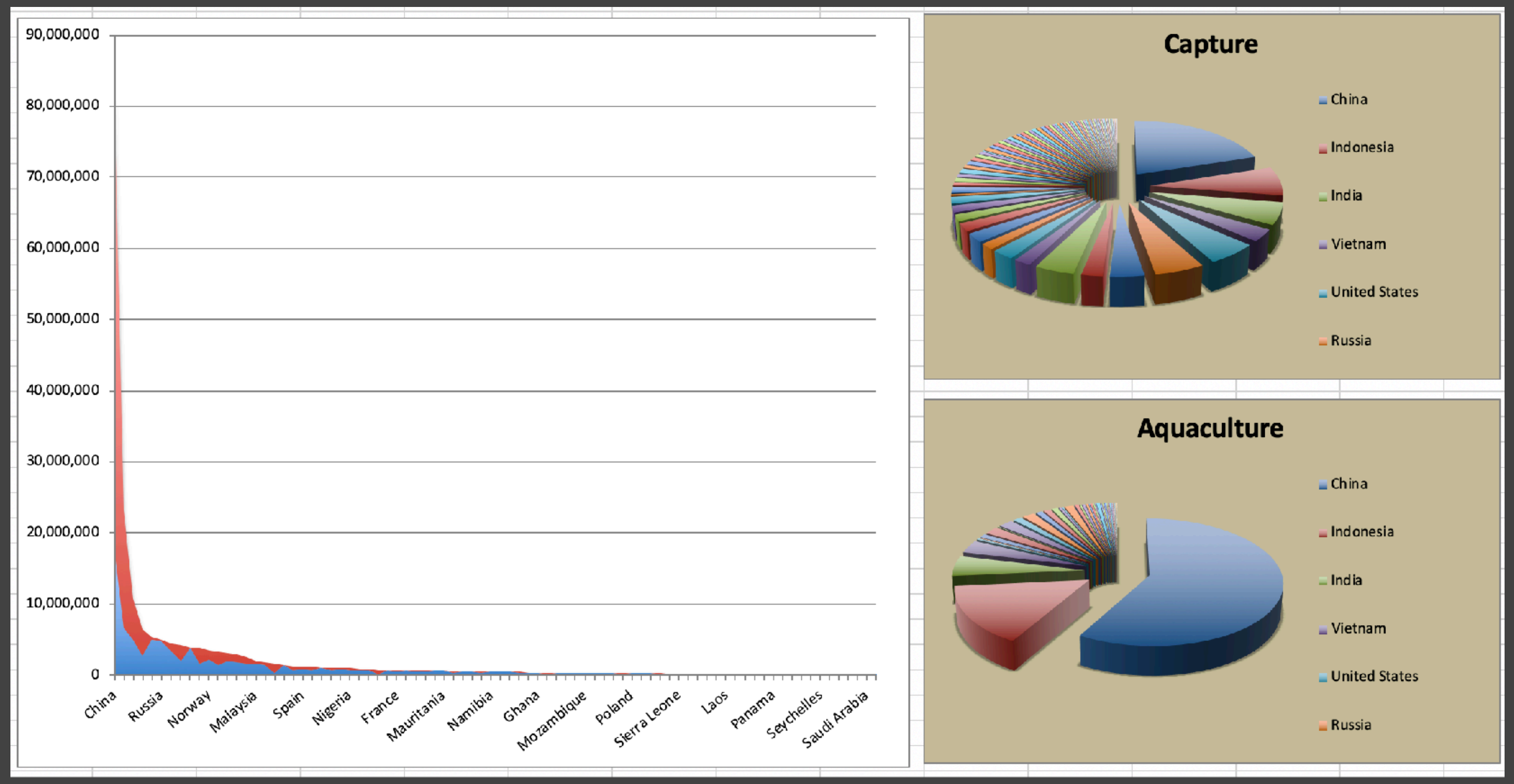
```
continents
#> # A tibble: 245 x 2
#>    country                continent
#>    <chr>                  <chr>
#>  1 Afghanistan            Asia
#>  2 Åland Islands          Europe
#>  3 Albania                Europe
#>  4 Algeria                Africa
#>  5 American Samoa         Oceania
#>  6 Andorra                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"
```

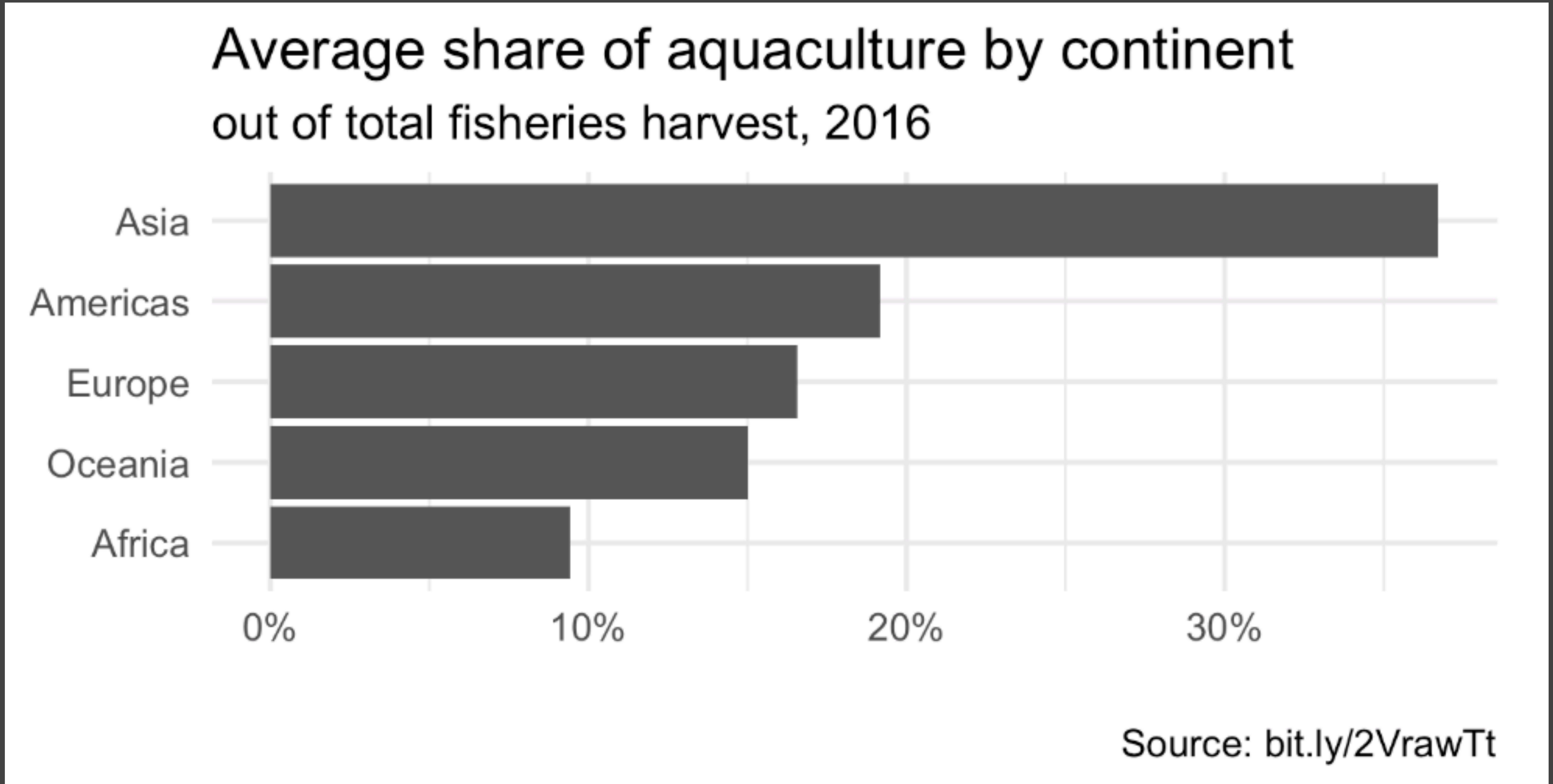
✓ data joins

✓ 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
```



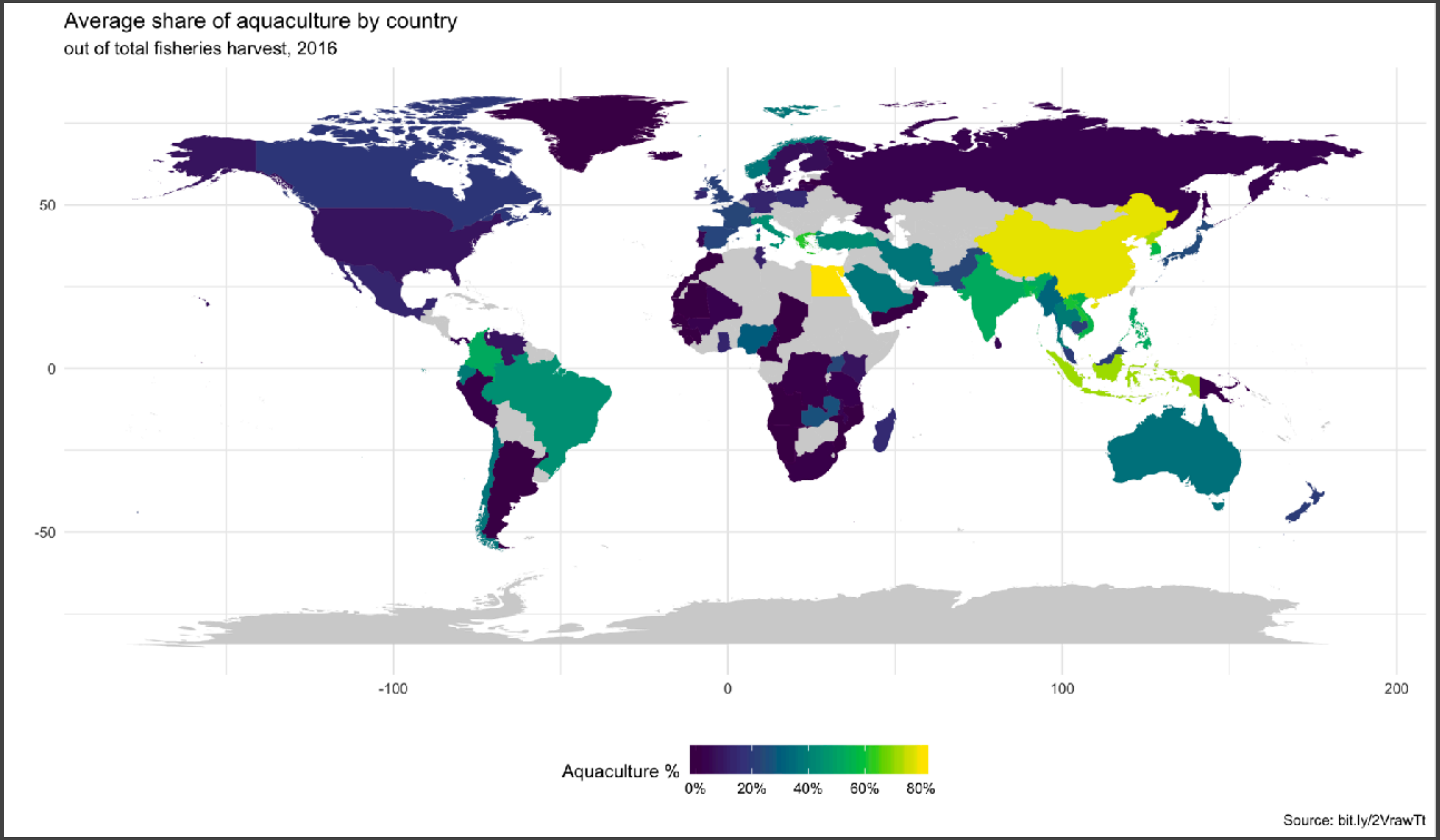
- ✓ data joins
- ✓ ethics
- ✓ critique
- ✓ improving visualisations



Average share of aquaculture by continent

out of total fisheries harvest, 2016

Source: bit.ly/2VrawTt

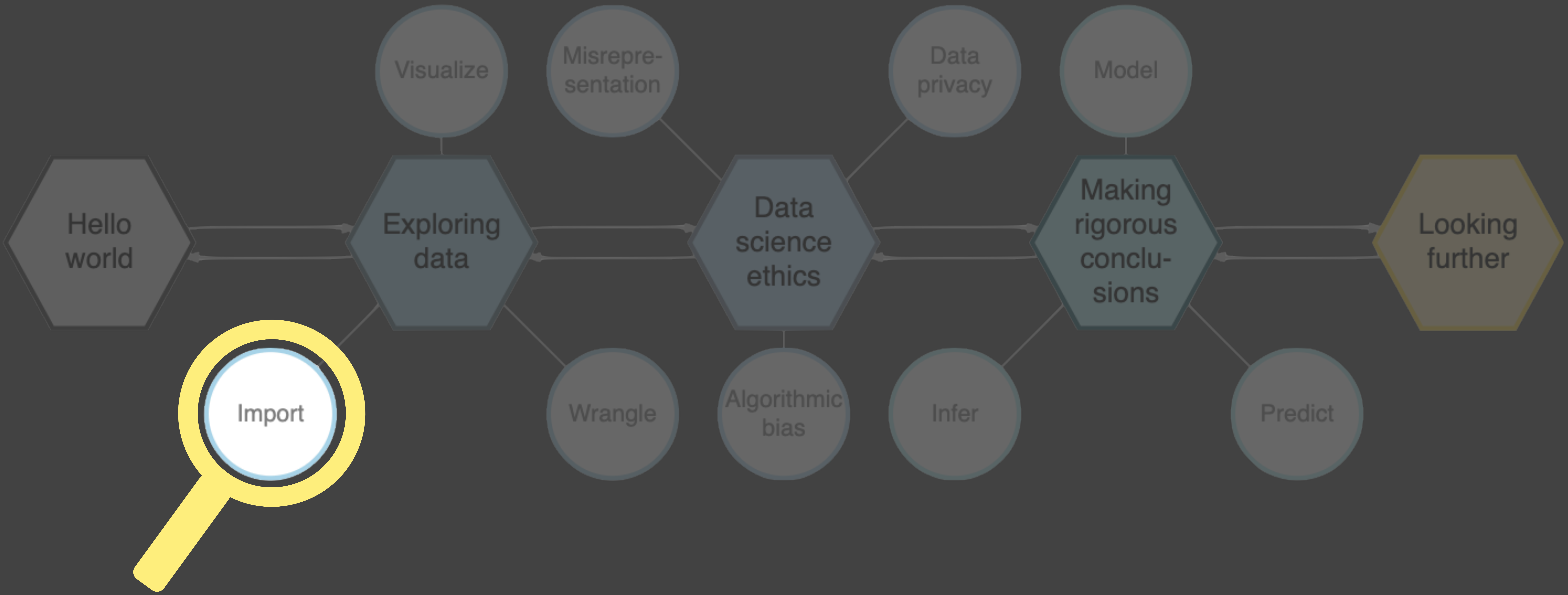


Average share of aquaculture by country

out of total fisheries harvest, 2016

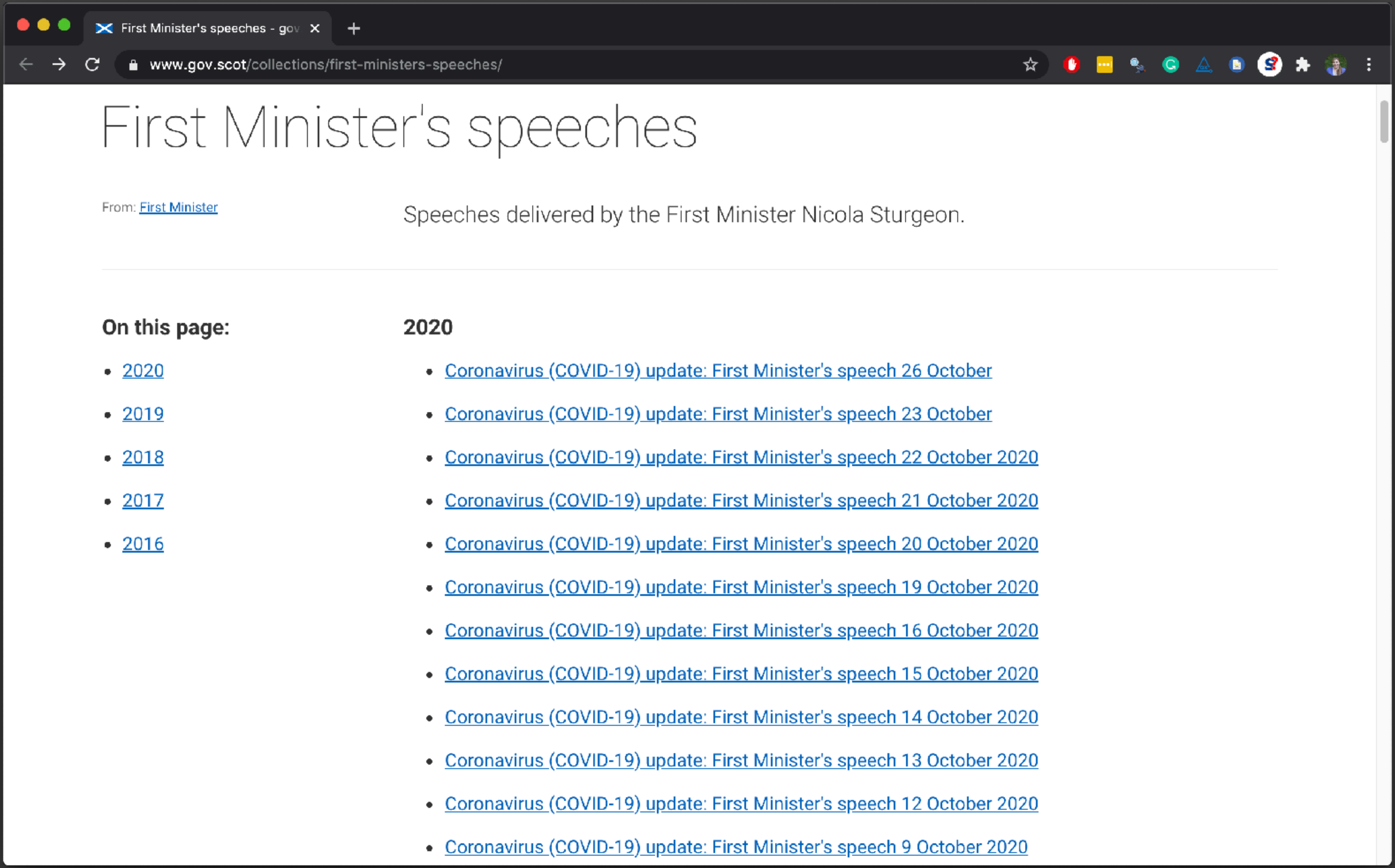
Source: bit.ly/2VrawTt

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


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. 3

First Minister’s COVID briefings


First Minister's speeches - gov
www.gov.scot/collections/first-ministers-speeches/

First Minister's speeches

From: First Minister

Speeches delivered by the First Minister Nicola Sturgeon.

On this page:

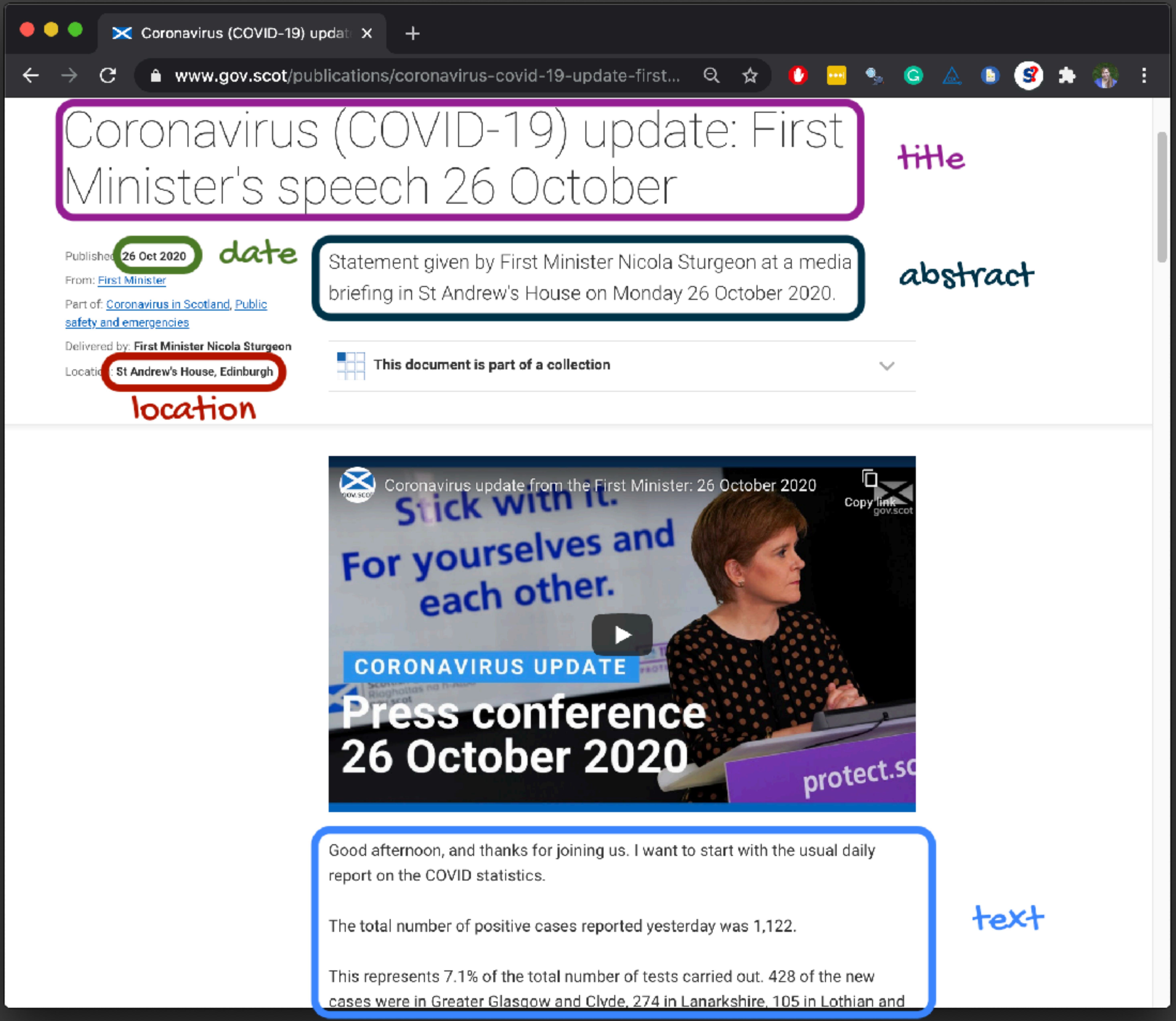
- 2020
- 2019
- 2018
- 2017
- 2016

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
```



Coronavirus (COVID-19) updat ×

www.gov.scot/publications/coronavirus-covid-19-update-first...

Coronavirus (COVID-19) update: First Minister's speech 26 October

Published: 26 Oct 2020

From: First Minister

Part of: Coronavirus in Scotland, Public safety and emergencies

Delivered by: First Minister Nicola Sturgeon

Location: St Andrew's House, Edinburgh

Statement given by First Minister Nicola Sturgeon at a media briefing in St Andrew's House on Monday 26 October 2020.

This document is part of a collection

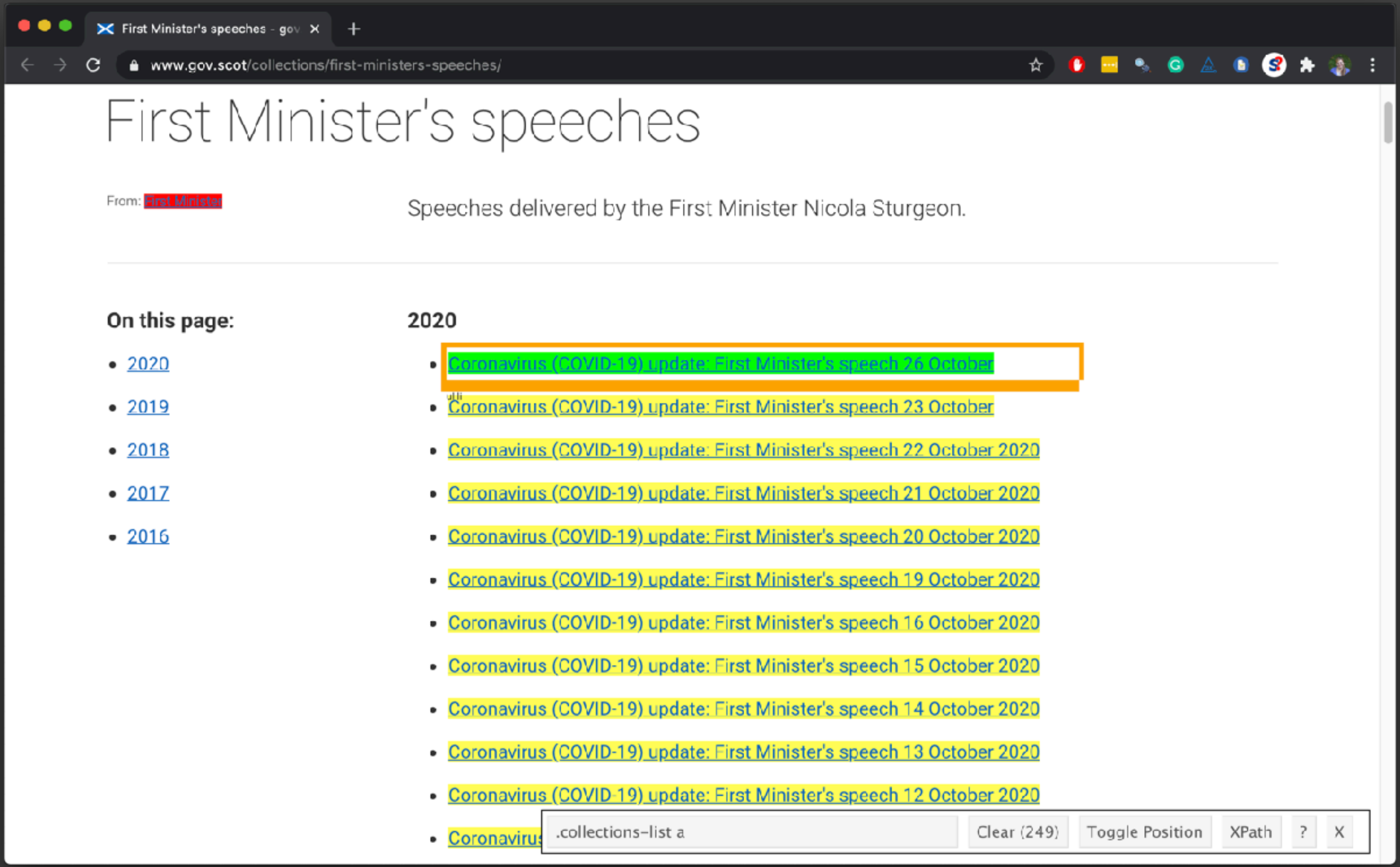
Coronavirus update from the First Minister: 26 October 2020

Good afternoon, and thanks for joining us. I want to start with the usual daily report on the COVID statistics.

The total number of positive cases reported yesterday was 1,122.

This represents 7.1% of the total number of tests carried out. 428 of the new cases were in Greater Glasgow and Clyde, 274 in Lanarkshire, 105 in Lothian and

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


First Minister's speeches - gov
www.gov.scot/collections/first-ministers-speeches/

First Minister's speeches

From: First Minister

Speeches delivered by the First Minister Nicola Sturgeon.

On this page:

- 2020
- 2019
- 2018
- 2017
- 2016

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
- Coronaviru

.collections-list a

✓ ethics

✓ web scraping

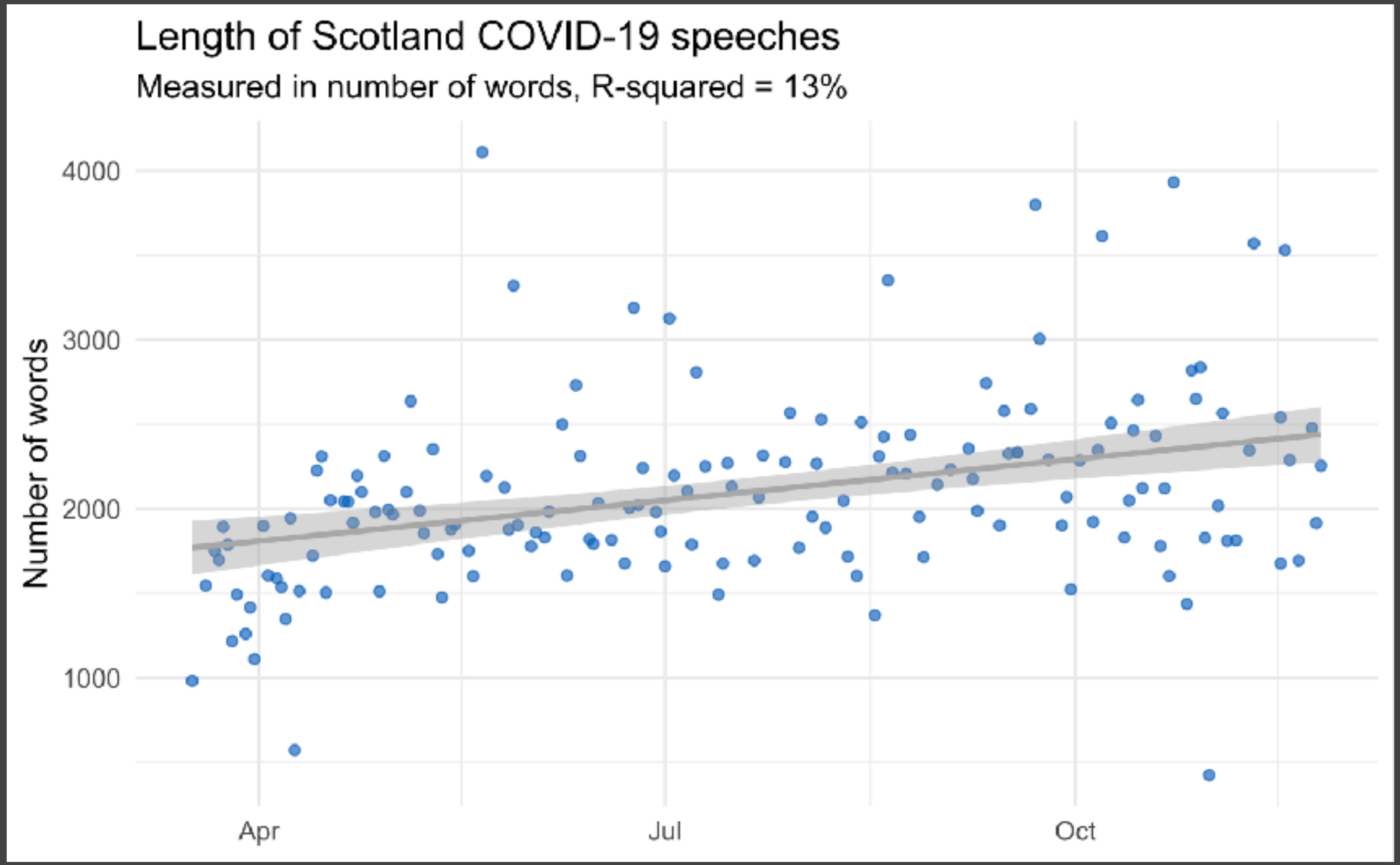
✓ text parsing

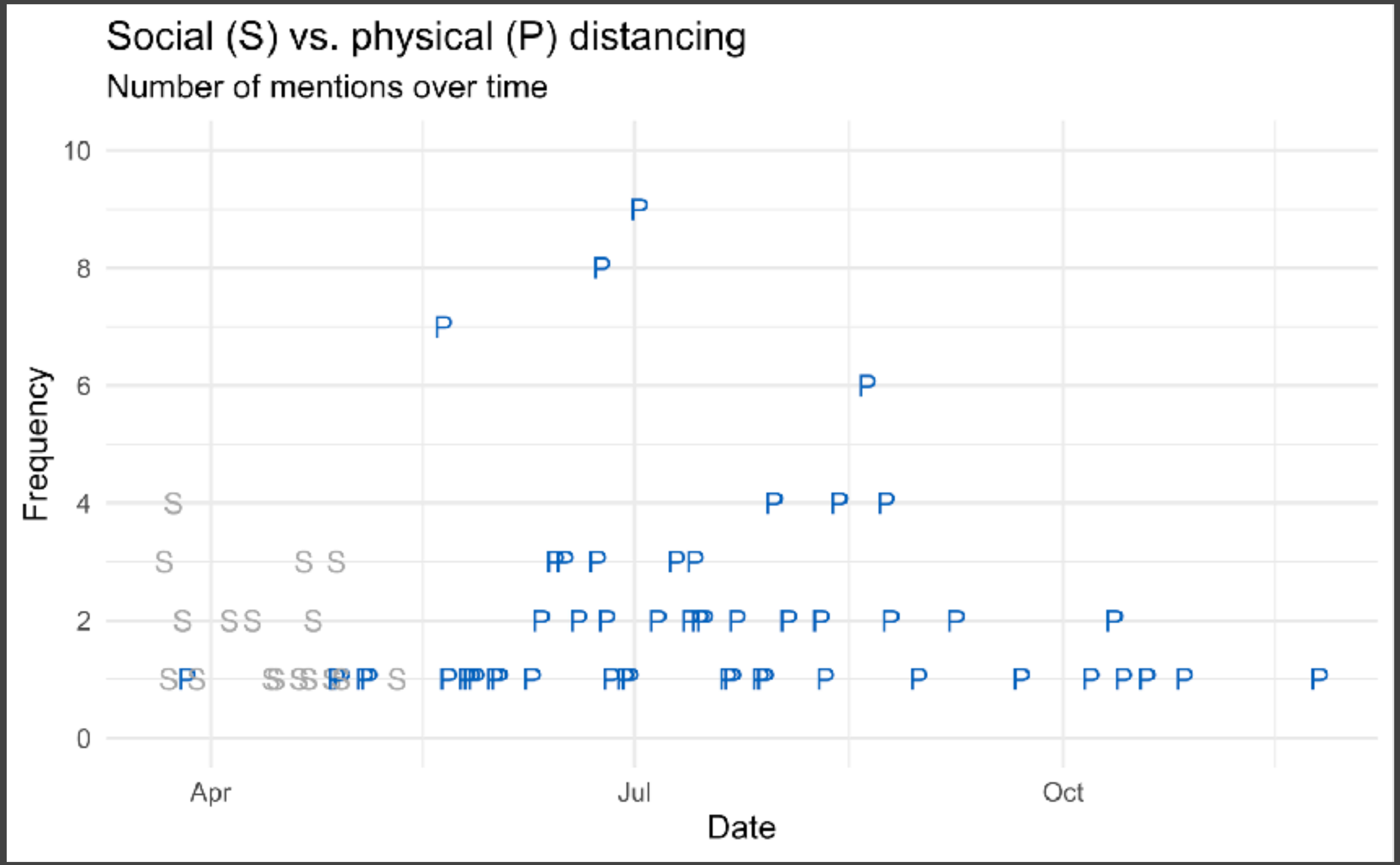
✓ data types

✓ regular expressions

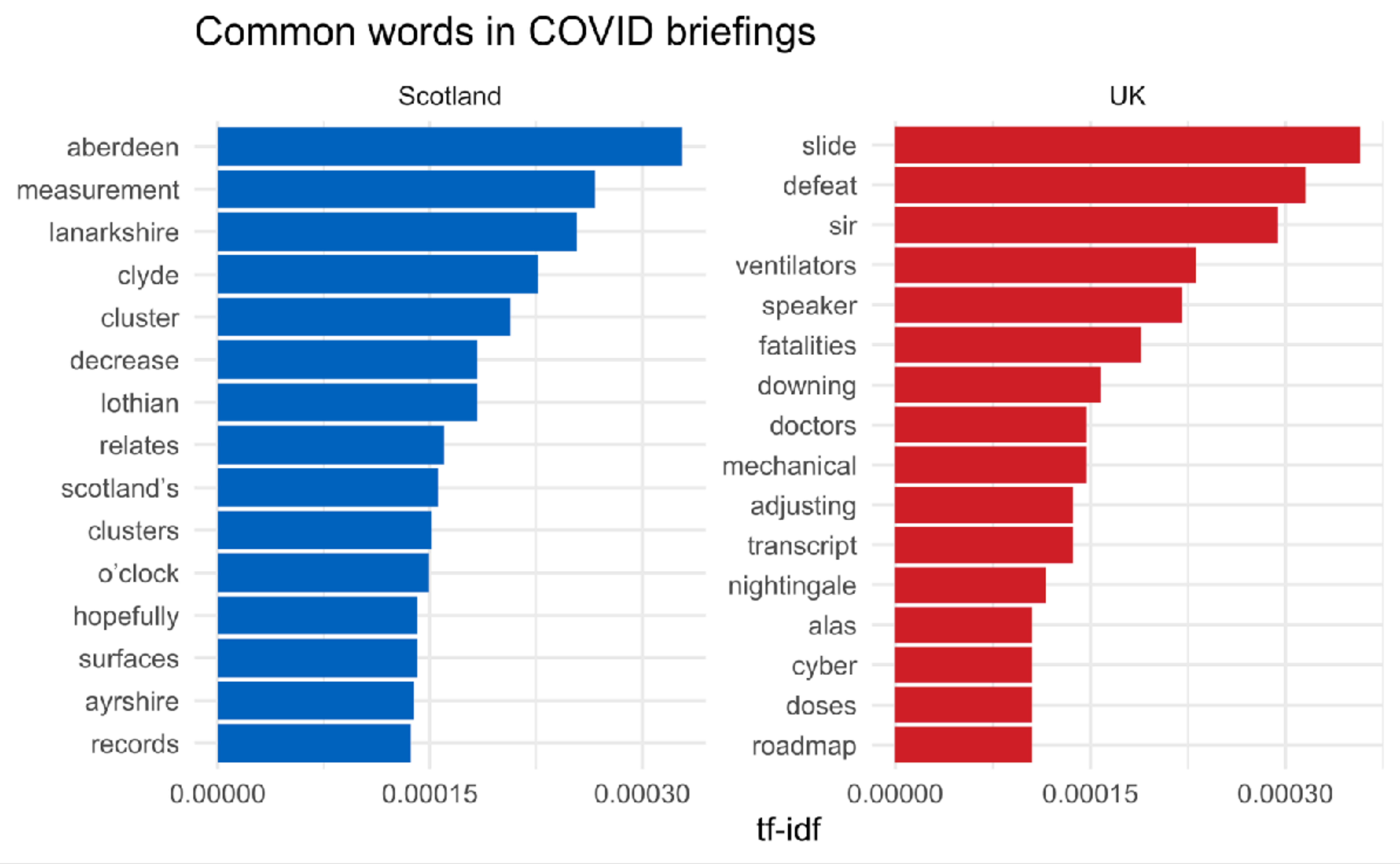
✓ functions

✓ iteration

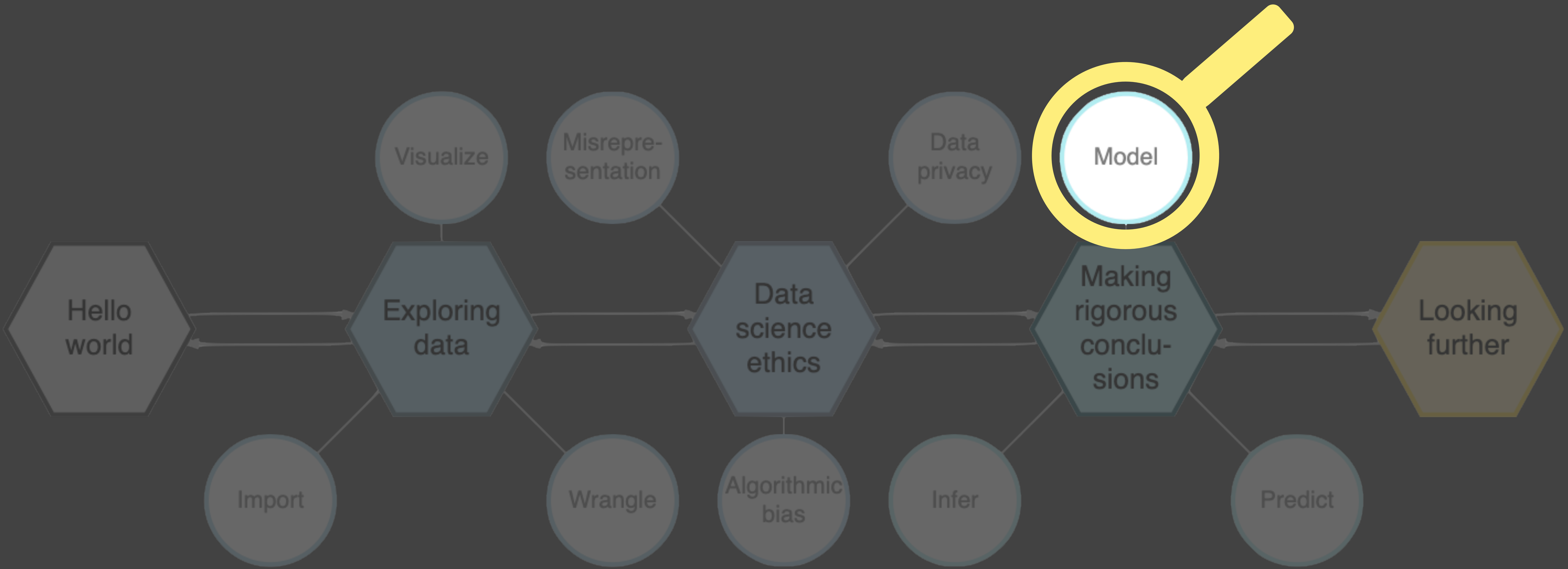




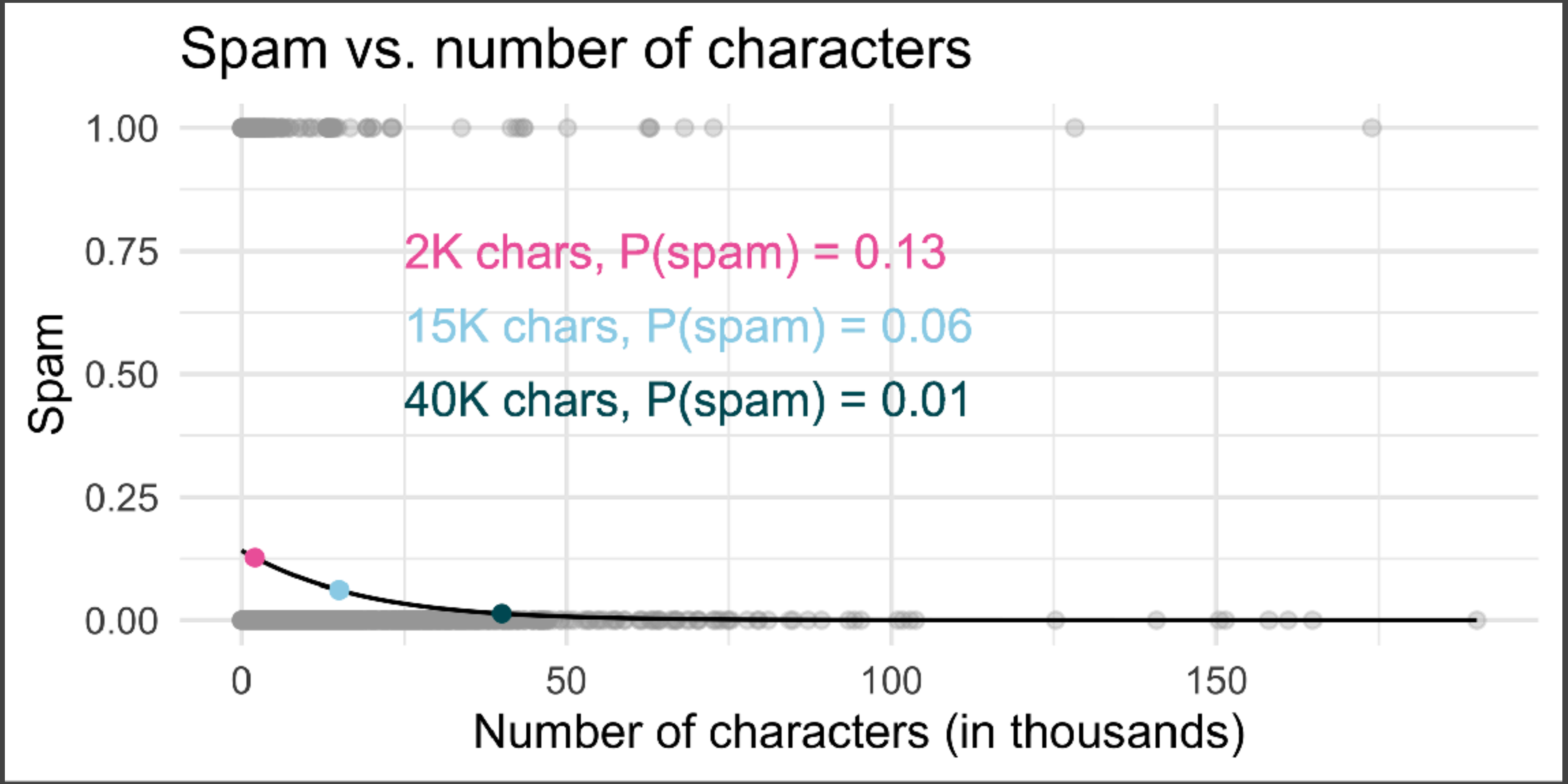
- ✓ ethics
- ✓ web scraping
- ✓ text parsing
- ✓ data types
- ✓ regular expressions
- ✓ functions
- ✓ iteration
- ✓ visualisation
- ✓ interpretation



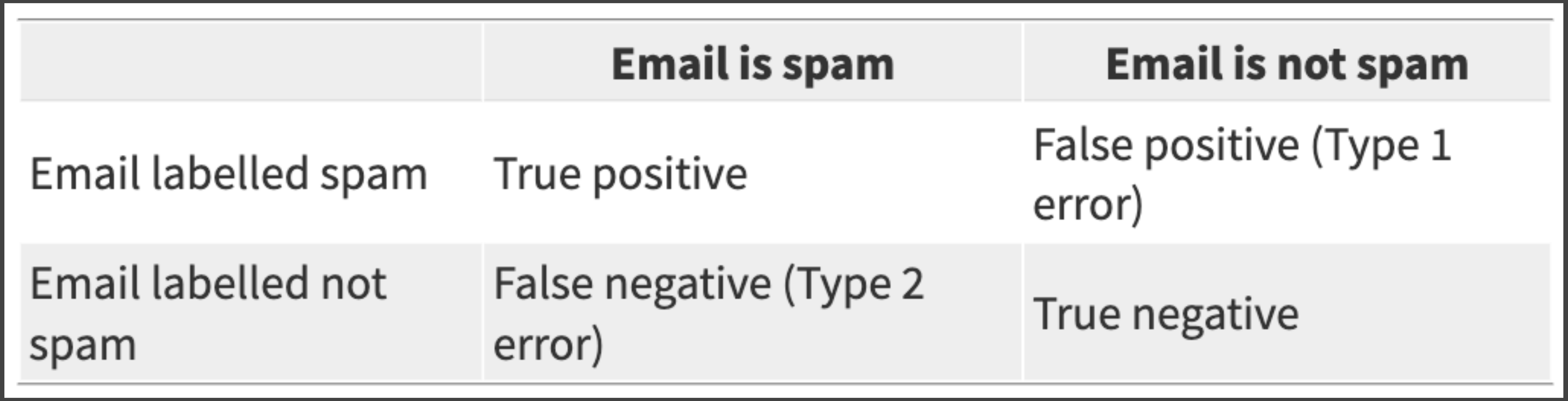
- ✓ ethics
- ✓ web scraping
- ✓ text parsing
- ✓ data types
- ✓ regular expressions
- ✓ functions
- ✓ iteration
- ✓ visualisation
- ✓ interpretation
- ✓ text analysis


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. 3
spam filters

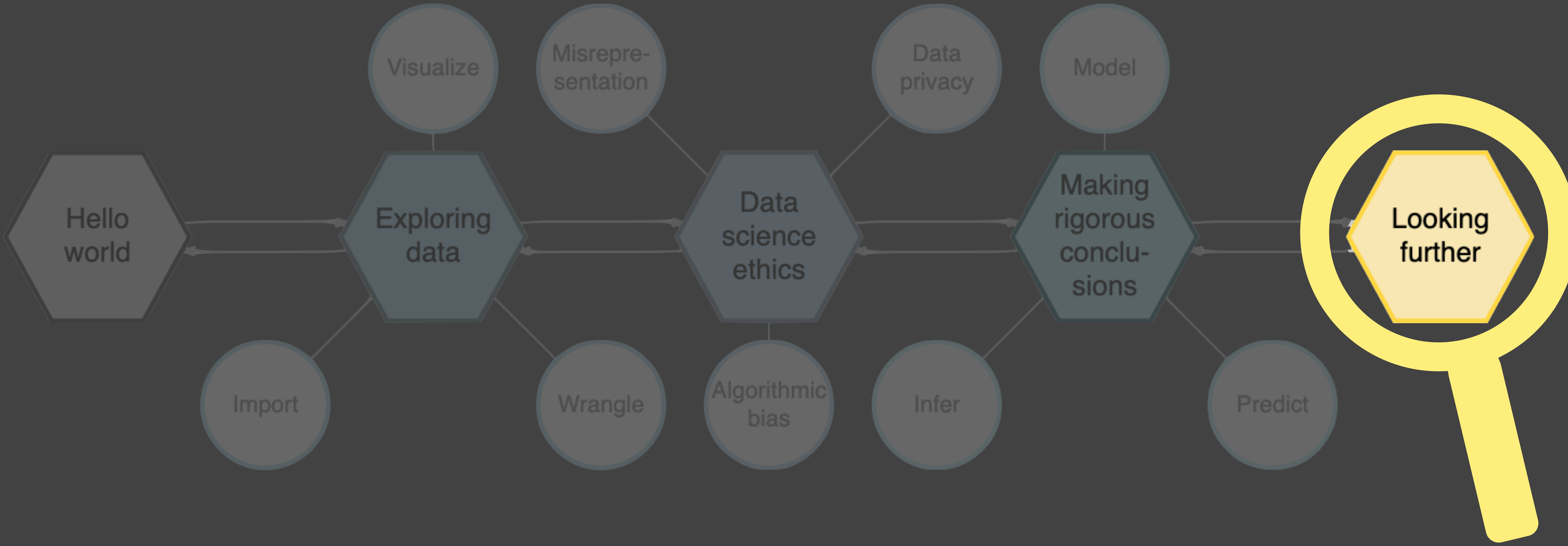


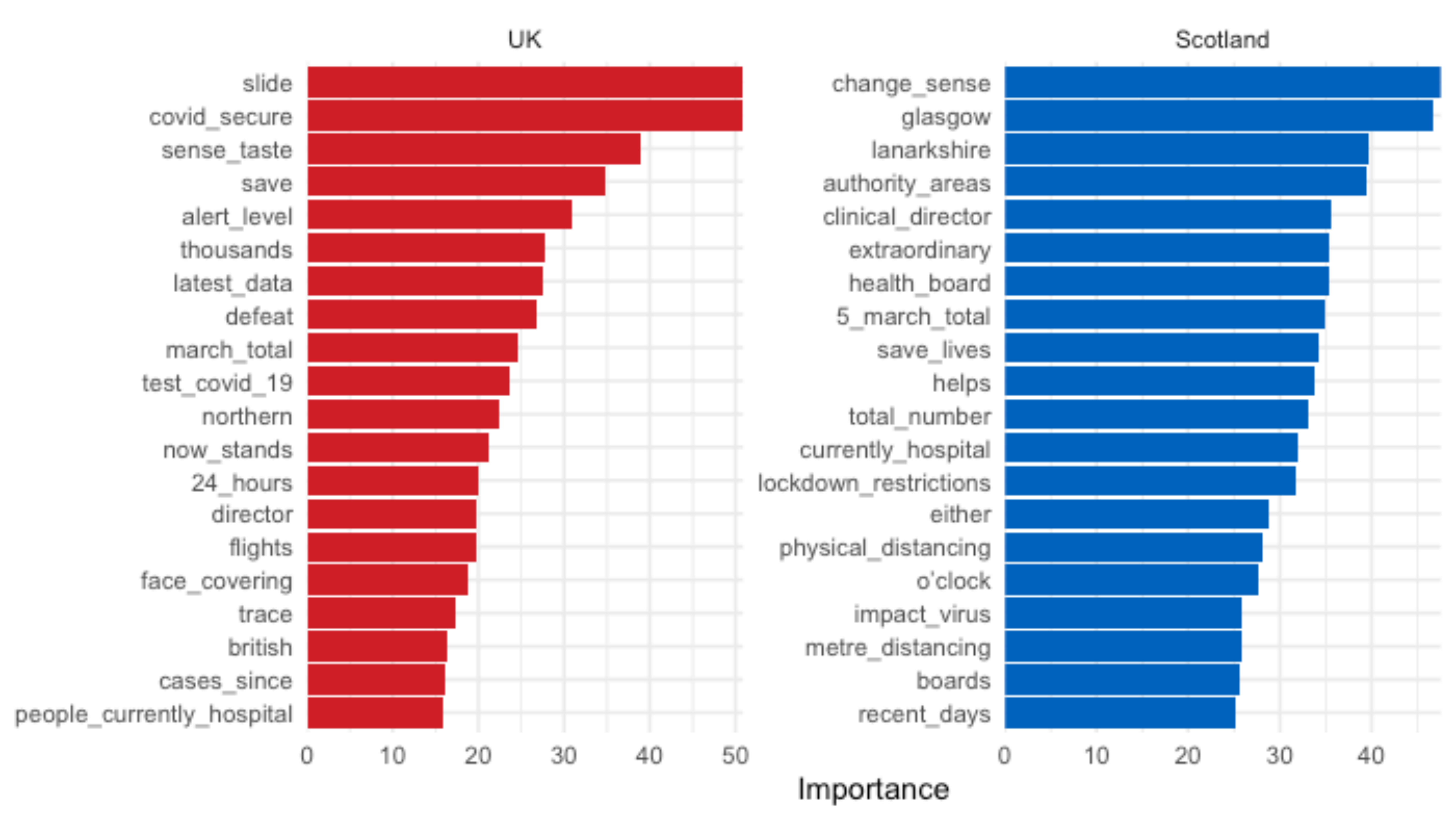
- ✓ logistic regression
- ✓ prediction



	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

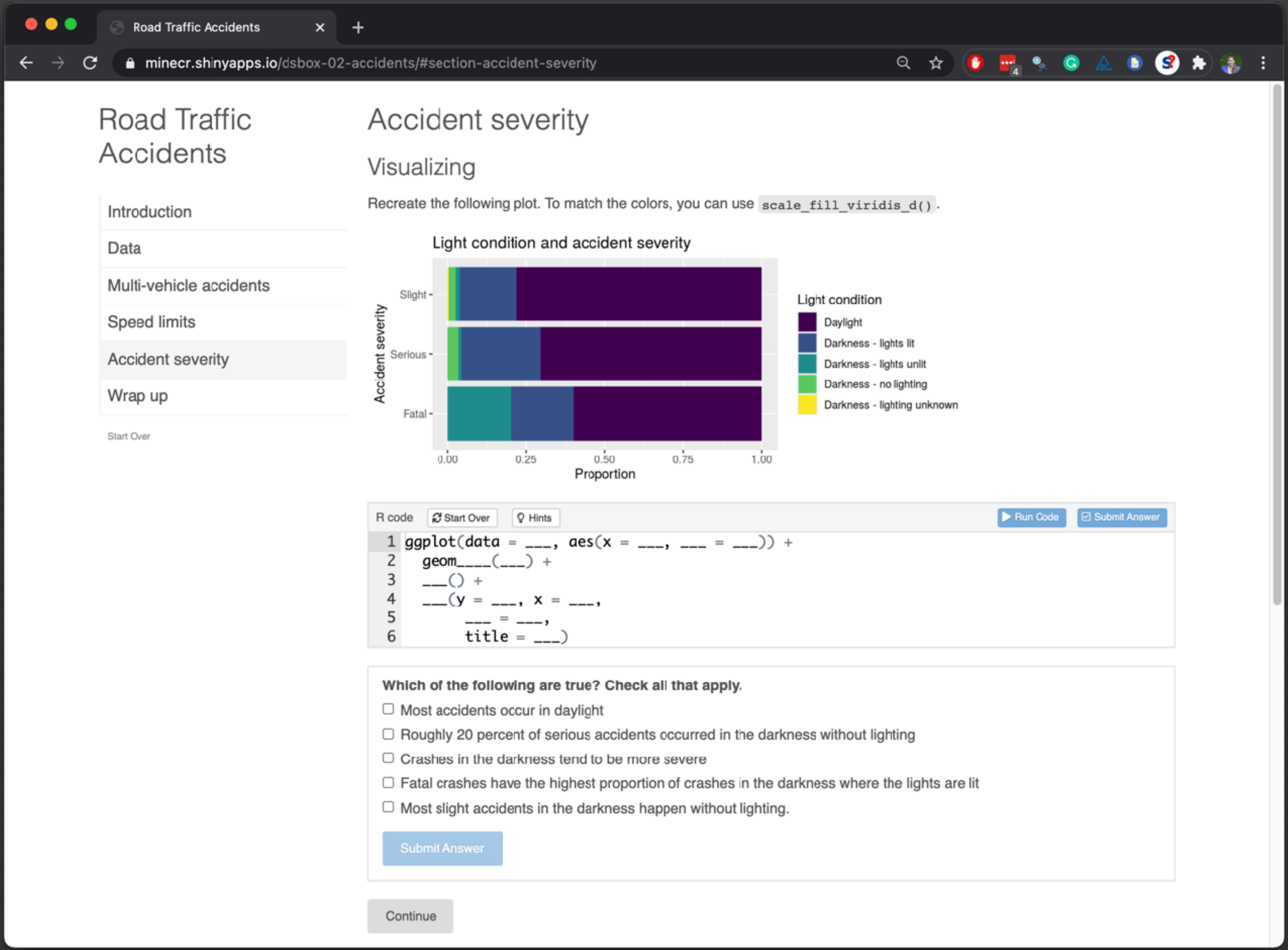
- ✓ logistic regression
- ✓ prediction
- ✓ decision errors
- ✓ sensitivity / specificity
- ✓ intuition around loss functions


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



✓ machine learning for text data

✓ repetition



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()`.

```
ggplot(data = ___, aes(x = ___, ___ = ___)) +
  geom____(___) +
  ___() +
  ___(y = ___, x = ___,
       ___ = ___,
       title = ___)
```

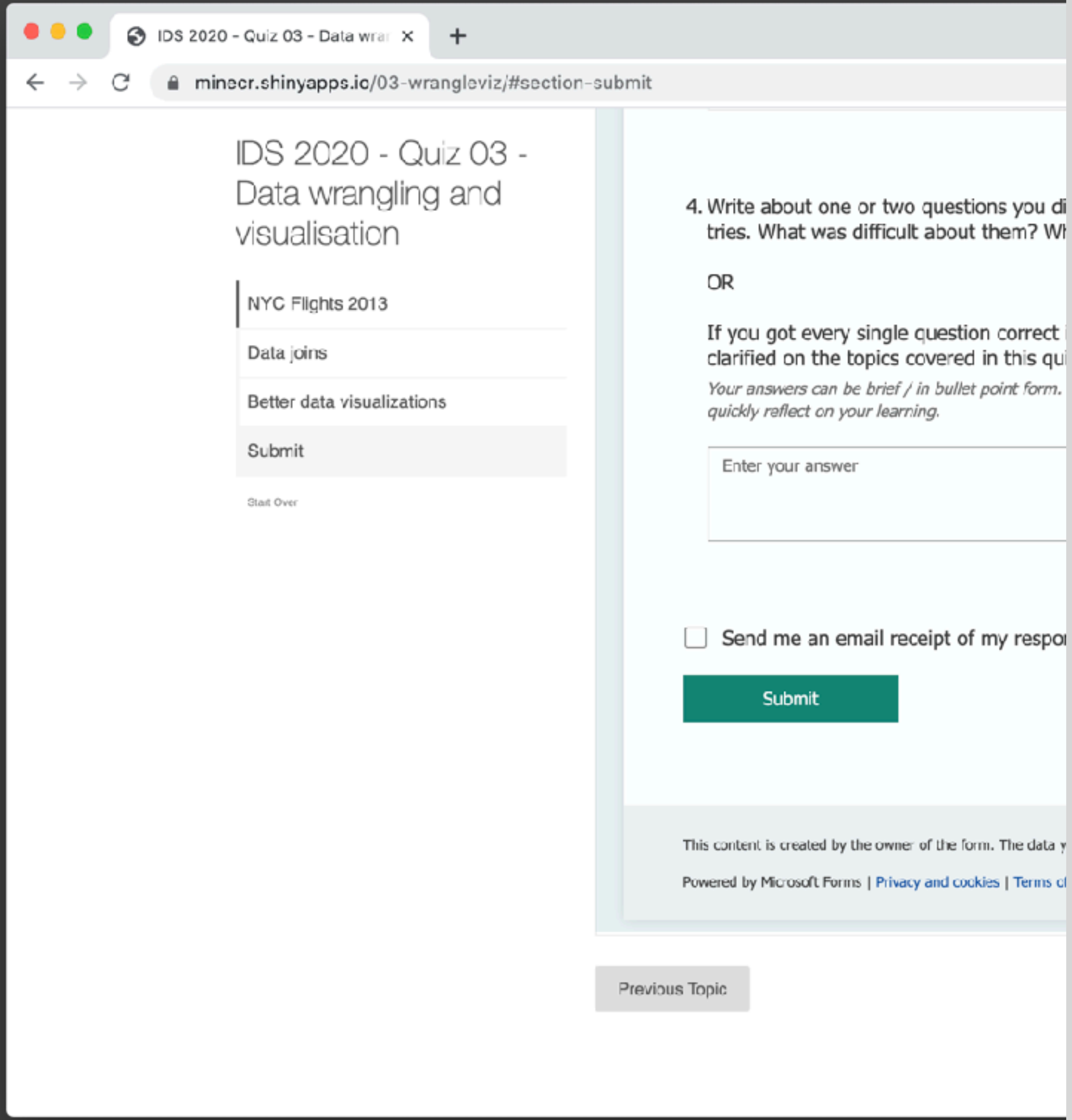
Which of the following are true? Check all that apply.

- ☐ 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



IDS 2020 - Quiz 03 - Data wra
minecr.shinyapps.io/03-wrangleviz/#section-submit

IDS 2020 - Quiz 03 - Data wrangling and visualisation

NYC Flights 2013
Data joins
Better data visualizations
Submit
Start Over

4. Write about one or two questions you di
tries. What was difficult about them? Wh

OR

If you got every single question correct
clarified on the topics covered in this qui

Your answers can be brief / in bullet point form.
quickly reflect on your learning.

Enter your answer

□ Send me an email receipt of my respo

This content is created by the owner of the form. The data y

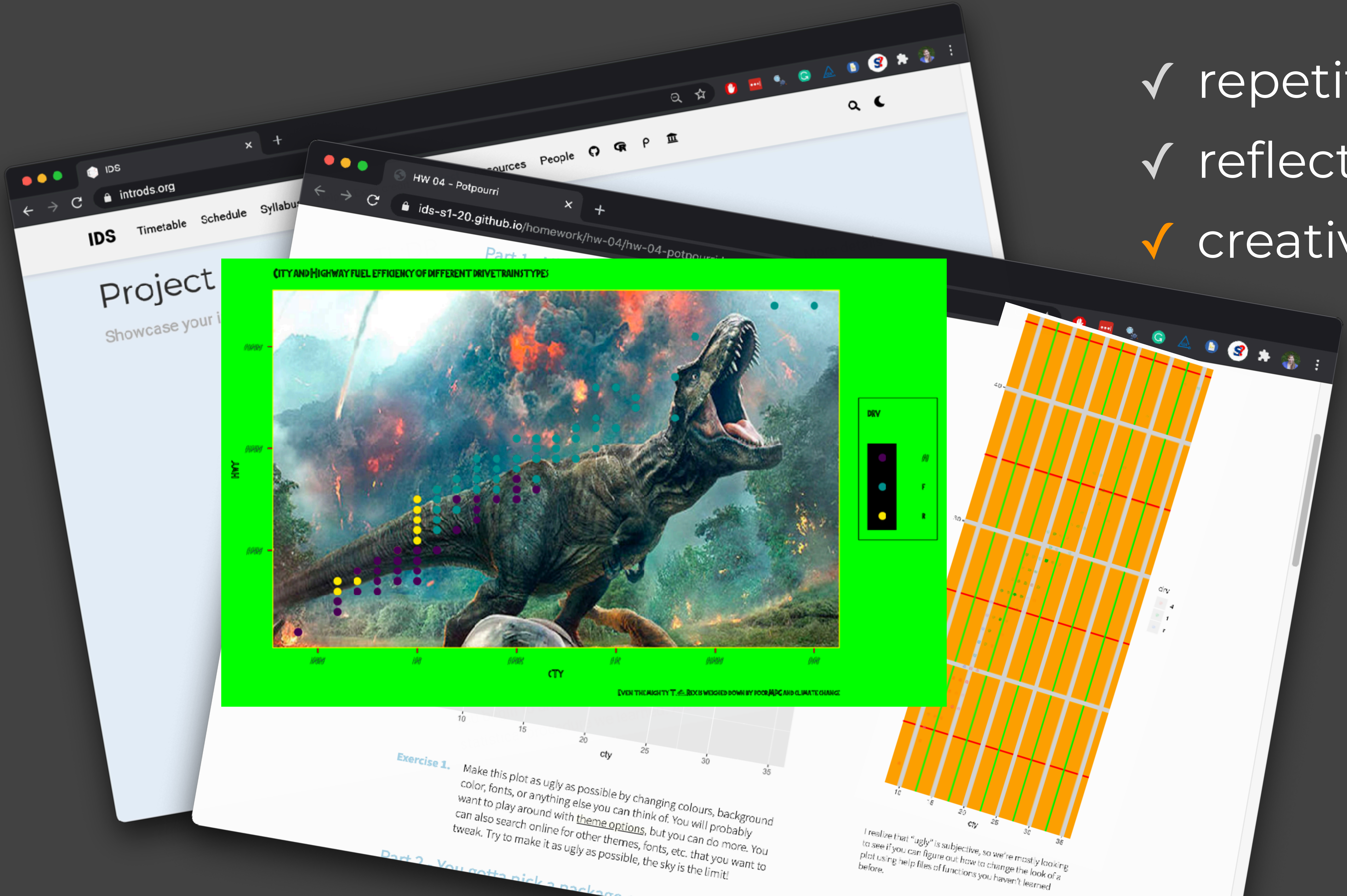
Powered by Microsoft Forms | Privacy and cookies | Terms o

```
# 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



IDS
introds.org
IDS
Timetable
Schedule
Syllabus
Resources
People
Project
Showcase your i
HW 04 - Potpourri
ids-s1-20.github.io/homework/hw-04/hw-04-potpourri
Part 1
CITY AND HIGHWAY FUEL EFFICIENCY OF DIFFERENT DRIVETRAIN TYPES
HWY
CTY
DRV
4
F
R
EVEN THE MIGHTY T-REX IS WEIGHED DOWN BY POOR MPG AND CLIMATE CHANGE
10
15
20
25
30
35
cty
40
30
drv
4
f
r

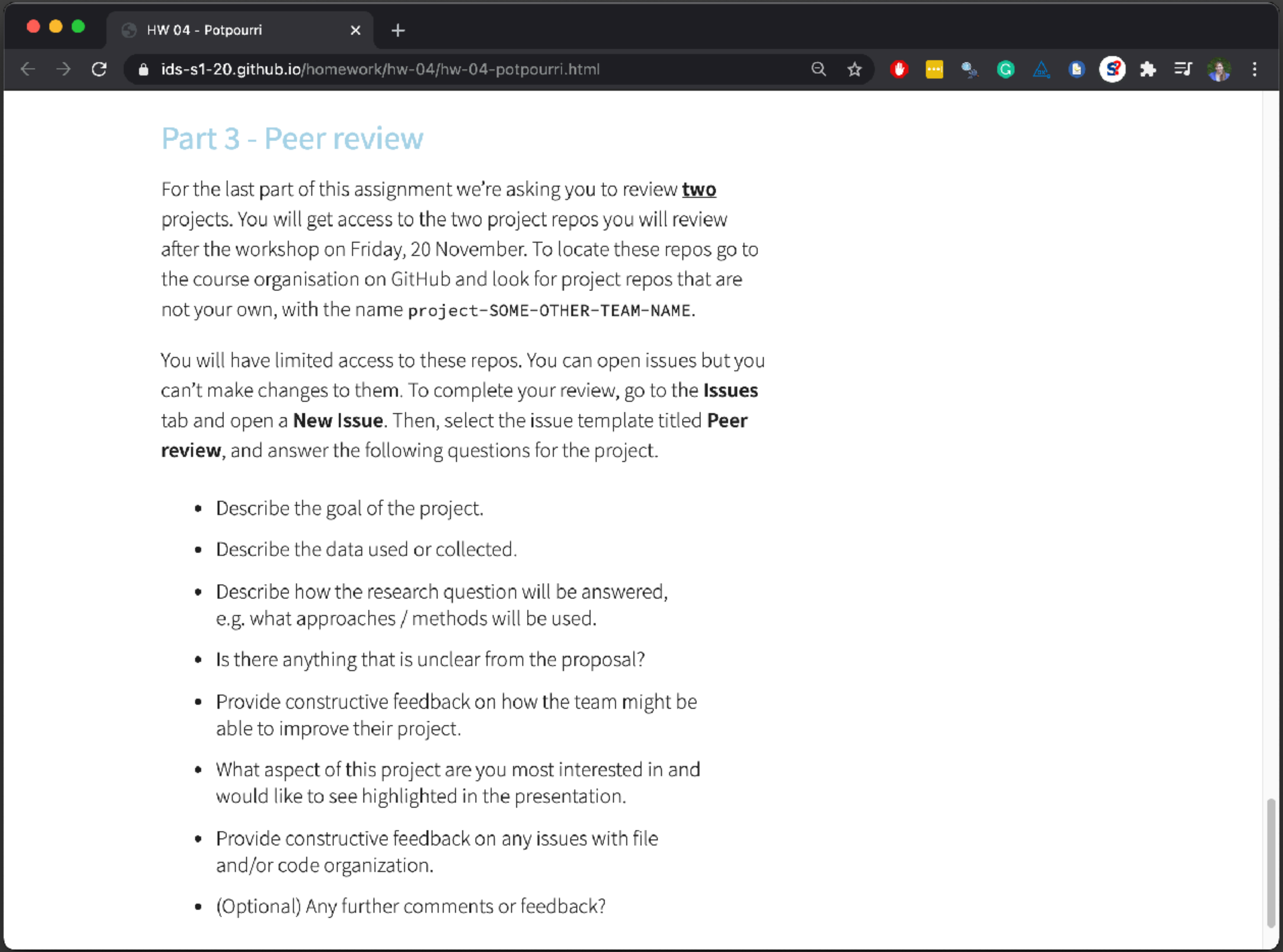
Exercise 1. Make this plot as ugly as possible by changing colours, background color, fonts, or anything else you can think of. You will probably want to play around with theme options, but you can do more. You can also search online for other themes, fonts, etc. that you want to tweak. Try to make it as ugly as possible, the sky is the limit!

Part 2 - You gotta pick a package

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.

- ✓ repetition
- ✓ reflection
- ✓ creativity

tips



Part 3 - Peer review

For the last part of this assignment we're asking you to review **two** projects. You will get access to the two project repos you will review after the workshop on Friday, 20 November. To locate these repos go to the course organisation on GitHub and look for project repos that are not your own, with the name `project-SOME-OTHER-TEAM-NAME`.

You will have limited access to these repos. You can open issues but you can't make changes to them. To complete your review, go to the **Issues** tab and open a **New Issue**. Then, select the issue template titled **Peer review**, and answer the following questions for the project.

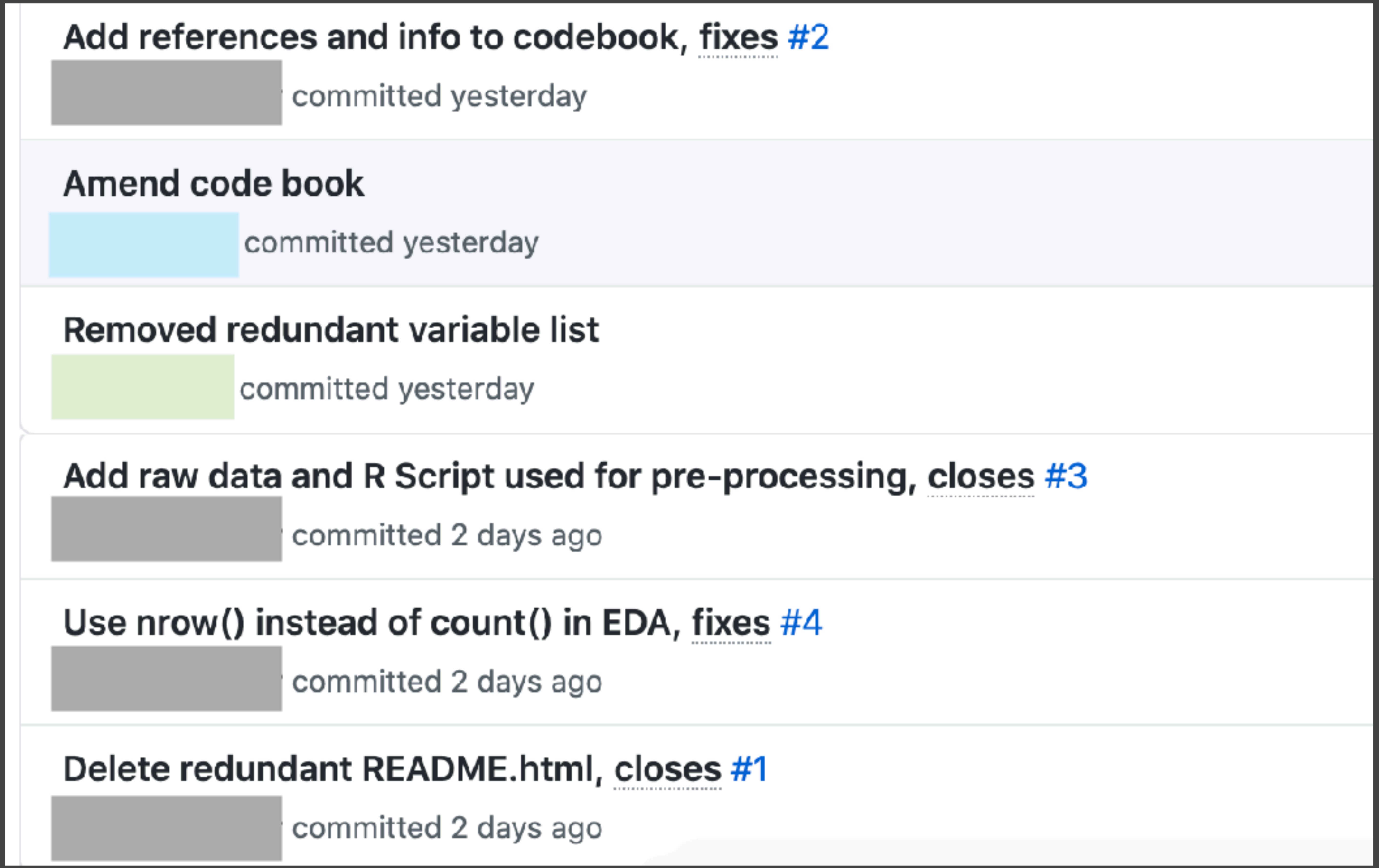
- Describe the goal of the project.
- Describe the data used or collected.
- Describe how the research question will be answered, e.g. what approaches / methods will be used.
- Is there anything that is unclear from the proposal?
- Provide constructive feedback on how the team might be able to improve their project.
- What aspect of this project are you most interested in and would like to see highlighted in the presentation.
- Provide constructive feedback on any issues with file and/or code organization.
- (Optional) Any further comments or feedback?

✓ reflection

✓ creativity

✓ peer review

tips



Add references and info to codebook, fixes #2
[redacted] committed yesterday

Amend code book
[redacted] committed yesterday

Removed redundant variable list
[redacted] committed yesterday

Add raw data and R Script used for pre-processing, closes #3
[redacted] committed 2 days ago

Use nrow() instead of count() in EDA, fixes #4
[redacted] committed 2 days ago

Delete redundant README.html, closes #1
[redacted] committed 2 days ago

✓ repetition

✓ reflection

✓ creativity

✓ peer review

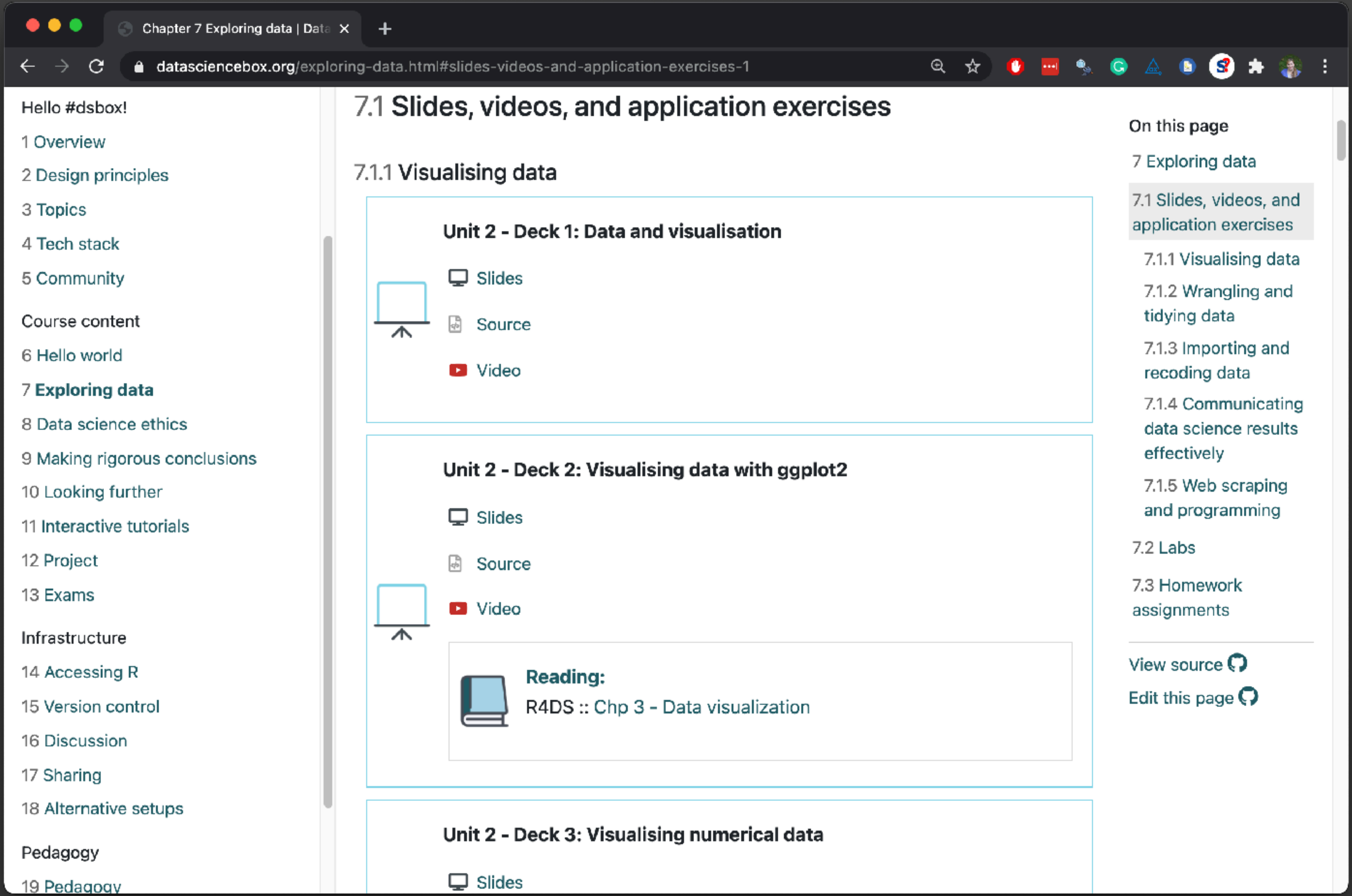
✓ real workflows

tips

RStudio
tidyverse
www.rstudio.com
tidymodels
www.rstudio.com
rmarkdown
www.rstudio.com

student toolbox

instructor toolbox



Chapter 7 Exploring data | Data
datasciencebox.org/exploring-data.html#slides-videos-and-application-exercises-1

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.

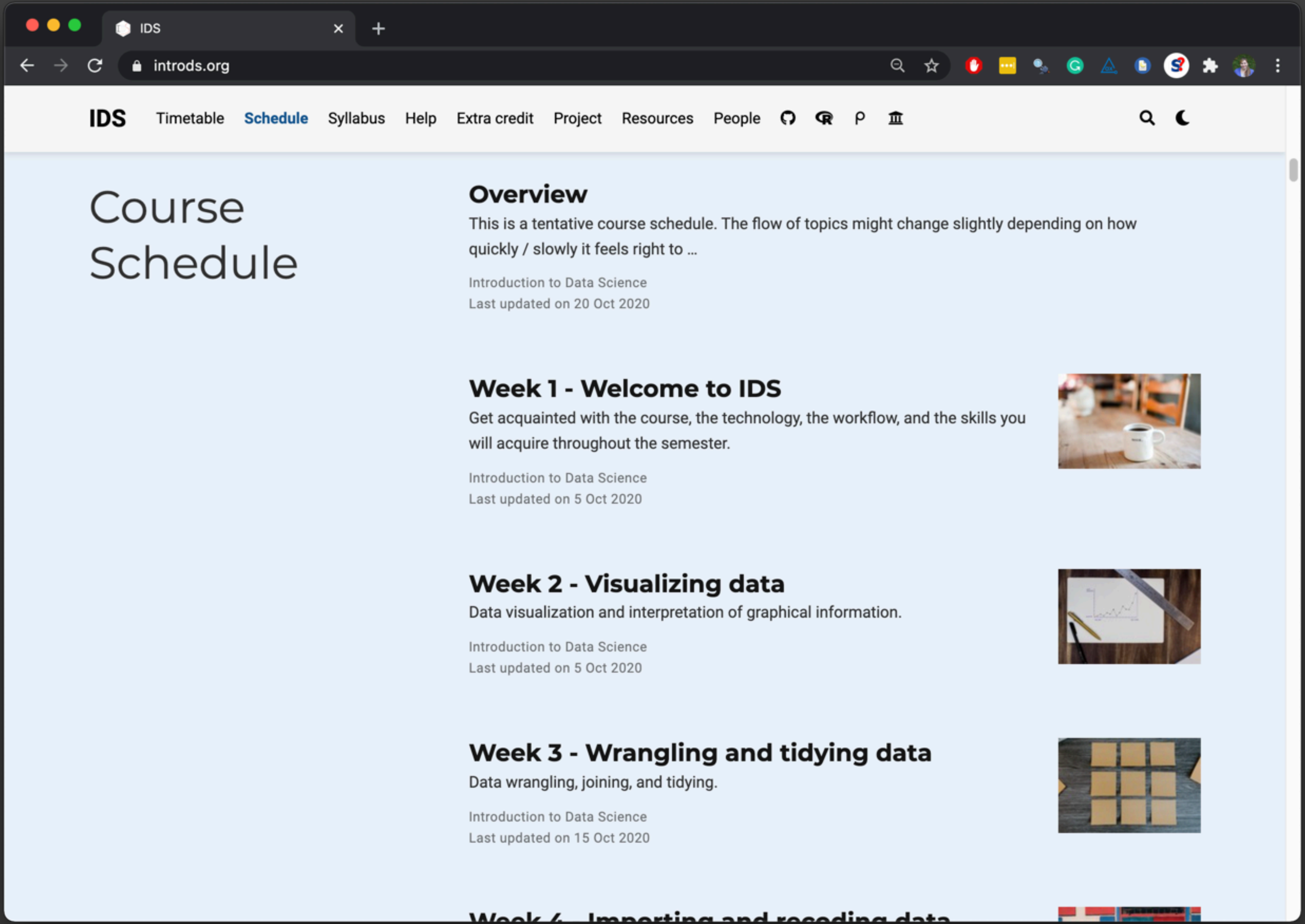
© 2020 The Author(s). Published with license by Taylor and Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the named author(s) have been asserted.

Mine Çetinkaya-Rundel &
Victoria Ellison (2020)

A Fresh Look at Introductory Data Science

Journal of Statistics Education

DOI: 10.1080/10691898.2020.1804497



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/fresh-ds-jmm

🔗 datasciencebox.org

🐦 minebocek

mine-cetinkaya-rundel

✉ cetinkaya.mine@gmail.com