introductory data science
a fresh look
mine çetinkaya-rundel

bit.ly/fresh-ds-jmm
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
data visualisation
data wrangling, tidying, acquisition
exploratory data analysis
predictive modeling + uncertainty quantification
effective communication of results

focus on

interactive visualizations
text analysis
machine learning
Bayesian inference
...

foray into

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

emphasise
Exploring data

Visualize

Data science ethics

Making rigorous conclusions

Hello world

Import

Wrangle

Algorithmic bias

Data privacy

Model

Infer

Predict

Looking further
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
- 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
‣ 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
ex. 2

fisheries of the world
```r
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"
```
```r
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
Average share of aquaculture by continent out of total fisheries harvest, 2016

Asia: 40%  
Americas: 25%  
Europe: 20%  
Oceania: 15%  
Africa: 10%

Source: bit.ly/2VrawTt

✓ data joins  
✓ ethics  
✓ critique  
✓ improving visualisations
✓ data joins
✓ ethics
✓ critique
✓ improving
✓ visualisations
✓ mapping
First Minister's COVID briefings
First Minister's speeches

Speeches delivered by the First Minister Nicola Sturgeon.

On this page:

- 2020
  - Coronavirus (COVID-19) update: First Minister's speech 26 October 2020
- 2019
  - Coronavirus (COVID-19) update: First Minister's speech 23 October 2019
- 2018
  - Coronavirus (COVID-19) update: First Minister's speech 22 October 2018
- 2017
  - Coronavirus (COVID-19) update: First Minister's speech 21 October 2017
- 2016
  - Coronavirus (COVID-19) update: First Minister's speech 20 October 2016
  - Coronavirus (COVID-19) update: First Minister's speech 19 October 2016
  - Coronavirus (COVID-19) update: First Minister's speech 16 October 2016
  - Coronavirus (COVID-19) update: First Minister's speech 15 October 2016
  - Coronavirus (COVID-19) update: First Minister's speech 14 October 2016
  - Coronavirus (COVID-19) update: First Minister's speech 13 October 2016
  - Coronavirus (COVID-19) update: First Minister's speech 12 October 2016
  - Coronavirus (COVID-19) update: First Minister's speech 9 October 2016
robotstxt::paths_allowed("https://www.gov.scot/")

www.gov.scot

[1] TRUE
Coronavirus (COVID-19) update: First Minister's speech 26 October

Statement given by First Minister Nicola Sturgeon at a media briefing in St Andrew's House on Monday 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, 224 in Lanarkshire, 165 in Lothian and...
- **ethics**
- **web scraping**
- **text parsing**
- **data types**
- **regular expressions**
- **functions**
- **iteration**
- **visualisation**
- **interpretation**
Common words in COVID briefings

- aberdeen
- measurement
- lanarkshire
- clyde
- cluster
- decrease
- lothian
- relates
- scotland's
- clusters
- o'clock
- hopefully
- surfaces
- ayrshire
- records

Scotlands:

- slide
- defeat
- sir
- ventilators
- speaker
- fatalities
- downing
- doctors
- mechanical
- adjusting
- transcript
- nightingale
- alas
- cyber
- doses
- roadmap

UK:

- 0.0000
- 0.00015
- 0.00030

tf-idf

- ethics
- web scraping
- text parsing
- data types
- regular expressions
- functions
- iteration
- visualisation
- interpretation
- text analysis
ex. 3
spam filters
✓ logistic regression
✓ prediction

Spam vs. number of characters

2K chars, P(spam) = 0.13
15K chars, P(spam) = 0.06
40K chars, P(spam) = 0.01
<table>
<thead>
<tr>
<th></th>
<th>Email is spam</th>
<th>Email is not spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email labelled spam</td>
<td>True positive</td>
<td>False positive (Type 1 error)</td>
</tr>
<tr>
<td>Email labelled not spam</td>
<td>False negative (Type 2 error)</td>
<td>True negative</td>
</tr>
</tbody>
</table>

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

Visualizing

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

```
1 ggplot(data = ____, aes(x = ____ , ____ = ____ ) ) +
2 geom____(____ ) +
3 ____ ) +
4 ____ (v = ____ , 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.
1. Write about one or two questions you didn’t like today. What was difficult about them? Write your answer below.

2. If you got every single question correct, did you feel like you were doing well on the topic? Use your answer above to justify your position.

3. You can use more than one format to clarify the questions and answers, such as a list or an article. Quick reflect on your learning.

4. x-axis

# 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
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?
<table>
<thead>
<tr>
<th>Task</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add references and info to codebook, fixes #2</td>
<td>committed yesterday</td>
</tr>
<tr>
<td>Amend code book</td>
<td>committed yesterday</td>
</tr>
<tr>
<td>Removed redundant variable list</td>
<td>committed yesterday</td>
</tr>
<tr>
<td>Add raw data and R Script used for pre-processing, closes #3</td>
<td>committed 2 days ago</td>
</tr>
<tr>
<td>Use nrow() instead of count() in EDA, fixes #4</td>
<td>committed 2 days ago</td>
</tr>
<tr>
<td>Delete redundant README.html, closes #1</td>
<td>committed 2 days ago</td>
</tr>
</tbody>
</table>
DATA SCIENCE IN A BOX
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
A Fresh Look at Introductory Data Science

Mine Çetinkaya-Rundel & Victoria Ellison

1. Introduction

How can we effectively and efficiently teach data science to students with little or no background in computing or statistical thinking? How can we equip them with the skills and tools for reasoning with various types of data and data-driven insights? This article describes an introductory data science course that is well-suited to these goals.

As in any course of its kind, the course focuses on data acquisition and management, exploratory data analysis, data visualization, inferential statistical modeling, and effective communication of results. To this end, the course also provides various hands-on exercises to help students learn, practice, and master the tools used in this course.

In Section 3, we describe an overview of the course curriculum and introduce students to the concepts of data science. In Section 4, we introduce students to the concepts of data science and show how they can be used to understand and analyze data. In Section 5, we describe how to use data science tools and techniques to analyze data. In Section 6, we introduce students to the concepts of data science and show how they can be used to understand and analyze data. In Section 7, we describe how to use data science tools and techniques to analyze data.
Course Schedule

Overview
This is a tentative course schedule. The flow of topics might change slightly depending on how quickly/slower 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 3 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 reading data

---

introds.org