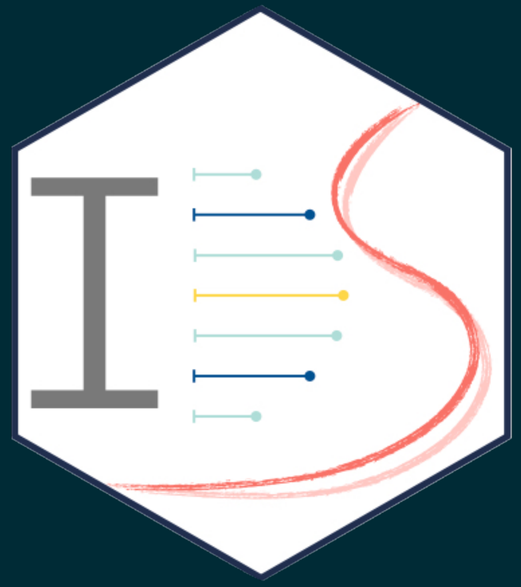
Wrapping up



What did you learn?

1. Welcome to IDS
2. Visualising data
3. Wrangling and tidying data
4. Importing and recoding data
5. Communicating data science results effectively
6. Web scraping and programming
7. Data science ethics
8. Modelling data
9. Classification and model building
10. Model validation and uncertainty
11. Text analysis, interactive web apps, and machine learning

What did I learn?

you asked...

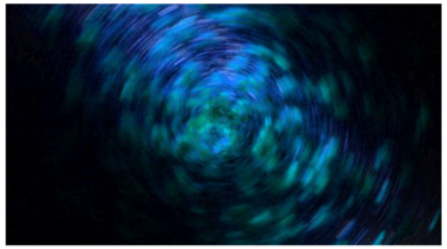
"What kind of research might someone do in the field of data science?"

just one example...

turing.ac.uk/research/research-programmes/data-science-science

Programme projects

Masking clouds in satellite imagery →
Using machine learning techniques to classify cloudy pixels from satellite images to aid more accurate monitoring of remotely sensed environmental and climate variables



Neutron scattering and machine learning →
Using neural networks to interpret magnetic structure from inelastic neutron scattering experiments

Diffuse multiple scattering analysis →
Analysing complex multi-phase materials using X-ray scattering and machine learning

Radiation detectors and machine learning →
Developing a machine learning approach to discriminating between different types of radiation in cutting edge scintillator detectors

Data science for climate resilience in East Africa →
Combining satellite and social data to measure and grow the impact of tree planting schemes in East Africa

Risk and uncertainty in peatland carbon emissions →
Assessing and managing the risk of abrupt greenhouse gas emissions from peatlands

you asked...

"Before the US election, the polls indicated that Biden will win by a landslide but after the results it was not as big as polls suggested. Is there a reason why this happens? Should we rely on polls before elections?"

Comparing election outcomes to our forecast and to the previous election

Posted by Andrew on 6 November 2020, 9:16 am

by Andrew Gelman and Elliott Morris

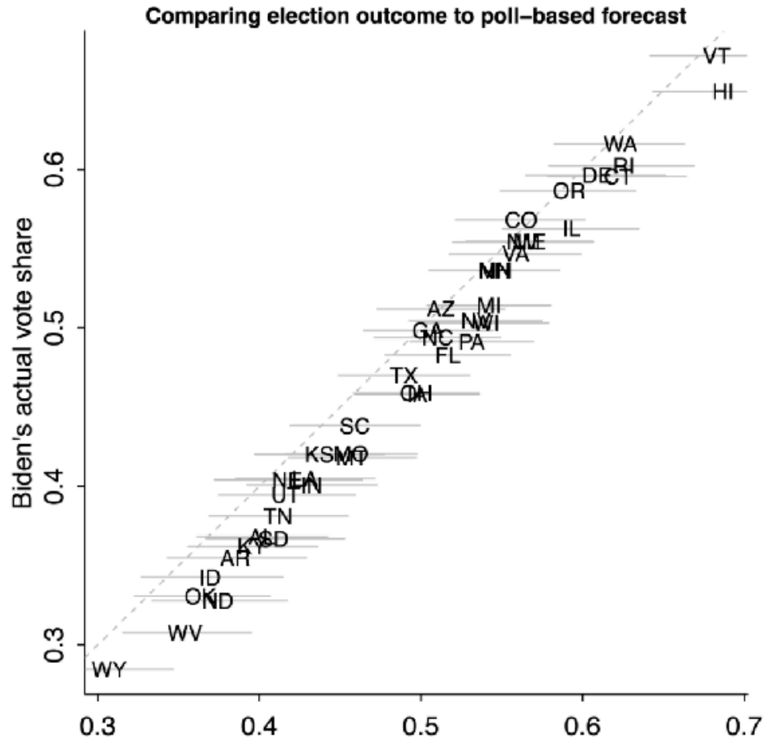
Now that we have almost all the votes from almost all the states, we can step back and answer two questions:

1. How far off were our predictions?

2. How did Joe Biden's performance compare to Hillary Clinton's four years earlier?

How far off were our predictions?

Here's what we have so far:



Search

RECENT COMMENTS

› Anoneuoid on Understanding Janet Yellen

› Anoneuoid on Understanding Janet Yellen

› Anonymous on Unfair to James Watson?

› Joseph Delaney on Understanding Janet Yellen

› David J. Littleboy on Understanding Janet Yellen

› Dave C. on Unfair to James Watson?

› David J. Littleboy on Understanding Janet Yellen

› Joshua on Understanding Janet Yellen

› Azar on Considerate Swedes only die during the week.

› Ewan Cameron on Are female scientists worse mentors? This study pretends to know

› Commenter on Understanding Janet Yellen

› Kien on Basbøll's Audenesque paragraph on science writing, followed by a resurrection of a 10-year-old debate on Gladwell

› John N-G on Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond

statmodeling.stat.columbia.edu/2020/11/06/comparing-election-outcomes-to-our-forecast-and-to-the-previous-election

Can we stop talking about how we're better off without election forecasting?

Posted by Jessica Hullman on 12 November 2020, 3:08 pm

This is a public service post of sorts, meant to collect some reasons why getting rid of election forecasts is a non-starter in one place.

First to set context: what are the reasons people argue we should give them up? This is far from an exhaustive list (and some of these reasons overlap) but a few that I've heard over the last week are:

- If the polls are right, we don't need forecasters. If polls are wrong, we don't need forecasters.
- Forecasts are hard to evaluate, therefore subject to influences of the forecaster's goals, e.g. to not appear too certain that they can be blamed. Hence we can't trust them as unbiased aggregations of evidence.
- Forecasters may have implicit knowledge from experience, such as a sense of approximately what the odds should be, but it's hard to transparently and systematically incorporate that knowledge. When a forecaster 'throws in a little error here, throws in a little error there' to get the uncertainty they want at a national level, they can end up with model predictions that defy common sense in other ways, calling into question how coherent the predictions are. The ways we want forecasts to behave may sometimes conflict with probability theory.
- There's too much at stake to take chances on forecasts that may be wrong but influence behavior.

I don't think these questions are unreasonable. But it's worth considering the implications of a suggestion that forecasts have no clear value, or even do more harm than good, since I suspect some people may jump to this conclusion without recognizing the subtext it entails. Here are some things I think of when I hear people questioning the value of election forecasts:

#1 – A carefully constructed forecast is (very likely to be) better than the alternative. Or to quote a Bill James line Andrew has used, "The alternative to good statistics is not no statistics, it's bad statistics."

What would happen if there were no professional forecasts from groups like the Economist team or professional forecasters like Nate Silver? A deep stillness as we all truly acknowledge the uncertainty of the situation does not strike me as the most likely scenario. Instead, people may look to the sorts of overreactions to polls that we already see in the media to tell them what will happen, without referring back to previous elections. Or maybe they anxiously query friends and neighbors (actually there's probably some valuable information there, but only if we aggregate across people!), or extrapolate from the attention paid to candidates on public television, or how many signs they see in nearby yards or windows, or examine tea leaves, look at entrails of dead animals, etc.

One alternative that already exists is prediction markets. But it's hard to argue that they are more accurate than a carefully constructed forecast. For instance, it's not clear we can really interpret the prices in a market as aggregating information about win probabilities in any straightforward way, and there's reasons to think they don't make the best use of new data. They can produce strange predictions too at times, like giving Trump a >10% chance of winning Nevada even after it's been called by some outlets.

Even in seemingly "extreme" cases like 2016 or 2020, where bigger than anticipated poll errors led to forecasts seeming overconfident about

Search

RECENT COMMENTS

› John N-G on Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond

› Andrew on Understanding Janet Yellen

› Renzo Alves on Understanding Janet Yellen

› Curious on Unfair to James Watson?

› Andrew on Unfair to James Watson?

› rm bloom on Unfair to James Watson?

› Bradley Stiritz on Unfair to James Watson?

› Andrew on Unfair to James Watson?

› Joshua on Unfair to James Watson?

› Poincare on Unfair to James Watson?

› rm bloom on "We've got to look at the analyses, the real granular data. It's always tough when you're looking at a press release to figure out what's going on."

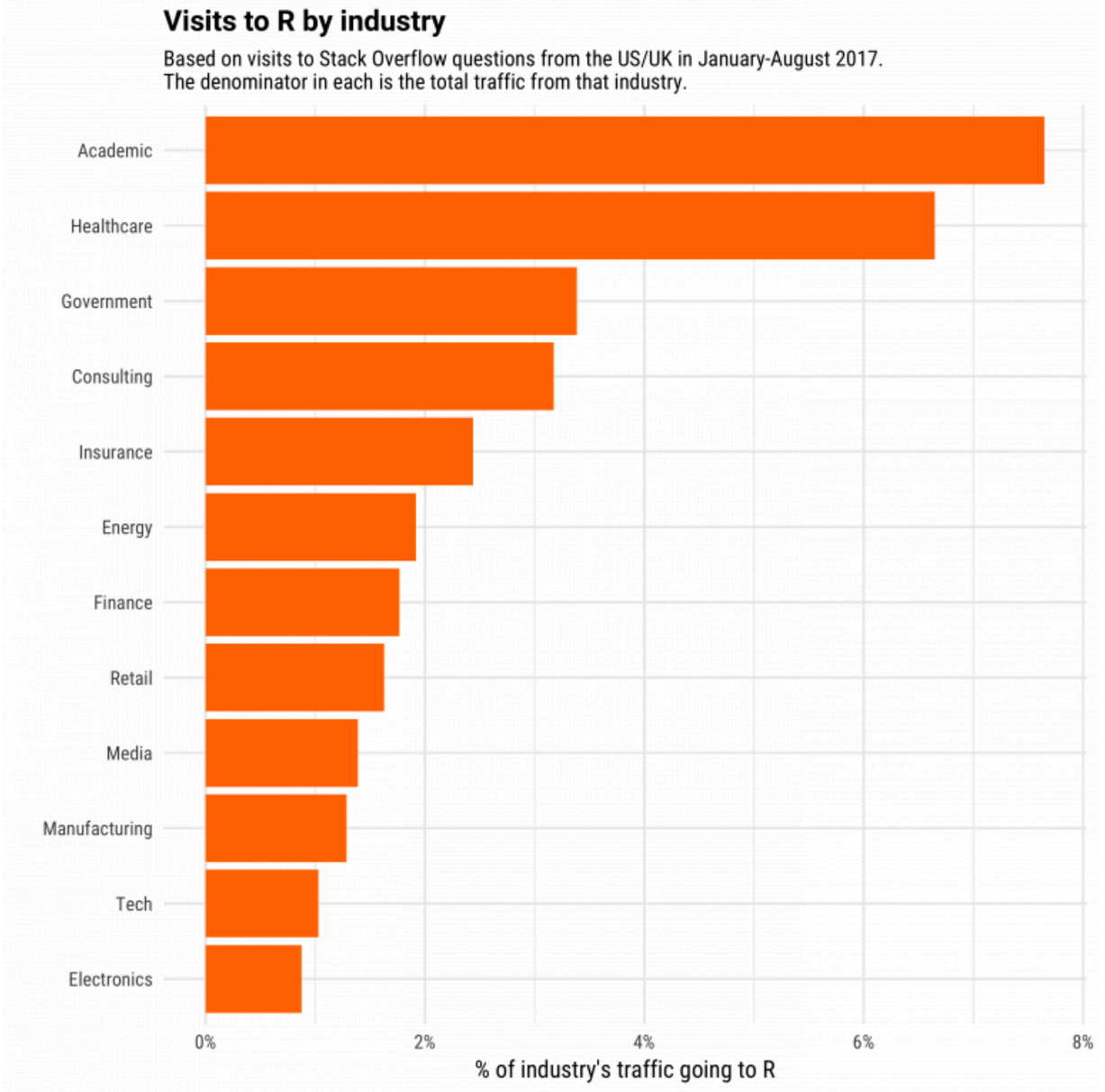
› Curious on Unfair to James Watson?

› Nitpicker general on Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent

statmodeling.stat.columbia.edu/2020/11/12/can-we-stop-talking-about-how-were-better-off-without-election-forecasting

you asked...

"Which area I can use R in my future?"

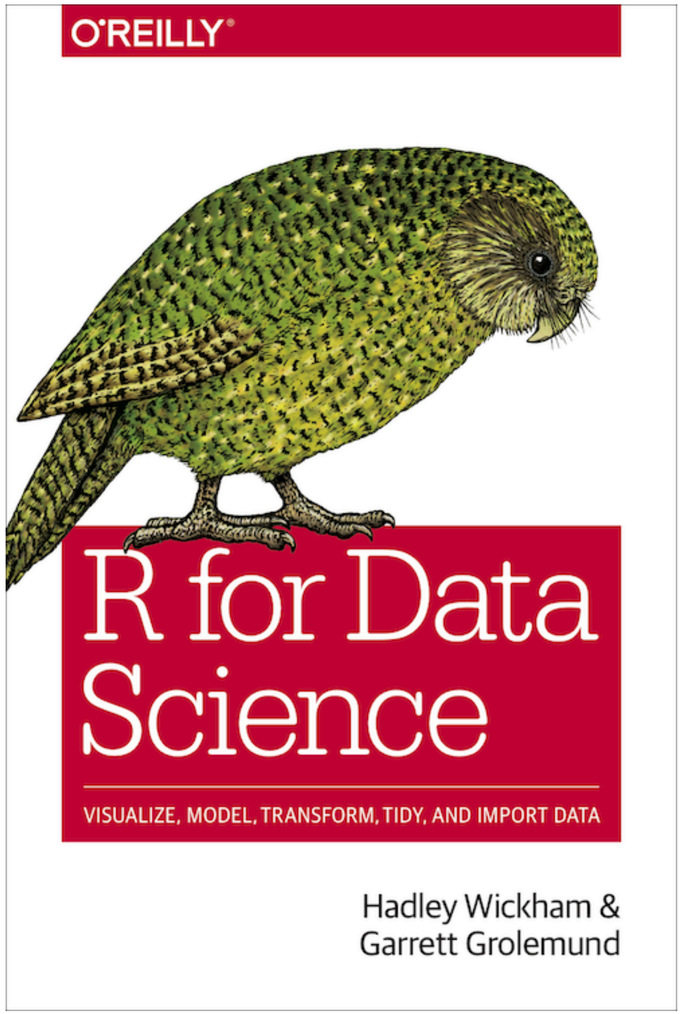


stackoverflow.blog/2017/10/10/impressive-growth-r

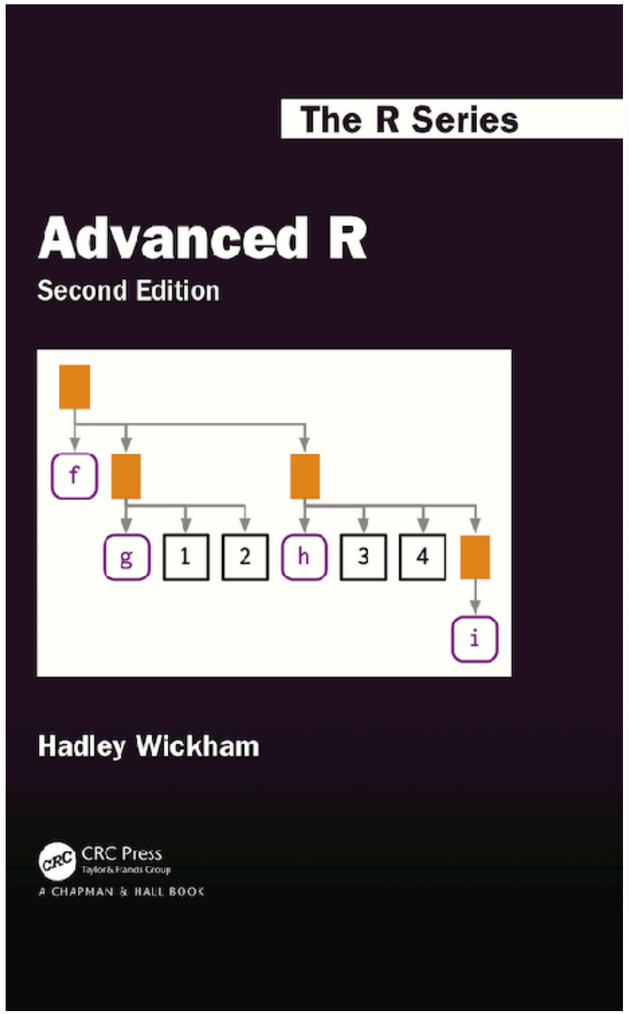
you asked...

"How can I improve my data science skills on my own? Any recommendations for websites, books etc.?"

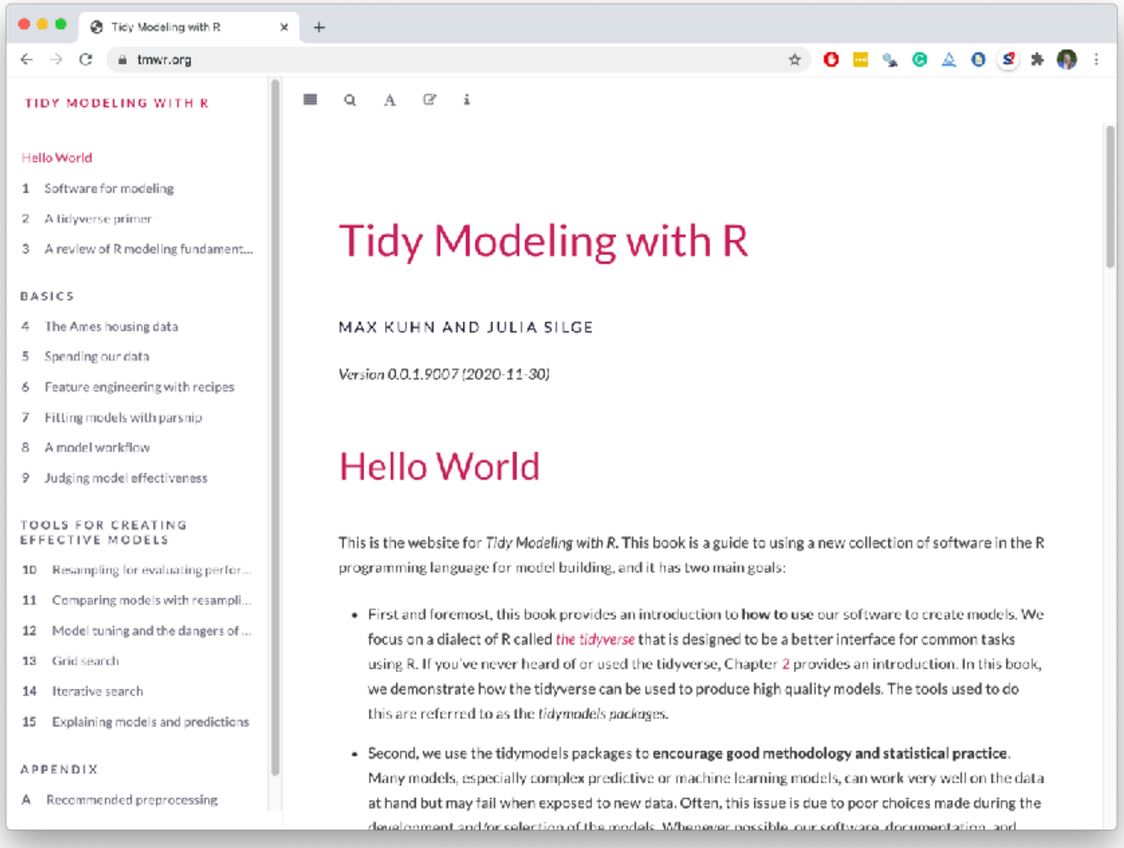
r4ds.had.co.nz

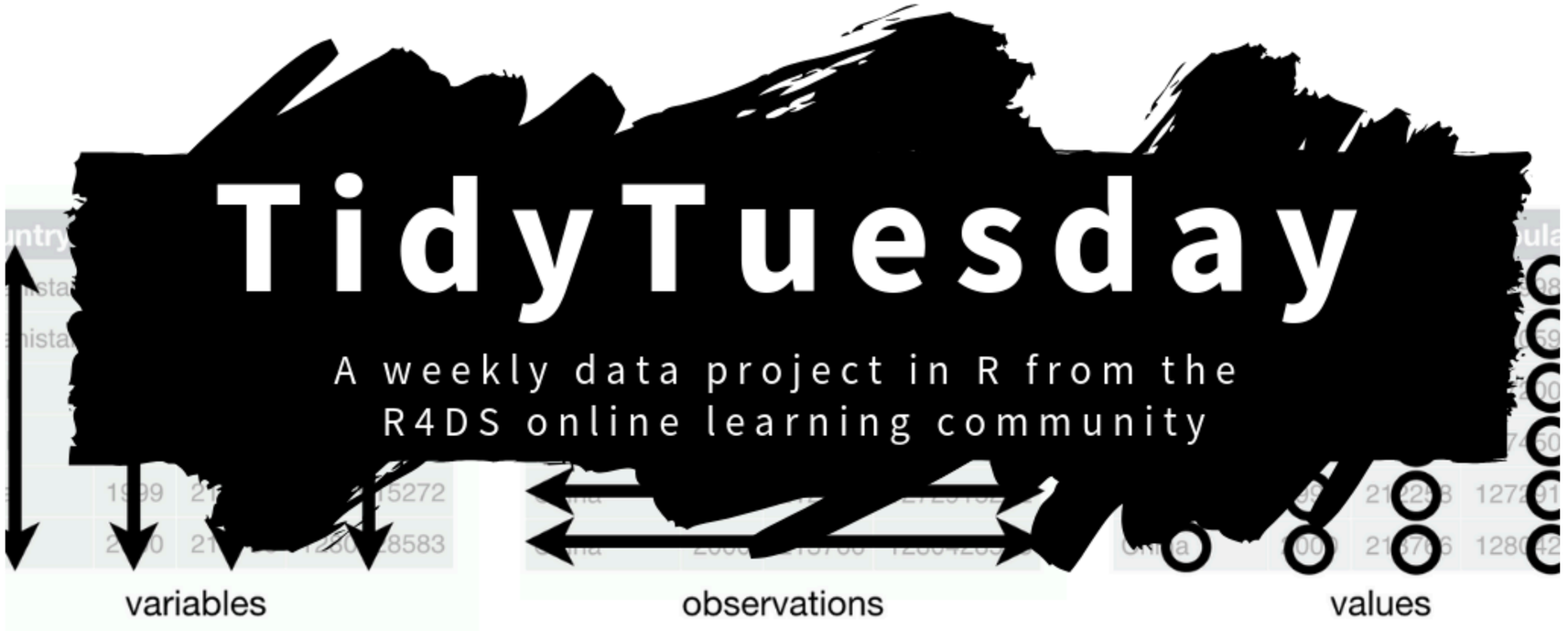


adv-r.hadley.nz



tmwr.org




TidyTuesday
A weekly data project in R from the
R4DS online learning community
variables
observations
values

rweekly.org

Live

- Advent of 2020, Day 2 – How to get started with Azure Databricks (tomaztsql.wordpress.com)
- Introducing ggrgl - a 3d extension to ggplot (coolbutuseless.github.io)
- Advent of Code 2020-02 with R & JavaScript (colinfay.me)

More

R Weekly 2020-48 Your first R package, magrittr, engineering Shiny

30 Nov 2020

type to filter

Release Date: 2020-11-30

This week's release was curated by Maëlle Salmon, with help from the R Weekly team members and contributors. (twitter.com)

- How to have (my) content shared by R Weekly? (github.com)

Highlights

- Your first R package in 1 hour (pipinghotdata.com)
- magrittr 2.0 is here! (tidyverse.org)
- Use case from "Engineering Production-Grade Shiny Apps" - Building an App, from Start to Finish (engineering-shiny.org)

you asked...

"I really enjoyed taking data science this semester, are there any other courses I can take semester 2 or next year that develop on what we've been learning?"

Undergraduate Course: Statistics (Year 2) (MATH08051)

Course Outline			
School	School of Mathematics	**College**	College of Science and Engineering
Credit level (Normal year taken)	SCQF Level 8 (Year 2 Undergraduate)	**Availability**	Available to all students
SCQF Credits	10	**ECTS Credits**	5
Summary	This course provides an introduction to the basic concepts of Statistics. The underlying rigorous mathematical framework is provided for analyzing different forms of data and the interpretation of the corresponding results discussed in detail. By the end of the course, students will be able to estimate parameter values for different statistical models and conduct a range of hypothesis tests.		
Course description	Topics will include : Sampling distributions Estimators, including MLEs Interval estimation Hypothesis testing Regression Analysis of Variance (ANOVA) In addition the statistical package R will be introduced and used within the course.		

Undergraduate Course: Statistical Computing (MATH10093)

Course Outline			
School	School of Mathematics	**College**	College of Science and Engineering
Credit level (Normal year taken)	SCQF Level 10 (Year 3 Undergraduate)	**Availability**	Available to all students
SCQF Credits	10	**ECTS Credits**	5
Summary	This course provides an introduction to programming within the statistical package R. Various computer intensive statistical algorithms will be discussed and their implementation in R will be investigated.		
Course description	Topics to be covered include : - basic commands of R (including plotting graphics); - data structures and data manipulation; - writing functions and scripts; - optimising functions in R; and - programming statistical techniques and interpreting the results (including bootstrap algorithms).		

Undergraduate Course: Statistical Learning (MATH10094)

Course Outline			
School	School of Mathematics	**College**	College of Science and Engineering
Credit level (Normal year taken)	SCQF Level 10 (Year 4 Undergraduate)	**Availability**	Available to all students
SCQF Credits	10	**ECTS Credits**	5
Summary	This course will give an introduction to modern machine learning, from a statistical perspective. NB. This course is normally delivered *biennially* but the next instance in 2020-21 has been cancelled.		
Course description	Likely topics include:- - supervised and unsupervised learning - classification - regression - discriminant analysis - regularisation - support vector machines - deep learning - and-random forests		

Undergraduate Course: Informatics 1 - Cognitive Science (INFR08020)

Course Outline			
School	School of Informatics	College	College of Science and Engineering
Credit level (Normal year taken)	SCQF Level 8 (Year 1 Undergraduate)	Availability	Available to all students
SCQF Credits	20	ECTS Credits	10
Summary	This course is designed as a first introduction to Cognitive Science. It will provide a selective but representative overview of the subject, suitable for all interested students, including students on the Cognitive Science degrees and external students. The aim of the lecturing team is to present a unified view of the field, based on a computational approach to analysing cognition. The material is organized by cognitive function (e.g., language, vision), rather than by subdiscipline (e.g., psychology, neuroscience). The course covers language, vision and attention, memory, motor control and action, and reasoning and generalization. All topics will be presented from a computational point of view, and this perspective will be reinforced by lab sessions in which students use implementations of cognitive models. The course will also provide a basic grounding in the methods of Cognitive Science, focusing on computational modelling and experimental design.		
Course description	Course Description The syllabus covers the following topics. They are listed separately here, but in some cases they will be presented in an interleaved fashion: 1. Language - the language faculty - models of linguistic data, words and rules theory - Connectionist models of language - language acquisition: speech segmentation, word learning, learning syntactic categories - categorization and models of word meaning - understanding sentences		

projects...

CfS spotlight

Innovative Teaching in the Time of COVID-19

By co-developing online data science tutorials and autom
students, Mine Çetinkaya-Rundel from the Centre for Stat
undergraduates can still receive high-quality learning exp

With many university staff across the UK experiencing stress and extreme w
associated with moving teaching online, and a number of students compla
quality of their learning experiences during the pandemic, online teaching
subject of intense debate for everyone in university education.

A large part of what has made online teaching so challenging is the lack of
pandemic provided. Suddenly and urgently, teaching staff had to prepare a
classes online from home, introducing a host of practical and technical pro
are still grappling with even now.

Opportunity for innovation

Mine Çetinkaya-Rundel from the Centre for Statistics has delivered Introdu
Science, an introductory level course on data science and statistical thinkir
School of Mathematics, since she joined the University in 2019.

Not only did Mine face the same hurdles everyone in university teaching ha
she also had to deliver it to up to 300 students – three times as many as the
reduce the marking, as the core staff would not grow proportional to the nu

Mine immediately flipped this obstacle on its head, seeing it as "an opportu
to deliver the same amount of exercises and feedback to students without
realised that the key was to introduce as much automated feedback on stu

bit.ly/df-edi

DataFest 2020 Results

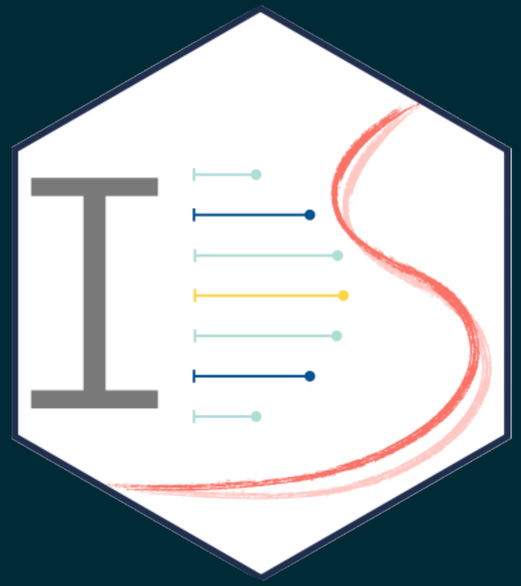
Eight teams participated in the DataFest @ EDI 2020 - COVID-19 Virtual Data Challenge over two weeks between 30 May and 13 June, 2020.

For this competition, we challenged participants to explore the societal impacts of the COVID-19 pandemic other than its direct health outcomes. Participants were allowed to explore everything from the effects on pollution levels, transportation levels, or working from home. They could investigate changes in the number of people posting on TikTok with their families or do an analysis on online education. We left the focus up to them and urged them to be thoughtful and creative as they analyzed data and communicated their insights about some of pandemic's impacts on society.

Specifically for this challenge the participants were asked to

- choose an outcome such as one of the ones listed above,
- find the appropriate data to explore the effect of the COVID-19 pandemic on this issue, and
- present their findings (along with the dataset they used).

THANK YOU!



introds.org