

Notes for data science students

Dr. Mine Dogucu

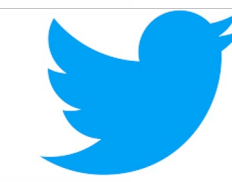
bit.ly/dogucu-talks

2024-04-04

HELLO
my name is



mdogucu@uci.edu



[@MineDogucu](https://twitter.com/MineDogucu)



[MineDogucu](https://www.linkedin.com/company/MineDogucu)



- Where are right now?
- What year are you (e.g. junior, senior, grad)?
- What has been your favorite course EVER?

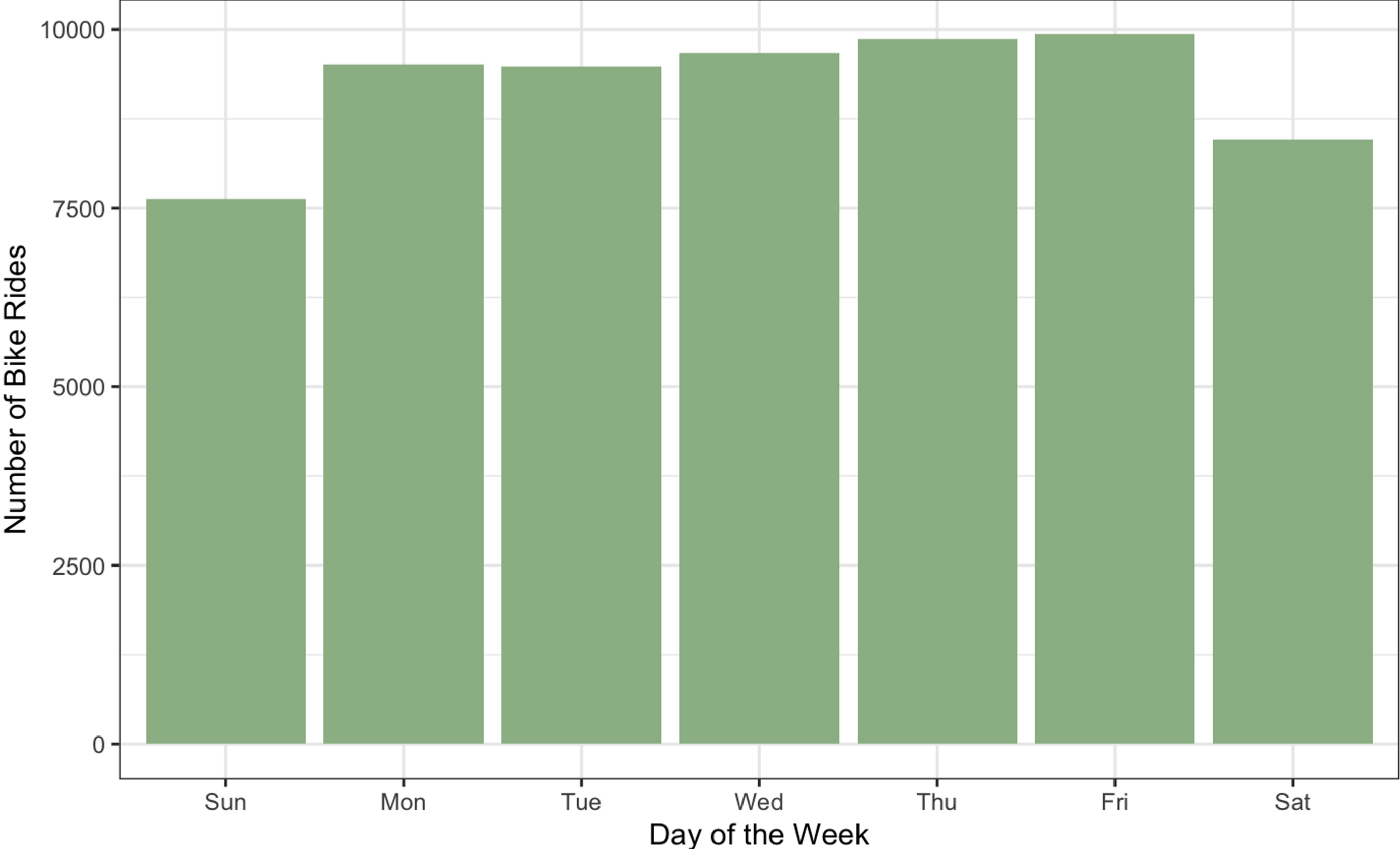
Introduction to Data Science

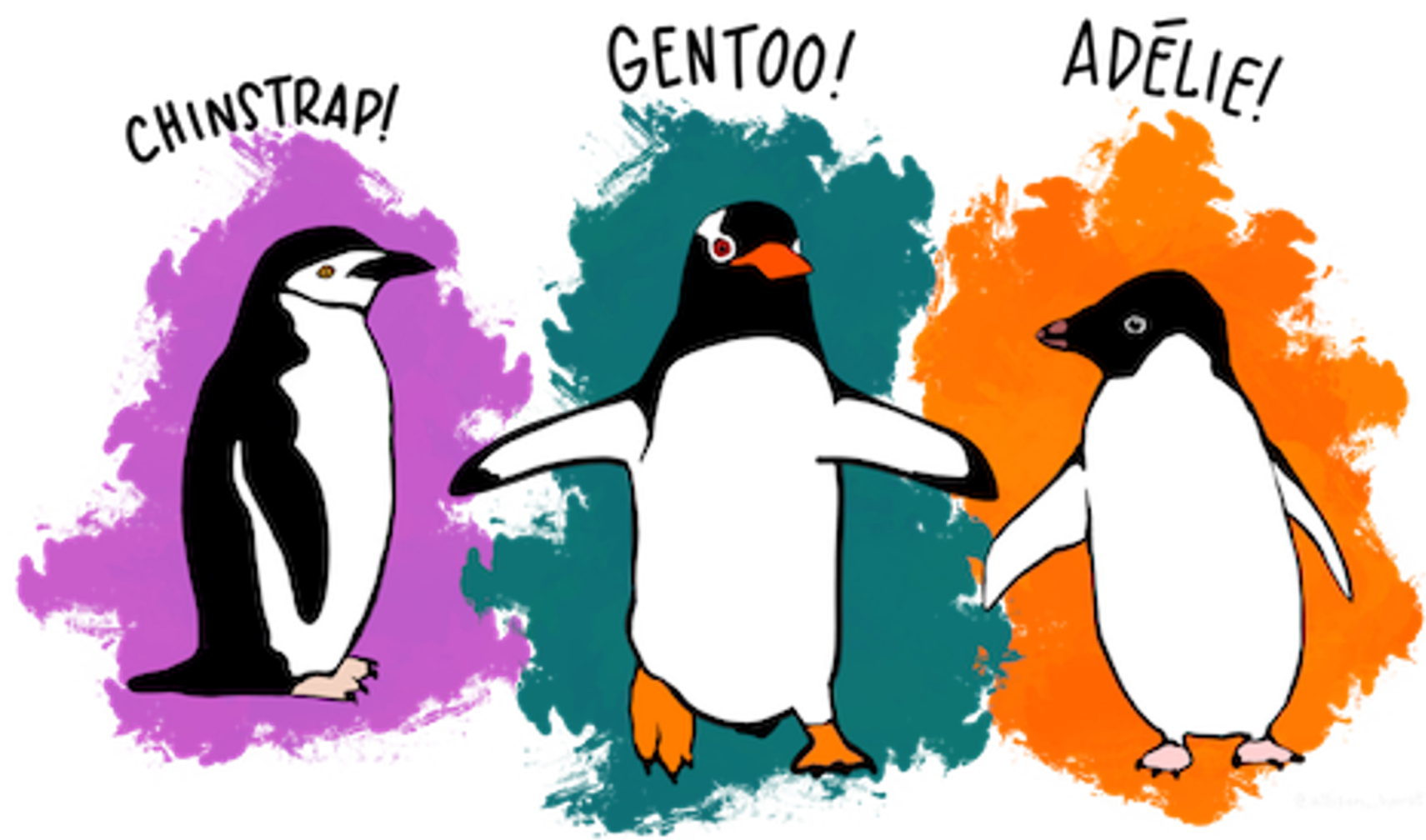
Course Goals:

- explore data using descriptive statistics and visualizations;
- read, write, and tidy up datasets;
- make predictions and conclusions using models;
- consider impact of decisions related to data on humans, other livings, and the planet;
- write human- and machine-readable code using R.

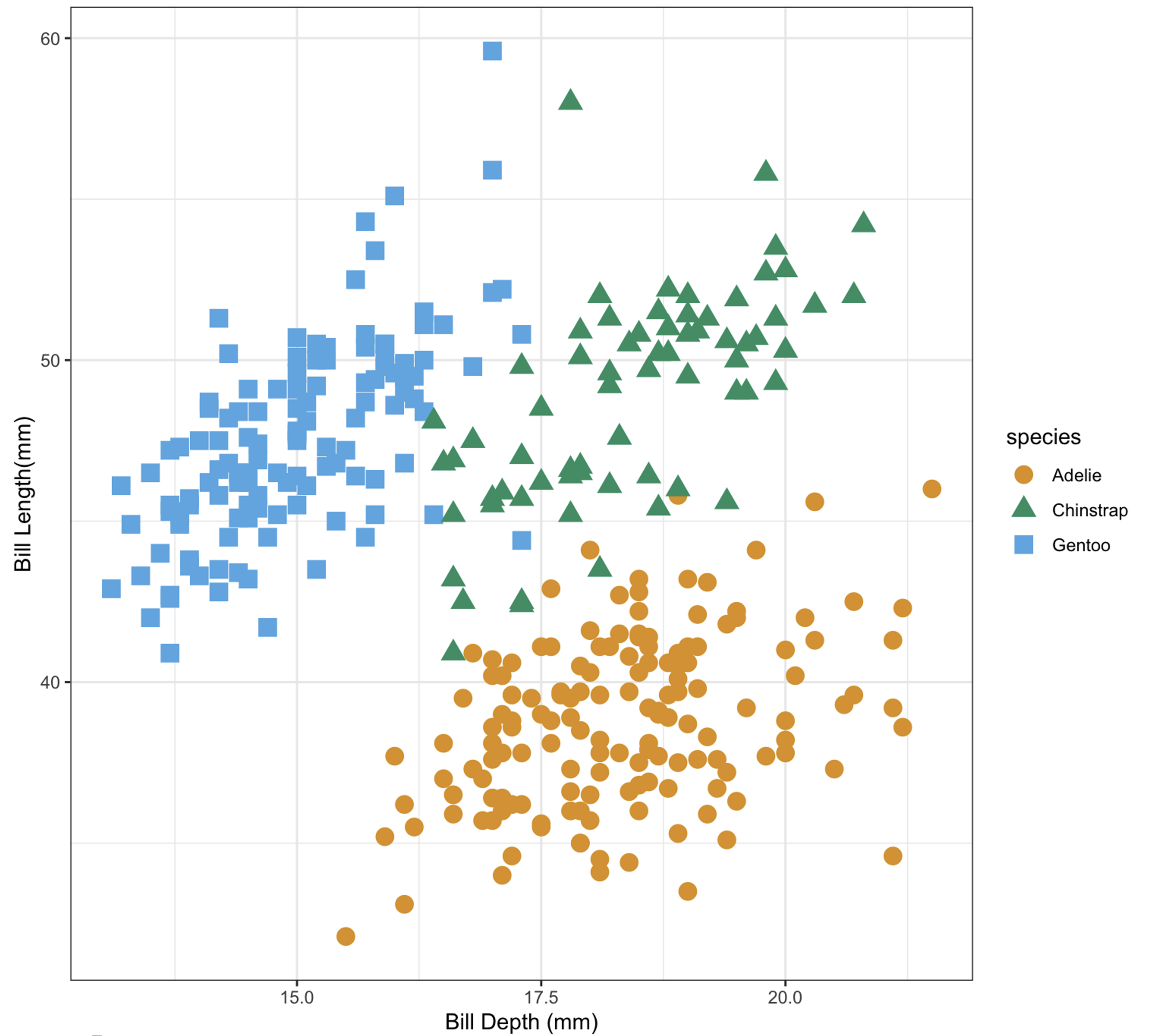


explore data using descriptive statistics and visualizations;





Artwork by [@allison_horst](#)



read, write, and tidy up datasets;

The screenshot shows the CDC website's navigation bar with links for 'Diseases & Conditions', 'Healthy Living', 'Travelers' Health', 'Emergency Preparedness', and 'More'. Below the navigation is a search bar with a magnifying glass icon and a link to 'Advanced Search'. The main content area is titled 'Data & Statistics' and features three featured articles:

- Age Without Injury**: You can take simple steps to prevent injuries, so you can stay healthy and independent longer.
- Injuries Among Children**: Child unintentional injury death rates decreased, but injury is still the leading cause of death.
- The National ALS Registry**: This October marks 11 years since the National ALS Registry was established. Learn more.

The screenshot shows the Los Angeles Open Data website. The main heading is 'LOS ANGELES OPEN DATA' in large white letters against a background of the city skyline at sunset. Below the heading is the text 'Explore the City of Los Angeles' Open Data'. At the bottom, there is a search bar with the placeholder text 'Search for Data'.

The screenshot shows the Data.gov website. The main heading is 'The home of the U.S. Government's open data'. Below the heading is the text 'Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and [more](#).' and 'For information regarding the Coronavirus/COVID-19, please visit [Coronavirus.gov](#).' At the bottom, there is a 'GET STARTED' button with the text 'SEARCH OVER 321,982 DATASETS' and a search bar containing the text 'Health Care Provider Charge Data'.



TOP 100

VIDEO GAMES OF ALL TIME

Start Over ↑

Comments

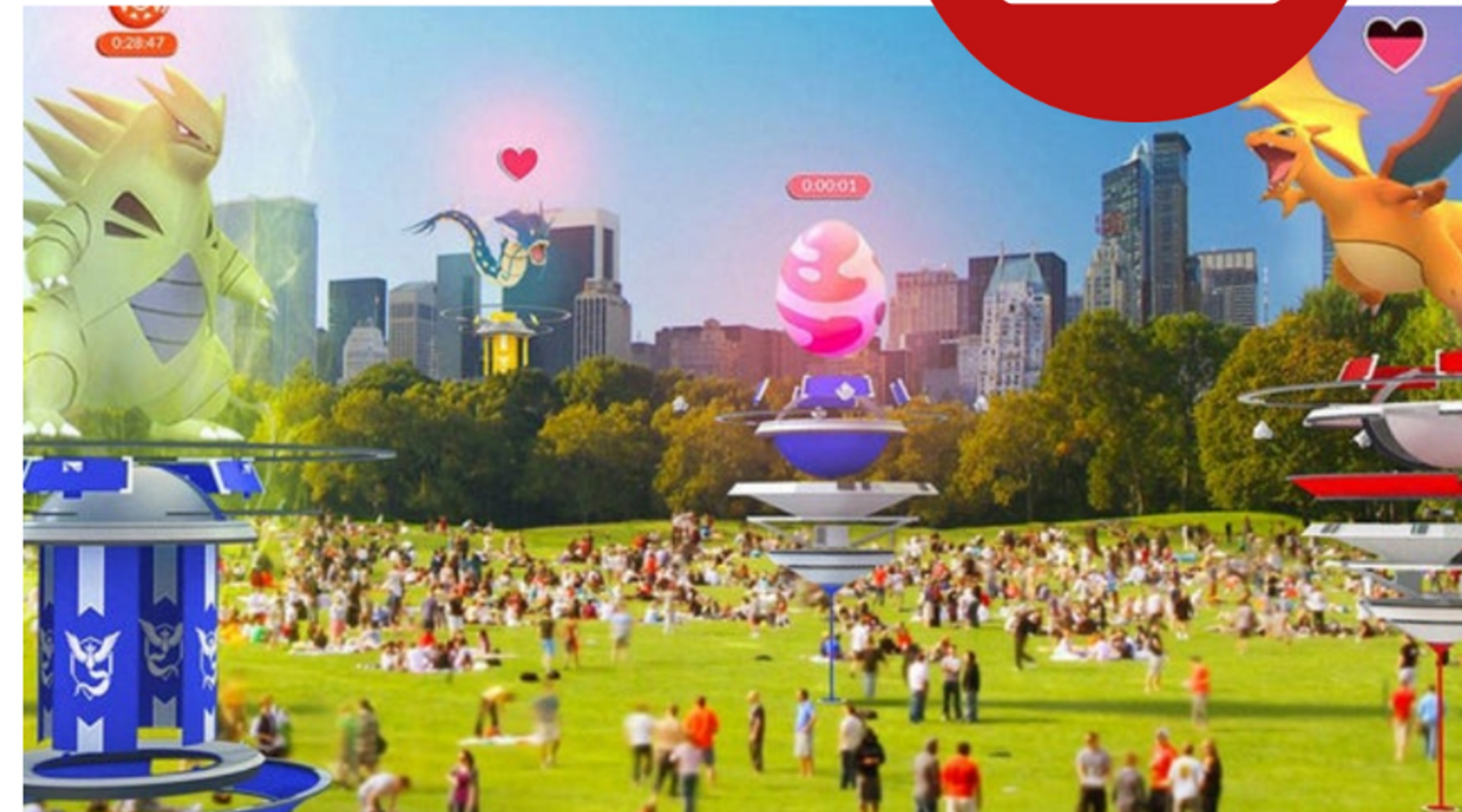
- 100 Pokémon Go
- 99 Borderlands 2
- 98 Divinity: Original Sin 2
- 97 Dishonored
- 96 Final Fantasy VII
- 95 Assassin's Creed IV: Black Flag
- 94 Monkey Island 2: LeChuck's Revenge
- 93 Burnout 3: Takedown

If you'd like to know more, check out this [list of changes to the Best 100 Games of All Time list -- and why we made those changes.](#)

Pokémon Go

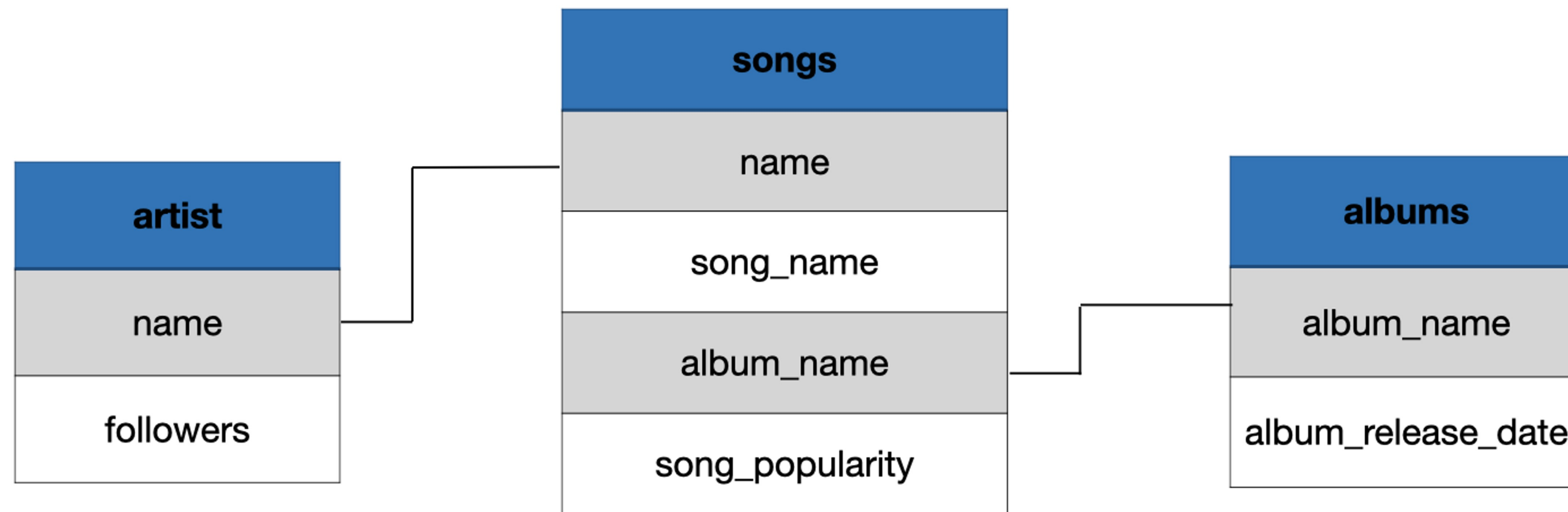
RELEASED
2016

JUSTIN VACHON → Pokemon GO in 2019 is a game I shouldn't care about. When it launched in 2016 it was in a lot of ways a mediocre experience. Outside of catching the original 151 Pokemon the game itself relied heavily on the nostalgia of the Pokemon franchise and augmented reality gimmick of having them show up in the real world. If you didn't care about the IP, the game itself was very lacking. In 2019, the game is flooded with a multitude of tasks, activities, and events that can involve anyone from yourself to a large group of people. These additions create an experience that incentivizes users to be more dedicated to daily play without feeling like a grind. Friendship has been introduced and allows users to now exchange gifts, trade or even battle each other. Quests (research tasks as they are referred to in-

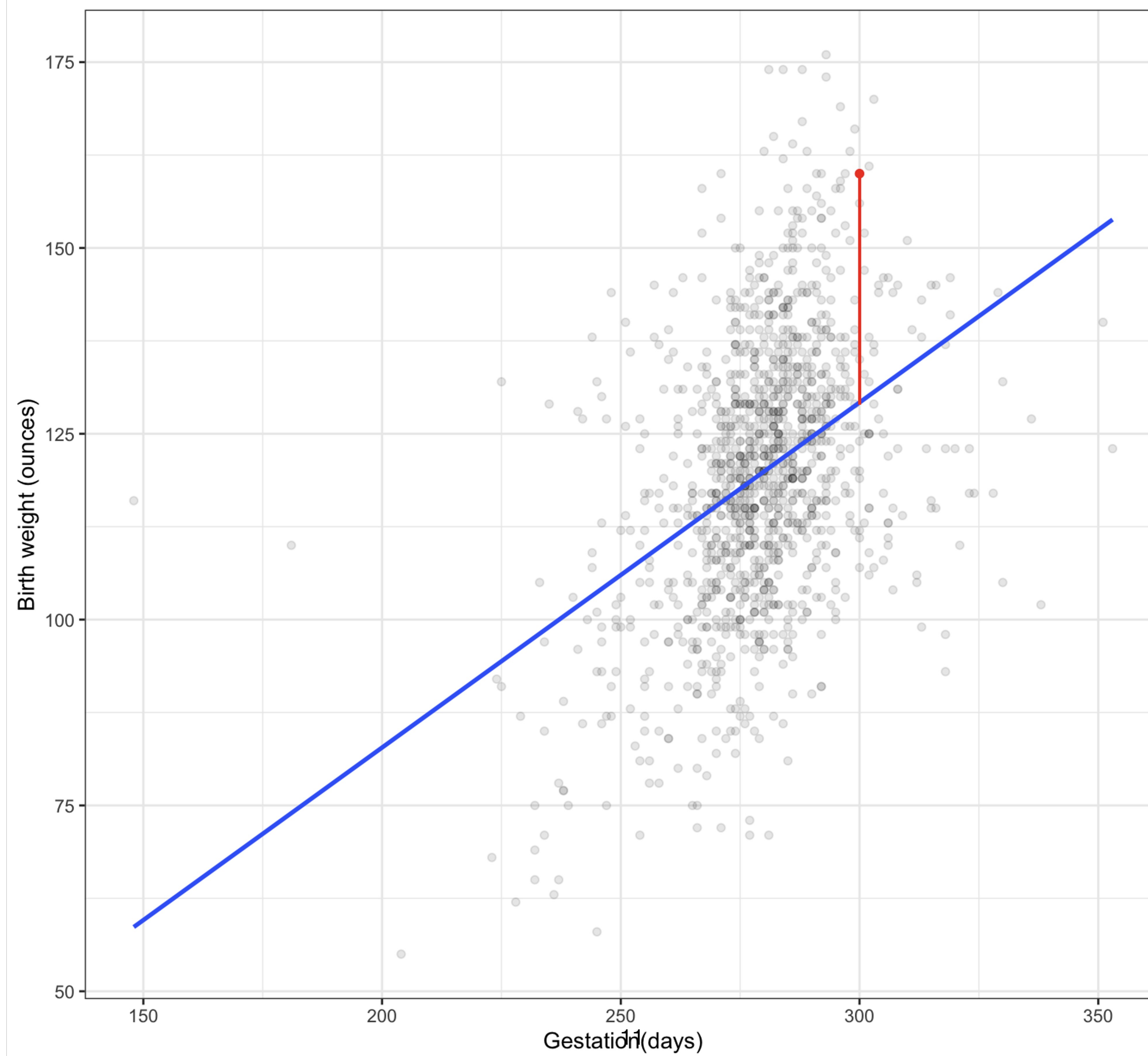


DID YOU KNOW?

- ⊕ There are currently five (of eight) generations of Pokemon in the game, with more being added each year.



make predictions and conclusions using models;



logistic regression example:

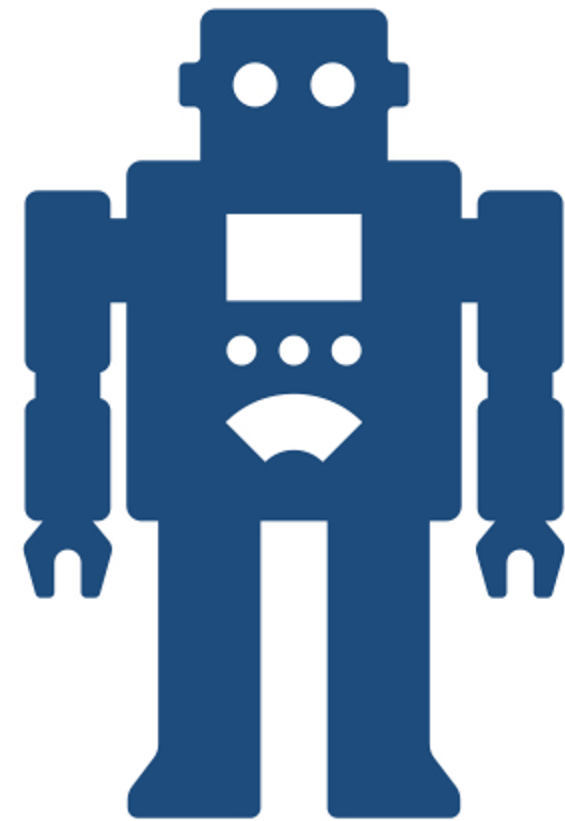
**Are Emily and Greg More Employable than Lakisha and Jamal? A
Field Experiment on Labor Market Discrimination**

consider impact of decisions related to data on humans, other livings, and the planet;

Finding data online does not grant permission to use



Human subjects



Bot access



Commercial use

write human- and machine-readable code using R.



master 3 branches 64 tags Go to file Add file Code

john-halderman Automatic monthly update. 8840794 on Dec 31 79 commits

.gitattributes	May 16th update; moving datasets to large file storage; added do...	6 years ago
Artists.csv	Automatic monthly update.	11 months ago
Artists.json	Automatic monthly update.	11 months ago
Artworks.csv	Automatic monthly update.	11 months ago
Artworks.json	Automatic monthly update.	11 months ago
README.md	Automatic monthly update.	10 months ago

☰ README.md

The Museum of Modern Art (MoMA) Collection

The Museum of Modern Art (MoMA) acquired its first artworks in 1929, the year it was established. Today, the Museum's evolving collection contains almost 200,000 works from around the world spanning the last 150 years. The collection includes an ever-expanding range of visual expression, including painting, sculpture, printmaking, drawing, photography, architecture, design, film, and media and performance art.

MoMA is committed to helping everyone understand, enjoy, and use our collection. The Museum's [website](#)

cderv Add class to book cover to size using CSS

Latest commit ebf0d09 17 days ago History

6 contributors

76 lines (45 sloc) 4.68 KB

Raw Blame

rmarkdown



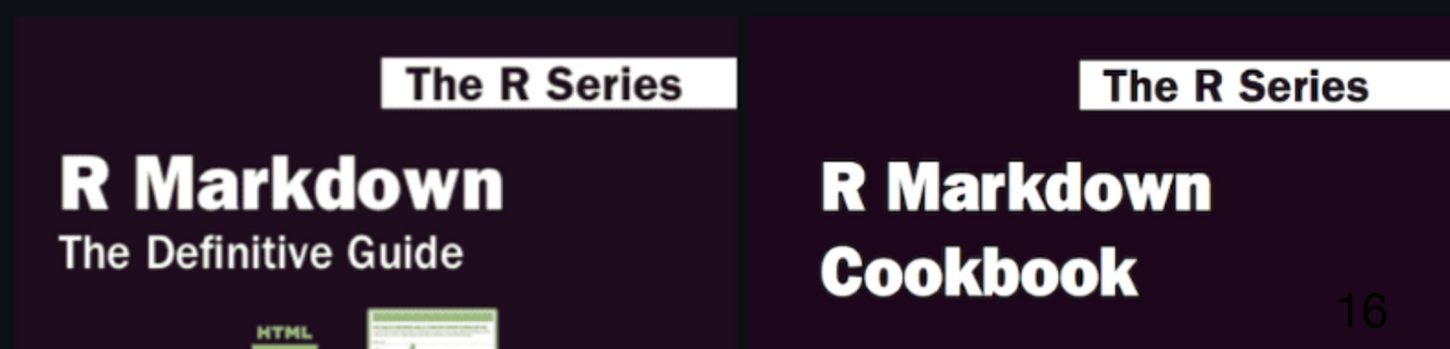
R-CMD-check passing CRAN 2.11 codecov 60%

The rmarkdown package helps you create dynamic analysis documents that combine code, rendered output (such as figures), and prose. You bring your data, code, and ideas, and R Markdown renders your content into a polished document that can be used to:

- Do data science interactively within the RStudio IDE,
- Reproduce your analyses,
- Collaborate and share code with others, and
- Communicate your results with others.

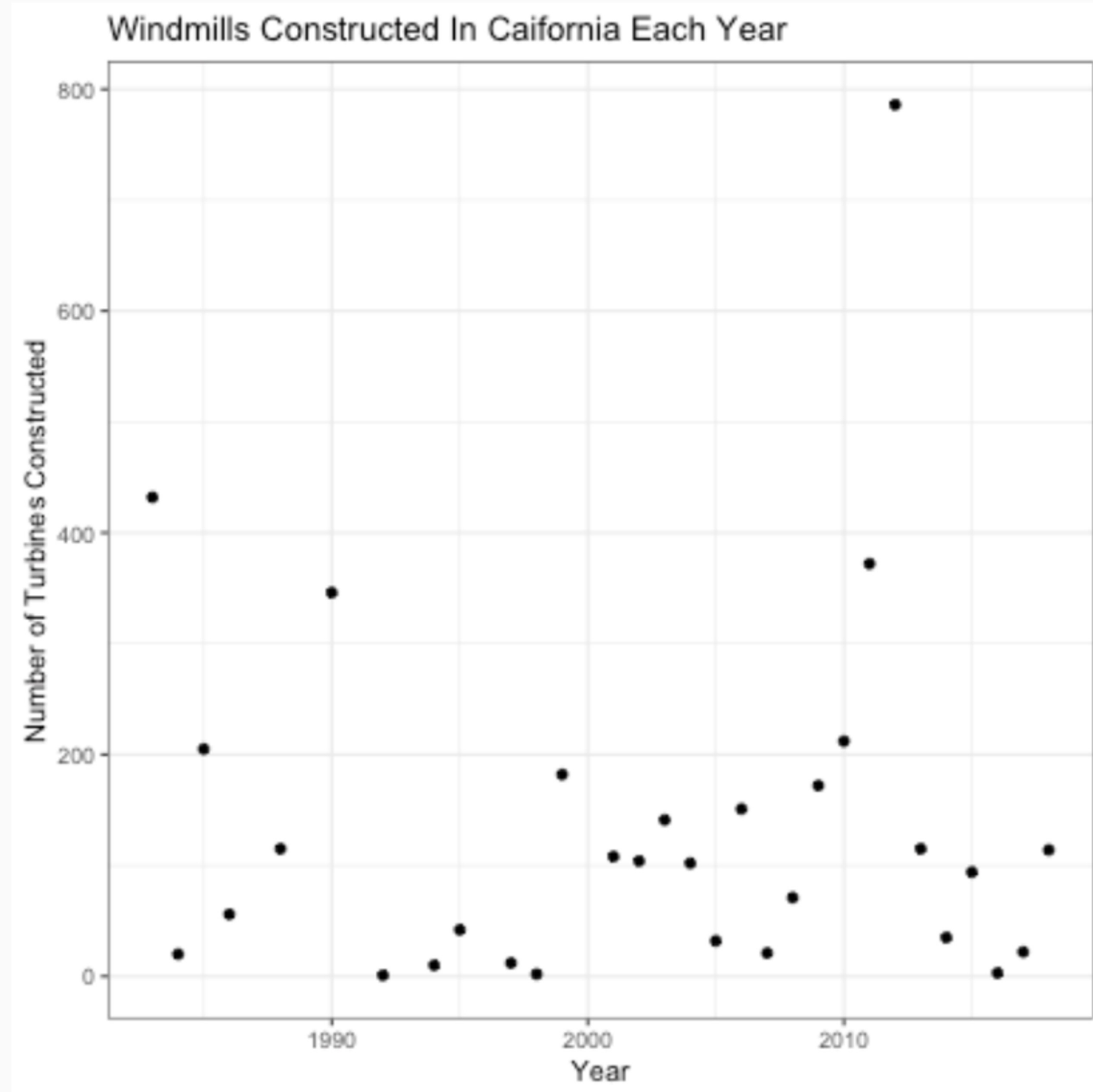
R Markdown documents can be rendered to many output formats including HTML documents, PDFs, Word files, slideshows, and more, allowing you to focus on the content while R Markdown takes care of your presentation.

Books



Final Projects

Windmills in California Over Time



4078 turbines in California.

What's next?

Exploratory Data Analysis -> Data Visualization

Datasets -> Databases

Communication -> Writing + Speaking

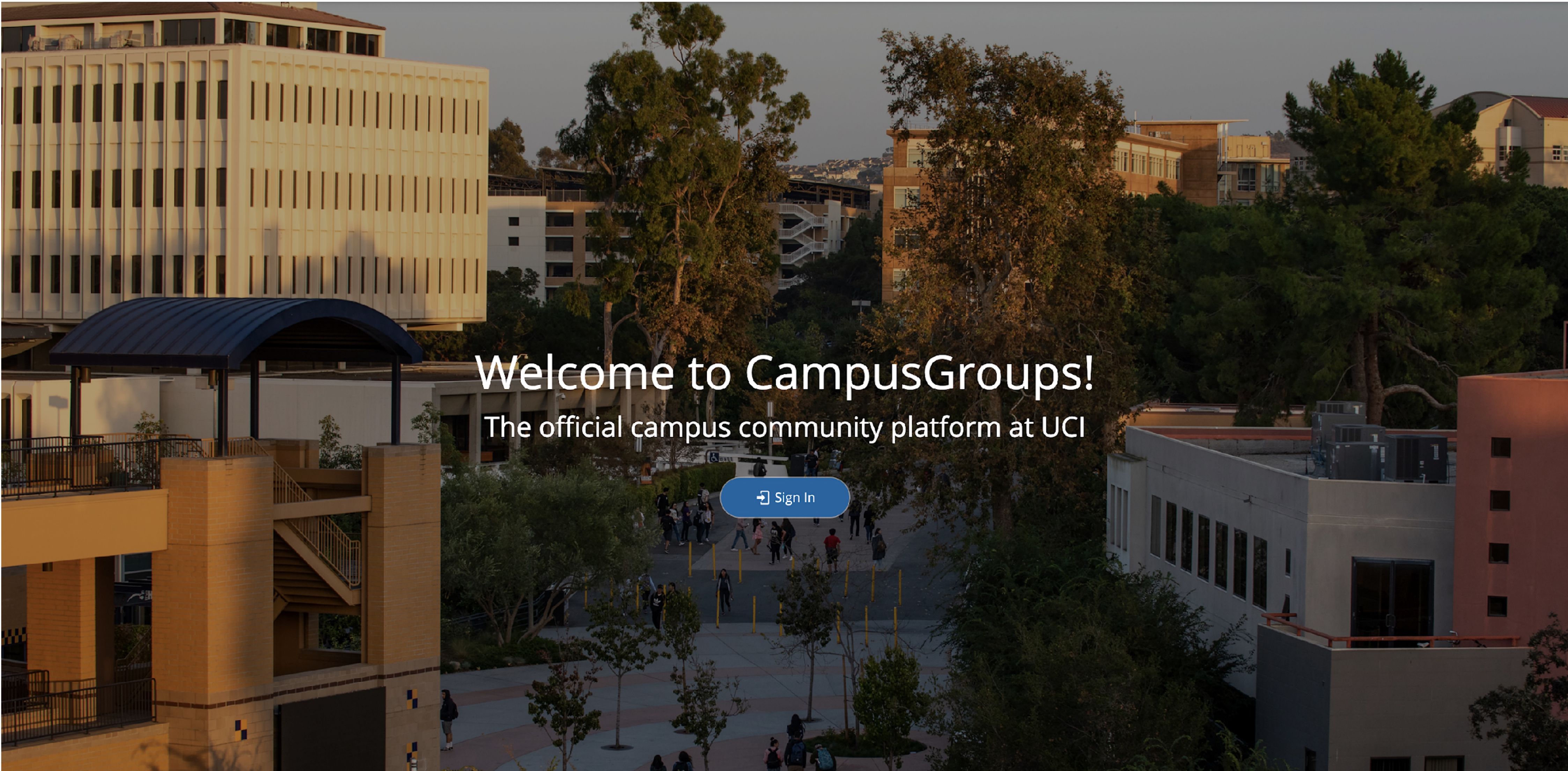
Statistical Modeling -> Probability + Statistics + Machine Learning

Ethics -> Philosophy + Human Computer Interaction

On Campus Resources

- Attend webinars/seminars related to data science.
- Utilize services offered by career office.
- Use writing center.

~~Networking~~
Connecting with people



Welcome to CampusGroups!

The official campus community platform at UCI

[Sign In](#)

meetup



[R Ladies Irvine](#)



[PyData SoCal](#)



[SoCal R Users Group](#)



ASA



ACM



IEEE

Consider Using Social Media for Professional Reasons

- Twitter (now X)
- LinkedIn
- Mastodon

Resource:

[Twitter for R programmers](#)

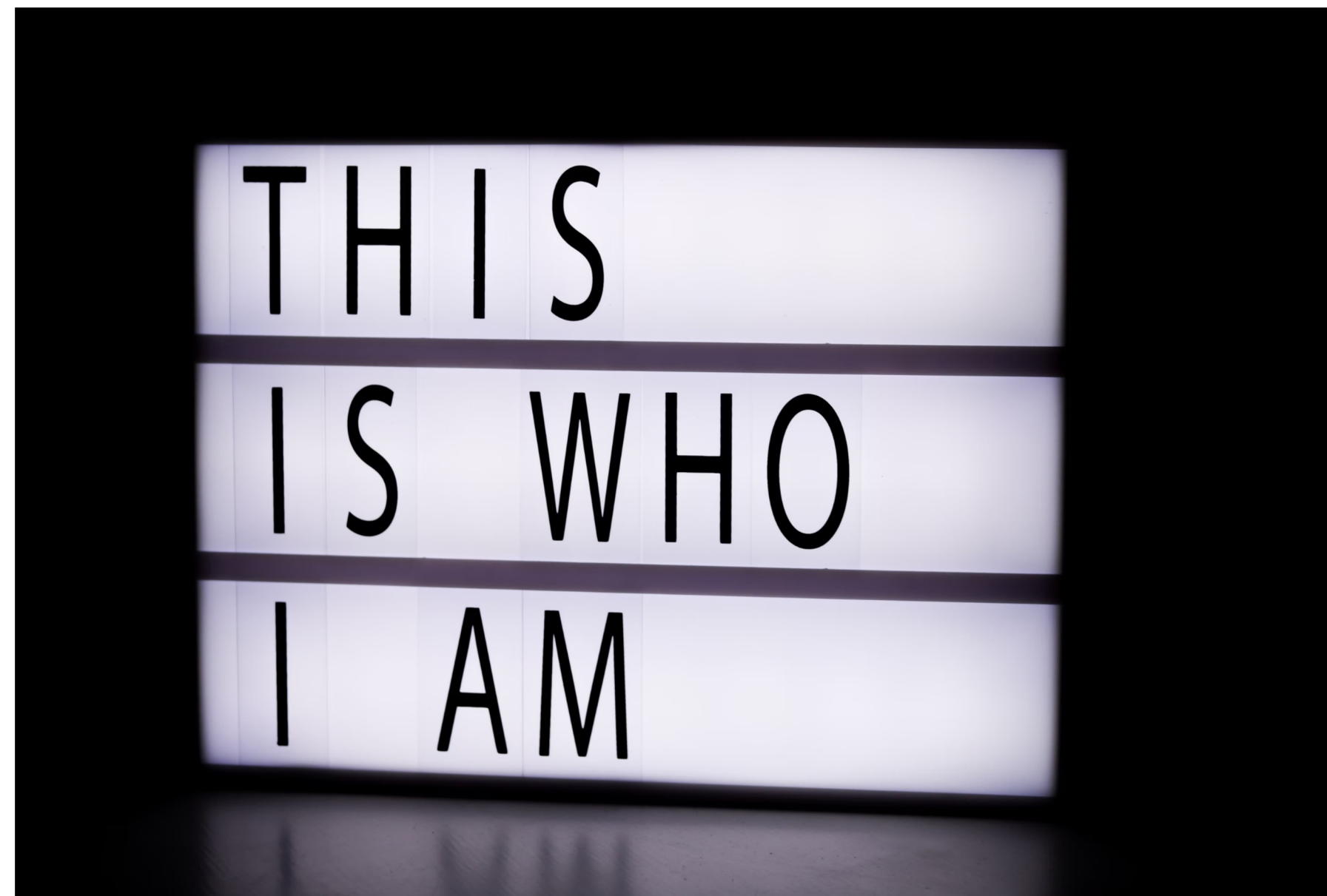
Portfolio

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

variables observations values

The image features a central graphic with the text 'TidyTuesday' in a large, white, sans-serif font. Below it, in a smaller white font, is the subtitle 'A weekly data project in R from the R4DS online learning community'. The background is a dark, textured brushstroke. To the left and right, there are faint, semi-transparent data tables. The left table has columns for 'country', 'year', and 'population', with rows for 1999 and 2000. The right table has columns for 'country', 'year', and 'population', with rows for 1999 and 2000. Below the graphic, three labels are positioned: 'variables' under the left table, 'observations' under the middle, and 'values' under the right table. Arrows point from these labels to the corresponding parts of the data tables. The 'variables' label has a vertical arrow pointing to the 'country' column. The 'observations' label has a horizontal arrow pointing to the 'year' column. The 'values' label has a vertical arrow pointing to the 'population' column. There are also some small circles and lines connecting the labels to the data points.

<https://github.com/rfordatascience/tidytuesday>



Create a website

Quarto
distill
blogdown
Wordpress

Deploy your website

GitHub pages
Netlify

Contests

Undergraduate Statistics Project Competition

Undergraduate Statistics Class Project Competition

Undergraduate Statistics
Research Project Competition

Introductory Statistics

Intermediate Statistics

Projects from courses with
no statistics/data science
prerequisite course

Submission: A short report
(up to 3 pages)

Projects from non-intro and
non-capstone courses

Submission: A short report
(up to 3 pages)

Research projects from activities
like summer research or a capstone
course

Submission: A report (up to 20
pages)

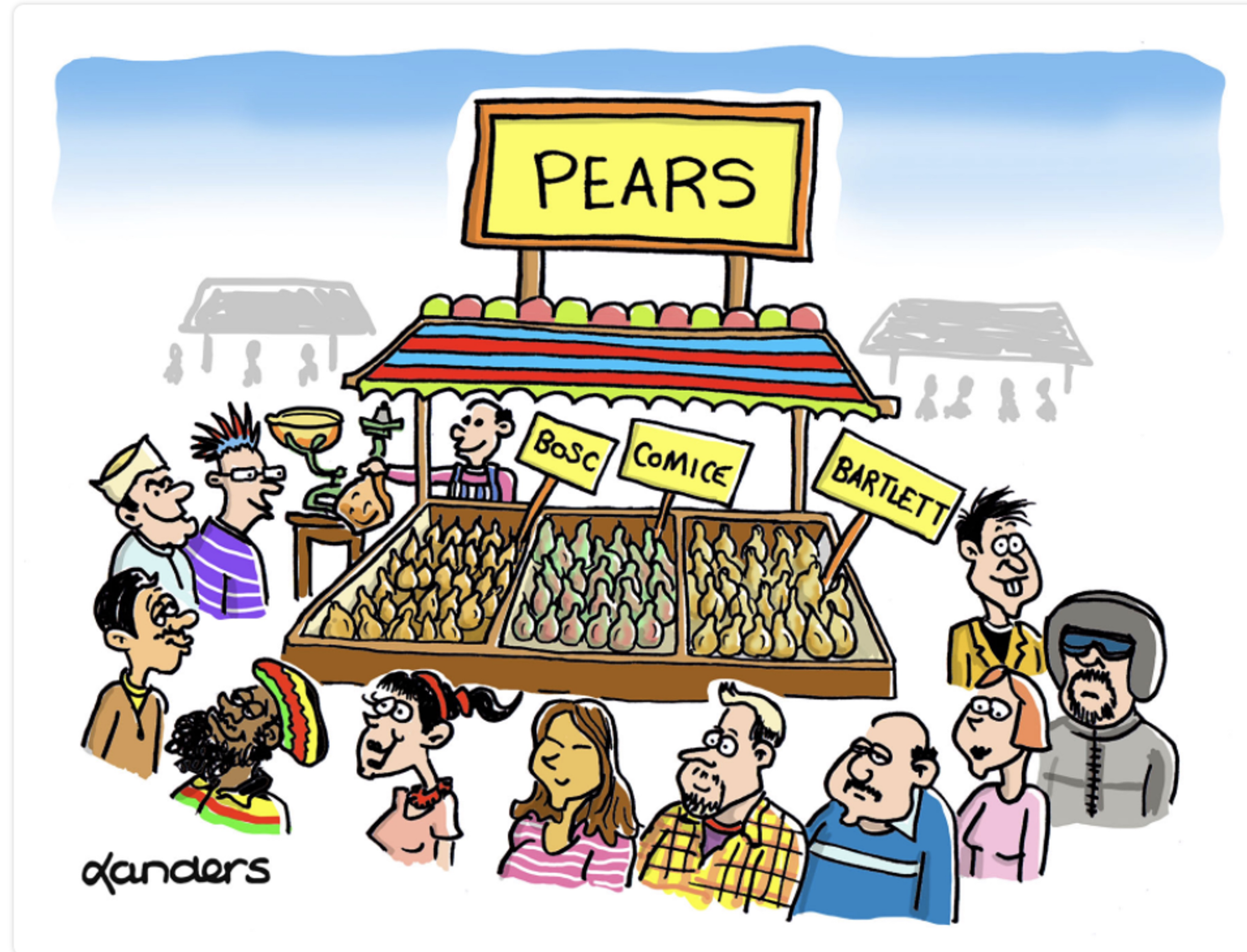


Next Deadline: Dec 22nd

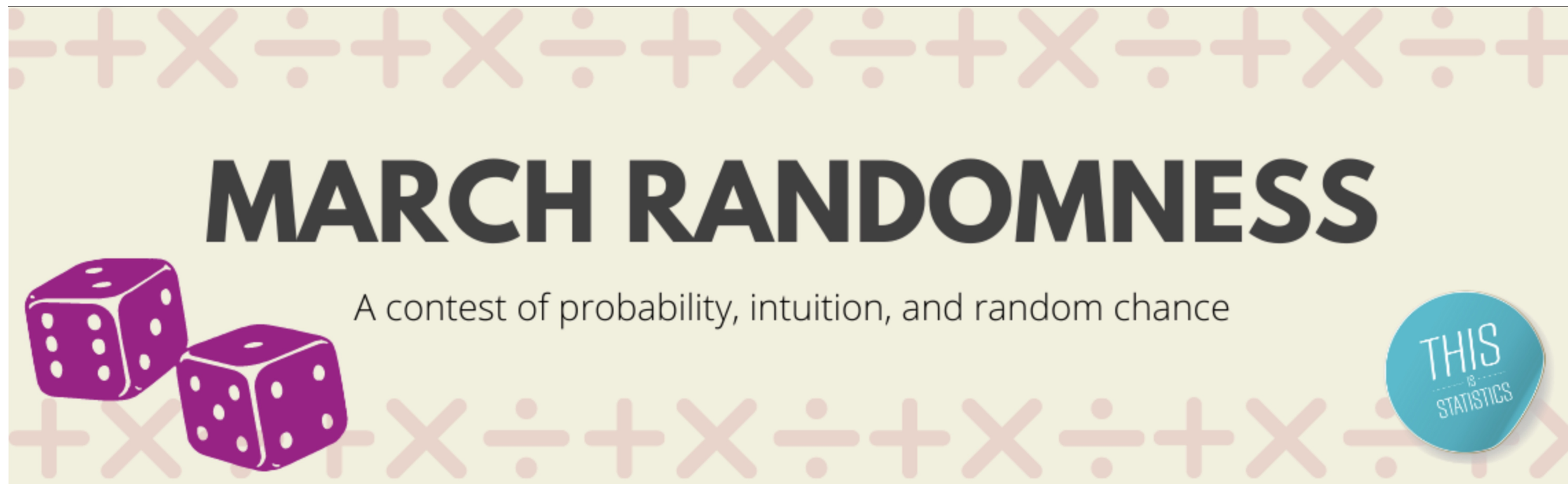


Details:
www.causeweb.org/usproc

CAUSEweb Cartoon Caption Contest





March Randomness

A decorative banner with a light beige background and a repeating pattern of mathematical symbols: plus (+), multiplication (x), and division (÷).

MARCH RANDOMNESS

A contest of probability, intuition, and random chance

Two purple dice with white pips, one slightly behind the other.A blue circular logo with the text "THIS & STATISTICS" in white, tilted slightly to the right.

Career Preparation Resources

Electronic Undergraduate Statistics Research Conference

FRIDAY, NOVEMBER 5TH, 2021

Watch

Undergraduate Statistics Project Competition (USPROC) award winners will present their work.

Learn

Info session on graduate school and panel on careers in statistics & data science.

Be Inspired

Keynote address by Sydeaka Watson, Senior Research Scientist (Data Scientist) at Eli Lilly and Company.

Get Involved

To share your own statistical work and register for the conference, submit an abstract by Friday, October 15th. For more information, go here! Prizes will be given for the Best Video Presentations!

Register

To attend this FREE conference, click here. All undergraduate students and faculty are invited.

Build a Career in Data Science Podcast



12 episodes

Build a Career in Data Science

Jacqueline Nolis and Emily Robinson

Careers

★★★★★ 4.9 • 26 Ratings

[Listen on Apple Podcasts ↗](#)



FEB 11, 2021

Chapter 11: Data Science in Production



Putting data science into production can mean a ton of things: from customer-facing models run millions of times a day to continuously live dashboards for stakeholders. But writing code for production and getting it to work can be intimidating for many data scientists, and lots of us have

[▶ PLAY](#) 49 min

NSF Research Experiences for Undergraduates (REU) search



ASA Internships and Fellowships





Find your next job in data science in the Federal Government

Start your search

<https://usajobs.github.io/microsite-data-science>

Keywords:

Statistics

Data

Biostatistics

Bioinformatics

Psychometrics

Econometrics

▪

▪

▪



A First-Gen's Guide to Grad School: How to Get in, Survive, and Thrive

Helping first-gen students along their PhD journey!

<https://first-gen-guide.com/resources>



[Tips on Recommendation Letters for Students and Instructors](#)

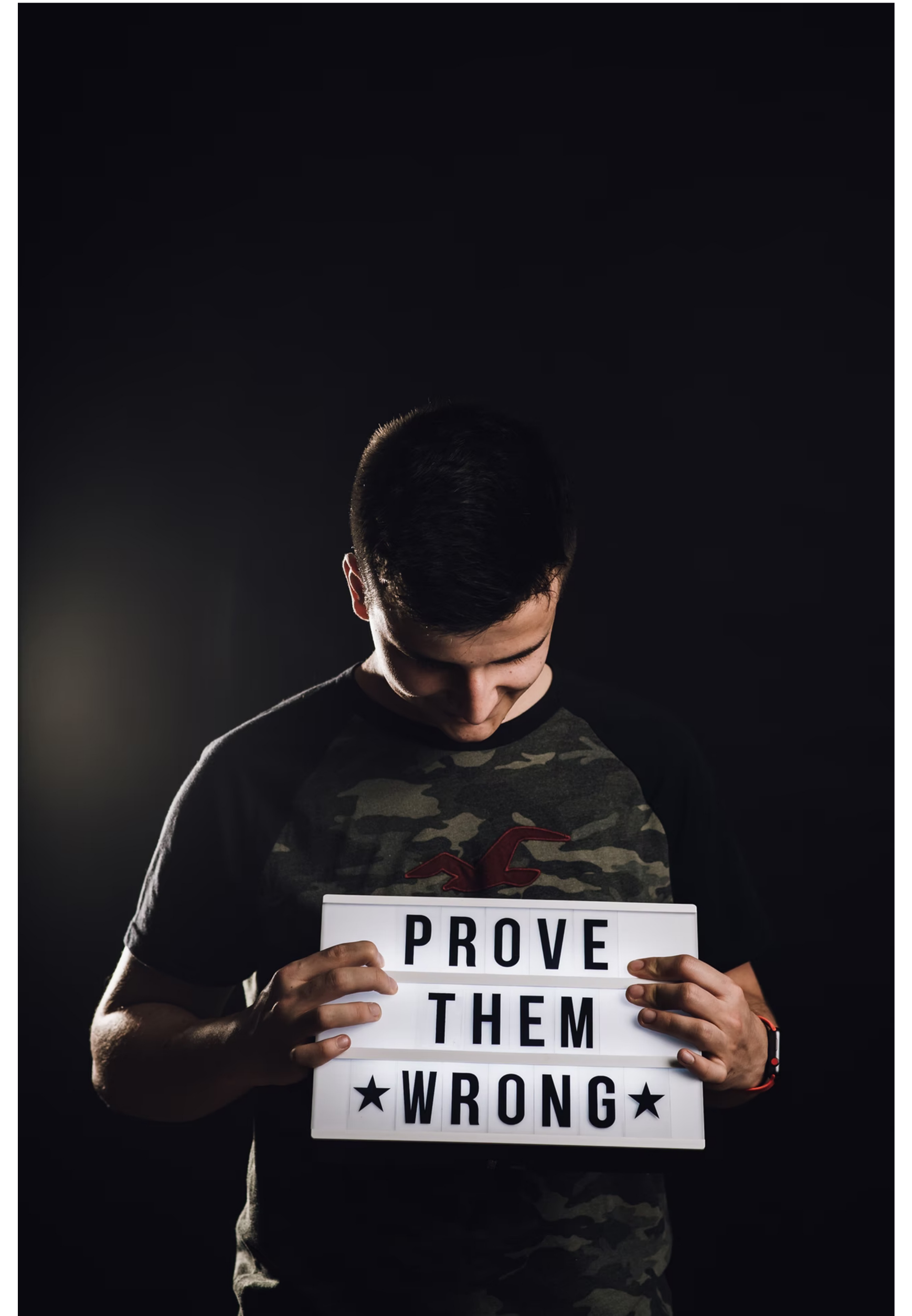
Applying to the NSF Graduate Research Fellowship (GRFP)

<https://www.simonpcouch.com/blog/2021-08-02-apply-to-nsf-grfp/>

Last but not least

You might hear things like

- You are not a data scientist because you are not strong in math/stats/cs.
- What you do is not data science because it does not involve advanced computing.
- You cannot be a data scientist because you took courses on X and Y but not Z.





Questions?

bit.ly/dogucu-talks



mdogucu@uci.edu



[@MineDogucu](https://twitter.com/MineDogucu)



[MineDogucu](https://www.linkedin.com/company/MineDogucu)