Data Science Portfolios 101

Dr. Rachael Tatman
September 19, 2018

kaggle

What are you showing off in a data science portfolio?

What you can do for someone who hires you.

What are you showing off in a data science portfolio?

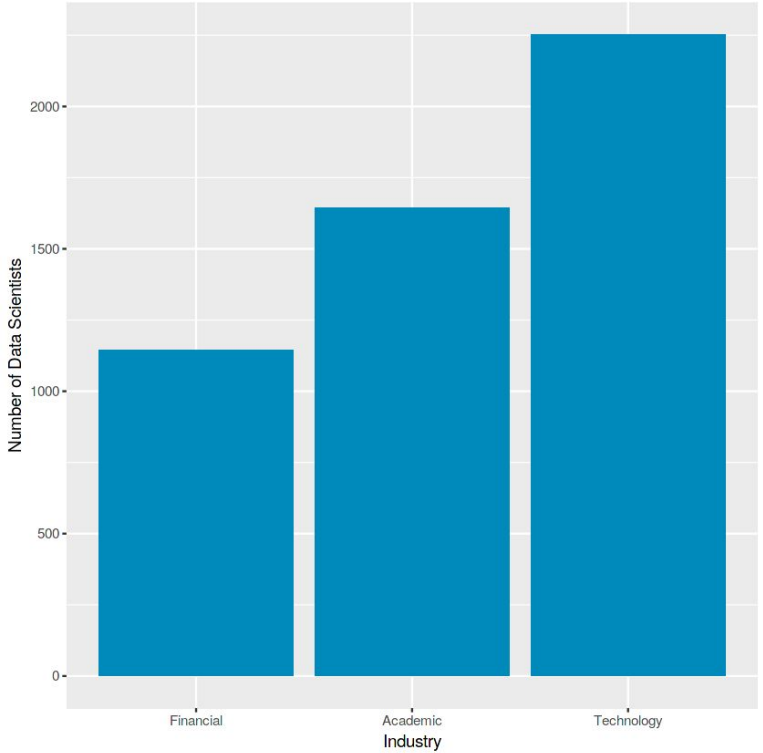
What you can do for someone who hires you.

1) Research data science jobs & identify the type of role you'd be well-sutied for.
2) Show off skills relevent to that job.

Business stuff

Professional data scientists tend to work in a few main industries (obviously dependent on location). Top three:

- Technology
- Academia
- Finance
- (In DC, probably also government work/defense contractors)



Business stuff:

Startup vs. established company

Startup:

- You wear a lot of different hats
- “Agile”, quickly-changing environment
- Job ***in***security
- More **generalist**
- May not have access to senior mentors

Larger company:

- You do one job as part of a team
- More structure/process
- Job security
- More **specialist**
- Will probably have access to senior mentors in your role

Business stuff:

Related job titles/career paths

- **Machine Learning Engineer**: You'll be building & deploying ML systems, probably using some flavor of deep learning.
- **Data engineer**: You'll be building and maintaining databases & pipelines. (Think: industrial strength data cleaning.)
- **UX Researcher**: You'll be running behavioural experiments and analyzing the results.
- **Developer relation roles**: You'll be talking to people about products (internally & externally).
- **Technical program/project manager**: You'll be helping things get done (planning/organizing). These roles usually require certificationscertificaitons.

Domain knowledge

Show that you can do the sorts of things you want to be hired to do.

Domain knowledge

Types of data:

- Tabular/relational
- Text
- Image/video
- Time series
- Geospatial
- Domain specific that uses techniques from multiple other (e.g. genetic data)

Domain knowledge

Types of data:

- Tabular/relational
- Text
- Image/video
- Time series
- Geospatial
- Domain specific that uses techniques from multiple other (e.g. genetic data)

Option 1: Show **deep** expertise in a specific type of data (generally not tabular/relational: all data scientists should know how to work with tables).

Option 2: Show that you can do work in a range of domains.

Domain knowledge

Types of data:

- Tabular/relational
- Text
- Image/video
- Time series
- Geospatial
- Domain specific that uses techniques from multiple other (e.g. genetic data)

Quick exercise:

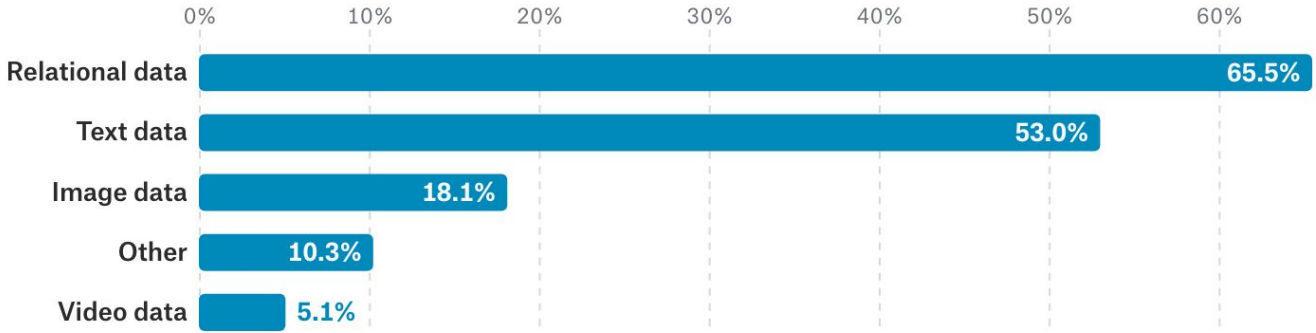
- Think about four projects you've already completed or could complete quickly
- What type of data is each on? Do these projects together show depth or range?

Domain knowledge

What type of data is used at work?

Relational data is the most commonly reported type of data used at work for all industries except for Academia and the Military and Security industry where text data's used more.

Company Size | Industry | Job Title



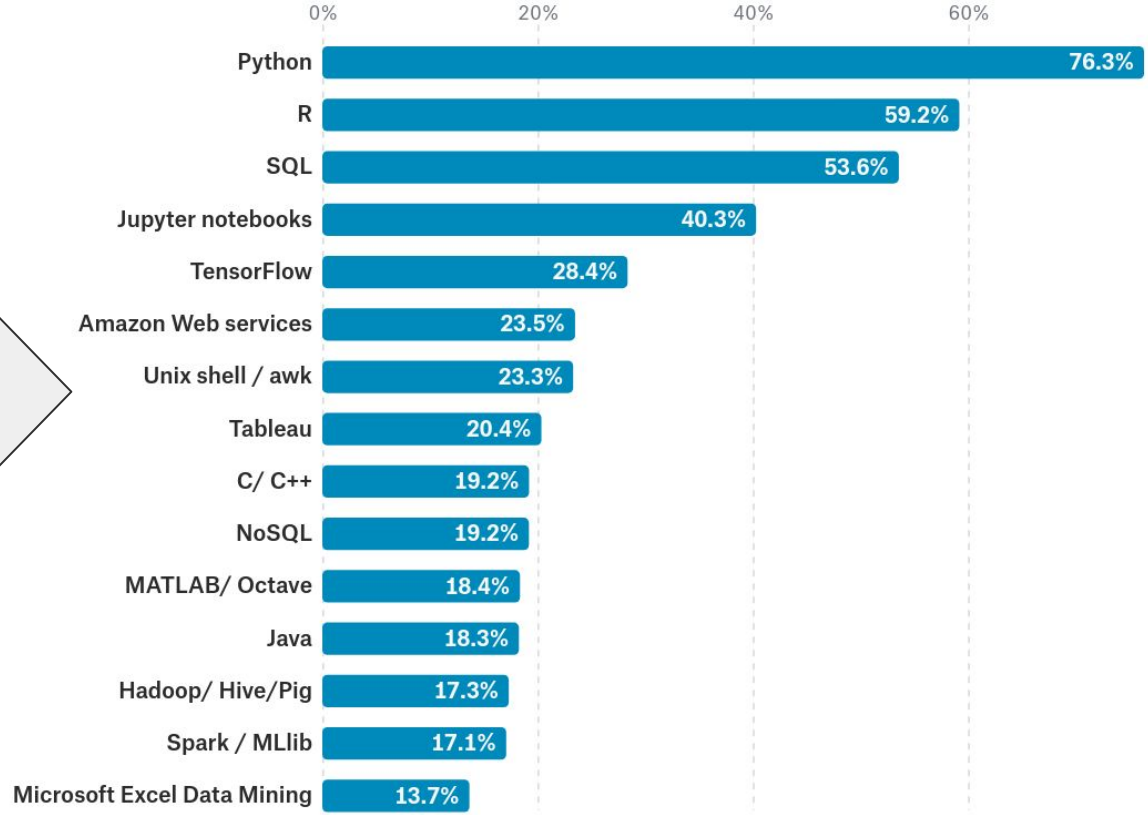
8,024 responses

Domain knowledge

Some skills I'd recommend you highlight:

- Programming language of choice (Python, R, Julia)
- Ability to interact with databases (SQL)
- Visualization
- "Storytelling" (Can someone with no background in whatever area your project is in read it and gain some new understanding?)
- Deploying small sample projects (e.g. a RESTful API for a ML model you trained or a nice Shiny dashboard)

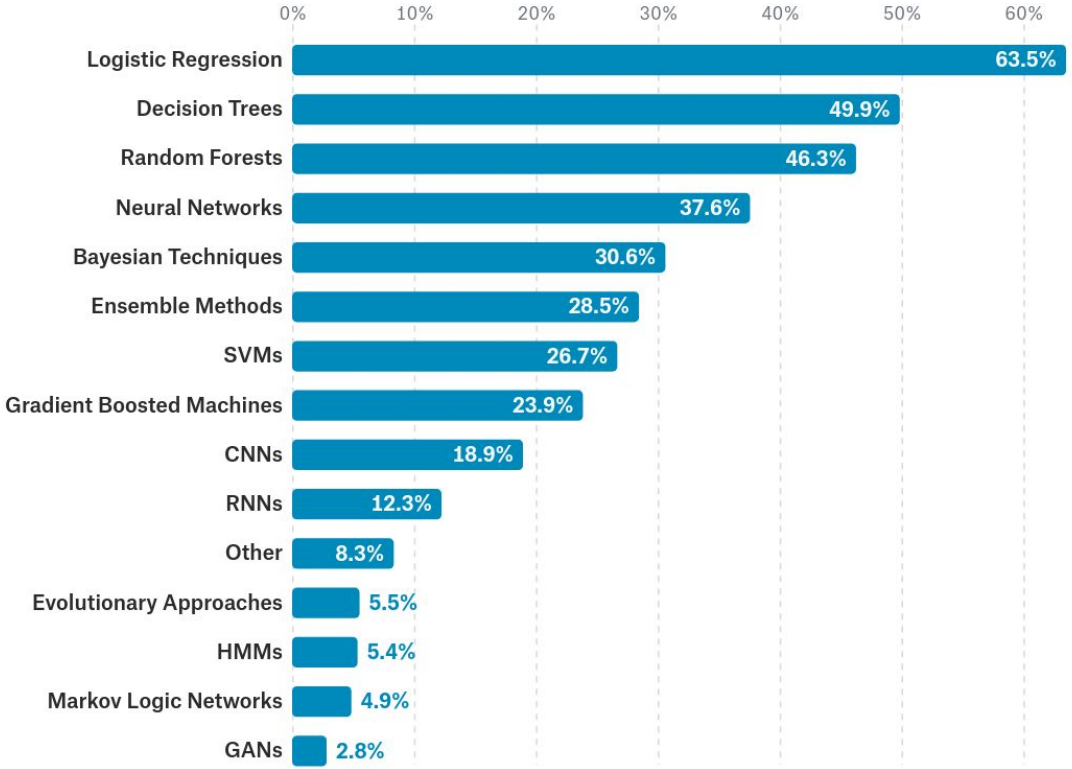
Currently relevant tools (subject to change!)



7,955 responses

@rctatman

What methods to focus on?



Methods used by professional data scientists (ranked by popularity)

@rctatman

Domain knowledge

If you're showing code (not just a dashboard or something) make sure it looks professional!

- Clean, readable code (remove all your "checking stuff out" bits, like printing out any parts of a dataframe)
- Use version control
- Pick a style guide and use it consistently! (A linter can help here)
- Break your project into multiple files. Example:
    1. Utility functions/package
    2. Data cleaning
    3. Modelling
    4. Model evaluation & visualizations

Portfolios are also about what you *don't* include

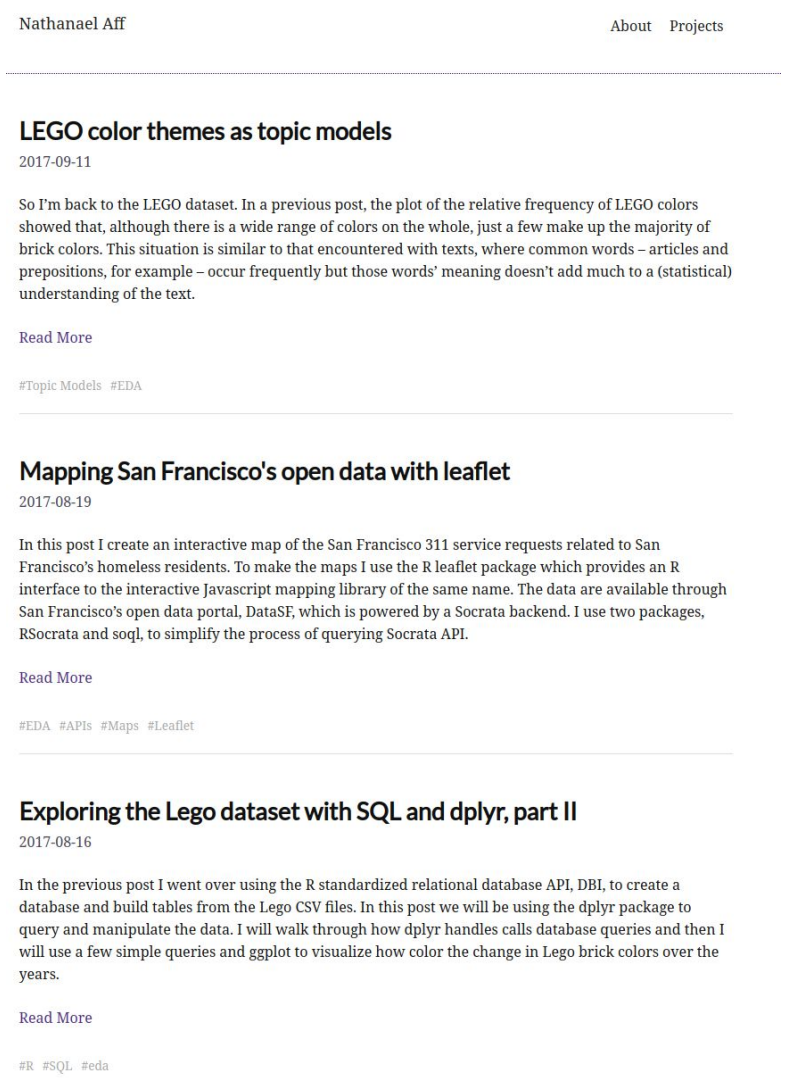
- **Quality over quantity**! Don't throw in every student project
- Avoid sharing data cleaning (just link to the file with the code)
- Avoid EDA, portfolio pieces should have a clear story
- Check for grammar errors/clarity

@rctatman

Sample Portfolios

Sample Portfolio

- Nathanael is a freelance data science consultant
- Portfolios don’t have to look fancy!
- Really nice mix of projects: GIS, novel applications of NLP techniques (!), mapping and working with databases



Nathanael Aff

About Projects

LEGO color themes as topic models

2017-09-11

So I’m back to the LEGO dataset. In a previous post, the plot of the relative frequency of LEGO colors showed that, although there is a wide range of colors on the whole, just a few make up the majority of brick colors. This situation is similar to that encountered with texts, where common words – articles and prepositions, for example – occur frequently but those words’ meaning doesn’t add much to a (statistical) understanding of the text.

Read More

#Topic Models #EDA

Mapping San Francisco's open data with leaflet

2017-08-19

In this post I create an interactive map of the San Francisco 311 service requests related to San Francisco’s homeless residents. To make the maps I use the R leaflet package which provides an R interface to the interactive Javascript mapping library of the same name. The data are available through San Francisco’s open data portal, DataSF, which is powered by a Socrata backend. I use two packages, RSocrata and soql, to simplify the process of querying Socrata API.

Read More

#EDA #APIs #Maps #Leaflet

Exploring the Lego dataset with SQL and dplyr, part II

2017-08-16

In the previous post I went over using the R standardized relational database API, DBI, to create a database and build tables from the Lego CSV files. In this post we will be using the dplyr package to query and manipulate the data. I will walk through how dplyr handles calls database queries and then I will use a few simple queries and ggplot to visualize how color the change in Lego brick colors over the years.

Read More

#R #SQL #eda

Sample Portfolio

JULIA SILGE

BLOG ABOUT RESUME

BLOG

Sep, 2018
- Sep 8, 2018 Training, evaluating, and interpreting topic models

Jul, 2018
- Jul 19, 2018 Amazon Alexa and Accented English

Jun, 2018
- Jun 30, 2018 Punctuation in literature

May, 2018
- May 18, 2018 Understanding PCA using Stack Overflow data
- May 30, 2018 Public Data Release of Stack Overflow's 2018 Developer Survey

Apr, 2018
- Apr 11, 2018 Stack Overflow questions around the world

Jan, 2018
- Jan 10, 2018 tidytext 0.1.6
- Jan 25, 2018 The game is afoot! Topic modeling of Sherlock Holmes stories

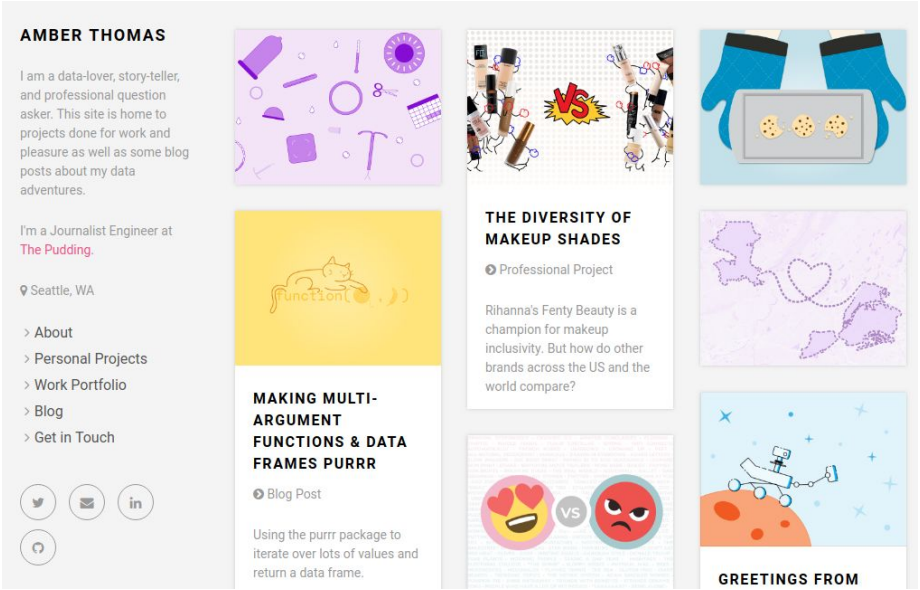
Dec, 2017
- Dec 27, 2017 One year as a data scientist at Stack Overflow

- Julia is a data scientist at StackOverflow
- Mix of projects for work & fun (Note: Don't feel pressured to do the thing you do professionally for fun!)
- Shows that she's got deep expertise in NLP & text processing

@rctatman

Sample Portfolio

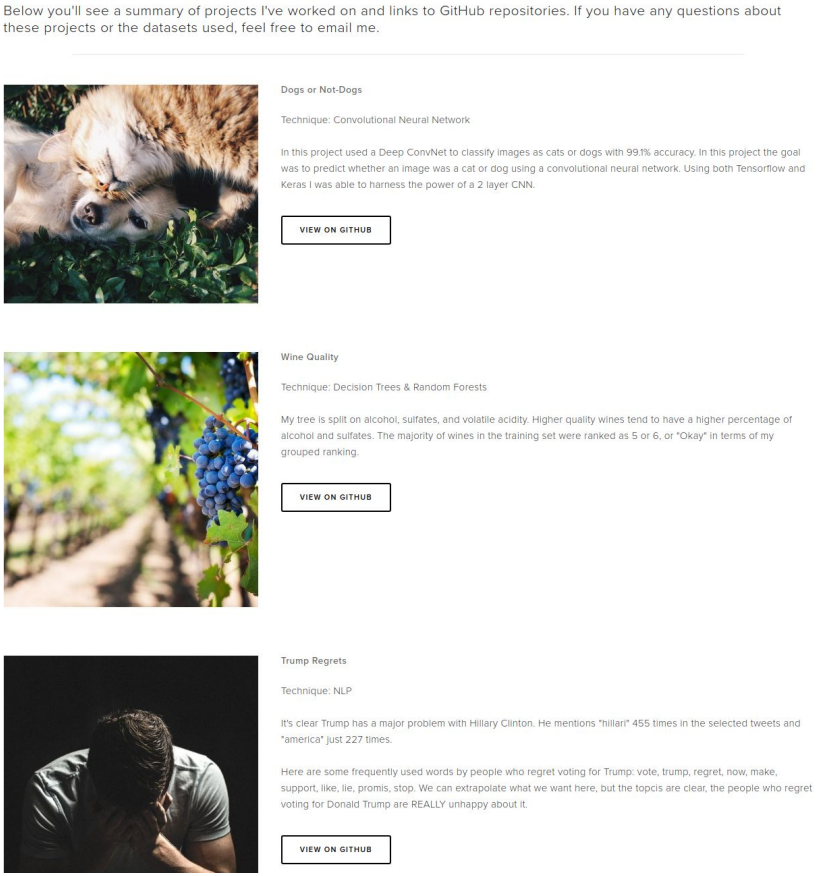
- Amber is a Journalist Engineer @ The Pudding
- She does visualization work, so having a beautiful portfolio is important to her brand
- A really nice mix of projects that I want to read more about
- Shows off ability to write compelling titles/blurbs for content (important for a journalist!)



@rctatman

Sample Portfolio

- Portfolio of Ayodele Odubela, currently, Director of Machine Learning at Astral_AR
- Potfiolio is from when she was a student
- Good mix of projects with a nice, readable summary of each:
  - Image data (CNN)
  - Tabular data (Decision trees & random forests)
  - Text data (variety of NLP techniques)



@rctatman

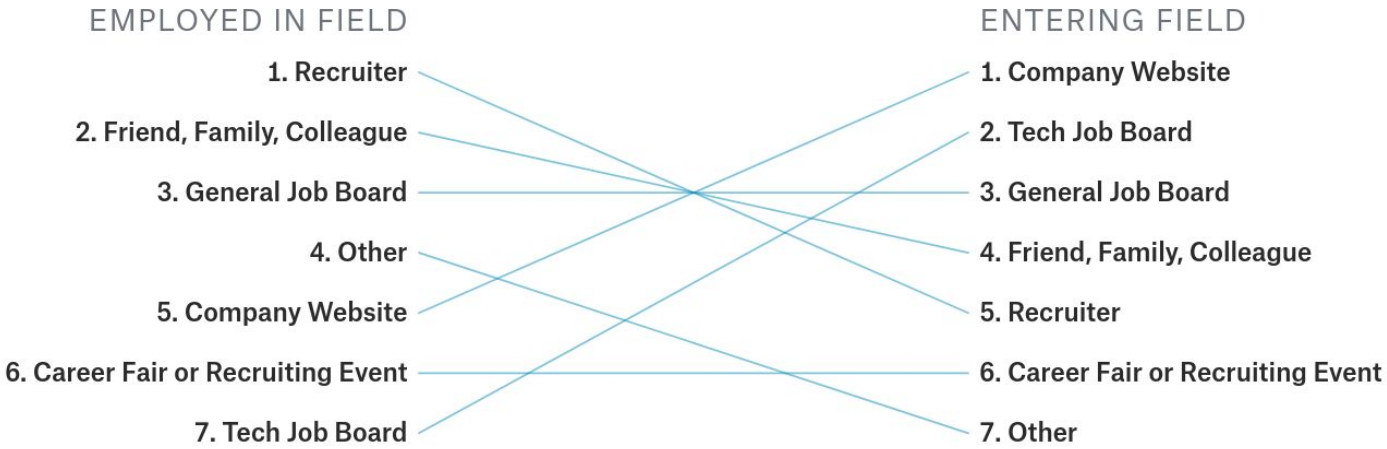
General tips:

- Include projects you're genuinely excited about
- Unique projects are better: folks are more likely to remember them & it shows you can work independently
- Write for humans, not machines (no code-only notebooks!)
- Make sure you know what's going on in your code **and can explain it** (you'll definitely be asked during interviews!)
- If you update your portfolio/blog semi-regularly, you won't have to scramble at the last minute
- Get you a website! GitHub pages are free & you can write your site in R (I recommend Blogdown)

One last tip: don't look @ job boards for jobs!

How do you look for or find work?

When you're job hunting, it may be tempting to look for work on company websites or tech-specific job boards, but according to people who are employed in the data science realm, these are among the least helpful ways to find work. Instead, try to contact recruiters or build up your network to break into the field.



@rctatman

Thanks!
Questions?

@rctatman

kaggle™