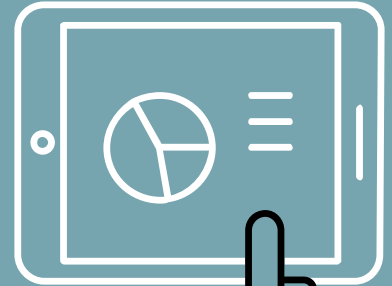
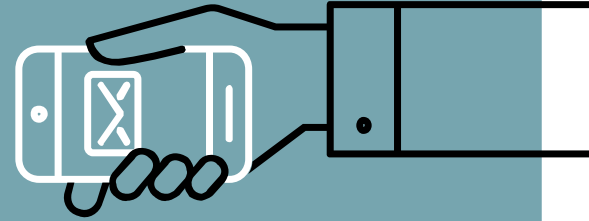
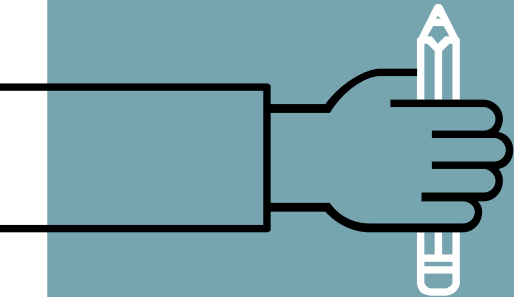
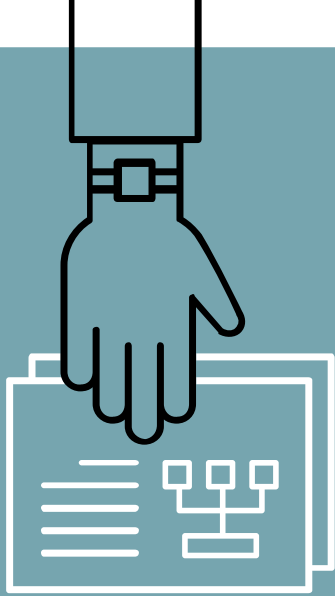


BST 270

Reproducible Data Science

Winter 2021
Session 6



Module 6 Part 1 Discussion

Can you use Git to do version control for datasets that contain sensitive information?

Yes. [GitLab](#) is recommended for this. Google Cloud Platform can also be made HIPAA compliant and [host Git repositories](#) with version control.

How much time typically do you end up splitting between organizing all the data and files, understanding the field/context you're working in, and actually doing analyses?

When I start a project I spend a lot of time discussing the goals/purpose of the project with my team and doing a lit review and overview of what data is available. I keep notes during each meeting and update any project progress on a team Confluence page. When I start to code I organize my files as I go. This saves me a lot of time in the long run. I spend most of my time cleaning the data, exploring the data and figuring out what kinds of analyses are appropriate. I spend the least amount of time on actually running the analyses. All of the prep work before the analyses takes up ~60-70% of my time. Data analysis ~10% and writing it all up ~20-30%.



Module 6 Part 1 Discussion

I understand the master and development branches in Git, but what is the use of the bug branch? I assume this is where you can keep documentation of past bugs in your code, but does it contain actual code as well? Why is this branch necessary if you can just add a new version to your master branch that fixes the bug?

Bug branches are used to explicitly track the difference between bug development and feature development. Bug branches will be created when there is a bug on the live site that should be fixed and merged into the next deployment.

Are there ever issues where data-sharing is hindered due to a paywall? For example, do organizations that collect data ever force users to pay some kind of monetary fee to access the data? If so, does this have significant effects on what must be done to make research reproducible?

Yep. One example is Twitter data. You can pay less (or nothing in some cases) for a limited number of tweets. You can pay hundreds of thousands of dollars for the Twitter “firehose” - all of the Tweets matching your keyword criteria in real time. Data that is paid for can make analyses very difficult to reproduce, but since Twitter data is technically public, I believe you can share the data you received. If the data is sensitive and can't be shared, I think a researcher would have to gain access from the paper author(s) to access the data.



Module 6 Part 1 Discussion

How many fold should we use in the training set?

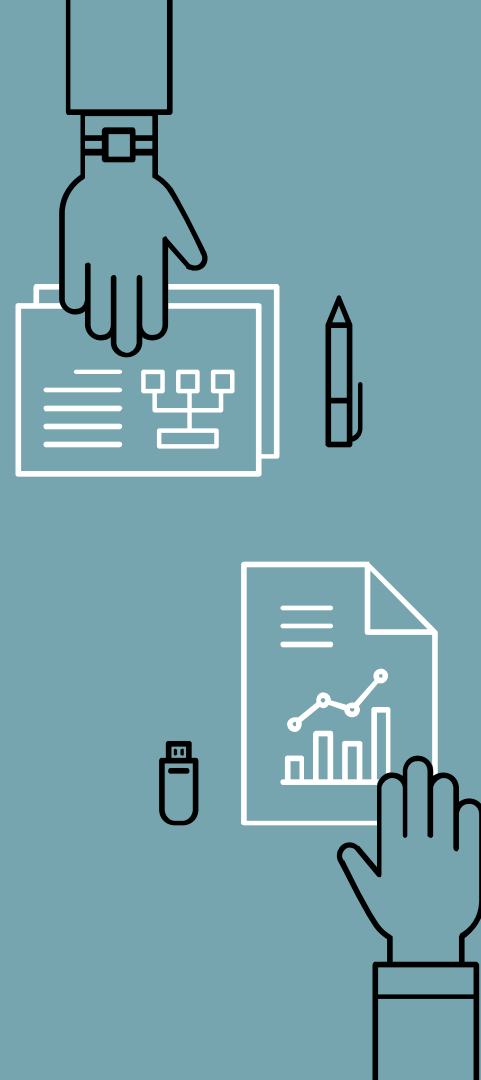
This is one of the arbitrary machine learning parameters every coder faces. It's really up to you and there is no "right" answer. The most common numbers are 5 and 10.

Is it worth documenting everything you when writing code/performing data analysis, including things that get completely scrapped in the end? Or do you just include the relevant information for what is included in the final product?

I tend to keep track of everything, even if it didn't work out. That way, later on if I forget that we tried something and it didn't work, I don't waste time trying it again. It also makes me feel like I worked a lot more.

I found the Github modules very interesting as I had never used Github before prior to this class. What are some of the best practices / key considerations for Github to be successful in larger teams?

I recommend GitLab for larger teams. And usually, the data science and/or engineering teams will have specific protocols when using GitLab. Like a process/rules for creating/naming repositories, merging branches, checking code, etc.



Module 6 Part 1 Discussion

Is there any shortage of Git and Github when we do coding and analysis?

I'm not sure exactly what this is asking but GitHub now allows unlimited private repositories. This is a fairly recent change and you used to have to pay for private repositories when the videos were filmed.

Is there a movement to make pharmaceutical companies also make their data/results more transparent and reproducible?

Yes. [This paper](#) is a great resource. I think reproducibility is more important to some pharmaceutical companies than it is to academic researchers.

If you collected the data yourself, how do you obtain a DOI number for it to use for citations? Does the repository that you are publishing your data to generate a DOI for you?

I've never done it before but I found [this resource](#).



Module 6 Part 1 Discussion

I'm a bit confused as to what gitignore actually does; what happens if you don't use it? Why exactly do you not want your R workspace to be pushed to Git?

gitignore is used when you don't want to track certain files, meaning the different versions aren't subject to version control. This can be for files with sensitive information, or files that aren't necessary to your analysis that you don't want to track. I use gitignore for my intro to data science course so students can copy my lecture files and take notes without disrupting the repository. [See details](#) at the bottom of the repo's README file.

The videos definitely stress the importance of data / code documentation, which I completely agree with. I wonder how well the field / academics currently do this. Are there requirements in journals for data / code to be well-documented?

For some journals, yes. For Nature: "Upon publication, Nature Research Journals consider it best practice to release custom computer code in a way that allows readers to repeat the published results."



Module 6 Part 1 Discussion

Would you mind share some of your experience with data management and data sharing?

I will chat about this in class because it's too much to type!

Is there any difference or similarity between dealing with small data set and large data set? Dr. Merce said that we do not need to download the large data set from the cloud for analysis but if the result of the analysis is still large or the outcome is needed for the research, how can we cope with such issues of handling big data set?

When working with big datasets you generally transform the data into something a lot smaller. For example, you may only need a few columns or rows, or you can aggregate the data in some way. When I work with big datasets I use SQL in Google Cloud Platform to query the rows/columns I need and then aggregate the data in some way if I can before I load it into a Jupyter notebook. From there I usually save the data as a csv file on my machine and proceed with coding as usual.



Module 6 Part 1 Discussion

Video 6.2.3 talks about different editors for Python - what has worked best for you, and what would you suggest?

Jupyter Notebook and Jupyter Lab via [Anaconda Navigator](#) have worked the best for me. I know a few colleagues who prefer Spyder and VS Code.

Would you suggest taking data management courses or is that something you learn through experience?

Only if you will be a data engineer/managing data a lot in the future. If you will occasionally pulling data from a database I would suggest learning SQL basics. The fundamentals are pretty easy to learn and a few commands will go far. You can watch free videos online.

What exactly is meant by data citation? What is the most appropriate way to cite data?

A data citation is just citing where you got your data from if you used data that was publicly available, or could be accessed in some way by other researchers. Think of it as another entry in your bibliography except for data instead of a paper or book. Usually the publishers responsible for the data will provide a citation they would prefer be used, but [this is a great resource](#).



Module 6 Part 1 Discussion

What about GitHub that makes it unique compared to other resources of project sharing resources?

Two big things that come to mind are version control (although it's not the only version control resource out there) and the community. The community of GitHub users is huge and they are usually very helpful if you have a bug or issue with anything.

Are there a few examples of situations where one would prefer an editor that's not IDE?

I think it has to do with preference. I know some people who learned how to code with plain scripts and don't like IDEs. I personally really hate coding without one. Sometimes you will need to save your code to a plain script in order to execute it. I've had to do this when using the School's computing cluster. But I always use an IDE to write and test the original code.

Are there ways for two people to collaborate on a project in different languages? For example, can Git or any other version control software convert python to R?

Git can't - at least not that I'm aware of. But you can add python code chunks and run them in an Rmarkdown file and you can add R code chunks and run them in a Jupyter Notebook. So, there are ways to collaborate in multiple languages. It can make things more difficult at times but it's doable.





In-Class Project



We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). [Relation between Optimism and Lipids in Midlife](#). The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-3
- Critique reproducibility



We will attempt to reproduce the results from and critique the reproducibility of:

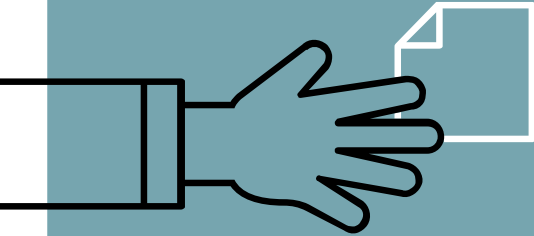
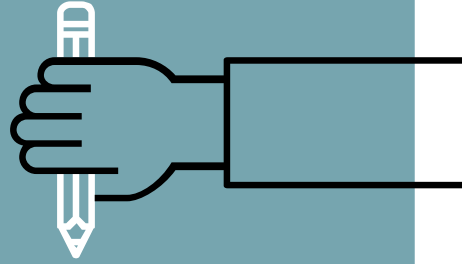
[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). [Relation between Optimism and Lipids in Midlife](#). The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-3
- Critique reproducibility



Individual Project



Reproducing a NYT Post

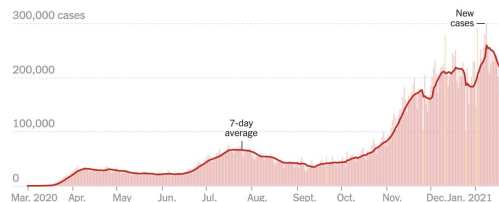
Data visualization is an incredibly powerful tool that can affect health policy decisions. Ensuring they are easy to interpret, and more importantly, showcase accurate insights from data is paramount for scientific transparency and the health of individuals. For this assignment you are tasked with reproducing COVID-19 visualizations and tables published by the [New York Times](#).

Specifically, you will attempt to reproduce the following for January 17th, 2021:

1. New cases as a function of time with a rolling average plot - the first plot on the page (you don't need to recreate the colors or theme)
2. Table of cases, hospitalizations and deaths - the first table on the page

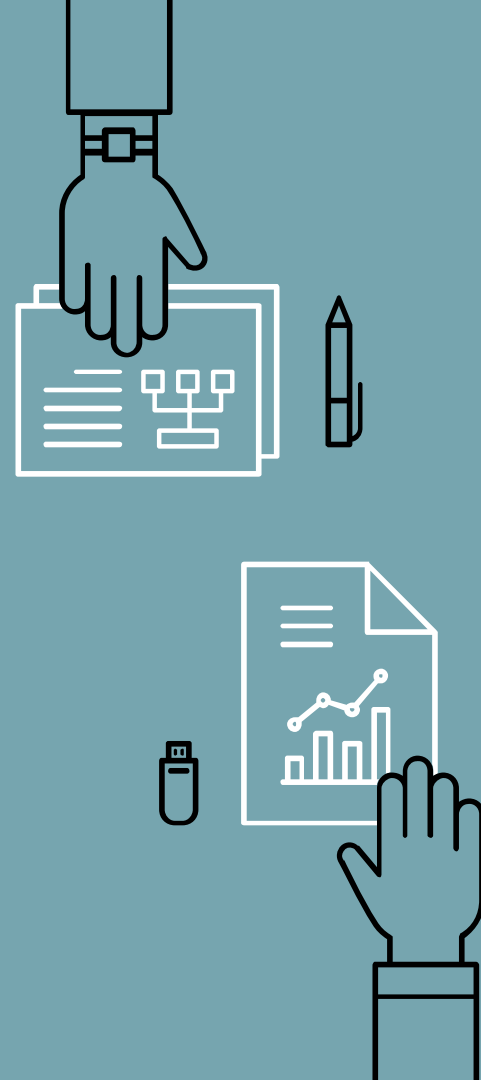
Coronavirus in the U.S.: Latest Map and Case Count

Updated January 18, 2021, 7:56 A.M. E.T.
[Leer en español](#)



	TOTAL REPORTED	ON JAN. 17	14-DAY CHANGE
Cases	23.9 million+	169,641	+3% ↗
Deaths	397,612	1,730	+26% ↗
Hospitalized		124,387	+3% ↗

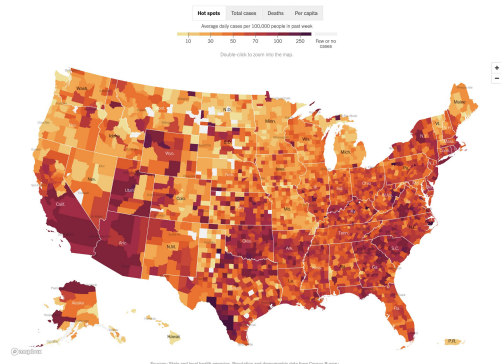
Day with reporting anomaly. Hospitalization data from the Covid Tracking Project; 14-day change trends use 7-day averages.



Reproducing a NYT Post











3. The county-level map for previous week ('Hot spots') - the second plot on the page (only the 'Hot Spots' plot)

4. Table of cases by state - the second table on the page (do not need to include per 100,000 or per capita columns)



Cases and deaths by state and county

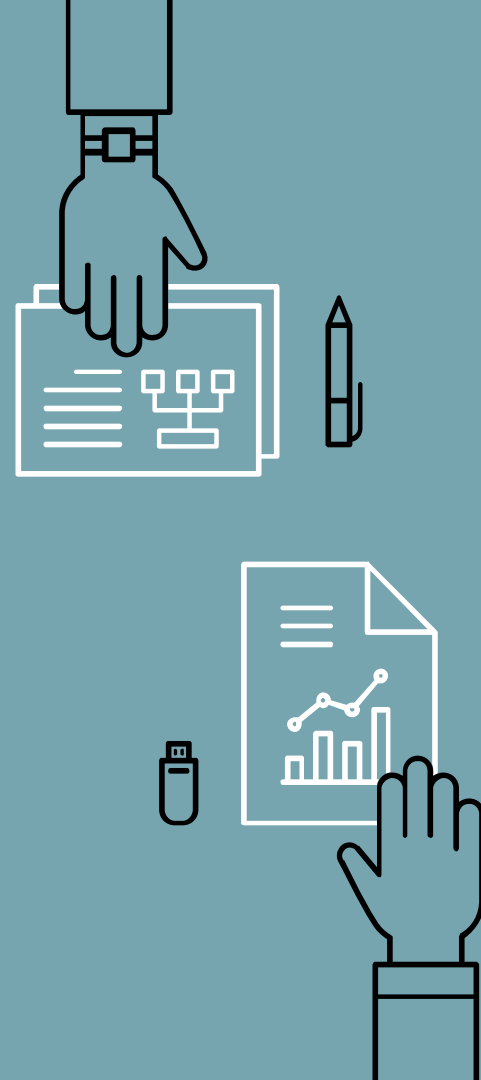
This table is sorted by places with the most cases per 100,000 residents in the last seven days. Charts are colored to reveal when outbreaks emerged.

Cases	Deaths	Search counties							
			TOTAL CASES	PER 100,000	DAILY AVG. IN LAST 7 DAYS*	PER 100,000	WEEKLY CASES PER CAPITA		
							FEWER	MORE	
→ Arizona MAP →			673,882	9,258	7,905	109		March 1	June 1
→ California MAP →			3,006,583	7,609	39,580	100			
→ South Carolina MAP →			388,194	7,539	4,808	93			
→ Rhode Island MAP →			104,443	9,859	976	92			
→ Oklahoma MAP →			354,979	8,971	3,374	85			
→ Georgia MAP →			791,322	7,453	8,457	80			
→ Utah MAP →			323,837	10,101	2,548	79			
→ Texas MAP →			2,127,334	7,337	22,782	79			
→ New York MAP →			1,242,818	6,389	15,281	79			
→ Massachusetts MAP →			470,140	6,821	5,336	77			

Reproducing a NYT Post

Notes:

- ▶ You can download the files [here](#). You will need to create a repository (with README file) [here](#).
- ▶ You don't need to make the plots look exactly the same as the post - just showcase the most important information
- ▶ You don't need to make the tables look pretty, but you do need to print out at least some of the rows to check the numbers against the NYT post
- ▶ Due: Monday, January 25th by 11:59pm EST
- ▶ You will have class time today, tomorrow and Thursday to work on it.
 - You don't have to stay on Zoom and can work on it at any time, but Matt and I will stay on Zoom until 11am each day if you have any questions or need help



Homework

- Watch Module 6 part 2 videos
 - 6.3.11 – 6.5.1
- [Submit Module 6 part 2 discussion points](#)

