

SMOOTH SAILING

a seaborn primer using
open health data

Kaitlin Throgmorton, MLIS

Quick-start, low-code methods for visualizing open health data.

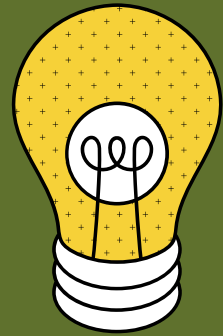
Today, we're going to walk through the **lifecycle** of **selecting**, **filtering**, and **visualizing** a dataset with seaborn, a Python library.

This workshop is aimed at **anyone who wants to quickly get up and running with data analysis and exploration**, and is meant to be approachable for most levels.



Access the workshop GitHub repo:

https://github.com/kthrog/dataviz_workshop



OUTCOMES FOR TODAY'S SESSION

1 Access open health data

Learn where to find and how to access open health data created by U.S. government sources such as Centers for Disease Control and Prevention (CDC)

2 Use an API

Learn what an API is and how to use one, including functions like filtering and querying

3 Create visualizations

Visualize data with **seaborn**, an open-source Python library for attractive, informative statistical graphs



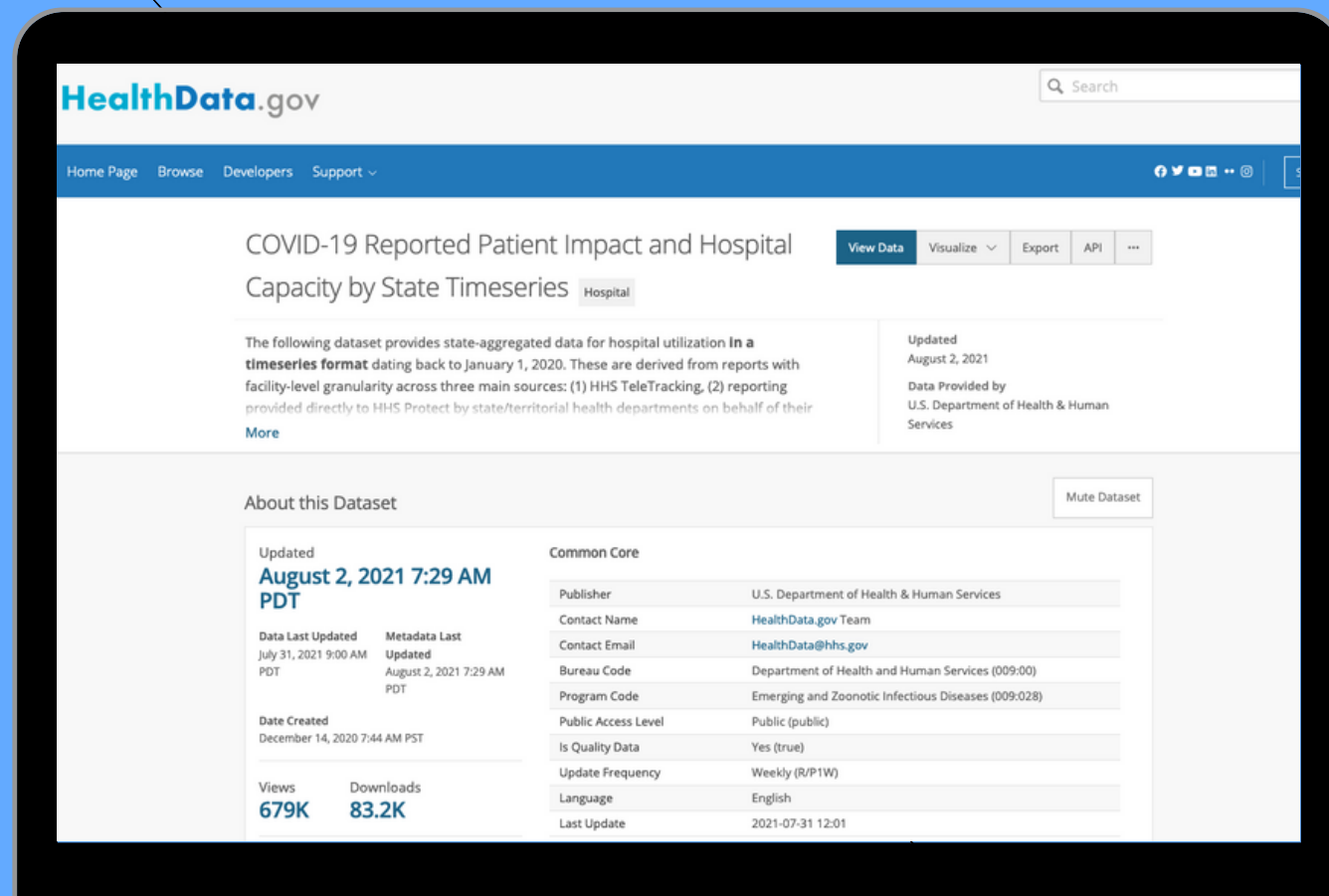
FINDING & ACCESSING OPEN HEALTH DATA

OPEN HEALTH DATA RESOURCES FROM THE U.S. GOVERNMENT

- data.gov
- healthdata.gov
- data.cdc.gov
- [NIH-Supported Data Repositories](#)
- ...and many more!

(Kim, 2019; Powell, 2021)





TODAY'S DATASET

COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries

healthdata.gov

(U.S. Department of Health & Human Services, 2020; Sainato, 2021)



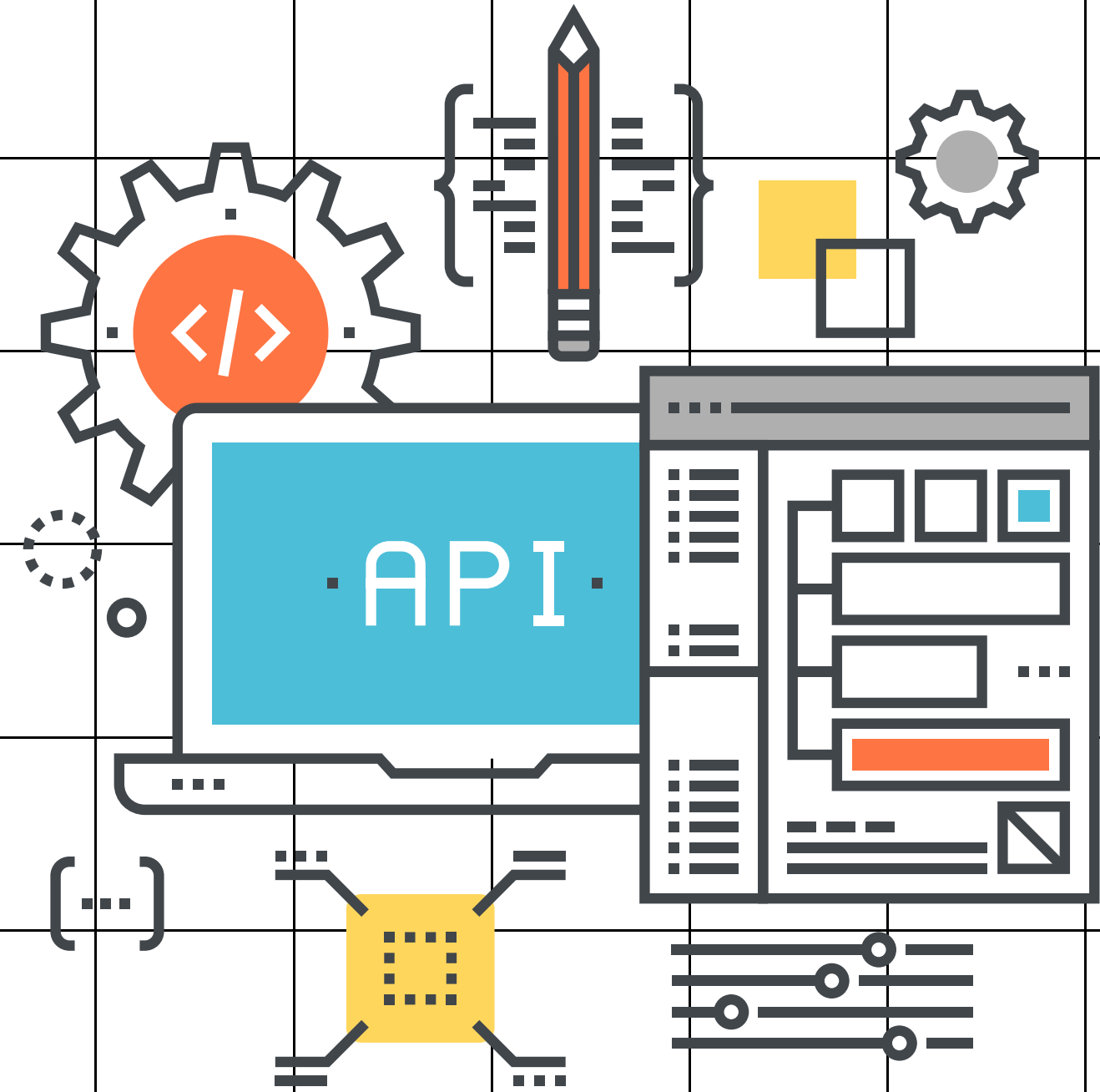
API FILTERING & QUERYING

WHAT'S AN API?

API stands for **A**pplication **P**rogramming **I**nterface.

APIs provide structured access to data; users then access the API's structured data via programmatic interfaces such as code or protocols.

When designed well, they can be a very powerful **data retrieval** and **manipulation** tool.



AN EXAMPLE: SOCRATA OPEN DATA API



The Socrata Platform

Many open government datasets are accessed via the Socrata data platform, including data.gov, cdc.data.gov, healthdata.gov, data.ct.gov

Socrata Open Data API (SODA)

Datasets on Socrata platforms can be filtered and queried using SODA; all communication with this API is conducted via HTTPS

(Tarkowska et al., 2018)

See resources + documentation section at end (or GitHub repo)
for links to SODA documentation.

ACCESSING DATA VIA SODA

The screenshot shows a data portal interface with a top navigation bar containing 'View Data', 'Visualize', 'Export', 'API', and a menu icon. A green circle highlights the 'API' button. A green arrow points from the 'API' button to a yellow callout box on the left that says 'can also export to CSV'. Below the navigation bar, there is a section with 'Updated July 26, 2021' and 'Data Provided by U.S. Department of Health & Services'. A modal window titled 'Access this Dataset via SODA API' is open, showing a description of the SODA API and two buttons: 'Data Docs' and 'Developer Portal'. Below these buttons, the 'API Endpoint' is displayed as a table with two columns: the endpoint URL and the format. A green circle highlights the entire 'API Endpoint' section, and another green circle highlights the 'Copy' button.

can also export to CSV

Updated
July 26, 2021

Data Provided by
U.S. Department of Health & Services

Access this Dataset via SODA API

The Socrata Open Data API (SODA) provides programmatic access to this dataset including the ability to filter, query, and aggregate data.

Data Docs Developer Portal

API Endpoint	
<code>https://healthdata.gov/resource/g62h-syeh.js</code>	JSON

Copy

See resources + documentation section at end (or GitHub repo)
for links to SODA documentation.

ACCESSING DATA VIA SODA

View Data

Visualize

Export

API

...

Updated
July 26, 2021

Data Provided by
U.S. Department of Health & Human Services

Access this Dataset via SODA API

The Socrata Open Data API (SODA) provides programm access to this dataset including the ability to filter, quer aggregate data.

API Docs

Developer Portal

API Endpoint

https://healthdata.gov/resource/g62h-syeh.jso

JSON

Copy

Fields

Each column in [the dataset](#) is represented by a single `field` in its SODA API. Using [filters](#) and [SoQL queries](#), you can search for records, limit your results, and change the way the data is output. For example, you could filter this dataset by its `state` field using a query like the following:

try it docs copy json

https://healthdata.gov/resource/g62h-syeh.json?state=AL

For richer filtering, you can combine filters together by stacking parameters on your URL or by using [SoQL queries](#). Learn more about about each of the fields in this dataset by clicking the headers below, or read more about the SODA API using the navigation at the top of the page.

Learn more about:

- Simple Filters
- SoQL Queries
- Available SoQL Functions
- Paging Through Data

state	text	state
date	floating_timestamp	date
critical_staffing_shortage_toda...	number	critical_staffing_shortage_today_yes
critical_staffing_shortage_toda...	number	critical_staffing_shortage_today_no
critical_staffing_shortage_toda...	number	critical_staffing_shortage_today_not_reported
critical_staffing_shortage_anti...	number	critical_staffing_shortage_anticipated_within_week_yes
critical_staffing_shortage_anti...	number	critical_staffing_shortage_anticipated_within_week_no

See resources + documentation section at end (or GitHub repo) for links to SODA documentation.

CONSTRUCTING API ENDPOINTS & ADDING FILTERS

<https://healthdata.gov/resource/g62h-syeh.json>

DATASET IDENTIFIER RESPONSE TYPE (DATA FORMAT)

<https://healthdata.gov/resource/g62h-syeh.json?state=CT>

KEY = VALUE

SIMPLE FILTER

See resources + documentation section at end (or GitHub repo)
for links to SODA documentation.

WRITING A SOQL QUERY

- Retrieve more (or less) rows than default (=1000) with **\$limit** parameter

[https://healthdata.gov/resource/g62h-syeh.json?\\$limit=50000](https://healthdata.gov/resource/g62h-syeh.json?$limit=50000)

- Retrieve rows based on **date** variable with a **\$where** parameter

[https://healthdata.gov/resource/g62h-syeh.json?\\$where=date%20between%20%272020-11-01T12:00:00%27%20and%20%272021-07-28T12:00:00%27](https://healthdata.gov/resource/g62h-syeh.json?$where=date%20between%20%272020-11-01T12:00:00%27%20and%20%272021-07-28T12:00:00%27)

KEY = VALUE RANGE

DATE RANGE = NOV' 2020 - JUL' 2021

See resources + documentation section at end (or GitHub repo)
for links to SODA documentation.

WRITING A SOQL QUERY

- Chain these all together (use & to combine) to narrowly slice the data

`https://healthdata.gov/resource/g62h-syeh.json? $limit=50000&state=CT&$where=date%20between%20%272020-11-01T12:00:00%27%20and%20%272021-07-28T12:00:00%27`

WHY USE AN API TO ACCESS DATA?

- "APIs...enhance adherence to FAIR data principles..."
(Tarkowska et al., 2018)
- Automated dataset updates
- Reduced storage and processing issues if filtering and querying performed in advance
- Sometimes, it's the only option a data repository offers!



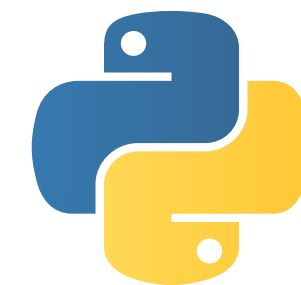


VISUALIZING DATA WITH SEABORN

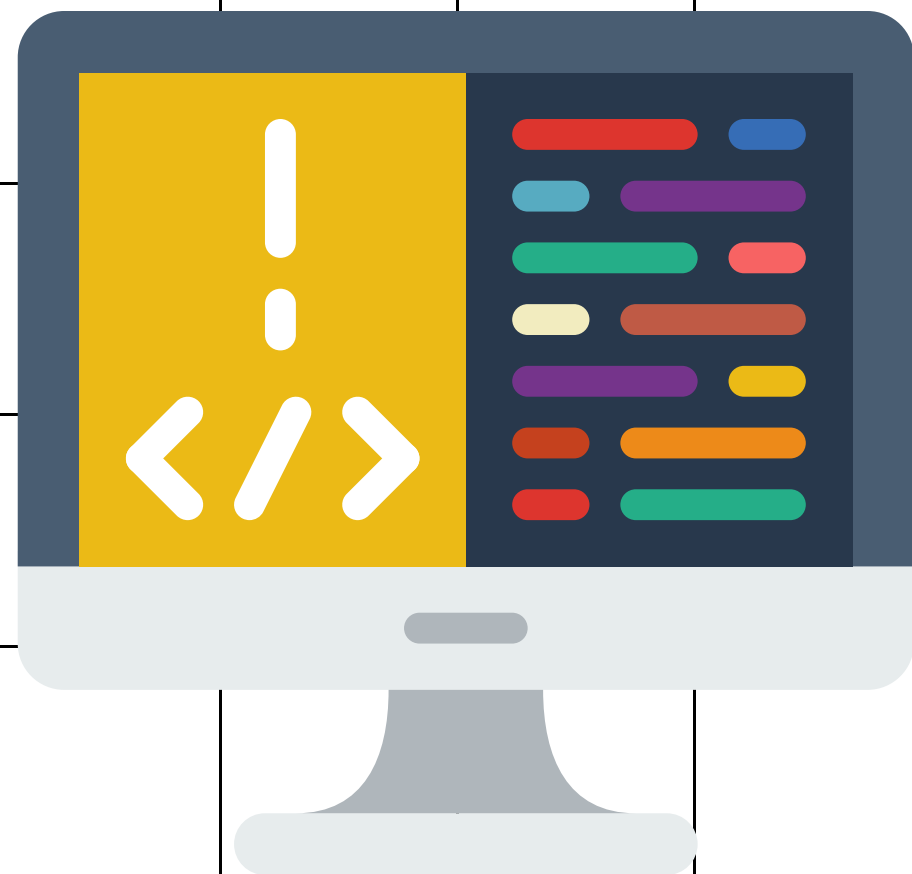


WHY USE SEABORN

- "Makes it easy to translate questions about data into graphics that can answer them"
- Excellent documentation
- Extensive gallery of attractive statistical graphs with many customization options
- Python library, built on pandas, scipy, and numpy, and meant as high-level interface for matplotlib



(Waskom, 2021)



WHAT WE'LL BE DOING

In a Jupyter notebook (a tool we'll be using to write, test, and run our code), we'll be using Python and several of its libraries to:

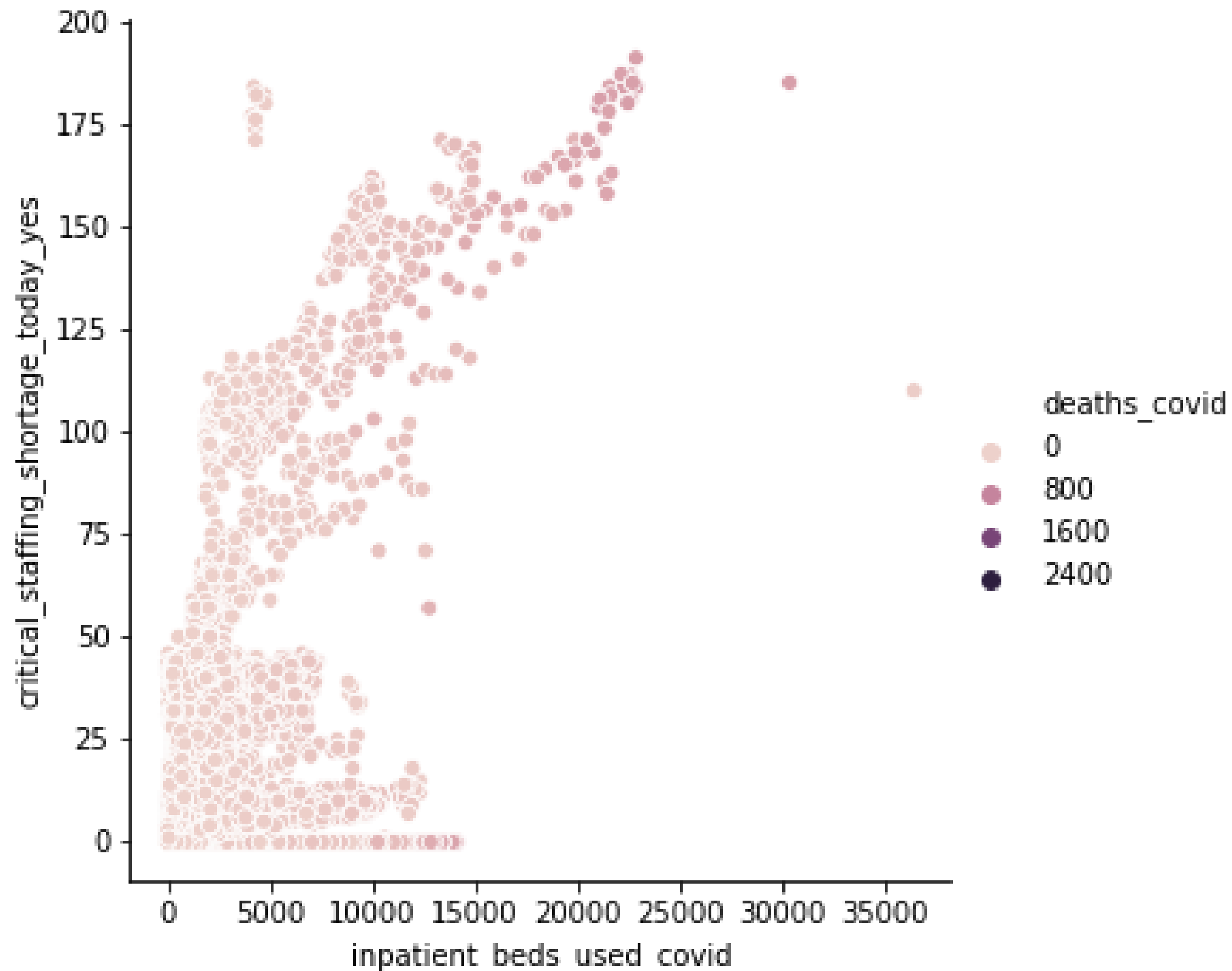
- Read data (API endpoint in JSON format) in as dataframe via pandas `pd.read_json()` function.
- Graph several charts in seaborn.

To clone the workshop GitHub repo, navigate to a directory where you want the files to be on your device, and type the following into a terminal window:

```
git clone https://github.com/kthrog/dataviz_workshop
```


LIVE DATA RESOURCE & CODE DEMO

Follow along by downloading and using this Jupyter notebook:
https://github.com/kthrog/dataviz_workshop/blob/main/materials/seaborn_data_viz_blank_for_demo.ipynb

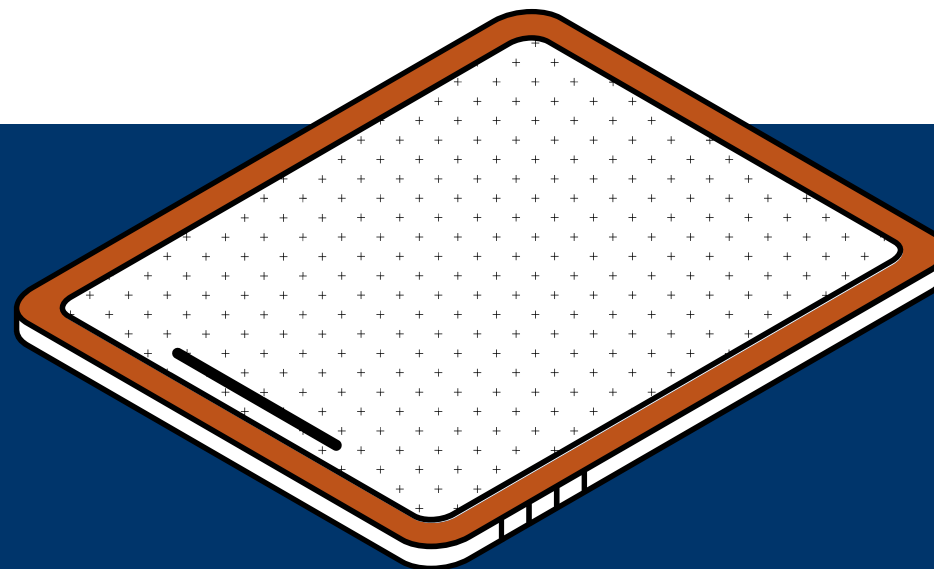
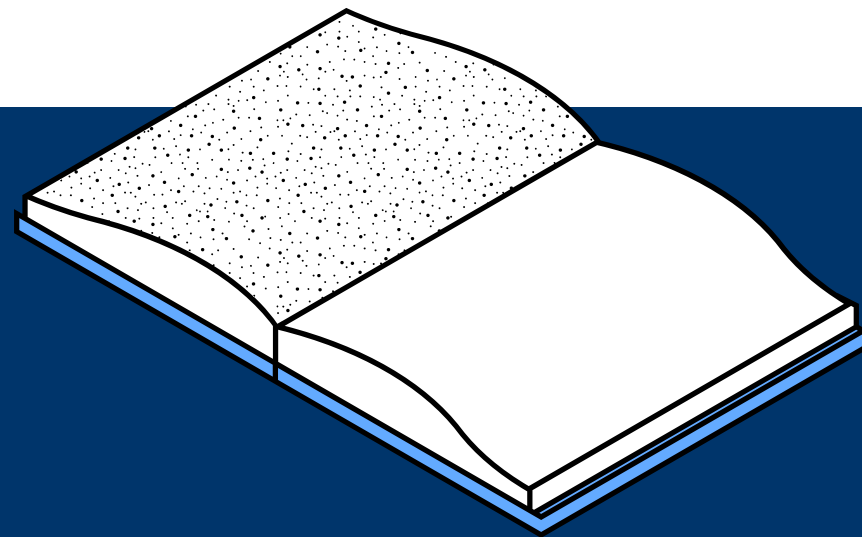


Final Figure.

Using DHHS data on U.S. hospitals, we've plotted inpatient beds occupied by COVID-19 patients versus whether a critical staffing shortage is occurring, with dots shaded by incidence of COVID-19 deaths.

POTENTIAL USE CASES

- Comparing your own data that you've generated to a larger dataset — e.g., if you had access to Yale Hospital's staffing and COVID inpatient bed usage, you could compare those numbers to these national ones
- Exploring public data you're interested in, quickly, to see if you actually want to interact with / download the full dataset
- Practicing data viz!





**HOW ELSE MIGHT
YOU USE THIS
KNOWLEDGE IN
YOUR STUDIES,
RESEARCH, OR
WORK?**

THANK YOU!

REFERENCES

Kim, H. (2019). Data.gov at Ten and the OPEN Government Data Act. Data.gov. <https://www.data.gov/meta/data-gov-at-ten-and-the-open-government-data-act/>

Powell, K. (2021). The broken promise that undermines human genome research. *Nature* 590, 198-201. <https://doi.org/10.1038/d41586-021-00331-5>

Sainato, M. (2021). 'We went from heroes to zeroes': US nurses strike over work conditions. *The Guardian*. <https://www.theguardian.com/society/2021/jul/30/us-nurses-strike-covid-coronavirus-conditions-understaffing>

Tarkowska, A., Carvalho-Silva, D., Cook, C.E., Turner, E., Finn, R.D., Yates, A.D. (2018). Eleven quick tips to build a usable REST API for life sciences. *PLoS Computational Biology* 14(12): e1006542. <https://doi.org/10.1371/journal.pcbi.1006542>

U.S. Department of Health & Human Services. (2020). COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries. <https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh>

Waskom, M. L., (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>

RESOURCES & DOCUMENTATION

RESOURCES:

Workshop GitHub Repo

https://github.com/kthrog/dataviz_workshop

Data

<https://www.data.gov/>

<https://www.healthdata.gov/>

<https://data.cdc.gov/>

<https://data.cdc.gov/browse>

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Visualization

10 Simple Rules for Better Figures | *PLOS Comp Bio*

<https://doi.org/10.1371/journal.pcbi.1003833>

How to Choose the Right Data Visualization | *Chartio*

<https://chartio.com/learn/charts/how-to-choose-data-visualization/>

DOCUMENTATION:

Python

<https://www.python.org/>

Jupyter Notebook

[https://jupyter-](https://jupyter-notebook.readthedocs.io/en/stable/notebook.html#introduction)

[notebook.readthedocs.io/en/stable/notebook.html#introduction](https://jupyter-notebook.readthedocs.io/en/stable/notebook.html#introduction)

Socrata Open Data API (SODA)

<https://dev.socrata.com/>

Seaborn

<https://seaborn.pydata.org/>

Pandas

<https://pandas.pydata.org/>

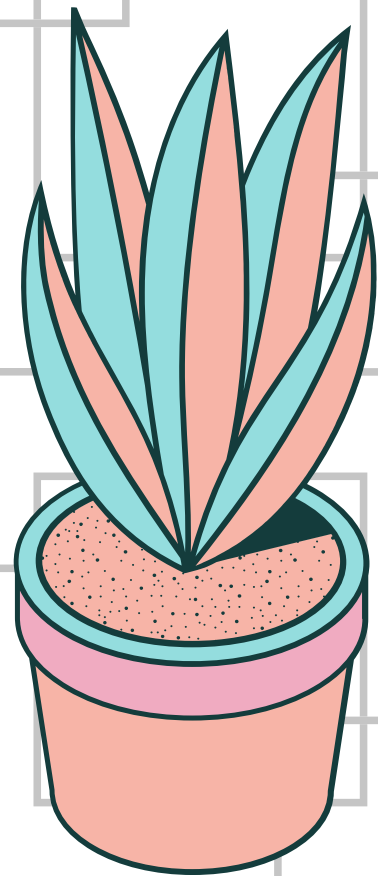
For full list of resources, references, and documentation:

https://github.com/kthrog/dataviz_workshop/edit/main/materials/resources.md

RATIONALE FOR WORKSHOP DECISIONS



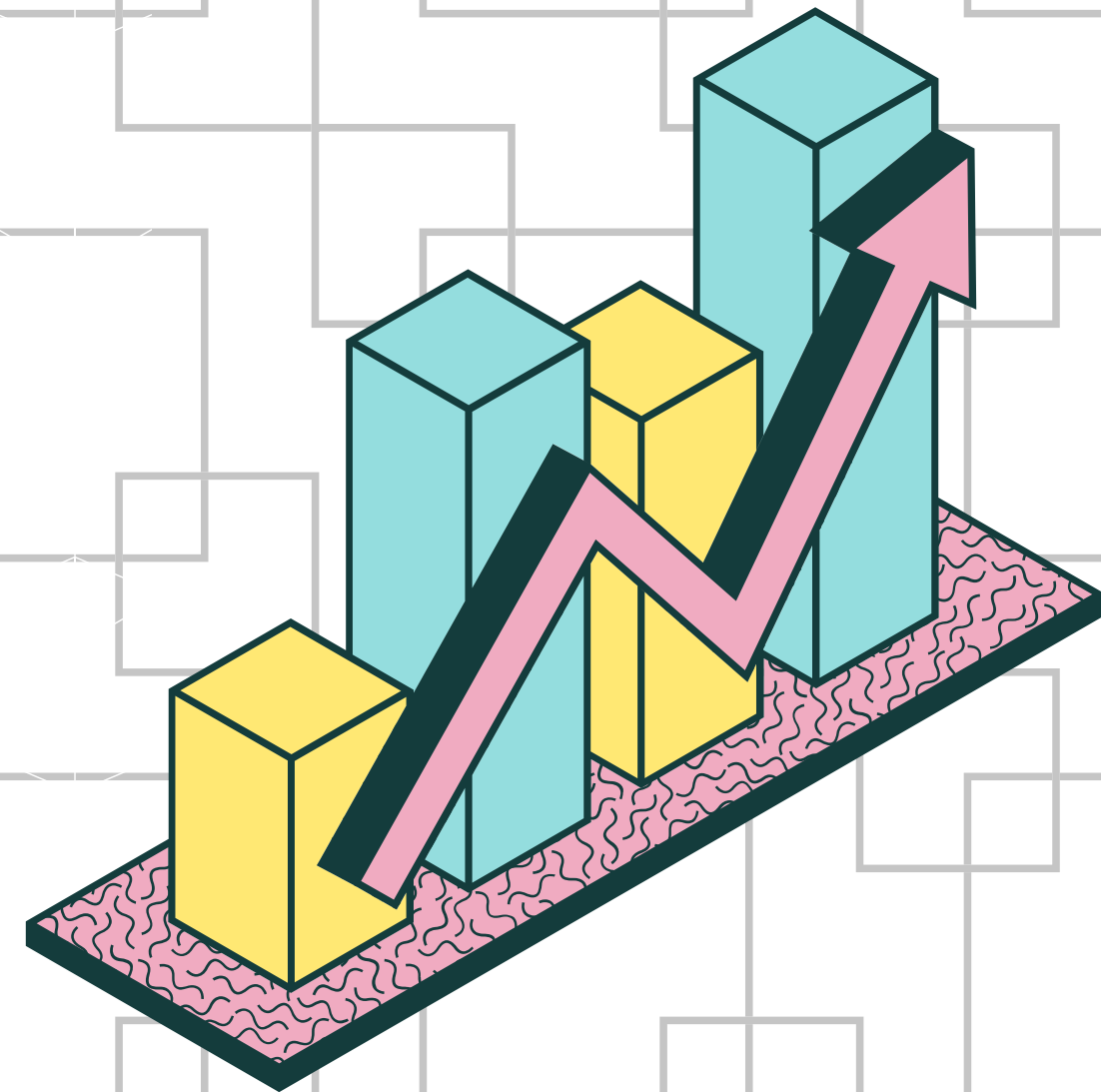
MAIN ETHOS



- **Keep it approachable, applicable, and user-focused**
- **Cultivate a growth mindset**
- **Take any opportunity to boost literacy**
- **Stoke curiosity and enthusiasm**

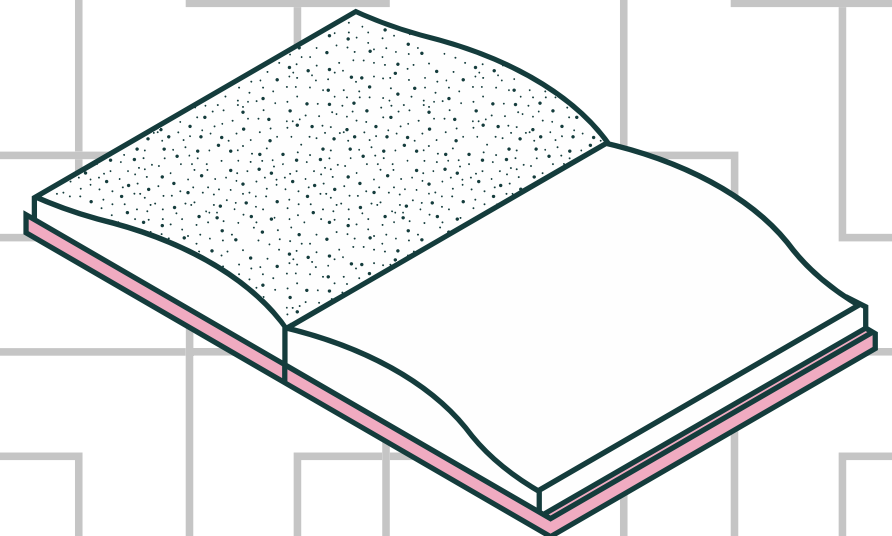
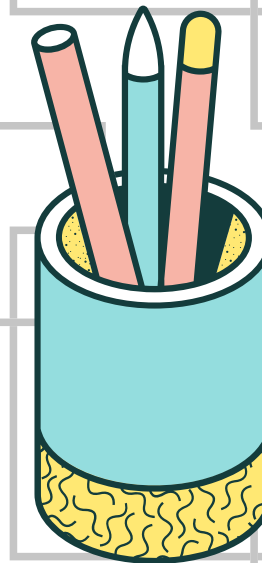
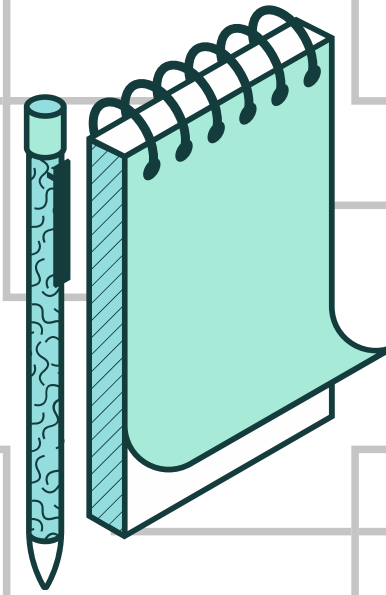
WHY: TOPIC

- Chose multi-part topic in order to give learners full snapshot of process, from dataset acquisition to visualization
- seaborn seemed perfect for this short format – with its single function call for plotting



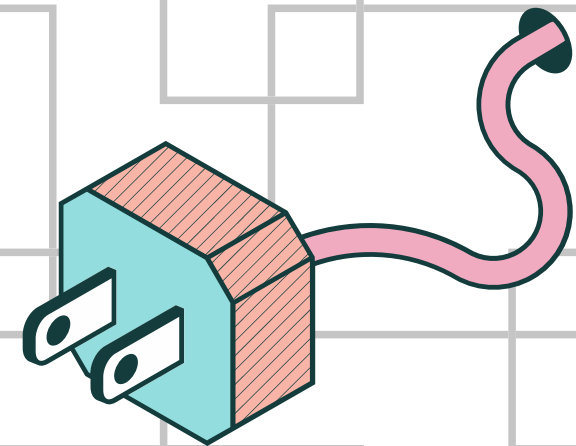
METHODS & DEVELOPMENT

- Used backwards course design to develop
- Wanted multi-layered session approachable to multiple levels
- Live coding inspired by Software Carpentry approach



OTHER NOTES

- In a non-interview scenario, I would:
 - Send tool installation instructions in advance
 - Assess before and after session
 - Incorporate multiple activities and more discussion questions (though I might need a bit more time to do this well)
 - Test session (especially notebook demos) on multiple device types (e.g., PC, Chromebook, iPad, etc.)



THANK YOU!

QUESTIONS?