

# Kickoff

Peter Ganong and Maggie Shi

October 16, 2024

## Ahead of today's lecture

- ▶ If you haven't already: `pip install palmerpenguins`

## New student repo

- ▶ New student repo: <https://github.com/uchicago-harris-dap/student30538> (link)
- ▶ You will have to fork it again (see instructions in README.md)
- ▶ Simpler structure: just one branch, with separate folders for `before_lecture` and `after_lecture`

## Tip from Ed

- ▶ Jakub found a cool extension for viewing and cleaning data in VSCode (similar to RStudio)
- ▶ <https://code.visualstudio.com/docs/datascience/data-wrangler> (link)

The screenshot shows the Data Wrangler interface within VS Code. The left sidebar contains the 'OPERATIONS' panel with a search bar and a list of operations: Find and replace (4), Format (7), Formulas (4), Numeric (4), Schema (5), Sort and filter (2), Custom operation, Group by and aggregate, and New column by example. Below this is the 'DATA SUMMARY' panel showing the data shape (3,818 rows x 12 columns), the number of columns (12), the total number of rows (3,818), and the number of missing values by column (4,809). At the bottom is the 'CLEANING STEPS' panel.

The main panel displays a table with 3,818 rows and 12 columns. The columns are: #, index, host\_is\_superhost, neighbourhood, property\_type, room\_type, and #. The table shows the first 10 rows of data. The 'neighbourhood' column has a value of 'Queen Anne' for all rows. The 'property\_type' column has values 'Apartment', 'House', and 'Cabin'. The 'room\_type' column has values 'Entire home/apt', 'Private room', and 'Shared room'.

#	index	host_is_superhost	neighbourhood	property_type	room_type	#
0	f	80%	Capitol Hill:	9%	House:	45%
1	t	20%	Ballard:	6%	Apartment:	45%
2	f		Belletown:	5%	Townhouse:	3%
3	f		Other:	69%	Other:	7%
4	f				Entire home/apt:	67%
5	f				Private room:	30%
6	t				Shared room:	3%
7	t					
8	f					
9	t					

The bottom status bar shows the 'DATA WRANGLER' tab is active, and the '3 New operation' message is displayed.

## MUD concerns about workload

- ▶ We removed a section from PS2. Revised version is in student repo.
- ▶ We are adding office hours this week
  - ▶ Akbar over zoom on Friday
  - ▶ Joaquin in person Thursday
  - ▶ *You will get priority if you post your question to the office hours thread on Ed before you come*
- ▶ Looking forward, PS3 is shorter than PS1 or PS2 as best as we can tell

## MUD other

- ▶ Will we require that you transform your dataset in altair or pandas? *Answer: no, use what works best for you*
- ▶ Can altair functions (e.g., `transform_filter()`) be used outside of altair? *Answer: no, altair only transforms within the plot it's making*
- ▶ Final project: why just one member present? *Answer: presentation slots are 8 minutes and we want to reduce transition time.*
- ▶ Challenges debugging – we will do a mini-lesson on this next week
- ▶ Good questions we will have answers to by Monday (but don't have yet, sorry!)
  - ▶ What is a concrete example of a time/setting when winsorizing would be the best path for handling outliers
  - ▶ Adding better axis labels (motivated by the diamonds histogram for carat)

## Final Project Optional Meetings

- ▶ Preferred but not required: send questions in advance via private Ed post tagged “Final project Ganong”
- ▶ Ganong’s sections: Next week *during* lab in room 0049. Will have a signup at the start of lab.
- ▶ Shi’s sections: office hours.