Materials
- **the worksheet:** tinyurl.com/d8discussion2
- **slide template:** kevin-miao.com (under Teaching)

DATA 8
Spring 2021

Discussion 2

Introduction to Tables

Introduction

- 👋 I'm **Kevin Miao**, nice to meet you!
  - Senior in *Computer Science*
  - Email: kevinmiao@berkeley.edu
  - Office Hours:
    - **Thursday** from **1-2 PM** or email me!
    - Ask questions after lab & discussion

    Wednesdays, Fridays from 9-9:30 AM
- From Eindhoven, The Netherlands 🇳🇱
- I am really interested in researching Explainable Deep Learning
- In my free time, I cook, workout, take photos or listen to Taylor Swift music.
- Reach out to me if you have any questions or just want to chat!

Break out rooms

Introduce yourself to each other

Some suggestions:

- What is your year/major/hometown?
- What ice breaker do you dislike the most?

Assignment:

➔ Report something that you all have in common.

I encourage you to exchange contact info to form study groups, if you feel comfortable doing so!

Today

- Introductions & Icebreakers
- Administrivia
- Mini-review
- Worksheet
  - Questions on:
    1. Tables
    2. Programming & Causality

Administrivia: Announcements

- **Vitamin 1** will be **due today**
  - The last question will be given to you sometime during section
- **Homework 1** is **due tomorrow**
  - For a bonus point, submit by tonight!
- DSP students: Make sure to talk to your DSP advisor and have them send a letter through AIM
- Tutoring Sections (~4-5 students) signups are/will be released on Piazza. Great way to learn the material in a different setting.
- Please complete the **welcome survey** in the email
- *How are you feeling?*

Administrivia: Course Policy

- Discussion attendance/participation is mandatory (5% of your grade; 2 drops)
  - No credit for attending another section
  - Let me know if you are going to be late
- Lab Credit
  - (Accuracy) Submit lab by Wednesday 9AM

  OR
  - (Effort) Attend your assigned lab section
- Final Exam held on May 11, 3-6PM & Midterm TBD
- Academic Honesty

For more info, see syllabus

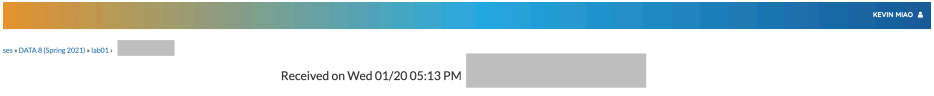
Administrivia: Set-Up

- **DataHub** is your notebook for this class
- **OkPy** is for your submissions of your notebooks
- **Gradescope** for vitamins & written homework questions/Exam regrades
- **Piazza** for questions and posts on course logistics
- **OH.data8.org** for (M-F) Office Hours Queue
- **Frequent Technical Issues**
  - OkPy
  - Jupyter Notebook

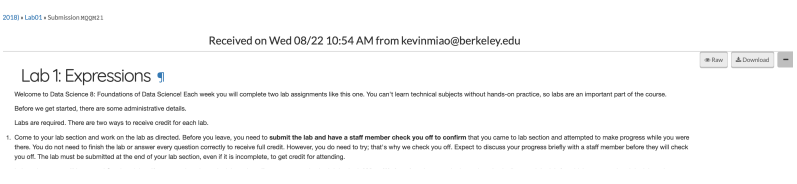
Are you experiencing any issues? Please stay at the end of discussion.

FTI: Empty OkPy submission

- **Problem:** When I run the cell `_ = ok.submit()` and follow the link, it shows up like this

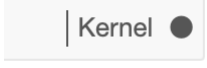


instead of



- **Solution**: Try `save and checkpoint`, `submit` again. If that does not work, please create a Piazza post.

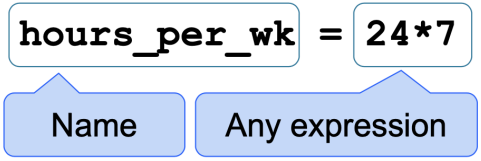
FTI: Jupyter

- **Problem:** Jupyter Notebook froze and the kernel symbol in the right corner looks like this 

  - **Solution:** `Save and checkpoint` the notebook, restart the kernel, run the notebook from top to bottom again.
- **Problem**: I removed code created by staff
- **Solution:** You can rename the notebook and re-click the link on the hw/project/lab link

Tables & Programming

- Python programming relies heavily on **assignment statements**



- **Tables** is a representation of **data**
  - Each **row** represents one individual
  - Each **column** represents one attribute
    - A **label** is that attribute

Causality

- What are **associations**?

  Relationship between variables ; link

- Is an **association** automatically a **causation**?

  no, it's not automatically a causal relationship

- How do we test for **causation**?

  Randomized Control Experiment

- What are **confounding factors**?

  Any variable that distracts you from the variable that you are analyzing

To the worksheet! ✍️

tinyurl.com/d8discussion2

1. Ready, Willing and Table

Let's look at an example table called `staff` (*shown in two parts)*

Name	Year	Semesters on Staff
Anuja	3	3
Tam	4	6
Angela G	4	6
Pulkit	2	2
Sam	4	7
Carlos	3	2
Yanay	4	7
Margaret	3	5
Greg	4	6
Natalie	4	4
Nicole	2	2
Aanika	4	5
Thomas	4	3
Will	2	1
Ritvik	3	4
Rujula	2	1
Angela Z	4	5
Stephanie D	4	5
Rithvik	2	1
Ruhi	3	4
Joyce	3	2
Stephanie X	2	2
Meghan	3	5
Parham	4	6
King	3	2
Ellen	3	5
Eddie	2	1
Josh	3	1
Kevin	4	4

The table has 29 rows, each corresponding to one member of Data 8 Staff. Each row has three attributes, the staff member's name, year, and how many semesters they have been on staff. Using just the information from the staff table, do we have enough information to generate the following by hand? If not, what additional information do you need? *(You don't need to worry about how you'd do it in Python.)*

1a. Ready, Willing and Table

Using just the information from the staff table, do we have enough information to generate the following by hand? If not, what additional information do you need? *(You don't need to worry about how you'd do it in Python.)*

Year	Semesters on Staff average
2	1.42857
3	3.3
4	5.33333

T / F

You seperately calculate the average for 2nd, 3rd, 4th years

1b. Ready, Willing and Table

Name	Year
Anuja	Junior
Carlos	Junior
Margaret	Junior
Ritvik	Junior
Ruhi	Junior
Joyce	Junior
Meghan	Junior
King	Junior
Ellen	Junior
Josh	Junior

Using just the information from the staff table, do we have enough information to generate the following by hand? If not, what additional information do you need? *(You don't need to worry about how you'd do it in Python.)*

T / F

2nd => Sophomore

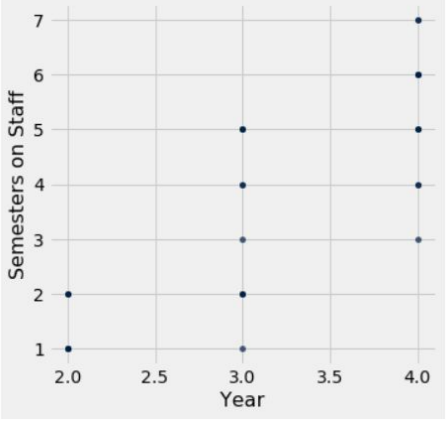
3rd => Junior

...

↳ This labeled data is not present in our 'Staff' table.

1c. Ready, Willing and Table

Using just the information from the staff table, do we have enough information to generate the following by hand? If not, what additional information do you need? *(You don't need to worry about how you'd do it in Python.)*



T / F

You have enough info!

graph a plot; each person is a dot!

1d. Ready, Willing and Table

Using just the information from the staff table, do we have enough information to generate the following by hand? If not, what additional information do you need? *(You don't need to worry about how you'd do it in Python.)*

Semesters on Staff	2	3	4
1	4	1	0
2	3	3	0
3	0	1	1
4	0	2	2
5	0	3	3
6	0	0	4
7	0	0	2

(T) / F

We have enough data!
We group people together on 'semesters' & 'year' and count.

2a. Causality, Coworkers and Coffee

Divyesh collected the following information about his coworkers' methods of getting to work and their coffee consumption.

Method	Number of Coworkers	Average Cups of Coffee per Day
Take the Bus to Work	12	1.1
Drive to Work	15	1.9

A. Divyesh is trying to compute the absolute value of the difference between the total number of cups drank by driving coworkers per year vs the total number of cups drank by busing coworkers per year. He will do all of this in a single cell. Identify the errors in the following cell and correct them. *Make sure that the code cell outputs a single positive number.*

```
number_cups_bus = 12(1.1)
number_cups_drive = 15(1.9)
number_cups_day_difference = ((number_cups_bus - number_cups_drive)
number_cups_week_difference = number_cups_difference * 7
yearly cups = number_cups_week_difference * 52
```

2a. Causality, Coworkers and Coffee

Divyesh collected the following information about his coworkers' methods of getting to work and their coffee consumption.

Method	Number of Coworkers	Average Cups of Coffee per Day
Take the Bus to Work	12	1.1
Drive to Work	15	1.9

A. Divyesh is trying to compute the absolute value of the difference between the total number of cups drank by driving coworkers per year vs the total number of cups drank by busing coworkers per year. He has written the following code, but it contains a few mistakes. Find and correct them. *Make sure that the code cell outputs a single positive number.*

```
number_cups_bus = 12(1.1)
number_cups_drive = 15(1.9)
number_cups_day_difference = ((number_cups_bus - number_cups_drive)
number_cups_week_difference = number_cups_difference * 7
yearly_cups = number_cups_week_difference * 52
```

Vitamin 1

Question 5

What is the number of unique* mistakes in the code in question 2A.

2a. Causality, Coworkers and Coffee

1 no abs
2 asterisk
3 extra c
4 underscae
5 "_day"
6 outputting

A. Divyesh is trying to compute the absolute value of the difference between the total number of cups drank by driving coworkers per year vs the total number of cups drank by busing coworkers per year. He will do all of this in a single cell. Identify the errors in the following cell and correct them. *Make sure that the code cell outputs a single positive number.*

```
number_cups_bus = 12(1.1)
number_cups_drive = 15(1.9)
number_cups_day_difference = (number_cups_bus - number_cups_drive)
number_cups_week_difference = number_cups_difference * 7
yearly_cups = number_cups_week_difference * 52
print(yearly_cups)
```

abs

_day

6

2b. Causality, Coworkers and Coffee

Divyesh collected the following information about his coworkers' methods of getting to work and their coffee consumption.

Method	Number of Coworkers	Average Cups of Coffee per Day
Take the Bus to Work	12	1.1
Drive to Work	15	1.9

B. Is there a relationship between transportation method and coffee consumption—an association, a causal relationship or something else? Why?

There's a link!

People who drive to work ⟹ more coffee per day

We don't know if it's causal!

↳ It could be confounding

↳ need to do a Randomized Control Experiment

How did I do?

https://tinyurl.com/kevind8feedback