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!
        - From **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?
- Two truths, one lie!
- What ice breaker do you dislike the most?

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 2** will be **due today**
  - The answer to the last question will be given to you today
- **Homework 1** is **due tomorrow**
  - For a bonus point, submit by tonight!
- DSP students: Make sure to talk to your DSP advisor and having them send a letter through AIM
- Tutoring Sections (~4-5 students) signups are released on Piazza. Great way to learn the material in a different setting.
- 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
  - (Effort) Attend your assigned lab section
- Final Exam held on May 11, 3-6 PM & Midterm TBD
- Academic Honesty

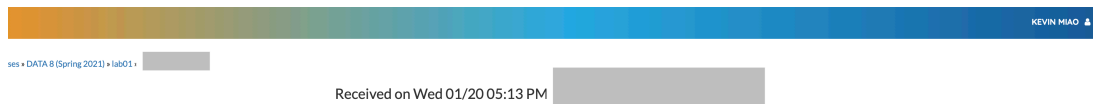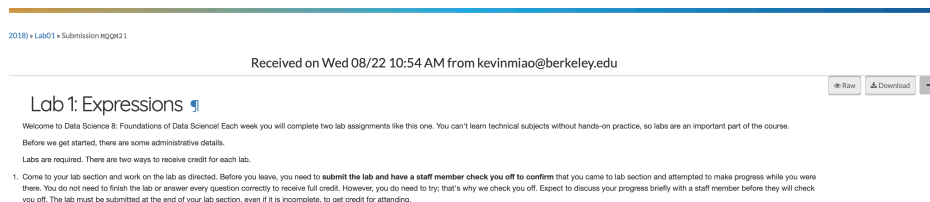For more info, see syllabus

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

FAQ: Empty OkPy submission

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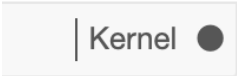

KEVIN MIAO

ses » DATA 8 (Spring 2021) » lab01

Received on Wed 01/20 05:13 PM

instead of

2018) » Lab01 » Submission MQQK21

Received on Wed 08/22 10:54 AM from kevinmiao@berkeley.edu

Lab 1: Expressions ¶

Welcome to Data Science 8: Foundations of Data Science! Each week you will complete two lab assignments like this one. You can't learn technical subjects without hands-on practice, so labs are an important part of the course.

Before we get started, there are some administrative details.

Labs are required. There are two ways to receive credit for each lab.

1. Come to your lab section and work on the lab as directed. Before you leave, you need to **submit the lab and have a staff member check you off to confirm** that you came to lab section and attempted to make progress while you were there. You do not need to finish the lab or answer every question correctly to receive full credit. However, you do need to try; that's why we check you off. Expect to discuss your progress briefly with a staff member before they will check you off. The lab must be submitted at the end of your lab section, even if it is incomplete, to get credit for attending.

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

FAQ: 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 is a language that heavily relies on **assignment statements**


```
hours_per_wk = 24*7
```
Name   Any expression
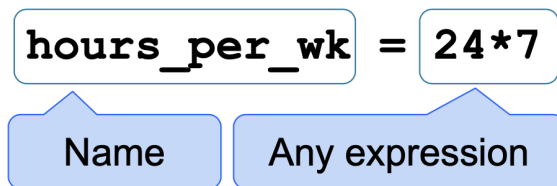
- **Tables** have
  - Each **row** represents one individual
  - Each **column** represents one attribute
    - A **label** is that attribute

Causality

- What are **associations**?

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

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

- What are **confounding factors**?

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

Name	Year	Semesters on Staff
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 inform ___ eed? *(You don't need to worry about how you'd do it in Python.)*

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

T / F

Year	Semesters on Staff average
2	1.42857
3	3.3
4	5.33333

Semesters on Staff	2	3	4
	1	4	1	0

Discussion | [Kevin Miao](#)

7

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

age

857

3.3

333

Semesters on Staff   2   3   4

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

1c. ...ing ...d Table

Year	Semesters on Staff average
2	1.42857
3	3.3
4	5.33333

Using jus... ...ble, ...nough information to generate the following ...nform... need? *(You don't need to worry about how you'd do it in Python.)*

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

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

I. ...ly, Willing and Table

Staff average
1.42857
3.3
5.33333

Margaret	Junior
Ritvik	Junior
Ruhi	Junior
Joyce	Junior
Meghan	Junior
King	Junior
Ellen	Junior
Josh	Junior

... just ... on from the staff table, do we have enough information to generate the ...ing b... ...t, 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

4.0

2a. Causality, Coworkers and Coffee

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

ffee

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
```

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?

How did I do?

https://tinyurl.com/kevind8feedback