Data 8 Spring 2024

Data Types, Extending Tables

Lab 03

Jan/Feb 2024

In lecture, you have been introduced to various *data types* in Python such as integers, strings, and arrays. These data types are particularly important for manipulating and extracting useful information out of data, an important skill for data science. In this section, we'll be analyzing some of the behavior that Python displays when dealing with particular data types.

1. ”Fun” with Arrays

Suppose we have executed the following lines of code. Answer each question with the appropriate output associated with each line of code, or write ERROR if you think the operation is not possible.

```
odd_array = make_array(1, 3, 5, 7)

even_array = np.arange(2, 10, 2)

an_array = make_array(‘1’, ‘2’, ‘3’, ‘4’)
```

a. `odd_array + even_array`

b. `odd_array + an_array`

c. `even_array.item(3) * odd_array.item(1)`

d. `odd_array * 3`

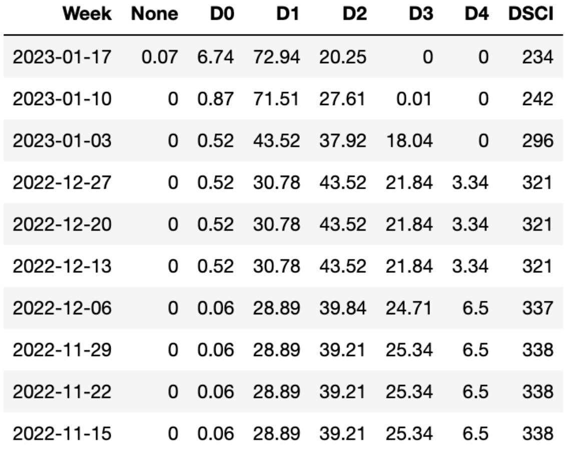
e. `(odd_array + 1) == even_array`

f. `an_array.item(3) + ‘abcd’`

In this next section, we will practice working with tables. In particular, we'll be focusing on table methods and what data types they return. This will help in understanding how to effectively manipulate tables. Remember to make use of the Python Reference guide when working through these questions – A similar guide will be provided on exams.

2. Drought

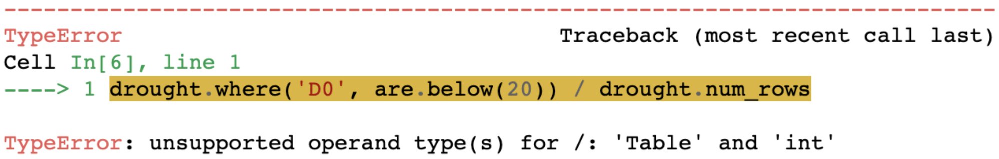
Your friend Sarah is interested in the level of drought experienced in California over time. She obtained data on the percentage of the population experiencing each level of drought (D0 being the lowest severity, D4 being the highest), as well as the DSCI score (higher = more prevalence and severity of drought). The table below is called `drought` and was obtained from the National Drought Mitigation Center at the University of Nebraska-Lincoln.



Week	None	D0	D1	D2	D3	D4	DSCI
2023-01-17	0.07	6.74	72.94	20.25	0	0	234
2023-01-10	0	0.87	71.51	27.61	0.01	0	242
2023-01-03	0	0.52	43.52	37.92	18.04	0	296
2022-12-27	0	0.52	30.78	43.52	21.84	3.34	321
2022-12-20	0	0.52	30.78	43.52	21.84	3.34	321
2022-12-13	0	0.52	30.78	43.52	21.84	3.34	321
2022-12-06	0	0.06	28.89	39.84	24.71	6.5	337
2022-11-29	0	0.06	28.89	39.21	25.34	6.5	338
2022-11-22	0	0.06	28.89	39.21	25.34	6.5	338
2022-11-15	0	0.06	28.89	39.21	25.34	6.5	338

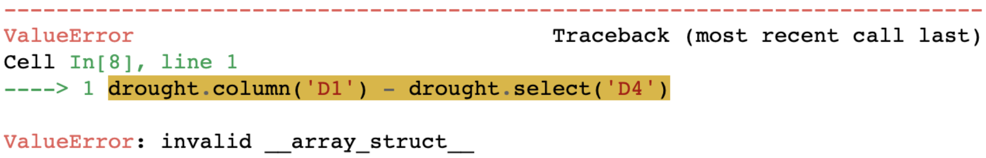
Unfortunately, the code Sarah wrote to analyze the data has some bugs. Below are some error messages that appeared, along with what Sarah was trying to calculate; describe the bugs and how you would fix them.

a. The proportion of weeks with less than 20% of people experiencing D0 drought



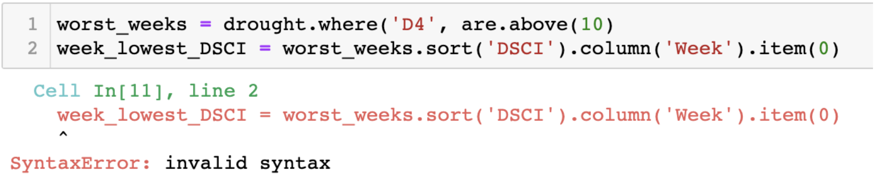
```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[6], line 1
----> 1 drought.where('D0', are.below(20)) / drought.num_rows

TypeError: unsupported operand type(s) for /: 'Table' and 'int'
```

b. The difference each week of the percent of people in D1 drought and the percent in D4 drought



```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[8], line 1
----> 1 drought.column('D1') - drought.select('D4')

ValueError: invalid __array_struct__
```

c. The week with the lowest DSCI score out of weeks with more than 10% experiencing D4 drought



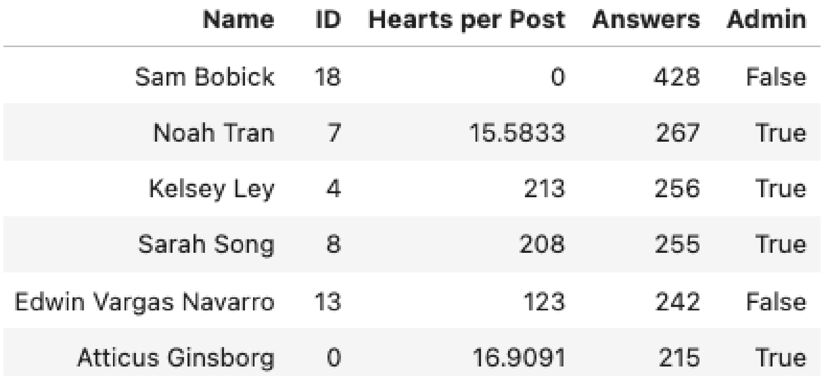
```
1 worst_weeks = drought.where('D4', are.above(10)
2 week_lowest_DSCI = worst_weeks.sort('DSCI').column('Week').item(0)

  Cell In[11], line 2
    week_lowest_DSCI = worst_weeks.sort('DSCI').column('Week').item(0)
    ^
SyntaxError: invalid syntax
```

3. EdStem is Our Friend

Data 8 staff loves to answer questions on Ed! The table `staff` contains information about the staff's Fall 2023 Data 8 Ed statistics. There are 5 columns:

- **Name**: string, name of the staff member
- **ID**: int, the ID of the staff member
- **Hearts per Post**: float, the amount of hearts awarded per post made by a staff member (0 is also given to those with 0 posts)
- **Answers**: int, number of answers to Ed questions and comments
- **Admin**: boolean, whether or not the staff member was an admin of the Data 8 FA23 Ed

Some rows are shown below:



Name	ID	Hearts per Post	Answers	Admin
Sam Bobick	18	0	428	False
Noah Tran	7	15.5833	267	True
Kelsey Ley	4	213	256	True
Sarah Song	8	208	255	True
Edwin Vargas Navarro	13	123	242	False
Atticus Ginsborg	0	16.9091	215	True

a. For each of the columns in `staff`, identify if the data contained in that column is numerical or categorical.

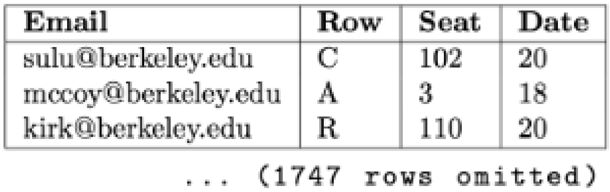
b. Professor Sahai wants to award a very lucky staff member a prize based on their Ed performance; however, he has a few things he's looking for in them: The staff member have more than 0 hearts per post, they must be an admin, and then out of all the staff members left, they must have the highest number of answers. Fill in the code blanks to find out who gets a prize.

```
good_staff = staff.where(_________________________).where(_________________________)


lucky_staff_name = ________________.sort(_____________________).column(________________).item(0)
```

4. Fa17 Midterm Q2 Modified

A table named `seat` contains a row for each time a student submitted the attendance form in lecture on September 18th, 20th, or 22nd. The table contains four columns.

- **Email**: a string, the email address of the student
- **Row**: a string, the letter of the row in which they claim to be seated
- **Seat**: an int, the number of the seat in which they claim to be seated
- **Date**: an int, the date of the submission, either 18, 20, or 22.



Email	Row	Seat	Date
sulu@berkeley.edu	C	102	20
mccoy@berkeley.edu	A	3	18
kirk@berkeley.edu	R	110	20

```
... (1747 rows omitted)
```

Fill in the blanks of the Python expressions to compute the described values. You must use *all* and *only* the lines provided. The last (or only) line of each answer should evaluate to the value described.

a. The largest seat number in the `seat` table.

Method 1:

```
max(________________________________)
```

Method 2:

```
_____________________.sort(_____________________, _____________________)
      ._____________________._____________________
```

b. The total number of attendance submissions for September 20th in rows A, B, C, D, or E.

Hint: You can use Table.where predicates to compare letters lexicographically (e.g. A is below B), or use a different Table.where predicate

```
u = seat._____________________(_____________________, _____________________)
u._____________________(_____________________, _____________________)._____________________
```