

# Discussion 3

---

## Data Types & Table Manipulations

**Materials:** [tinyurl.com/d8-disc03](https://tinyurl.com/d8-disc03)  
or access through [kevin-miao.com](https://kevin-miao.com)  
under teaching



# Today

---

- Check-In
- Announcements
- Review: Data Types & Table Manipulations
- Worksheet
  - The worksheet is purposely made longer so you have more practice that you can do on your own
  - Challenge Question (last question on worksheet) is optional

# Announcements

---

- **Vitamin 3** will be **due today**
- **Homework 2** is **due tomorrow**
  - For a bonus point, submit by tonight!
- Make sure you have the most recent version of **lab 3**
- Autograder issues
  - Grades for homework and lab will be delayed until further notice (likely 1-2 weeks)
  - This does not affect deadlines and normal workflow
- Tutoring Sections (~4-5 students) signups are filling up quickly.
  - If you couldn't sign up, check back weekly!

# Data Types

---

- The most important types:
  - **Numerical**
    - `Int` : integers (1, 2)
    - `Float` : decimals (1.0, 1.100..)
    - If you divide ints, you will get floats
  - **Text**
    - Strings ('Hello' or "Hello")
  - **Array**
    - A sequence or list of items/objects of the same type
      - e.g. array of ints, floats or strings
  - **Tables**
- Check the type of an item with `type (<name of object>)`

# Arrays

---

- **Ranges**

- `np.arange(start, end [, step])`
  - An array of increasing integers
  - `np.arange(0,3) → [0,1,2]`

- **Array Methods**

- `array1 = make_array('Gamestop', 'AMC', 'BlackBerry')`
- `array2 = make_array(1, 5, 10)`
- Finding the length of an array? `len(array1)`
- Finding the first item of both arrays? `array.item(0)`
- How do I sum all the items? `np.sum(array)`
- `array2 * 3 ⇒ {3, 15, 30}`
- How do I sum all the items? Many more `min`, `max`, `cumsum` ...

# Table Manipulations

---

- **Table Methods**

## Changing a table

- Adding in a column: `tbl.with_column(...)`
- Creating a new table with certain columns `tbl.select(column)`  
or `tbl.drop(column)`
- Change column names `tbl.relabeled(...)`

## Obtaining values

- Finding the size: `tbl.num_rows`, `tbl.num_columns`
- Obtaining a column as an array: `tbl.column(<name> or <number>)`
- Obtaining a row: `tbl.row(<number(s)>)`

# Table Manipulations

## Don't memorize. Your best friend is the Python Reference Sheet!

Data 8 Spring 2021

FAQ

Office Hours

Policies

Staff

Resources

Links

### Python Reference

Created by Nishant Kheterpal and Jessica Hu

Materials

Python Reference

Supplemental Videos

#### Table Functions and Methods

In the examples in the left column, `np` refers to the NumPy module, as usual. Everything else is a function, a method, an example of an argument to a function or method, or an example of an object we might call the method on. For example, `tbl` refers to a table, `array` refers to an array, and `num` refers to a number. `array.item(0)` is an example call for the method `item`, and in that example, `array` is the name previously given to some array.

Name	Chapter	Description	Input	Output
<code>Table()</code>	6	Create an empty table, usually to extend with data	None	An empty <b>Table</b>
<code>Table().read_table(filename)</code>	6	Create a table from a data file	<b>string</b> : the name of the file	<b>Table</b> with the contents of the data file
<code>tbl.with_columns(name, values)</code> <code>tbl.with_columns(n1, v1, n2,</code>	6	A table with an additional or replaced column or columns. <b>name</b> is a string for the name of a column, <b>values</b> is an array	1. <b>string</b> : the name of the new column; 2. <b>array</b> : the values in that column	<b>Table</b> : a copy of the original

**To the worksheet!** 🖋️

[tinyurl.com/d8-disc03](https://tinyurl.com/d8-disc03)

# 1. What Would Python Do?

---

## 1. What Would Python Do?

For each of the following examples, presume that the code was run in an empty cell. Write down what Python would output. If the code results in an Error, explain why an error would occur.

a) "I love " + "Data 8"

" I love Data 8"

b) "I love Data " + 8

ERROR!

[1, 2, 3]      [2, 6, 8]  
          ↑

c) np.arange(1, 4) + np.arange(2, 7, 2)

array([3, 6, 9])

d) make\_array(3, 4, 5) + np.arange(2, 7, 1)

ERROR!

[3]      ↔      [2, 3, 4, 5, 6]  
          Do not  
          match up!

## 2a. Causality, Coworkers and Coffee

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

### Vitamin 2

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

What does Python output after running the code from 2C of the discussion worksheet?

*The question is only one number. Do not enter any spaces, words or sentences. These answers will be marked wrong; regrades won't be allowed.*

```
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
```

## 2. 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.

```
odd_array = make_array(1, 3, 5, 7)
even_array = np.arange(2, 10, 2)
```

↳ [2, 4, 6, 8]

a) `odd_array + even_array`

[3, 7, 11, 15]

b) `odd_array.item(1)`

3

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

24

d) `odd_array*3`

[3, 9, 15, 21]

# 3. eBay Auctions

---

Your friend AI is curious to see whether or not it's cheaper to buy his favorite items on eBay rather than through some other platform! AI stumbles upon some auction data from eBay, and decides to use his newly developed Table skills to do some data-crunching. However, AI is making a few mistakes and needs your help. For the following questions, identify why the code won't work as is.

The table below is called `ebay` and contains more than just the 3 rows displayed.

*ebay*  
→

Auction_ID	Item	Opening_Bid	Closing_Price
1	Jacket	50	75
2	Smartwatch	100	150
3	Tablet	350	600

# 3a. eBay Auctions

Auction_ID	Item	Opening_Bid	Closing_Price
1	Jacket	50	75
2	Smartwatch	100	150
3	Tablet	350	600

a) # Use comments to describe the code you write  
`ebay.where('Opening_Bid', are.above(60))/ ebay.num_rows`

Can't divide this b/c it's a table!



`ebay.where('opening-bid', are.above(60))` num\_rows / `ebay.num_rows`  
↓ number  
NUMBER

## 3b. eBay Auctions

---

Auction_ID	Item	Opening_Bid	Closing_Price
1	Jacket	50	75
2	Smartwatch	100	150
3	Tablet	350	600

b) `ebay.column('Closing_Price') - ebay.select("Opening_Bid")`

array of numbers

table

instead → `ebay.column('closing_price')`

`- ebay.column("opening_bid")`

`ebay.select("opening_bid").column(...)`

# 3c. eBay Auctions

Al really wants a new jacket, but his budget is only \$150. To see whether eBay has had good deals on jackets historically, Al tries to filter the auction data such that it only contains jackets that were sold for a closing price less than \$150. He writes the following code to do so, but hasn't realized the mistake he's making. Help Al fix it so that your friend can hopefully get the jacket he deserves!

```
only_jackets = ebay.where('Item', are.equal_to('Jacket'))  
jackets_under_price = ebay.where('Closing_Price', are.below(150))
```

Ⓘ

only-jackets

Ⓣ

→ filtering out only jackets

→ filtering out only under \$150

Auction_ID	Item	Opening_Bid	Closing_Price
1	Jacket	50	75
2	Smartwatch	100	150
3	Tablet	350	600

# 4. Violence in California

---

You are working on a project related to arrest-related violence in California. After a bit of searching, you finally find some relevant data and import it into your Jupyter Notebook. The table arrests is shown below. (Dataset source: <https://www.kaggle.com/sohier/arrest-related-violence-in-california>)



	<b>Incident_Date_Str</b>	<b>City</b>	<b>Zip_Code</b>	<b>Num_Involved_Civilians</b>	<b>Num_Involved_Officers</b>
<b>0</b>	7/3/2016	Hayward	94544	1	2
<b>1</b>	11/20/2016	San Leandro	94578	1	1
<b>2</b>	5/30/2016	Dublin	94568	1	2
<b>3</b>	6/30/2016	Dublin	94568	1	2
<b>4</b>	5/25/2016	Hayward	94545	1	1

# 4a. Violence in California

---

Before you start your exploration, you want to understand your data better. For each of the columns **City**, **Num\_Involved\_Civilians** and **Zip\_Code**, identify if the data contained in that column is numerical or categorical.

Numerical

- Num  
Involved  
Civilians

Categorical

- City  
- Zip Code

	Incident_Date_Str	City	Zip_Code	Num_Involved_Civilians	Num_Involved_Officers
0	7/3/2016	Hayward	94544	1	2
1	11/20/2016	San Leandro	94578	1	1
2	5/30/2016	Dublin	94568	1	2
3	6/30/2016	Dublin	94568	1	2
4	5/25/2016	Hayward	94545	1	1

## 4b. Violence in California

---

Suppose you were only interested in arrests that involved two, three or four civilians. Assign `range_civilians` to a table that only contains rows from the `arrests` table that correspond to arrests involving two, three or four civilians.

`range_civilians` = `arrests`. where ("num-involved-civilians",  
are-between(2,5))

	Incident_Date_Str	City	Zip_Code	Num_Involved_Civilians	Num_Involved_Officers
0	7/3/2016	Hayward	94544	1	2
1	11/20/2016	San Leandro	94578	1	1
2	5/30/2016	Dublin	94568	1	2
3	6/30/2016	Dublin	94568	1	2
4	5/25/2016	Hayward	94545	1	1

# 4c. Violence in California

---

You're curious about finding out when the arrest involving the most officers occurred in Hayward. Assign `arrest_date` to a string that represents the date of the arrest that occurred in Hayward that involved the most officers. For this question, assume that the maximum number of officers involved in an arrest in Hayward is unique. It is okay for your code to wrap along the next line for `hay_arrests_sorted`.

`hay_arrests_sorted` = \_\_\_\_\_

`arrest_date` = \_\_\_\_\_

	<b>Incident_Date_Str</b>	<b>City</b>	<b>Zip_Code</b>	<b>Num_Involved_Civilians</b>	<b>Num_Involved_Officers</b>
<b>0</b>	7/3/2016	Hayward	94544	1	2
<b>1</b>	11/20/2016	San Leandro	94578	1	1
<b>2</b>	5/30/2016	Dublin	94568	1	2
<b>3</b>	6/30/2016	Dublin	94568	1	2
<b>4</b>	5/25/2016	Hayward	94545	1	1

# 5. End of an Error

For the following code snippet identify what errors are possible. The line of code pointed to in the output is the line that python will identify as having an error, and is not necessarily the only code that is in the cell. Why is the error possible or not possible? *As an example, in the following screenshot, python is identifying the third line as the error, and the error message is “NameError: ‘yanay’ is not defined”*

```
: king = 1
  meghan = 2
  king + meghan + yanay
  king
```

```
-----
NameError                                 Traceback
<ipython-input-23-7f79e91a3cc8> in <module>
     1 king = 1
     2 meghan = 2
----> 3 king + meghan + yanay
     4 king

NameError: name 'yanay' is not defined
```

Letter	Error Message
A	<code>AttributeError: 'NoneType' object has no attribute 'sort'</code>
B	<code>ValueError: The column "y" is not in the table.</code>
C	<code>ValueError: Column length mismatch. New column does not have the same number of rows as table.</code>
D	<code>TypeError: with_column() missing 1 required positional argument: 'values'</code>
E	<code>SyntaxError: invalid syntax</code>
F	<code>TypeError: unsupported operand type(s) for +: 'int' and 'str'</code>
G	<code>ValueError: operands could not be broadcast together with shapes (3,) (2,)</code>

# 5a. End of an Error

---

a)  $z = x + y$

Letter	Possible?	Why?
A		
B		
C		
D		
E		
F		
G		

Letter	Error Message
A	<code>AttributeError: 'NoneType' object has no attribute 'sort'</code>
B	<code>ValueError: The column "y" is not in the table.</code>
C	<code>ValueError: Column length mismatch. New column does not have the same number of rows as table.</code>
D	<code>TypeError: with_column() missing 1 required positional argument: 'values'</code>
E	<code>SyntaxError: invalid syntax</code>
F	<code>TypeError: unsupported operand type(s) for +: 'int' and 'str'</code>
G	<code>ValueError: operands could not be broadcast together with shapes (3,) (2,)</code>

# 5b. End of an Error

---

b) `x.sort(1, descending=True)`

Letter	Possible?	Why?
A		
B		
C		
D		
E		
F		
G		

Letter	Error Message
A	<code>AttributeError: 'NoneType' object has no attribute 'sort'</code>
B	<code>ValueError: The column "y" is not in the table.</code>
C	<code>ValueError: Column length mismatch. New column does not have the same number of rows as table.</code>
D	<code>TypeError: with_column() missing 1 required positional argument: 'values'</code>
E	<code>SyntaxError: invalid syntax</code>
F	<code>TypeError: unsupported operand type(s) for +: 'int' and 'str'</code>
G	<code>ValueError: operands could not be broadcast together with shapes (3,) (2,)</code>

**How did I do?**

<https://tinyurl.com/kevind8feedback>