Lab 2: Table operations

Welcome to Lab 2! This week, we'll learn how to import a module and practice table operations!

Recommended Reading:

- [Introduction to tables (https://www.inferentialthinking.com/chapters/03/4/Introduction_to_Tables)](https://www.inferentialthinking.com/chapters/03/4/Introduction_to_Tables)

First, set up the tests and imports by running the cell below.

In [2]:
```
# Just run this cell

import numpy as np
from datascience import *


# These lines load the tests.
# When you log-in please hit return (not shift + return) after typing in your e
mail
from client.api.notebook import Notebook
ok = Notebook('lab02.ok')
_ = ok.submit()
```

```
=====================================================================
Assignment: Tables
OK, version v1.14.19
=====================================================================



Saving notebook... No valid file sources found
Submit... 100% complete
Backup... 100% complete
Submission successful for user: austenzhu@berkeley.edu
URL: https://okpy.org/cal/data8/sp20/lab02/submissions/r2Gjk6
```

1. Review: The building blocks of Python code

The two building blocks of Python code are *expressions* and *statements*. An **expression** is a piece of code that

- is self-contained, meaning it would make sense to write it on a line by itself, and
- usually evaluates to a value.

Here are two expressions that both evaluate to 3:

```
3
5 - 2
```

One important type of expression is the **call expression**. A call expression begins with the name of a function and is followed by the argument(s) of that function in parentheses. The function returns some value, based on its arguments. Some important mathematical functions are listed below.

Function	Description
`abs`	Returns the absolute value of its argument
`max`	Returns the maximum of all its arguments
`min`	Returns the minimum of all its arguments
`pow`	Raises its first argument to the power of its second argument
`round`	Rounds its argument to the nearest integer

Here are two call expressions that both evaluate to 3:

```
abs(2 - 5)
max(round(2.8), min(pow(2, 10), -1 * pow(2, 10)))
```

The expression `5 - 2` and the two call expressions given above are examples of **compound expressions**, meaning that they are actually combinations of several smaller expressions. `5 - 2` combines the expressions `5` and `2` by subtraction. In this case, `5` and `2` are called **subexpressions** because they're expressions that are part of a larger expression.

A **statement** is a whole line of code. Some statements are just expressions. The expressions listed above are examples.

Other statements *make something happen* rather than *having a value*. For example, an **assignment statement** assigns a value to a name.

A good way to think about this is that we're **evaluating the right-hand side** of the equals sign and **assigning it to the left-hand side**. Here are some assignment statements:

```
height = 1.3
the_number_five = abs(-5)
absolute_height_difference = abs(height - 1.688)
```

An important idea in programming is that large, interesting things can be built by combining many simple, uninteresting things. The key to understanding a complicated piece of code is breaking it down into its simple components.

For example, a lot is going on in the last statement above, but it's really just a combination of a few things. This picture describes what's going on.

Question 1.1. In the next cell, assign the name `new_year` to the larger number among the following two numbers:

1. the **absolute value** of $2^5 - 2^{11} - 2^1 + 1$, and
2. $5 \times 13 \times 31 + 5$.

Try to use just one statement (one line of code). Be sure to check your work by executing the test cell afterward.

```
BEGIN QUESTION
name: q11
```

In [5]:
```
new_year = max(abs(2**5 - 2**11 - 2 ** 1 + 1), 5*13*31 + 5) #SOLUTION
new_year
```

Out[5]: 2020

In [7]:
```
# TEST
new_year
```

Out[7]: 2020

We've asked you to use one line of code in the question above because it only involves mathematical operations. However, more complicated programming questions will more require more steps. It isn't always a good idea to jam these steps into a single line because it can make the code harder to read and harder to debug.

Good programming practice involves splitting up your code into smaller steps and using appropriate names. You'll have plenty of practice in the rest of this course!

2. Importing code

[source (https://www.reddit.com/r/ProgrammerHumor/comments/cgtk7s/theres_no_need_to_reinvent_the_wheel_oc/)](https://www.reddit.com/r/ProgrammerHumor/comments/cgtk7s/theres_no_need_to_reinvent_the_wheel_oc/)

Most programming involves work that is very similar to work that has been done before. Since writing code is time-consuming, it's good to rely on others' published code when you can. Rather than copy-pasting, Python allows us to **import modules**. A module is a file with Python code that has defined variables and functions. By importing a module, we are able to use its code in our own notebook.

Python includes many useful modules that are just an `import` away. We'll look at the `math` module as a first example. The `math` module is extremely useful in computing mathematical expressions in Python.

Suppose we want to very accurately compute the area of a circle with a radius of 5 meters. For that, we need the constant π, which is roughly 3.14. Conveniently, the `math` module has `pi` defined for us:

In [8]:
```
import math
radius = 5
area_of_circle = radius**2 * math.pi
area_of_circle
```
Out[8]: 78.53981633974483

In the code above, the line `import math` imports the math module. This statement creates a module and then assigns the name `math` to that module. We are now able to access any variables or functions defined within `math` by typing the name of the module followed by a dot, then followed by the name of the variable or function we want.

```
<module name>.<name>
```

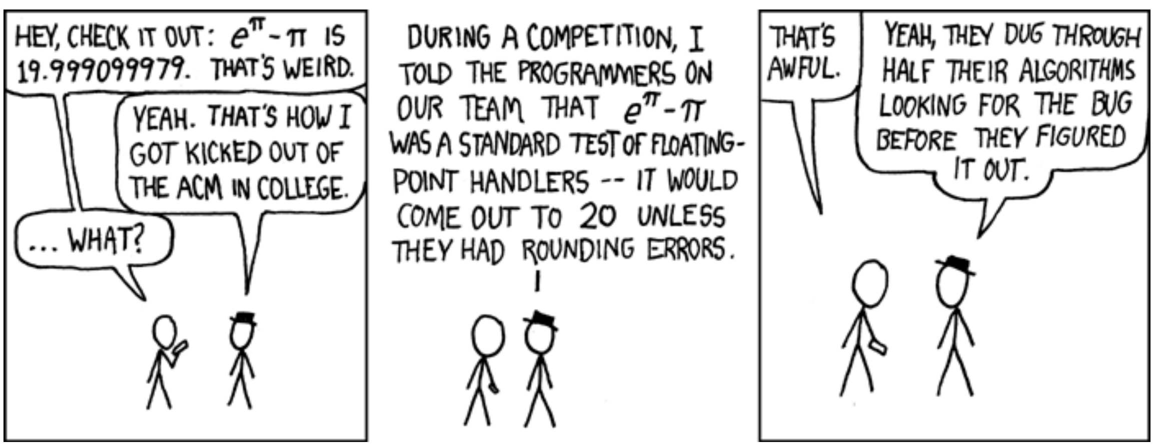
Question 2.1. The module `math` also provides the name `e` for the base of the natural logarithm, which is roughly 2.71. Compute $e^{\pi} - \pi$, giving it the name `near_twenty`.

Remember: You can access `pi` from the `math` module as well!

```
BEGIN QUESTION
name: q21
```

In [9]:
```
near_twenty = math.e ** math.pi - math.pi # SOLUTION
near_twenty
```
Out[9]: 19.99909997918947

In [10]:
```
# TEST
round(near_twenty, 8)
```
Out[10]: 19.99909998



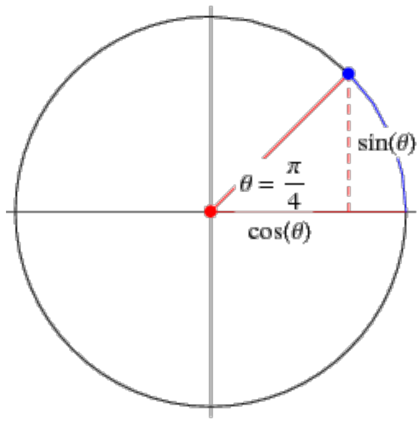
[Source (http://imgs.xkcd.com/comics/e_to_the_pi_minus_pi.png)](http://imgs.xkcd.com/comics/e_to_the_pi_minus_pi.png) [Explaination (https://www.explainxkcd.com/wiki/index.php/217:_e_to_the_pi_Minus_pi)](https://www.explainxkcd.com/wiki/index.php/217:_e_to_the_pi_Minus_pi)

2.1. Accessing functions

In the question above, you accessed variables within the `math` module.

Modules also define **functions**. For example, `math` provides the name `sin` for the sine function. Having imported `math` already, we can write `math.sin(3)` to compute the sine of 3. (Note that this sine function considers its argument to be in [radians (https://en.wikipedia.org/wiki/Radian)](https://en.wikipedia.org/wiki/Radian), not degrees. 180 degrees are equivalent to π radians.)

Question 2.1.1. A $\frac{\pi}{4}$-radian (45-degree) angle forms a right triangle with equal base and height, pictured below. If the hypotenuse (the radius of the circle in the picture) is 1, then the height is $\sin(\frac{\pi}{4})$. Compute that value using `sin` and `pi` from the `math` module. Give the result the name `sine_of_pi_over_four`.



[Source (http://mathworld.wolfram.com/images/eps-gif/TrigonometryAnglesPi4_1000.gif)](http://mathworld.wolfram.com/images/eps-gif/TrigonometryAnglesPi4_1000.gif)

```
BEGIN QUESTION
name: q211
```

In [11]:
```
sine_of_pi_over_four = math.sin(math.pi / 4) #SOLUTION
sine_of_pi_over_four
```

Out[11]: 0.7071067811865475

In [12]:
```
# TEST
# Hint: You can write the sine of 1.5*pi as:
#   math.sin(1.5 * math.pi)
import math
round(sine_of_pi_over_four, 8)
```

Out[12]: 0.70710678

For your reference, below are some more examples of functions from the `math` module.

Notice how different functions take in different numbers of arguments. Often, the [documentation (https://docs.python.org/3/library/math.html)](https://docs.python.org/3/library/math.html) of the module will provide information on how many arguments are required for each function.

Hint: If you press `shift+tab` while next to the function call, the documentation for that function will appear

In [13]:
```
# Calculating logarithms (the logarithm of 8 in base 2).
# The result is 3 because 2 to the power of 3 is 8.
math.log(8, 2)
```

Out[13]: 3.0

In [14]:
```
# Calculating square roots.
math.sqrt(5)
```

Out[14]: 2.23606797749979

There are various ways to import and access code from outside sources. The method we used above — `import <module_name>` — imports the entire module and requires that we use `<module_name>.<name>` to access its code.

We can also import a specific constant or function instead of the entire module. Notice that you don't have to use the module name beforehand to reference that particular value. However, you do have to be careful about reassigning the names of the constants or functions to other values!

In [15]:
```
# Importing just cos and pi from math.
# We don't have to use `math.` in front of cos or pi
from math import cos, pi
print(cos(pi))

# We do have to use it in front of other functions from math, though
math.log(pi)
```

```
-1.0
```

Out[15]: 1.1447298858494002

Or we can import every function and value from the entire module.

In [16]:
```
# Lastly, we can import everything from math using the *
# Once again, we don't have to use 'math.' beforehand
from math import *
log(pi)
```

Out[16]: 1.1447298858494002

Don't worry too much about which type of import to use. It's often a coding style choice left up to each programmer. In this course, you'll always import the necessary modules when you run the setup cell (like the first code cell in this lab).

Let's move on to practicing some of the table operations you've learned in lecture!

3. Table operations

The table `farmers_markets.csv` contains data on farmers' markets in the United States (data collected [by the USDA (https://apps.ams.usda.gov/FarmersMarketsExport/ExcelExport.aspx)](https://apps.ams.usda.gov/FarmersMarketsExport/ExcelExport.aspx)). Each row represents one such market.

Run the next cell to load the `farmers_markets` table.

In [17]:
```
# Just run this cell

farmers_markets = Table.read_table('farmers_markets.csv')
```

Let's examine our table to see what data it contains.

Question 3.1. Use the method `show` to display the first 5 rows of `farmers_markets`.

Note: The terms "method" and "function" are technically not the same thing, but for the purposes of this course, we will use them interchangeably.

Hint: `tbl.show(3)` will show the first 3 rows of `tbl`. Additionally, make sure not to call `.show()` without an argument, as this will crash your kernel!

In [18]:
```
farmers_markets.show(5) # SOLUTION
```

FMID	MarketName	street	city	County	State	zip	x	y	
1012063	Caledonia Farmers Market Association - Danville	nan	Danville	Caledonia	Vermont	05828	-72.1403	44.411	https: /site/caledo
1011871	Stearns Homestead Farmers' Market	6975 Ridge Road	Parma	Cuyahoga	Ohio	44130	-81.7286	41.3751	http://Stear
1011878	100 Mile Market	507 Harrison St	Kalamazoo	Kalamazoo	Michigan	49007	-85.5749	42.296	http://ww
1009364	106 S. Main Street Farmers Market	106 S. Main Street	Six Mile	nan	South Carolina	29682	-82.8187	34.8042	http://thetownofsixmil
1010691	10th Steet Community Farmers Market	10th Street and Poplar	Lamar	Barton	Missouri	64759	-94.2746	37.4956	

... (8541 rows omitted)

Notice that some of the values in this table are missing, as denoted by "nan." This means either that the value is not available (e.g. if we don't know the market's street address) or not applicable (e.g. if the market doesn't have a street address). You'll also notice that the table has a large number of columns in it!

`num_columns`

The table property `num_columns` returns the number of columns in a table. (A "property" is just a method that doesn't need to be called by adding parentheses.)

Example call: `<tbl>.num_columns`

Question 3.2. Use `num_columns` to find the number of columns in our farmers' markets dataset.

Assign the number of columns to `num_farmers_markets_columns` .

```
BEGIN QUESTION
name: q32
```

In [19]:
```
num_farmers_markets_columns = farmers_markets.num_columns #SOLUTION
print("The table has", num_farmers_markets_columns, "columns in it!")
```

```
The table has 59 columns in it!
```

In [20]:
```
# TEST
num_farmers_markets_columns == 59
```

Out[20]:
```
True
```

`num_rows`

Similarly, the property `num_rows` tells you how many rows are in a table.

In [21]:
```
# Just run this cell

num_farmers_markets_rows = farmers_markets.num_rows
print("The table has", num_farmers_markets_rows, "rows in it!")
```

```
The table has 8546 rows in it!
```

`select`

Most of the columns are about particular products -- whether the market sells tofu, pet food, etc. If we're not interested in that information, it just makes the table difficult to read. This comes up more than you might think, because people who collect and publish data may not know ahead of time what people will want to do with it.

In such situations, we can use the table method `select` to choose only the columns that we want in a particular table. It takes any number of arguments. Each should be the name of a column in the table. It returns a new table with only those columns in it. The columns are in the order *in which they were listed as arguments*.

For example, the value of `farmers_markets.select("MarketName", "State")` is a table with only the name and the state of each farmers' market in `farmers_markets`.

Question 3.3. Use `select` to create a table with only the name, city, state, latitude (`y`), and longitude (`x`) of each market. Call that new table `farmers_markets_locations`.

Hint: Make sure to be exact when using column names with `select`; double-check capitalization!

```
BEGIN QUESTION
name: q33
```

In [22]:
```
farmers_markets_locations = farmers_markets.select("city", "State", "y", "Marke
tName","x") #SOLUTION
farmers_markets_locations
```

Out[22]:

city	State	y	MarketName	x
Danville	Vermont	44.411	Caledonia Farmers Market Association - Danville	-72.1403
Parma	Ohio	41.3751	Stearns Homestead Farmers' Market	-81.7286
Kalamazoo	Michigan	42.296	100 Mile Market	-85.5749
Six Mile	South Carolina	34.8042	106 S. Main Street Farmers Market	-82.8187
Lamar	Missouri	37.4956	10th Steet Community Farmers Market	-94.2746
New York	New York	40.7939	112st Madison Avenue	-73.9493
Nashville	Tennessee	36.1184	12 South Farmers Market	-86.7907
New York	New York	40.809	125th Street Fresh Connect Farmers' Market	-73.9482
Wilmington	Delaware	39.7421	12th & Brandywine Urban Farm Market	-75.5345
Washington	District of Columbia	38.917	14&U Farmers' Market	-77.0321

... (8536 rows omitted)

In [23]:
```
# TEST
sorted(farmers_markets_locations.labels)
```

Out[23]: `['MarketName', 'State', 'city', 'x', 'y']`

In [24]:
```
# TEST
farmers_markets_locations.num_rows == 8546
```

Out[24]: `True`

`drop`

`drop` serves the same purpose as `select`, but it takes away the columns that you provide rather than the ones that you don't provide. Like `select`, `drop` returns a new table.

Question 3.4. Suppose you just didn't want the `FMID` or `updateTime` columns in `farmers_markets`. Create a table that's a copy of `farmers_markets` but doesn't include those columns. Call that table `farmers_markets_without_fmid`.

```
BEGIN QUESTION
name: q34
```

In [25]:
```
farmers_markets_without_fmid = farmers_markets.drop("FMID", "updateTime") #SOLUTION
farmers_markets_without_fmid
```

Out[25]:

MarketName	street	city	County	State	zip	x	y	
Caledonia Farmers Market Association - Danville	nan	Danville	Caledonia	Vermont	05828	-72.1403	44.411	https:// /site/caledoni
Stearns Homestead Farmers' Market	6975 Ridge Road	Parma	Cuyahoga	Ohio	44130	-81.7286	41.3751	http://Stearns
100 Mile Market	507 Harrison St	Kalamazoo	Kalamazoo	Michigan	49007	-85.5749	42.296	http://www
106 S. Main Street Farmers Market	106 S. Main Street	Six Mile	nan	South Carolina	29682	-82.8187	34.8042	http://thetownofsixmile
10th Steet Community Farmers Market	10th Street and Poplar	Lamar	Barton	Missouri	64759	-94.2746	37.4956	
112st Madison Avenue	112th Madison Avenue	New York	New York	New York	10029	-73.9493	40.7939	
12 South Farmers Market	3000 Granny White Pike	Nashville	Davidson	Tennessee	37204	-86.7907	36.1184	http://www.12southfar
125th Street Fresh Connect Farmers' Market	163 West 125th Street and Adam Clayton Powell, Jr. Blvd.	New York	New York	New York	10027	-73.9482	40.809	http://www.125thStreetFar
12th & Brandywine Urban Farm Market	12th & Brandywine Streets	Wilmington	New Castle	Delaware	19801	-75.5345	39.7421	
14&U Farmers' Market	1400 U Street NW	Washington	District of Columbia	District of Columbia	20009	-77.0321	38.917	

... (8536 rows omitted)

In [26]:
```
# TEST
farmers_markets_without_fmid.num_columns
```

Out[26]: 57

In [27]:
```
# TEST
print(sorted(farmers_markets_without_fmid.labels))
```

```
['Bakedgoods', 'Beans', 'Cheese', 'Coffee', 'County', 'Crafts', 'Credit', 'Eggs
', 'Facebook', 'Flowers', 'Fruits', 'Grains', 'Herbs', 'Honey', 'Jams', 'Juices
', 'Location', 'Maple', 'MarketName', 'Meat', 'Mushrooms', 'Nursery', 'Nuts', '
Organic', 'OtherMedia', 'PetFood', 'Plants', 'Poultry', 'Prepared', 'SFMNP', 'S
NAP', 'Seafood', 'Season1Date', 'Season1Time', 'Season2Date', 'Season2Time', 'S
eason3Date', 'Season3Time', 'Season4Date', 'Season4Time', 'Soap', 'State', 'Tof
u', 'Trees', 'Twitter', 'Vegetables', 'WIC', 'WICcash', 'Website', 'WildHarvest
ed', 'Wine', 'Youtube', 'city', 'street', 'x', 'y', 'zip']
```

Now, suppose we want to answer some questions about farmers' markets in the US. For example, which market(s) have the largest longitude (given by the `x` column)?

To answer this, we'll sort `farmers_markets_locations` by longitude.

In [28]:
```
farmers_markets_locations.sort('x')
```

Out[28]:

city	State	y	MarketName	x
Trapper Creek	Alaska	53.8748	Trapper Creek Farmers Market	-166.54
Kekaha	Hawaii	21.9704	Kekaha Neighborhood Center (Sunshine Markets)	-159.718
Hanapepe	Hawaii	21.9101	Hanapepe Park (Sunshine Markets)	-159.588
Kalaheo	Hawaii	21.9251	Kalaheo Neighborhood Center (Sunshine Markets)	-159.527
Hanalei	Hawaii	22.2033	Hawaiian Farmers of Hanalei	-159.514
Hanalei	Hawaii	22.2042	Hanalei Saturday Farmers Market	-159.492
Koloa	Hawaii	21.9067	Kauai Culinary Market	-159.469
Koloa	Hawaii	21.9081	Koloa Ball Park (Knudsen) (Sunshine Markets)	-159.465
Poipu	Hawaii	21.8815	West Kauai Agricultural Association	-159.435
Kilauea	Hawaii	22.2112	Kilauea Neighborhood Center (Sunshine Markets)	-159.406

... (8536 rows omitted)

Oops, that didn't answer our question because we sorted from smallest to largest longitude. To look at the largest longitudes, we'll have to sort in reverse order.

In [29]:
```
farmers_markets_locations.sort('x', descending=True)
```

Out[29]:

city	State	y	MarketName	x
Saint Croix	Virgin Islands	17.7449	Christian "Shan" Hendricks Vegetable Market	-64.7043
Saint Croix	Virgin Islands	17.7322	La Reine Farmers Market	-64.7789
Saint Croix	Virgin Islands	17.7099	Anne Heyliger Vegetable Market	-64.8799
St. Thomas	Virgin Islands	18.3428	Rothschild Francis Vegetable Market	-64.9326
Luquillo	Puerto Rico	18.3782	Feria Agrícola de Luquillo	-65.7207
San Lorenzo	Puerto Rico	18.1871	El Mercado Familiar	-65.9674
Gurabo	Puerto Rico	18.2526	El Mercado Familiar	-65.9786
Patillas	Puerto Rico	18.0069	El Mercado Familiar	-66.0135
Caguas zona urbana	Puerto Rico	18.2324	El Mercado Familiar	-66.039
Arroyo zona urbana	Puerto Rico	17.9686	El Maercado Familiar	-66.0617

... (8536 rows omitted)

(The `descending=True` bit is called an *optional argument*. It has a default value of `False` , so when you explicitly tell the function `descending=True` , then the function will sort in descending order.)

`sort`

Some details about sort:

1. The first argument to `sort` is the name of a column to sort by.
2. If the column has text in it, `sort` will sort alphabetically; if the column has numbers, it will sort numerically.
3. The value of `farmers_markets_locations.sort("x")` is a *copy* of `farmers_markets_locations` ; the `farmers_markets_locations` table doesn't get modified. For example, if we called `farmers_markets_locations.sort("x")` , then running `farmers_markets_locations` by itself would still return the unsorted table.
4. Rows always stick together when a table is sorted. It wouldn't make sense to sort just one column and leave the other columns alone. For example, in this case, if we sorted just the `x` column, the farmers' markets would all end up with the wrong longitudes.

Question 3.5. Create a version of `farmers_markets_locations` that's sorted by **latitude (`y`)**, with the largest latitudes first. Call it `farmers_markets_locations_by_latitude` .

```
BEGIN QUESTION
name: q35
```

In [30]:
```
farmers_markets_locations_by_latitude = farmers_markets_locations.sort('y', des
cending=True) #SOLUTION
farmers_markets_locations_by_latitude
```

Out[30]:

city	State	y	MarketName	x
Fairbanks	Alaska	64.8628	Tanana Valley Farmers Market	-147.781
Ester	Alaska	64.8459	Ester Community Market	-148.01
Fairbanks	Alaska	64.8444	Fairbanks Downtown Market	-147.72
Nenana	Alaska	64.5566	Nenana Open Air Market	-149.096
Delta Junction	Alaska	64.0385	Highway's End Farmers' Market	-145.733
Talkeetna	Alaska	62.3231	MountainTraders	-150.118
Talkeetna	Alaska	62.3228	Talkeetna Farmers Market	-150.118
Anchorage	Alaska	62.3163	Denali Farmers Market	-150.234
Valdez	Alaska	62.1079	Kenny Lake Harvest II	-145.476
Copper Valley	Alaska	62.0879	Copper Valley Community Market	-145.444

... (8536 rows omitted)

In [31]:
```
# TEST
type(farmers_markets_locations_by_latitude) == tables.Table
```

Out[31]: True

In [32]:
```
# TEST
list(farmers_markets_locations_by_latitude.column('y').take(range(3)))
```

Out[32]: [64.86275, 64.8459, 64.844414]

Now let's say we want a table of all farmers' markets in California. Sorting won't help us much here because California is closer to the middle of the dataset.

Instead, we use the table method `where` .

In [33]:
```
california_farmers_markets = farmers_markets_locations.where('State', are.equal_to('California'))
california_farmers_markets
```

Out[33]:

city	State	y	MarketName	x
Victorville	California	34.5593	Adelanto Stadium Farmers Market	-117.405
Alameda	California	37.7742	Alameda Farmers' Market	-122.277
Salinas	California	36.6733	Alisal Certified Farmers' Market	-121.634
Altadena	California	34.2004	Altadena Farmers' Market	-118.158
San Jose	California	37.3678	Alum Rock Village Farmers' Market	-121.833
Jackson	California	38.3488	Amador Farmers' Market-- Jackson	-120.774
Pine Grove	California	38.3488	Amador Farmers' Market-- Pine Grove	-120.774
Sutter Creek	California	38.3488	Amador Farmers' Market-- Sutter Creek	-120.774
Anderson	California	40.4487	Anderson Happy Valley Farmers Market	-122.408
Angels Camp	California	38.0722	Angels Camp Farmers Market-Fresh Fridays	-120.543

... (745 rows omitted)

Ignore the syntax for the moment. Instead, try to read that line like this:

> Assign the name **`california_farmers_markets`** to a table whose rows are the rows in the **`farmers_markets_locations`** table **`where`** the **`'State'`** s **`are equal to California`** .

`where`

Now let's dive into the details a bit more. `where` takes 2 arguments:

1. The name of a column. `where` finds rows where that column's values meet some criterion.
2. A predicate that describes the criterion that the column needs to meet.

The predicate in the example above called the function `are.equal_to` with the value we wanted, 'California'. We'll see other predicates soon.

`where` returns a table that's a copy of the original table, but **with only the rows that meet the given predicate**.

Question 3.6. Use `california_farmers_markets` to create a table called `berkeley_markets` containing farmers' markets in Berkeley, California.

```
BEGIN QUESTION
name: q36
```

In [34]:
```
berkeley_markets = california_farmers_markets.where('city', are.equal_to("Berkeley")) #SOLUTION
berkeley_markets
```

Out[34]:

city	State	y	MarketName	x
Berkeley	California	37.8697	Downtown Berkeley Farmers' Market	-122.273
Berkeley	California	37.8802	North Berkeley Farmers' Market	-122.269
Berkeley	California	37.8478	South Berkeley Farmers' Market	-122.272

In [35]:
```
# TEST
berkeley_markets.num_rows == 3
```

Out[35]: True

In [36]:
```
# TEST
list(berkeley_markets.column('city')) == ['Berkeley', 'Berkeley', 'Berkeley']
```

Out[36]: True

Recognize any of them?

So far we've only been using `where` with the predicate that requires finding the values in a column to be *exactly* equal to a certain value. However, there are many other predicates. Here are a few:

Predicate	Example	Result
`are.equal_to`	`are.equal_to(50)`	Find rows with values equal to 50
`are.not_equal_to`	`are.not_equal_to(50)`	Find rows with values not equal to 50
`are.above`	`are.above(50)`	Find rows with values above (and not equal to) 50
`are.above_or_equal_to`	`are.above_or_equal_to(50)`	Find rows with values above 50 or equal to 50
`are.below`	`are.below(50)`	Find rows with values below 50
`are.between`	`are.between(2, 10)`	Find rows with values above or equal to 2 and below 10

4. Analyzing a dataset

Now that you're familiar with table operations, let’s answer an interesting question about a dataset!

Run the cell below to load the `imdb` table. It contains information about the 250 highest-rated movies on IMDb.

In [37]:
```
# Just run this cell

imdb = Table.read_table('imdb.csv')
imdb
```

Out[37]:

Votes	Rating	Title	Year	Decade
88355	8.4	M	1931	1930
132823	8.3	Singin' in the Rain	1952	1950
74178	8.3	All About Eve	1950	1950
635139	8.6	Léon	1994	1990
145514	8.2	The Elephant Man	1980	1980
425461	8.3	Full Metal Jacket	1987	1980
441174	8.1	Gone Girl	2014	2010
850601	8.3	Batman Begins	2005	2000
37664	8.2	Judgment at Nuremberg	1961	1960
46987	8	Relatos salvajes	2014	2010

... (240 rows omitted)

Often, we want to perform multiple operations - sorting, filtering, or others - in order to turn a table we have into something more useful. You can do these operations one by one, e.g.

```
first_step = original_tbl.where("col1", are.equal_to(12))
second_step = first_step.sort('col2', descending=True)
```

However, since the value of the expression `original_tbl.where("col1", are.equal_to(12))` is itself a table, you can just call a table method on it:

```
original_tbl.where("col1", are.equal_to(12)).sort('col2', descending=True)
```

You should organize your work in the way that makes the most sense to you, using informative names for any intermediate tables you create.

Question 4.1. Create a table of movies released between 2010 and 2016 (inclusive) with ratings above 8. The table should only contain the columns `Title` and `Rating` , **in that order**.

Assign the table to the name `above_eight` .

Hint: Think about the steps you need to take, and try to put them in an order that make sense. Feel free to create intermediate tables for each step, but please make sure you assign your final table the name `above_eight` !

```
BEGIN QUESTION
name: q41
```

In [38]:
```
above_eight = imdb.where('Year', are.between(2010, 2017)).where('Rating', are.a
bove(8)).select('Title', 'Rating') #SOLUTION
above_eight
```

Out[38]:

Title	Rating
Gone Girl	8.1
Warrior	8.2
Intouchables	8.5
Shutter Island	8.1
12 Years a Slave	8.1
Inside Out (2015/I)	8.5
Jagten	8.2
Toy Story 3	8.3
How to Train Your Dragon	8.1
Interstellar	8.6

... (10 rows omitted)

In [39]:
```
# TEST
type(above_eight) == tables.Table
```

Out[39]:
```
True
```

In [40]:
```
# TEST
above_eight.num_rows == 20
```

Out[40]:
```
True
```

In [41]:
```
# TEST
# Make sure your columns are in the right order!
# First column should be 'Title', second column should be 'Rating'
print(above_eight.sort(0).take([17]))
```

```
Title       | Rating
Toy Story 3 | 8.3
```

Question 4.2. Use `num_rows` (and arithmetic) to find the *proportion* of movies in the dataset that were released 1900-1999, and the *proportion* of movies in the dataset that were released in the year 2000 or later.

Assign `proportion_in_20th_century` to the proportion of movies in the dataset that were released 1900-1999, and `proportion_in_21st_century` to the proportion of movies in the dataset that were released in the year 2000 or later.

Hint: The *proportion* of movies released in the 1900's is the *number* of movies released in the 1900's, divided by the *total number* of movies.

```
BEGIN QUESTION
name: q42
```

In [42]:
```
num_movies_in_dataset = imdb.num_rows #SOLUTION
num_in_20th_century = imdb.where('Year', are.below(2000)).num_rows #SOLUTION
num_in_21st_century = imdb.where('Year', are.above_or_equal_to(2000)).num_rows
#SOLUTION
proportion_in_20th_century = num_in_20th_century / num_movies_in_dataset #SOLUT
ION
proportion_in_21st_century =  num_in_21st_century / num_movies_in_dataset #SOLU
TION
print("Proportion in 20th century:", proportion_in_20th_century)
print("Proportion in 21st century:", proportion_in_21st_century)
```

```
Proportion in 20th century: 0.684
Proportion in 21st century: 0.316
```

In [43]:
```
# TEST
proportion_in_20th_century == 0.684
```

Out[43]: True

In [44]:
```
# TEST
proportion_in_21st_century == 0.316
```

Out[44]: True

5. Summary

For your reference, here's a table of all the functions and methods we saw in this lab. We'll learn more methods to add to this table in the coming week!

Name	Example	Purpose
`sort`	`tbl.sort("N")`	Create a copy of a table sorted by the values in a column
`where`	`tbl.where("N", are.above(2))`	Create a copy of a table with only the rows that match some *predicate*
`num_rows`	`tbl.num_rows`	Compute the number of rows in a table
`num_columns`	`tbl.num_columns`	Compute the number of columns in a table
`select`	`tbl.select("N")`	Create a copy of a table with only some of the columns
`drop`	`tbl.drop("2*N")`	Create a copy of a table without some of the columns

Alright! You're finished with lab 2! Be sure to...

- run all the tests (the next cell has a shortcut for that),
- **Save and Checkpoint** from the `File` menu,
- **run the last cell to submit your work**,
- and **ask one of the staff members to check you off**.

In [45]:
```
# For your convenience, you can run this cell to run all the tests at once!
import os
_ = [ok.grade(q[:-3]) for q in os.listdir("tests") if q.startswith('q')]
```

In [46]:
```
_ = ok.submit()
```

```
Saving notebook... No valid file sources found
Submit... 100% complete
Submission successful for user: austenzhu@berkeley.edu
URL: https://okpy.org/cal/data8/sp20/lab02/submissions/jYwXYz
```