UFRN UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

IMD INSTITUTO METRÓPOLE DIGITAL

IMD0905 - Data Science I

Lesson #8 - Introduction to Numpy

Ivanovitch Silva
August, 2018

Agenda

- Reading CSV files with NumPy
- Boolean Arrays
- Boolean Index
- Assigning values
- Challenge

nyc_taxis.csv

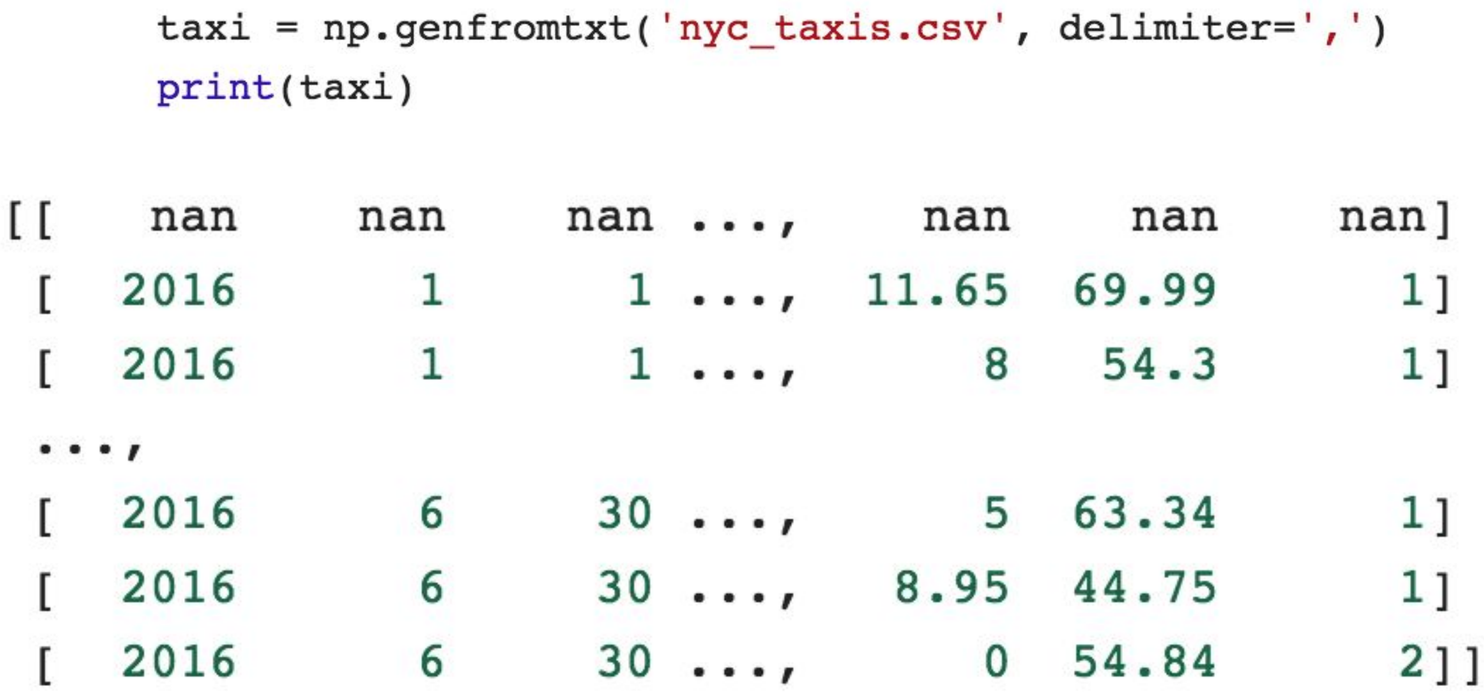
Update the repository

```
git clone https://github.com/ivanovitchm/IMD0905_datascience_one.git
```

Or ….

```
git pull
```

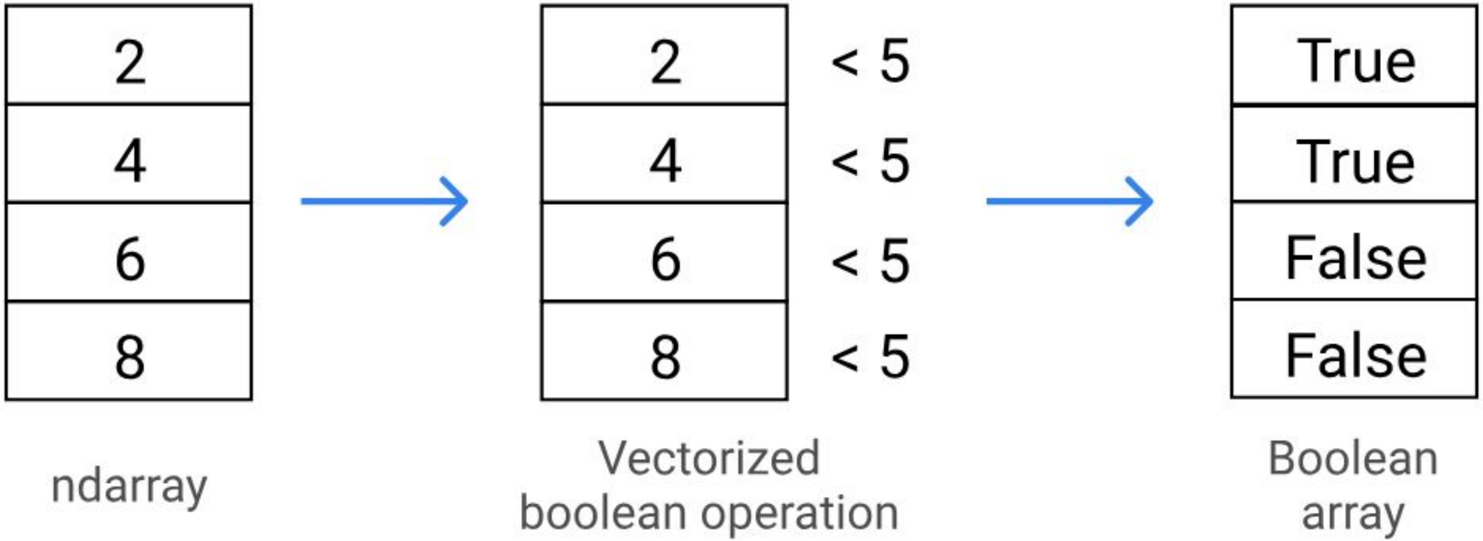
Reading CSV files from Numpy



```python
taxi = np.genfromtxt('nyc_taxis.csv', delimiter=',')
print(taxi)
```

```
[[    nan      nan      nan ...,     nan     nan     nan]
 [  2016        1        1 ...,  11.65   69.99       1]
 [  2016        1        1 ...,      8    54.3       1]
 ...,
 [  2016        6       30 ...,      5   63.34       1]
 [  2016        6       30 ...,   8.95   44.75       1]
 [  2016        6       30 ...,      0   54.84       2]]
```

Slicing from boolean arrays



2

4

6

8

ndarray

2 < 5

4 < 5

6 < 5

8 < 5

Vectorized boolean operation

True

True

False

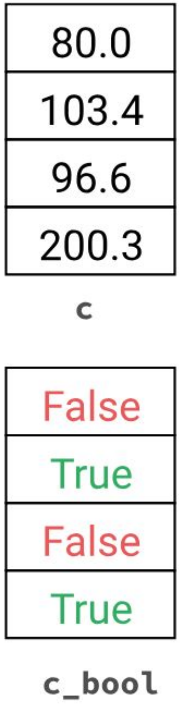
False

Boolean array

Boolean indexing with 1D ndarrays

```
c = np.array([80.0, 103.4,
              96.9, 200.3])
```

```
c_bool = c > 100
```

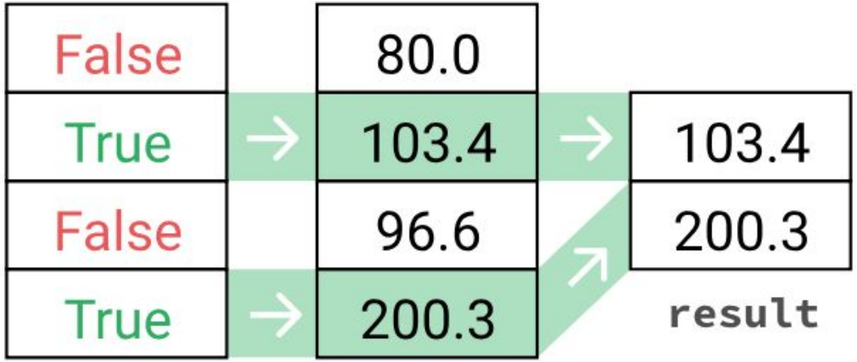


c
80.0
103.4
96.6
200.3

c_bool
False
True
False
True

Boolean indexing with 1D ndarrays

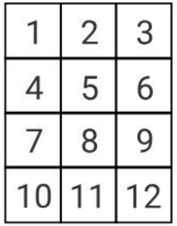
```
result = c[c_bool]
```



Boolean Indexing with 2D ndarrays

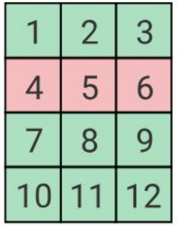
Code | **Visualization** | **Explanation**

```
arr = np.array([
                [ 1,  2,  3],
                [ 4,  5,  6],
                [ 7,  8,  9],
                [10, 11, 12]
               ])
print(arr)
```



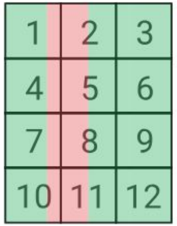
The original array

```
bool_1 = [True, False,
          True, True]
print(arr[bool_1])
```



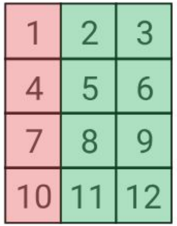
bool_1's shape (4) is the same as the shape of arr's first axis (4), so this selects the 1st, 3rd, and 4th rows.

```
print(arr[:,bool_1])
```



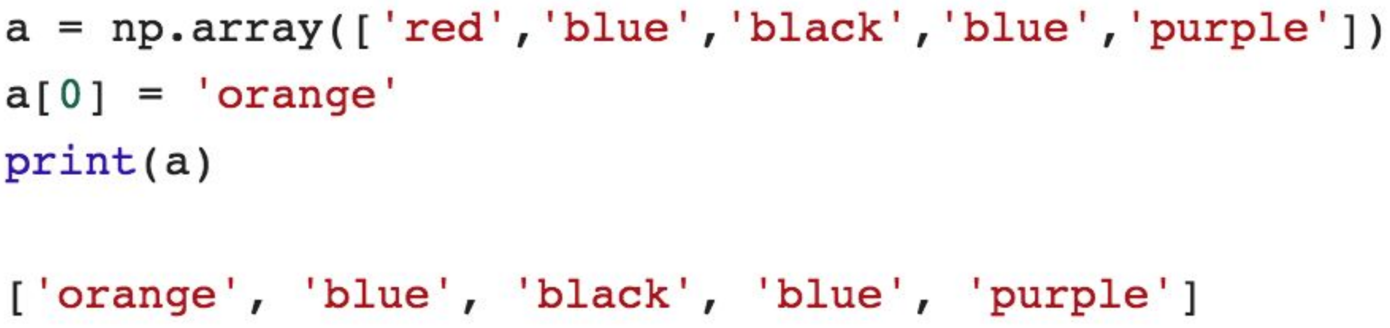
bool_1's shape (4) is not the same as the shape of arr's second axis (3), so it can't be used to index and produces an error

```
bool_2 = [False, True, True]
print(arr[:,bool_2])
```

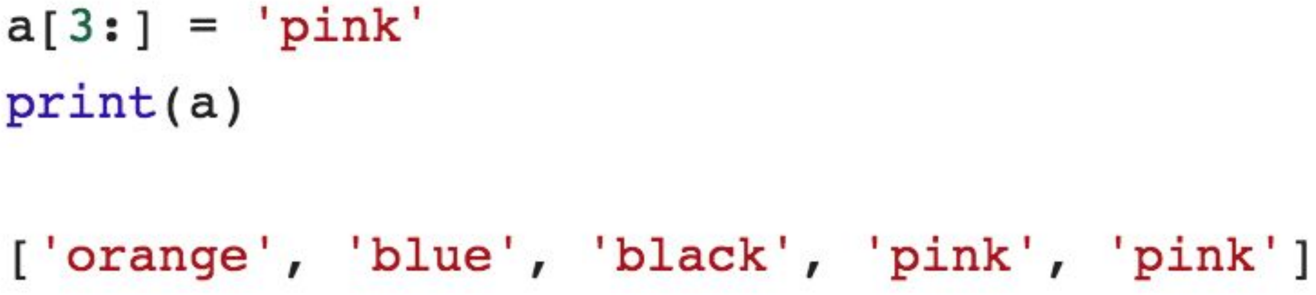


bool_2's shape (3) is the same as the shape of arr's second axis (3), so this selects the 2nd and 3rd columns.

Assigning values in 1D ndarray

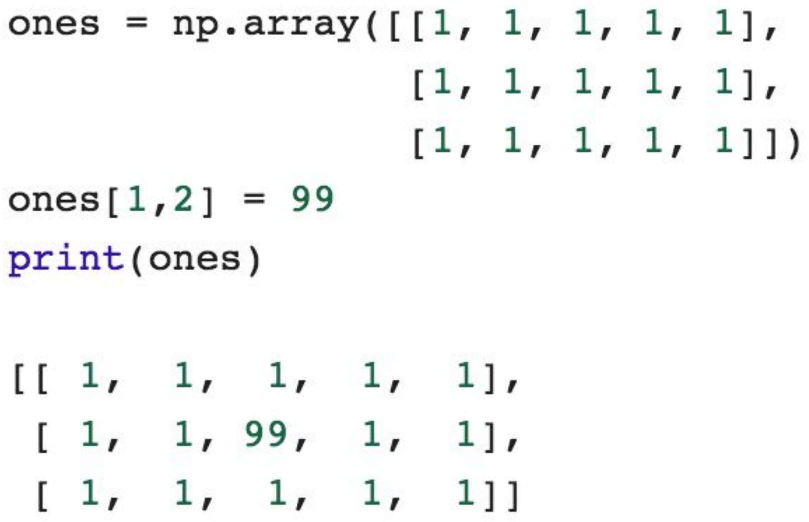


```
a = np.array(['red','blue','black','blue','purple'])
a[0] = 'orange'
print(a)

['orange', 'blue', 'black', 'blue', 'purple']
```

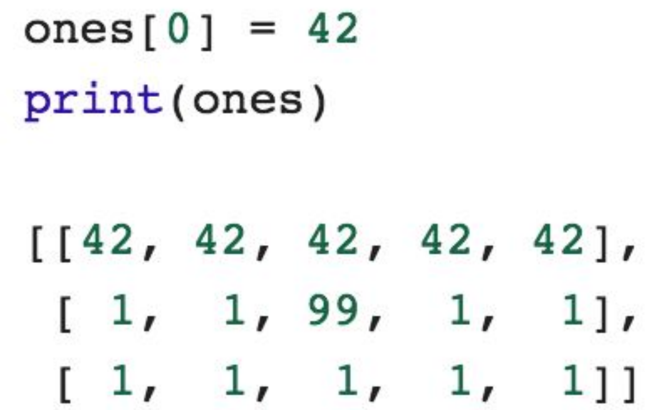


```
a[3:] = 'pink'
print(a)

['orange', 'blue', 'black', 'pink', 'pink']
```

Assigning values in 2D ndarray



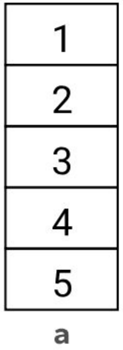
```
ones = np.array([[1, 1, 1, 1, 1],
                 [1, 1, 1, 1, 1],
                 [1, 1, 1, 1, 1]])
ones[1,2] = 99
print(ones)

[[ 1,  1,  1,  1,  1],
 [ 1,  1, 99,  1,  1],
 [ 1,  1,  1,  1,  1]]
```



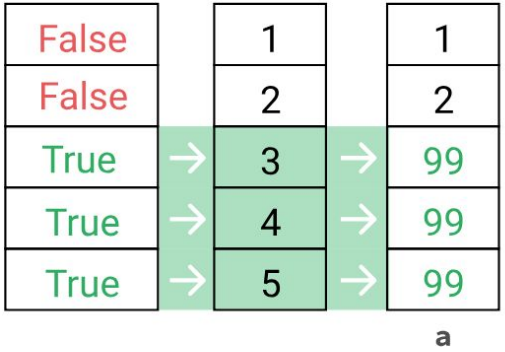
```
ones[0] = 42
print(ones)

[[42, 42, 42, 42, 42],
 [ 1,  1, 99,  1,  1],
 [ 1,  1,  1,  1,  1]]
```

Assignment Using Boolean Arrays

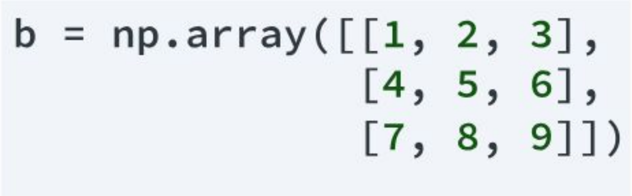
```
a = np.array([1, 2, 3, 4, 5])
```



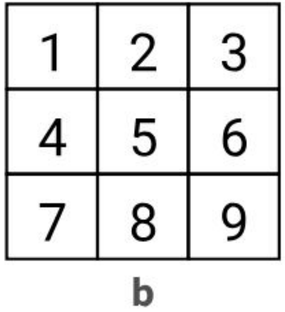
```
a[a > 2] = 99
```



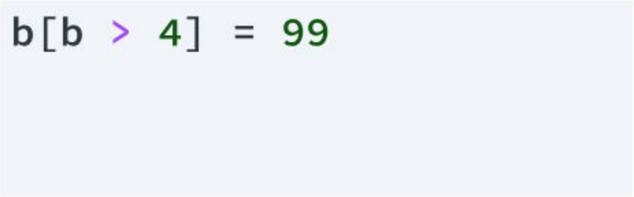
Assignment Using Boolean Arrays



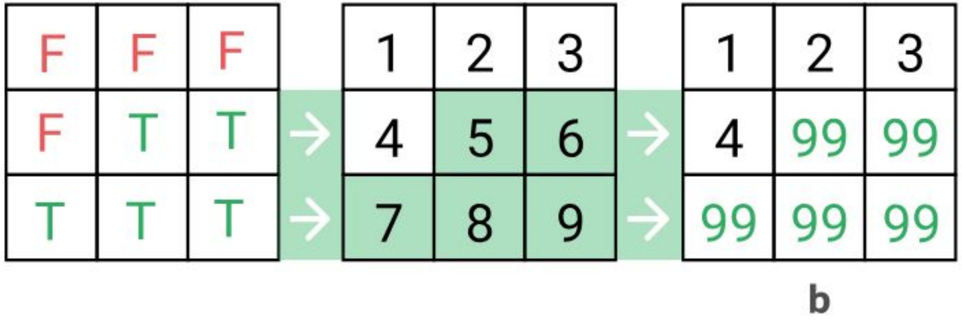
```
b = np.array([[1, 2, 3],
              [4, 5, 6],
              [7, 8, 9]])
```



1	2	3
4	5	6
7	8	9



```
b[b > 4] = 99
```



F	F	F
F	T	T
T	T	T

1	2	3
4	5	6
7	8	9

1	2	3
4	99	99
99	99	99

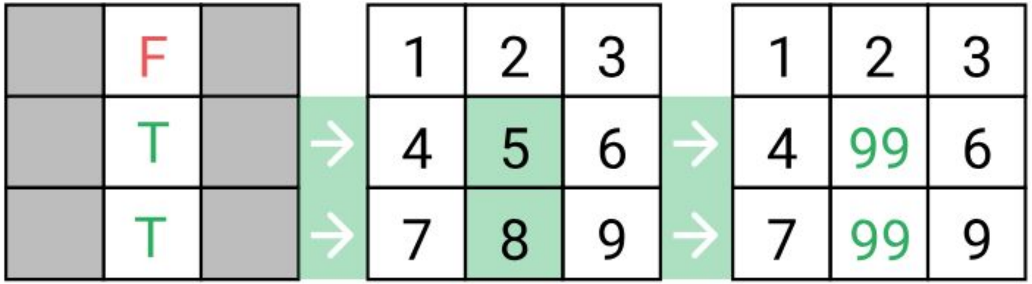
Assignment Using Boolean Arrays

```
c = np.array([[1, 2, 3],
              [4, 5, 6],
              [7, 8, 9]])
```

1	2	3
4	5	6
7	8	9

c

```
c[c[:, 1] > 2, 1] = 99
```



	F	
	T	
	T	

1	2	3
4	5	6
7	8	9

1	2	3
4	99	6
7	99	9

c

Challenges

Which is the most popular airport?
Calculating statistics for trip?

Lesson_08_Introduction_to_numpy.ipynb