

03.03-Operations-in-Pandas

June 27, 2018

This notebook contains an excerpt from the [Python Data Science Handbook](#) by Jake VanderPlas; the content is available [on GitHub](#).

The text is released under the [CC-BY-NC-ND license](#), and code is released under the [MIT license](#). If you find this content useful, please consider supporting the work by [buying the book](#)!

< [Data Indexing and Selection](#) | [Contents](#) | [Handling Missing Data](#) >

1 Operating on Data in Pandas

One of the essential pieces of NumPy is the ability to perform quick element-wise operations, both with basic arithmetic (addition, subtraction, multiplication, etc.) and with more sophisticated operations (trigonometric functions, exponential and logarithmic functions, etc.). Pandas inherits much of this functionality from NumPy, and the ufuncs that we introduced in [Computation on NumPy Arrays: Universal Functions](#) are key to this.

Pandas includes a couple useful twists, however: for unary operations like negation and trigonometric functions, these ufuncs will *preserve index and column labels* in the output, and for binary operations such as addition and multiplication, Pandas will automatically *align indices* when passing the objects to the ufunc. This means that keeping the context of data and combining data from different sources—both potentially error-prone tasks with raw NumPy arrays—become essentially foolproof ones with Pandas. We will additionally see that there are well-defined operations between one-dimensional Series structures and two-dimensional DataFrame structures.

1.1 Ufuncs: Index Preservation

Because Pandas is designed to work with NumPy, any NumPy ufunc will work on Pandas Series and DataFrame objects. Let's start by defining a simple Series and DataFrame on which to demonstrate this:

```
In [1]: import pandas as pd
import numpy as np
```

```
In [2]: rng = np.random.RandomState(42)
ser = pd.Series(rng.randint(0, 10, 4))
ser
```

```
Out[2]: 0    6
        1    3
        2    7
```

```
3      4
dtype: int64
```

```
In [3]: df = pd.DataFrame(rng.randint(0, 10, (3, 4)),
                        columns=['A', 'B', 'C', 'D'])

df
```

```
Out [3]:
```

	A	B	C	D
0	6	9	2	6
1	7	4	3	7
2	7	2	5	4

If we apply a NumPy ufunc on either of these objects, the result will be another Pandas object *with the indices preserved*:

```
In [4]: np.exp(ser)

Out [4]:
```

0	403.428793
1	20.085537
2	1096.633158
3	54.598150

```
dtype: float64
```

Or, for a slightly more complex calculation:

```
In [5]: np.sin(df * np.pi / 4)

Out [5]:
```

	A	B	C	D
0	-1.000000	7.071068e-01	1.000000	-1.000000e+00
1	-0.707107	1.224647e-16	0.707107	-7.071068e-01
2	-0.707107	1.000000e+00	-0.707107	1.224647e-16

Any of the ufuncs discussed in [Computation on NumPy Arrays: Universal Functions](#) can be used in a similar manner.

1.2 UFuncs: Index Alignment

For binary operations on two Series or DataFrame objects, Pandas will align indices in the process of performing the operation. This is very convenient when working with incomplete data, as we'll see in some of the examples that follow.

1.2.1 Index alignment in Series

As an example, suppose we are combining two different data sources, and find only the top three US states by *area* and the top three US states by *population*:

```
In [6]: area = pd.Series({'Alaska': 1723337, 'Texas': 695662,
                        'California': 423967}, name='area')
      population = pd.Series({'California': 38332521, 'Texas': 26448193,
                        'New York': 19651127}, name='population')
```

Let's see what happens when we divide these to compute the population density:

```
In [7]: population / area
```

```
Out[7]: Alaska          NaN
        California    90.413926
        New York       NaN
        Texas          38.018740
        dtype: float64
```

The resulting array contains the *union* of indices of the two input arrays, which could be determined using standard Python set arithmetic on these indices:

```
In [8]: area.index | population.index
```

```
Out[8]: Index(['Alaska', 'California', 'New York', 'Texas'], dtype='object')
```

Any item for which one or the other does not have an entry is marked with NaN, or "Not a Number," which is how Pandas marks missing data (see further discussion of missing data in [Handling Missing Data](#)). This index matching is implemented this way for any of Python's built-in arithmetic expressions; any missing values are filled in with NaN by default:

```
In [9]: A = pd.Series([2, 4, 6], index=[0, 1, 2])
        B = pd.Series([1, 3, 5], index=[1, 2, 3])
        A + B
```

```
Out[9]: 0    NaN
        1    5.0
        2    9.0
        3    NaN
        dtype: float64
```

If using NaN values is not the desired behavior, the fill value can be modified using appropriate object methods in place of the operators. For example, calling `A.add(B)` is equivalent to calling `A + B`, but allows optional explicit specification of the fill value for any elements in A or B that might be missing:

```
In [10]: A.add(B, fill_value=0)
```

```
Out[10]: 0    2.0
        1    5.0
        2    9.0
        3    5.0
        dtype: float64
```

1.2.2 Index alignment in DataFrame

A similar type of alignment takes place for *both* columns and indices when performing operations on DataFrames:

```
In [11]: A = pd.DataFrame(rng.randint(0, 20, (2, 2)),
                        columns=list('AB'))
```

A

```
Out[11]:    A  B
0    1  11
1    5   1
```

```
In [12]: B = pd.DataFrame(rng.randint(0, 10, (3, 3)),
                        columns=list('BAC'))
```

B

```
Out[12]:    B  A  C
0    4  0  9
1    5  8  0
2    9  2  6
```

```
In [13]: A + B
```

```
Out[13]:    A    B  C
0    1.0  15.0 NaN
1   13.0   6.0 NaN
2    NaN   NaN NaN
```

Notice that indices are aligned correctly irrespective of their order in the two objects, and indices in the result are sorted. As was the case with Series, we can use the associated object's arithmetic method and pass any desired `fill_value` to be used in place of missing entries. Here we'll fill with the mean of all values in A (computed by first stacking the rows of A):

```
In [14]: fill = A.stack().mean()
        A.add(B, fill_value=fill)
```

```
Out[14]:    A    B    C
0    1.0  15.0  13.5
1   13.0   6.0   4.5
2    6.5  13.5  10.5
```

The following table lists Python operators and their equivalent Pandas object methods:

Python Operator	Pandas Method(s)
+	<code>add()</code>
-	<code>sub()</code> , <code>subtract()</code>
*	<code>mul()</code> , <code>multiply()</code>
/	<code>truediv()</code> , <code>div()</code> , <code>divide()</code>
//	<code>floordiv()</code>
%	<code>mod()</code>
**	<code>pow()</code>

1.3 Ufuncs: Operations Between DataFrame and Series

When performing operations between a DataFrame and a Series, the index and column alignment is similarly maintained. Operations between a DataFrame and a Series are similar to operations between a two-dimensional and one-dimensional NumPy array. Consider one common operation, where we find the difference of a two-dimensional array and one of its rows:

```
In [15]: A = rng.randint(10, size=(3, 4))
        A
```

```
Out[15]: array([[3, 8, 2, 4],
                [2, 6, 4, 8],
                [6, 1, 3, 8]])
```

```
In [16]: A - A[0]
```

```
Out[16]: array([[ 0,  0,  0,  0],
                [-1, -2,  2,  4],
                [ 3, -7,  1,  4]])
```

According to NumPy's broadcasting rules (see [Computation on Arrays: Broadcasting](#)), subtraction between a two-dimensional array and one of its rows is applied row-wise.

In Pandas, the convention similarly operates row-wise by default:

```
In [17]: df = pd.DataFrame(A, columns=list('QRST'))
        df - df.iloc[0]
```

```
Out[17]:   Q  R  S  T
0  0  0  0  0
1 -1 -2  2  4
2  3 -7  1  4
```

If you would instead like to operate column-wise, you can use the object methods mentioned earlier, while specifying the axis keyword:

```
In [18]: df.subtract(df['R'], axis=0)
```

```
Out[18]:   Q  R  S  T
0 -5  0 -6 -4
1 -4  0 -2  2
2  5  0  2  7
```

Note that these DataFrame/Series operations, like the operations discussed above, will automatically align indices between the two elements:

```
In [19]: halfrow = df.iloc[0, ::2]
        halfrow
```

```
Out[19]: Q      3
        S      2
        Name: 0, dtype: int64
```

```
In [20]: df - halfrow
```

```
Out[20]:
```

	Q	R	S	T
0	0.0	NaN	0.0	NaN
1	-1.0	NaN	2.0	NaN
2	3.0	NaN	1.0	NaN

This preservation and alignment of indices and columns means that operations on data in Pandas will always maintain the data context, which prevents the types of silly errors that might come up when working with heterogeneous and/or misaligned data in raw NumPy arrays.

< [Data Indexing and Selection](#) | [Contents](#) | [Handling Missing Data](#) >