Introduction to Big Data

- Developed by Dr. Keungoui KIM
- https://awekim.github.io/portfolio/

Lecture 4. Data Manipulation with Pandas I

Numpy

```
from google.colab import drive
drive.mount('/content/drive')
```

```
import numpy as np
```

Create Array

+ 코드 + 텍스트

```
np_array = np.array([0,1,2,3,4,5,6,7,8,9])
np_array
```

```
np_arange = np.arange(10)
np_arange
```

```
np_arange = np.arange(1,10,2)
np_arange
```

```
np_linspace = np.linspace(1,10,8)
np_linspace
```

```
np_ones = np.ones(10)
np_ones
```

```
np_zeros = np.zeros(10)
np_zeros
```

```
np_full = np.full((5,2), 3)
np_full
```

```
np_eyes = np.eye(10)
np_eyes
```

```
np.array([0, 1, -1, 0]) / np.array([1, 0, 0, 0])
```

```
np.log(0)
```

```
np.nan
```

```
np.inf
```

```
np.percentile([1,2,3,4,5,6,7,8,9,10], 50)
```

```
np.percentile([1,2,3,4,5,6,7,8,9,10], 50,
              method = 'lower')
```

```
np.percentile([1,2,3,4,5,6,7,8,9,10],
              [0,25,50,75,100])
```

```
import numpy as np
```

```
a = np.arange(1,11)
a
```

```
np.where(a>5)
```

```
b = np.arange( 10, 110, 10)
b
```

```
np.where( b > 50)
```

```
np.where( b > 50, b, b/10 )
```

Dimension

```
# 0 dimension: scalar
np_0 = np.array(1)
np_0
```

```
np_0.shape
```

```
np_0.ndim
```

```
# 1 dimension: vector
np_1 = np.array([1])
np_1
```

```
np_1.shape
```

```
np_1.ndim
```

```
np_11 = np.array([1,3,5,7,9])
np_11
```

```
np_11.shape
```

```
np_11.ndim
```

```
# 2 dimension: matrix
np_2 = np.array([[1]])
np_2
```

```
np_2.shape
```

```
np_2.ndim
```

```
np_22 = np.array([[1,3,5], [7,9,11]])
np_22
```

```
np_22.shape
```

```
np_22.ndim
```

```
# More 3 dimensions: Tensor
np_3 =np.array([[[1]]])
np_3
```

```
np_3.shape
```

```
np_3.ndim
```

```
np_33 = np.array([[[1,3], [5,7]], [[9,11], [13,15]]])
np_33
```

```
np_33.shape
```

```
np_33.ndim
```

Reshape

```
# .arange
np.arange(1, 10)
```

```
np.arange(1, 10).reshape(3,3)
```

```
np.arange(1, 13).reshape(4,3)
```

```
np.arange(1, 13).reshape(3,4)
```

```
np.arange(1, 13).reshape(3,4).sum(axis=1)
```

```
np.arange(1, 13).reshape(3,2,2)
```

```
np.arange(1, 13).reshape(3,2,2).sum(axis=0)
# 1 + 5 + 9
```

```
np.arange(1, 13).reshape(3,2,2).sum(axis=1)
# 1 + 3, 2 + 4
```

```
np.arange(1, 13).reshape(3,2,2).sum(axis=2)
# 1 + 2, 3 + 4
```

Indexing

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

```
data_array[0]
```

```
data_array[0][0]
```

```
data_array[0,0]
```

```
data_array[0][0][0]
```

Operation

```
np_array_1 = np.arange(1, 10).reshape(3,3)
np_array_1
```

```
np_array_1 * 2
```

```
np_array_1 + np_array_1
```

```
np_array_1 - np_array_1
```

```
np_array_1 * np_array_1
```

```
np.dot(np_array_1, np_array_1)
```

```
np_array_1.sum()
```

```
np_array_1.sum(axis=0, keepdims=True)
```

```
np_array_1.sum(axis=1, keepdims=True)
```

```
np_array_1.max()
```

```
np_array_1.max(axis=0, keepdims=True)
```

```
np_array_1.max(axis=1, keepdims=True)
```

```
np_array_1.mean()
```

```
np_array_1.std()
```

np.nan

```
np.nan
```

```
np.nan + 10
```

```
np.nan * 10
```

Data type

```
data =[[1, 2, 3], [4, 5, 6], [7, 8, 9]]
```

```
data_array=np.array(data)
data_array
```

```
data_array.dtype
```

```
data =[[1, 2, 3], [4, 5, 6], [7, 8, 9], [10.0, 11, 12]]
data_array=np.array(data)
data_array
```

```
data_array.dtype
```

```
# convert data type

data_array_int = data_array.astype(np.int32)
# data_array_int = data_array.astype('int32')
data_array_int.dtype
```