Python을 활용한 데이터 분석 강의

Numpy & Pandas

HYUNDAI MOBIS

Numpy & Pandas

- Numpy와 Pandas는 수치분석 및 데이터 분석을 위한 쉬운 도구를 제공

```
import pandas as pd
```

```
train = pd.read_csv('titanic_train.csv')
```

```
train
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S

Numpy

- Numpy는 Numerical Python의 약자로, 수치연산기능을 제공
- Numpy는 자체적인 array를 제공 - list와 비슷하지만 보다 다양한 연산 메소드를 가지고 있으며 하나의 array에는 같은 type의 데이터만 담을 수 있다는 점에서 차이

```
import numpy as np

array1 = np.array([1, 2, 3])   # 1차원 array를 만든다.
print(array1.shape)            # (3,) 이라고 출력. 요소가 3인 1차원 array를 의미함
print(array1)

# 특정 요소에 접근하기 위해서는 list와 같이 인덱스를 사용한다.
print(array1[0], array1[2])
array1[1] = 10
print(array1)

array2 = np.array([[1,2,3],[4,5,6]])   # 2차원 array를 만든다.
print(array2.shape)                    # (2, 3) 이라고 출력. (row, column)
print(array2)
print(array2[0, 0], array2[0, 1], array2[1, 0])   # Prints "1 2 4"
```

```
(3,)
[1 2 3]
1 3
[ 1 10  3]
(2, 3)
[[1 2 3]
 [4 5 6]]
1 2 4
```

Numpy

```
import numpy as np

a = np.zeros((2,3))    # 모두 zero로 채워진 2x3 array를 생성
print(a)

b = np.ones((1,2))     # 모두 1로 채워진 1x2 array를 생성
print(b)

c = np.random.random((2,2)) # 랜덤 넘버로 채워진 2x2 array를 생성
print(c)
```

```
[[0. 0. 0.]
 [0. 0. 0.]]
[[1. 1.]]
[[0.72852896 0.39863978]
 [0.02971612 0.78768585]]
```

Numpy

```python
import numpy as np

x = np.array([[1,2],[3,4]], dtype=np.float64)
y = np.array([[5,6],[7,8]], dtype=np.float64)
print(x)
print(y)

print(x + y)
print(np.add(x, y))

print(x - y)
print(np.subtract(x, y))

print(x * y)
print(np.multiply(x, y))

print(x / y)
print(np.divide(x, y))

print(np.sqrt(x))
```

```
[[1. 2.]
 [3. 4.]]
[[5. 6.]
 [7. 8.]]
[[ 6.  8.]
 [10. 12.]]
[[ 6.  8.]
 [10. 12.]]
[[-4. -4.]
 [-4. -4.]]
[[-4. -4.]
 [-4. -4.]]
[[ 5. 12.]
 [21. 32.]]
[[ 5. 12.]
 [21. 32.]]
[[0.2        0.33333333]
 [0.42857143 0.5       ]]
[[0.2        0.33333333]
 [0.42857143 0.5       ]]
[[1.         1.41421356]
 [1.73205081 2.        ]]
```

HYUNDAI MOBIS

Pandas

- Pandas는 numpy를 기반으로 개발된 자료구조

```
import pandas as pd
import numpy as np
```

Pandas는 read_csv()라는 CSV 파일을 읽어주는 함수를 제공한다.

(참고) Excel 파일을 읽는 함수도 제공한다. ExcelFile(), read_excel()

- df = pd.ExcelFile("dummydata.xlsx")
- df = pd.read_excel(open('your_xls_xlsx_filename','rb'), sheetname='Sheet 1')

```
data = pd.read_csv("data/weather_year.csv")
```

- 자료구조
  - Series : 객체를 담을 수 있는 1차원 배열
  - DataFrame : 스프레드시트의 표 같은 형식으로 여러 column으로 구성