데이터 전처리를 위한

Pandas (I)

박성호 (neowizard2018@gmail.com)

Contents

데이터프레임 개요

- 판다스(Pandas)는 데이터프레임(DataFrame)과 시리즈(Series) 라는 데이터타입(DataType)과 데이터 분석을 위한 다양한 기능을 제공 해주는 파이썬 라이브러리

데이터프레임(DataFrame)

	Name	Country	Age	Job
0	John	USA	31	Student
1	Sabre	France	33	Lawyer
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor
5	Smith	USA	55	CEO
6	David	USA	48	Banker

시리즈(Series)

데이터프레임 생성 (from dictionary data)

➢ 데이터프레임은 dictionary 데이터 또는 list 데이터를 이용해서 생성할 수 있음

```
import pandas as pd

data_dict = { 'Name' : ['John', 'Sabre', 'Kim', 'Sato', 'Lee', 'Smith', 'David'],
              'Country' : ['USA', 'France', 'Korea', 'Japan', 'Korea', 'USA', 'USA'],
              'Age' : [ 31, 33, 28, 40, 36, 55, 48],
              'Job' : ['Student', 'Lawyer', 'Developer', 'Chef', 'Professor', 'CEO', 'Banker']
            }


df = pd.DataFrame(data_dict)

df
```

	Name	Country	Age	Job
0	John	USA	31	Student
1	Sabre	France	33	Lawyer
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor
5	Smith	USA	55	CEO
6	David	USA	48	Banker

데이터프레임 생성 (from list data)

```
import pandas as pd

data_list = [ ['John', 'USA', 31, 'Student'],
              ['Sabre', 'France', 33, 'Lawyer'],
              ['Kim', 'Korea', 28, 'Developer'],
              ['Sato', 'Japan', 40, 'Chef'],
              ['Lee', 'Korea', 36, 'Professor'],
              ['Smith', 'USA', 55, 'CEO'],
              ['David', 'USA', 48, 'Banker'],
            ]

column_name = ['Name', 'Country', 'Age', 'Job']

df = pd.DataFrame(data_list, columns = column_name)

df
```

	Name	Country	Age	Job
0	John	USA	31	Student
1	Sabre	France	33	Lawyer
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor
5	Smith	USA	55	CEO
6	David	USA	48	Banker

데이터프레임 기본 정보 확인

df.head(), df.tail(), df.info(), df.describe(), df.index, df.columns

df.head()

	Name	Country	Age	Job
0	John	USA	31	Student
1	Sabre	France	33	Lawyer
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor

df.tail()

	Name	Country	Age	Job
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor
5	Smith	USA	55	CEO
6	David	USA	48	Banker

df.describe()

	Age
count	7.000000
mean	38.714286
std	9.724784
min	28.000000
25%	32.000000
50%	36.000000
75%	44.000000
max	55.000000

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7 entries, 0 to 6
Data columns (total 4 columns):
Name       7 non-null object
Country    7 non-null object
Age        7 non-null int64
Job        7 non-null object
dtypes: int64(1), object(3)
memory usage: 304.0+ bytes
```

데이터프레임 csv 파일 저장 df.to_csv(...)

```
# index => Yes

df.to_csv('./test_dataframe_with_index.csv', index=True)
```

```
# index => No

df.to_csv('./test_dataframe_without_index.csv', index=False)
```

```
# header => Yes

df.to_csv('./test_dataframe_with_header.csv', header=True)
```

```
# header => No

df.to_csv('./test_dataframe_without_header.csv', header=False)
```

```
# header => YES, index => NO

df.to_csv('./test_dataframe_with_header_without_index.csv', header=True, index=False)
```

```
# header => No,  index => No

df.to_csv('./test_dataframe_without_header_without_index.csv', header=False, index=False)
```

csv 파일로부터 데이터프레임 생성 pd.read_csv(...)

```
import pandas as pd
df = pd.read_csv('./test_dataframe_with_header_without_index.csv')
df
```

	Name	Country	Age	Job
0	John	USA	31	Student
1	Sabre	France	33	Lawyer
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor
5	Smith	USA	55	CEO
6	David	USA	48	Banker

```
import pandas as pd
df = pd.read_csv('./test_dataframe_without_header_without_index.csv')
df
```

	John	USA	31	Student
0	Sabre	France	33	Lawyer
1	Kim	Korea	28	Developer
2	Sato	Japan	40	Chef
3	Lee	Korea	36	Professor
4	Smith	USA	55	CEO
5	David	USA	48	Banker

← 첫번째 데이터를 header 인식함

csv 파일로부터 데이터프레임 생성 pd.read_csv(...)

```
import pandas as pd

df = pd.read_csv('./test_dataframe_without_header_without_index.csv', header=None)

df
```

	0	1	2	3
0	John	USA	31	Student
1	Sabre	France	33	Lawyer
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor
5	Smith	USA	55	CEO
6	David	USA	48	Banker

```
cols = [ 'Name', 'Country', 'Age', 'Job' ]

df.columns = cols

df
```

	Name	Country	Age	Job
0	John	USA	31	Student
1	Sabre	France	33	Lawyer
2	Kim	Korea	28	Developer
3	Sato	Japan	40	Chef
4	Lee	Korea	36	Professor
5	Smith	USA	55	CEO
6	David	USA	48	Banker