데이터 전처리를 위한

Pandas (III)

박성호 (neowizard2018@gmail.com)

Contents

Missing Data (NaN, None 등) 처리 1

➢ 판다스 read_csv(…) 이용하여 다음과 같은 데이터 읽어 옴 (Missing Data 확인)

Name	Country	Age	Job
John	USA	31	Student
Sabre	France	33	
Kim	Korea	28	Developer
Sato		40	Chef
Lee	Korea	36	Professor
Smith	USA	55	CEO
David	USA		Banker

test_missing_data.csv

```
import pandas as pd

try:
    df = pd.read_csv('./test_missing_data.csv')
except Exception as err:
    print(str(err))
```

```
df
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
1	Sabre	France	33.0	NaN
2	Kim	Korea	28.0	Developer
3	Sato	NaN	40.0	Chef
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	CEO
6	David	USA	NaN	Banker

Missing Data (NaN, None 등) 처리 2 – isnull(), dropna()

- Missing Data 개수 확인

  df.isnull().sum()

```
df.isnull().sum()

Name       0
Country    1
Age        1
Job        1
dtype: int64
```

- 각 열(column)에 있는 각각의 Data 개수 확인 (NaN 제외한 데이터 개수)

  df['Name'].value_counts(), df['Country'].value_counts(),…

- NaN 값이 있는 행(row) 모두 제거

  df.dropna()

```
df1 = df.dropna()

df1
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
2	Kim	Korea	28.0	Developer
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	CEO

Missing Data (NaN, None 등) 처리 3 – fillna()

- Missing Data 를 특정 값으로 변경하기 (각 열의 NaN)

  df['열이름'].fillna(변경값, inplace=True)

```
df['Country'].fillna('Spain')
df['Age'].fillna(100.0)
df['Job'].fillna('Reporter')

df
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
1	Sabre	France	33.0	NaN
2	Kim	Korea	28.0	Developer
3	Sato	NaN	40.0	Chef
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	CEO
6	David	USA	NaN	Banker

```
df['Country'].fillna('Spain', inplace=True)
df['Age'].fillna(100.0, inplace=True)
df['Job'].fillna('Reporter', inplace=True)

df
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
1	Sabre	France	33.0	Reporter
2	Kim	Korea	28.0	Developer
3	Sato	Spain	40.0	Chef
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	CEO
6	David	USA	100.0	Banker

Missing Data (NaN, None 등) 처리 4 – fillna()

- Missing Data 를 특정 값으로 변경하기 (모든 NaN)

  df.fillna(변경값, inplace=True)

```
df_test = pd.read_csv('./test_missing_data.csv')

df_test
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
1	Sabre	France	33.0	NaN
2	Kim	Korea	28.0	Developer
3	Sato	NaN	40.0	Chef
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	CEO
6	David	USA	NaN	Banker

```
df_test.fillna('AAA', inplace=True)

df_test
```

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

[appendix] mean(), median(), replace()

- fillna() 에서 NaN 을 특정 값으로 변경할때 mean() 또는 median() 등으로 바꾸는 경우가 많음 (통계의 오류는 감안 해야함)

```
df_stat = pd.read_csv('./test_missing_data.csv')

df_stat
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
1	Sabre	France	33.0	NaN
2	Kim	Korea	28.0	Developer
3	Sato	NaN	40.0	Chef
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	CEO
6	David	USA	NaN	Banker

```
print('Age mean = ', df_stat['Age'].mean())
print('Age median = ', df_stat['Age'].median())
```

```
Age mean =  37.166666666666664
Age median =  34.5
```

[appendix] mean(), median(), replace()

- replace() 함수 이용하여 NaN ⇒ 특정값 또는 특정값 ⇒ NaN 으로 변경하는 경우도 있음 (특정값은 일반적으로 outlier 경우가 일반적임)

```
import numpy as np

df_stat['Age'].replace(np.nan, 50, inplace=True)

df_stat
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
1	Sabre	France	33.0	NaN
2	Kim	Korea	28.0	Developer
3	Sato	NaN	40.0	Chef
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	CEO
6	David	USA	50.0	Banker

```
import numpy as np

df_stat['Job'].replace('CEO', np.nan, inplace=True)

df_stat
```

	Name	Country	Age	Job
0	John	USA	31.0	Student
1	Sabre	France	33.0	NaN
2	Kim	Korea	28.0	Developer
3	Sato	NaN	40.0	Chef
4	Lee	Korea	36.0	Professor
5	Smith	USA	55.0	NaN
6	David	USA	50.0	Banker