
세상의 속도를
따라잡고 싶다면
Do it!
데이터 분석을 위한
판다스 입문
7일 완성!
테슬라 주식 • 우버 택시 • 빌보드 차트 • 예볼라 바이러스 데이터 등
현실 데이터 기반의 86개 예제로 판다스 실무 기본기 완성!
Chen, Daniel Y. 지음 | 강경덕 옮김
이지스퍼블리싱
파이썬 3.9 판다스
최신 버전 반영!
데이터 분석을 위한 판다스 입문
이지스퍼블리싱

6장 누락값 처리하기

06-1 누락값이란?

누락값과 누락값 확인하기

누락값을 위해 numpy 라이브러리 사용

```
from numpy import NaN, NAN, nan
```

누락값은 0, ' '(빈 값)과는 다른 개념이라는 것에 주의

```
print(NaN == 0)
```

```
False
```

```
print(NaN == '')
```

```
False
```

```
print(NaN == NaN)
```

```
False
```

```
print(NaN == nan)
```

```
False
```

누락값과 누락값 확인하기

누락값 확인하기

```
print(pd.isnull(NAN))
```

```
True
```

isnull () : null이면 True

```
print(pd.notnull(NaN))
```

notnull () : null이면 False

누락값의 개수 구하기

```
ebola = pd.read_csv('../data/country_timeseries.csv')
```

```
         Date  Day  Cases_Guinea  Cases_Liberia  Cases_SierraLeone  #
0    1/5/2015  289        2776.0            NaN             10030.0
1    1/4/2015  288        2775.0            NaN              9780.0
2    1/3/2015  287        2769.0         8166.0              9722.0
3    1/2/2015  286           NaN         8157.0                 NaN
4  12/31/2014  284        2730.0         8115.0              9633.0

   Cases_Nigeria  Cases_Senegal  Cases_UnitedStates  Cases_Spain  Cases_Mali  #
0            NaN            NaN                 NaN          NaN         NaN
1            NaN            NaN                 NaN          NaN         NaN
2            NaN            NaN                 NaN          NaN         NaN
3            NaN            NaN                 NaN          NaN         NaN
4            NaN            NaN                 NaN          NaN         NaN

   Deaths_Guinea  Deaths_Liberia  Deaths_SierraLeone  Deaths_Nigeria  #
0         1786.0             NaN              2977.0             NaN
1         1781.0             NaN              2943.0             NaN
2         1767.0          3496.0              2915.0             NaN
3            NaN          3496.0                 NaN             NaN
4         1739.0          3471.0              2827.0             NaN

   Deaths_Senegal  Deaths_UnitedStates  Deaths_Spain  Deaths_Mali
0             NaN                  NaN           NaN          NaN
1             NaN                  NaN           NaN          NaN
2             NaN                  NaN           NaN          NaN
3             NaN                  NaN           NaN          NaN
4             NaN                  NaN           NaN          NaN
```

count 메소드는 각 column의
누락 값이 아닌 값들만 카운트한다.
누락 값 = 전체 행의 수 – 누락값이 아닌 값들

```
num_rows = ebola.shape[0]
num_missing = num_rows - ebola.count()
print(num_missing)
```

```
Date                   0
Day                    0
Cases_Guinea          29
Cases_Liberia         39
Cases_SierraLeone     35
Cases_Nigeria         84
Cases_Senegal         97
Cases_UnitedStates   104
Cases_Spain          106
Cases_Mali           110
Deaths_Guinea         30
Deaths_Liberia        41
Deaths_SierraLeone    35
Deaths_Nigeria        84
Deaths_Senegal       100
Deaths_UnitedStates  104
```

누락값 처리하기

1. 누락 값을 임의의 값으로 변경한다. fillna(채울 값)

```
print(ebola.iloc[0:5, 0:5])
print(ebola.fillna(0).iloc[0:5, 0:5])
```

```
         Date  Day  Cases_Guinea  Cases_Liberia  Cases_SierraLeone
0    1/5/2015  289        2776.0            NaN            10030.0
1    1/4/2015  288        2775.0            NaN             9780.0
2    1/3/2015  287        2769.0         8166.0             9722.0
3    1/2/2015  286           NaN         8157.0                NaN
4  12/31/2014  284        2730.0         8115.0             9633.0
         Date  Day  Cases_Guinea  Cases_Liberia  Cases_SierraLeone
0    1/5/2015  289        2776.0            0.0            10030.0
1    1/4/2015  288        2775.0            0.0             9780.0
2    1/3/2015  287        2769.0         8166.0             9722.0
3    1/2/2015  286           0.0         8157.0                0.0
4  12/31/2014  284        2730.0         8115.0             9633.0
```

누락값 처리하기

2. 데이터프레임에 이미 있는 값으로 대신 채운다.

fillna 메소드 사용

인자 method = 'ffill' 인 경우 누락 값의 앞의 값을 가져온다.

'ffill' = forward fill / 'bfill' = backward fill

```
print(ebola.fillna(method='ffill').iloc[0:10, 0:5])
```

	Date	Day	Cases_Guinea	Cases_Liberia	Cases_SierraLeone
0	1/5/2015	289	2776.0	NaN	10030.0
1	1/4/2015	288	2775.0	NaN	9780.0
2	1/3/2015	287	2769.0	8166.0	9722.0
3	1/2/2015	286	2769.0	8157.0	9722.0
4	12/31/2014	284	2730.0	8115.0	9633.0
5	12/28/2014	281	2706.0	8018.0	9446.0
6	12/27/2014	280	2695.0	8018.0	9409.0
7	12/24/2014	277	2630.0	7977.0	9203.0
8	12/21/2014	273	2597.0	7977.0	9004.0

0,1 행은 처음부터 누락 값이기 때문에 그대로 남아있다!

```
print(ebola.fillna(method='bfill').iloc[0:10, 0:5])
```

누락값 처리하기

3. 양쪽에 있는 값을 이용하여 중간값을 넣는다.

   interpolate 메소드 사용

```
print(ebola.interpolate().iloc[0:10, 0:5])
```

누락값 처리하기

4. 누락값 삭제하기 – 누락값이 필요 없을 경우 누락값을 삭제해도 된다. 하지만 무작정 삭제하면 데이터가 너무 편향되거나 데이터의 개수가 너무 적어질 수 있다. dropna 메소드 이용

```
print(ebola.shape)
```

```
(122, 18)
```

```
ebola_dropna = ebola.dropna()
print(ebola_dropna.shape)
```

```
(1, 18)
```

누락값이 포함된 데이터 계산하기

누락값이 포함된 채로 계산하면 계산 결과도 누락값이 된다.

인자 skipna = True 로 두면 누락값을 포함해 계산한다. (0으로 생각)

```
print(ebola.Cases_Guinea.sum(skipna = True))
```

```
0    2776.0
1    2775.0
2    2769.0
3       NaN
4    2730.0
Name: Cases_Guinea, dtype: float64
11050.0
```

```
print(ebola.Cases_Guinea.sum(skipna = False))
```

```
nan
```