# 누락값 처리하기

- 누락값이란?
- 누락값이 생기는 이유
- 누락값 처리하기

- 누락값이란?
  - 누락값 확인하기
    - NaN, NAN, nan 과 같은 방법으로 표기
    - 논리 자료로 처리 가능한 True, False, 0, " 과 달리 값 자체가 없는 상태

```
print(np.NaN == True)
print(np.NaN == False)
print(np.NaN == 0)
print(np.NaN == '')
```

False

False

False

False

```
print(np.NaN == np.NaN)
```

False

False

False

False

- 누락값이란?
  - 누락값 확인하기
    - 누락값을 확인하는 isnull / notnull 메소드

```
import pandas as pd

print(pd.isnull(np.NaN))
print(pd.isnull(np.nan))
print(pd.isnull(np.NAN))

print(pd.notnull(np.NaN))
print(pd.notnull(42))
print(pd.notnull('missing'))
```

True

True

True

False

True

True

- 누락값이 생기는 이유
  - 1. 누락값이 있는 데이터 집합을 연결할 때

```
visited = pd.read_csv('data/survey_visited.csv')
survey = pd.read_csv('data/survey_survey.csv')
```

#### visited (날짜)

	ident	site	dated
0	619	DR-1	1927-02-08
1	622	DR-1	1927-02-10
2	734	DR-3	1939-01-07
3	735	DR-3	1930-01-12
4	751	DR-3	1930-02-26
5	752	DR-3	NaN
6	837	MSK-4	1932-01-14
7	844	DR-1	1932-03-22

### survey (날씨정보)

	taken	person	quant	reading
0	619	dyer	rad	9.82
1	619	dyer	sal	0.13
2	622	dyer	rad	7.80
3	622	dyer	sal	0.09
4	734	pb	rad	8.41
5	734	lake	sal	0.05
6	734	pb	temp	-21.50
7	735	pb	rad	7.22
8	735	NaN	sal	0.06
9	735	NaN	temp	-26.00
10	751	pb	rad	4.35
11	751	pb	temp	-18.50
12	751	lake	sal	0.10
13	752	lake	rad	2.19
14	752	lake	sal	0.09
15	752	lake	temp	-16.00
16	752	roe	sal	41.60
17	837	lake	rad	1.46
18	837	lake	sal	0.21
19	837	roe	sal	22.50
20	844	roe	rad	11.25

- 누락값이 생기는 이유
  - 1. 누락값이 있는 데이터 집합을 연결할 때

```
vs = visited.merge(survey, left_on='ident', right_on='taken')
print(vs)
```

### visited (날짜) + survey (날씨정보)

	ident	site	dated	taken	person	quant	reading
0	619	DR-1	1927-02-08	619	dyer	rad	9.82
1	619	DR-1	1927-02-08	619	dyer	sal	0.13
2	622	DR-1	1927-02-10	622	dyer	rad	7.80
3	622	DR-1	1927-02-10	622	dyer	sal	0.09
4	734	DR-3	1939-01-07	734	pb	rad	8.41
5	734	DR-3	1939-01-07	734	lake	sal	0.05
6	734	DR-3	1939-01-07	734	pb	temp	-21.50
7	735	DR-3	1930-01-12	735	pb	rad	7.22
8	735	DR-3	1930-01-12	735	NaN	sal	0.06
9	735	DR-3	1930-01-12	735	NaN	temp	-26.00
10	751	DR-3	1930-02-26	751	pb	rad	4.35
11	751	DR-3	1930-02-26	751	pb	temp	-18.50
12	751	DR-3	1930-02-26	751	lake	sal	0.10
13	752	DR-3	NaN	752	lake	rad	2.19
14	752	DR-3	NaN	752	lake	sal	0.09
15	752	DR-3	NaN	752	lake	temp	-16.00
16	752	DR-3	NaN	752	roe	sal	41.60
17	837	MSK-4	1932-01-14	837	lake	rad	1.46
18	837	MSK-4	1932-01-14	837	lake	sal	0.21
19	837	MSK-4	1932-01-14	837	roe	sal	22.50
20	844	DR-1	1932-03-22	844	roe	rad	11.25

- 누락값이 생기는 이유
  - 2. 데이터를 입력할 때

```
num_legs = pd.Series({'goat': 4, 'amoeba': nan})
print(num legs)
goat
amoeba
            NaN
dtype: float64
scientists = pd.DataFrame({
    'Name': ['Rosaline Franklin', 'William Gosset'],
    'Occupation': ['Chemist', 'Statistician'],
    'Born': ['1920-07-25', '1876-06-13'],
    'Died': ['1958-04-16', '1937-10-16'],
    'missing': [NaN, nan]})
print(scientists)
```

```
Name Occupation Born Died missing
O Rosaline Franklin Chemist 1920-07-25 1958-04-16 NaN
1 William Gosset Statistician 1876-06-13 1937-10-16 NaN
```

- 누락값 처리하기
  - 누락값 개수 확인
    - 에볼라 바이러스 데이터 가져오기

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

	Date	Day	Cases_Guinea	Cases_Liberia	Cases_SierraLeone	Cases_Nigeria	Cases_Senegal	Cases_UnitedStates	Cases_Spain	Cases_Mali	Deaths_(
0	1/5/2015	289	2776.0	NaN	10030.0	NaN	NaN	NaN	NaN	NaN	
1	1/4/2015	288	2775.0	NaN	9780.0	NaN	NaN	NaN	NaN	NaN	
2	1/3/2015	287	2769.0	8166.0	9722.0	NaN	NaN	NaN	NaN	NaN	
3	1/2/2015	286	NaN	8157.0	NaN	NaN	NaN	NaN	NaN	NaN	
4	12/31/2014	284	2730.0	8115.0	9633.0	NaN	NaN	NaN	NaN	NaN	
4											<b>•</b>

ases_Spain	Cases_Mali	Deaths_Guinea	Deaths_Liberia	Deaths_SierraLeone	Deaths_Nigeria	Deaths_Senegal	Deaths_UnitedStates	Deaths_Spain	Deaths_Mali
NaN	NaN	1786.0	NaN	2977.0	NaN	NaN	NaN	NaN	NaN
NaN	NaN	1781.0	NaN	2943.0	NaN	NaN	NaN	NaN	NaN
NaN	NaN	1767.0	3496.0	2915.0	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	3496.0	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	1739.0	3471.0	2827.0	NaN	NaN	NaN	NaN	NaN
4									<b>•</b>

- 누락값 처리하기
  - 누락값 개수 확인 count()
    - count 메소드로 누락값이 아닌 값의 개수 구하기

## print(ebola.count())

Date	122
Day	122
Cases_Guinea	93
Cases_Liberia	83
Cases_SierraLeone	87
Cases_Nigeria	38
Cases_Senegal	25
Cases_UnitedStates	18
Cases_Spain	16
Cases_Mali	12
Deaths_Guinea	92
Deaths_Liberia	81
Deaths_SierraLeone	87
Deaths_Nigeria	38
Deaths_Senegal	22
Deaths_UnitedStates	18
Deaths_Spain	16
Deaths_Mali	12
dtype: int64	

- 누락값 처리하기
  - 누락값 개수 확인 count()
    - shape[0]에 전체 행 개수 값의 개수 = 누락값의 개수 (브로드캐스팅)

```
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
Deaths_Spain	106
Deaths_Mali	110
dtype: int64	

- 누락값 처리하기
  - 누락값 개수 확인 count\_nonzero(), isnull(), value\_counts()
    - count\_nonzero(): 배열에서 0(False)이 아닌 값의 개수 확인

```
import numpy as np
print(np.count_nonzero(ebola.isnull()))
print(np.count_nonzero(ebola['Cases_Guinea'].isnull()))
```

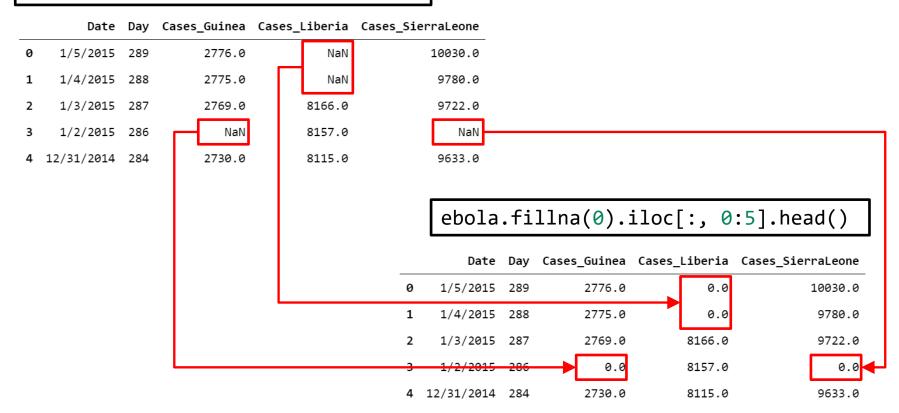
1214 29

- value\_counts() : 지정한 열의 빈도 확인

print(ebola['Cases\_Guinea'].value\_counts(dropna=False))

```
NaN 29
86.0 3
495.0 2
112.0 2
390.0 2
...
235.0 1
231.0 1
226.0 1
224.0 1
2776.0 1
Name: Cases Guinea, Length: 89, dtype: int64
```

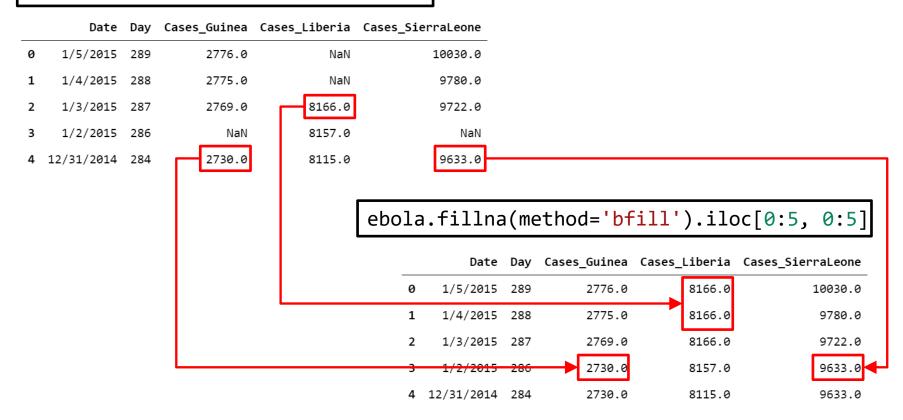
- 누락값 처리하기
  - 누락값 변경하기 fill\_na()
    - 누락값을 지정한 값으로 변경



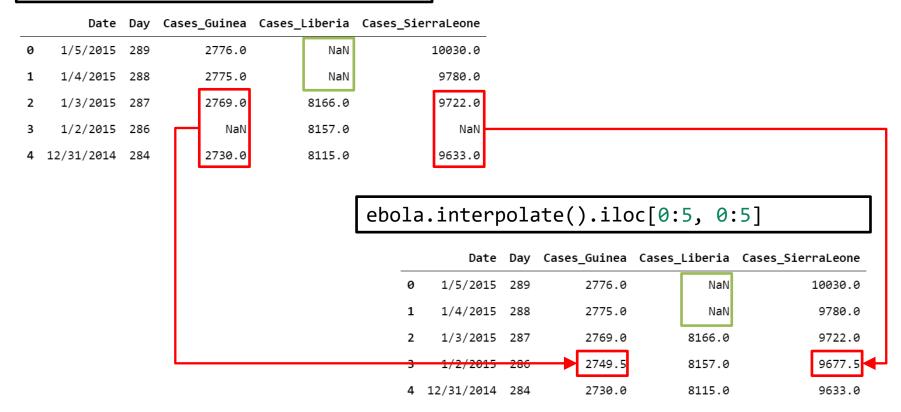
- 누락값 처리하기
  - 누락값 변경하기 fill\_na()
    - method='ffill' : 누락값이 나타나기 전 값으로 변경

	Date	Day	Cases_Guinea	Cases_Liberia	Cases_Sie	erraLeone				
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				
					ebola		•		•	OC[0:5, 0:5]
					0	1/5/2015		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
					<del>3</del>	1/2/2015	286	2769.0	8157.0	9722.0
					4	12/31/2014	284	2730.0	8115.0	9633.0

- 누락값 처리하기
  - 누락값 변경하기 fill\_na()
    - method='bfill' : 누락값이 나타난 이후 값으로 변경



- 누락값 처리하기
  - 누락값 변경하기 interpolate()
    - 누락값 양쪽에 있는 값을 이용하여 중간값을 구한 다음 변경



- 누락값 처리하기
  - 누락값 삭제하기 drop\_na()
    - 누락값이 포함된 데이터 삭제

데이터 상황에 따라 너무 많은 데이터가 삭제될 수도 있음