

レシビ48 等しくないインデックスの値を埋める (P127)

- 2つのSeriesを「加算演算子で足し合わせると」と、**片方にインデックスラベルがないと結果が常に欠損値になる**
- pandasには欠損値の値を埋める add メソッドがある
- このレシビでは、野球データセットで**インデックスが等しくない複数のSeriesを**
- add メソッドの fill_value パラメーターを使い、結果に欠損値がないように足し合わせる

(1) 3つの野球データセットを読み込み、playerIDをインデックスにセットする

```
In [4]: import pandas as pd

baseball_14 = pd.read_csv('baseball14.csv', index_col='playerID')
baseball_15 = pd.read_csv('baseball15.csv', index_col='playerID')
baseball_16 = pd.read_csv('baseball16.csv', index_col='playerID')

baseball_14.head()
```

```
Out[4]:
```

	yearID	stint	teamID	lgID	G	AB	R	H	2B	3B	...	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GDP
playerID																					
altuvjo01	2014	1	HOU	AL	158	660	85	225	47	3	...	59.0	56.0	9.0	36	53.0	7.0	5.0	1.0	5.0	20.0
cartech02	2014	1	HOU	AL	145	507	68	115	21	1	...	88.0	5.0	2.0	56	182.0	6.0	5.0	0.0	4.0	12.0
castrja01	2014	1	HOU	AL	126	465	43	103	21	2	...	56.0	1.0	0.0	34	151.0	1.0	9.0	1.0	3.0	11.0
corpoca01	2014	1	HOU	AL	55	170	22	40	6	0	...	19.0	0.0	0.0	14	37.0	0.0	3.0	1.0	2.0	3.0
dominna01	2014	1	HOU	AL	157	564	51	121	17	0	...	57.0	0.0	1.0	29	125.0	2.0	5.0	2.0	7.0	23.0

5 rows × 21 columns

```
In [5]: baseball_14.columns

Out[5]: Index(['yearID', 'stint', 'teamID', 'lgID', 'G', 'AB', 'R', 'H', '2B', '3B',
        'HR', 'RBI', 'SB', 'CS', 'BB', 'SO', 'IBB', 'HBP', 'SH', 'SF', 'GDP'],
        dtype='object')
```

(2)

df.index.difference(df2.index)

- インデックスの difference メソッドを使い、どのインデックスラベルがbaseball_14にあって
- baseball_15,16にないか、また逆のインデックスラベルは何か調べる

```
In [6]: baseball_14.index.difference(baseball_15.index)

Out[6]: Index(['corpoca01', 'dominna01', 'fowlede01', 'grossro01', 'guzmajeo01',
        'hoeslj01', 'krausma01', 'preslal01', 'singljo02'],
        dtype='object', name='playerID')
```

```
In [7]: baseball_14.index.difference(baseball_16.index)

Out[7]: Index(['cartech02', 'corpoca01', 'dominna01', 'fowlede01', 'grossro01',
        'guzmajeo01', 'hoeslj01', 'krausma01', 'preslal01', 'singljo02',
        'villajo01'],
        dtype='object', name='playerID')
```

(3) 多数の選手のインデックスが異なる。過去3年間の選手の全ヒット数を計算する。Hカラムにヒット数がある

```
In [23]: # cf
baseball_16['H']

Out[23]: playerID
altuvjo01    216
bregma101    53
castrja01    69
correca01    158
gattiev01    112
gomezca01    62
gonzama01    123
gourryu01    34
kempto01     26
marisja01    60
rasmusco01   76
reeda01     20
springe01    168
tuckepr01    22
vaibulu01    76
whitety01    54
Name: H, dtype: int64
```

```
In [10]: hits_14 = baseball_14['H']
hits_15 = baseball_15['H']
hits_16 = baseball_16['H']
```

```
In [15]: # cf
print(len(hits_14),len(hits_15),len(hits_16))

16 15 16
```

(4) まず加算演算子で2つのSeriesを足し合わす

```
In [20]: (hits_14 + hits_15)

Out[20]: playerID
altuvjo01    425.0
cartech02    193.0
castrja01    174.0
congeha01     NaN
corpoca01     NaN
correca01     NaN
dominna01     NaN
fowlede01     NaN
gattiev01     NaN
gomezca01     NaN
gonzama01    175.0
grossro01     NaN
guzmajeo01     NaN
hoeslj01     NaN
krausma01     NaN
lowrije01     NaN
marisja01    127.0
preslal01     NaN
rasmusco01     NaN
singljo02     NaN
springe01    175.0
tuckepr01     NaN
vaibulu01     NaN
villajo01     88.0
Name: H, dtype: float64
```

```
In [19]: # cf
print(len(hits_14 + hits_15))

24
```

(5)

baseball_14		baseball_15		baseball_16	
playerID		playerID		playerID	
altuvjo01	225	altuvjo01	200	altuvjo01	216
cartech02	115	cartech02	78	bregma101	53
castrja01	183	castrja01	71	castrja01	69
corpoca01	40	congeha01	46	correca01	158
dominna01	121	correca01	108	gattiev01	112
fowlede01	120	gattiev01	139	gomezca01	62
gonzama01	79	gomezca01	36	gonzama01	123
grossro01	84	gonzama01	96	gourryu01	34
guzmajeo01	31	lowrije01	51	kempto01	26
hoeslj01	21	marisja01	80	marisja01	60
krausma01	36	rasmusco01	103	rasmusco01	76
marisja01	47	springe01	107	reeda01	20
preslal01	62	tuckepr01	73	springe01	168
singljo02	52	vaibulu01	97	tuckepr01	22
springe01	68	villajo01	33	vaibulu01	76
villajo01	55			whitety01	54
Name: H, dtype: int64		Name: H, dtype: int64		Name: H, dtype: int64	

- 選手congeha01とcorpoca01correca01などが2015年にヒットを記録しているのに、結果は欠損値だ。

add、fill_value

- fill_value /パラメーターで欠損値がないようにして add メソッドを使う
- つまり、インデックスがない場合は0とみなして加算する、ってことか

```
In [24]: hits_14.add(hits_15, fill_value=0)

Out[24]: playerID
altuvjo01    425.0
cartech02    193.0
castrja01    174.0
congeha01    46.0
corpoca01    40.0
correca01    108.0
dominna01    121.0
fowlede01    120.0
gattiev01    139.0
gomezca01    36.0
gonzama01    175.0
grossro01    84.0
guzmajeo01    31.0
hoeslj01     21.0
krausma01    36.0
lowrije01    51.0
marisja01    127.0
preslal01    62.0
rasmusco01    76.0
singljo02     52.0
springe01    175.0
tuckepr01     22.0
vaibulu01     76.0
villajo01     88.0
Name: H, dtype: float64
```

(6) add メソッドをチェイニングして2016年のヒット数を加える

```
In [25]: hits_total = hits_14.add(hits_15, fill_value=0)\
        .add(hits_16, fill_value=0)

hits_total
```

```
Out[25]: playerID
altuvjo01    641.0
bregma101    53.0
cartech02    193.0
castrja01    243.0
congeha01    46.0
corpoca01    40.0
correca01    266.0
dominna01    121.0
fowlede01    120.0
gattiev01    251.0
gomezca01    98.0
gonzama01    298.0
gourryu01    34.0
grossro01    84.0
guzmajeo01    31.0
hoeslj01     21.0
kempto01     26.0
krausma01    36.0
lowrije01     51.0
marisja01    187.0
preslal01    62.0
rasmusco01    179.0
reeda01     20.0
singljo02     52.0
springe01    343.0
tuckepr01    95.0
vaibulu01    173.0
villajo01    88.0
whitety01    54.0
Name: H, dtype: float64
```

(7) 結果に欠損値がないかチェックする

```
In [26]: hits_total.hasnans

Out[26]: False
```

<解説>

- Seriesに欠損値のインデックスラベルが含まれていた場合、2つのSeriesをaddしたらそこは欠損値で埋められる

- ▼下記でやっていること（黄色箇所を+5で補填）

s	s1	s.add(s1, fill_value=5)	fill_valueなし
NaN	NaN	NaN	NaN
3.0	6.0	9.0	9.0
NaN	10.0	15.0	NaN
1.0		6.0	NaN

```
In [39]: import numpy as np
s = pd.Series(index=['a','b','c','d'], data=[np.nan, 3, np.nan, 1])
s1 = pd.Series(index=['a','b','c'], data=[np.nan, 6, 10])

print(s)
print()
print(s1)
```

```
a    NaN
b     3.0
c     NaN
d     1.0
dtype: float64

a    NaN
b     6.0
c    10.0
dtype: float64
```

```
In [40]: s.add(s1, fill_value=5)
```

```
Out[40]: a    NaN
b     9.0
c    15.0
d     6.0
dtype: float64
```

```
In [41]: s.add(s1)
```

```
Out[41]: a    NaN
b     9.0
c    NaN
d    NaN
dtype: float64
```

<補足>

- 上記はインデックスが1つだけのSeriesを足し合わせたときを例とした
- DataFrameではどうなるか？
- DataFrameを使つて high_light_null メソッドを使ってわかりやすくする
- 数値しないインデックスに欠損値を計算する仕様になっている

```
In [43]: # 2014より抜粋
df_14 = baseball_14[['G','AB','R','H']]
df_14.head()
```

```
Out[43]:
```

	G	AB	R	H
playerID				
altuvjo01	158	660	85	225
cartech02	145	507	68	115
castrja01	126	465	43	103
corpoca01	55	170	22	40
dominna01	157	564	51	121

```
In [44]: # 2015より抜粋
df_15 = baseball_15[['AB','R','H','HR']]
df_15.head()
```

```
Out[44]:
```

	AB	R	H	HR
playerID				
altuvjo01	638	86	200	15
cartech02	391	50	78	24
castrja01	337	38	71	11
congeha01	201	25	46	11
correca01	387	52	108	22

- 2つのDataFrameを足し合わせると行とカラムのラベルがアライメントして欠損値になる
- style属性を使って high_light_null メソッドを使ってわかりやすくする
- 両方のDataFrameに存在するplayerIDだけが非欠損値
- 欠損が列している箇所：
 - 行：共通しているカラム名はAB、R、H。それ以外は欠損となる
 - 行：いずれかに同じ行名がない箇所

```
In [45]: (df_14 + df_15).head(10).style.highlight_null('yellow')
```

```
Out[45]:
```

	AB	G	H	HR	R
playerID					
altuvjo01	1298.000000	nan	425.000000	nan	171.000000
cartech02	898.000000	nan	193.000000	nan	118.000000
castrja01	802.000000	nan	174.000000	nan	81.000000
congeha01	nan	nan	nan	nan	nan
corpoca01	nan	nan	nan	nan	nan
correca01	nan	nan	nan	nan	nan
dominna01	nan	nan	nan	nan	nan
fowlede01	nan	nan	nan	nan	nan
gattiev01	nan	nan	nan	nan	nan
gomezca01	nan	nan	nan	nan	nan

- add、fill_valueを使つても欠損値は残る箇所がある

```
In [48]: df_14.add(df_15, fill_value=0).head(10)\
        .style.highlight_null('yellow')
```

```
Out[48]:
```

	AB	G	H	HR	R
playerID					
altuvjo01	1298.000000	158.000000	425.000000	15.000000	171.000000
cartech02	898.000000	145.000000	193.000000	24.000000	118.000000
castrja01	802.000000	126.000000	174.000000	11.000000	81.000000
congeha01	201.000000	nan	46.000000	11.000000	25.000000
corpoca01	170.000000	55.000000	40.000000	nan	22.000000
correca01	387.000000	nan	108.000000	22.000000	52.000000
dominna01	564.000000	157.000000	121.000000	nan	51.000000
fowlede01	434.000000	116.000000	120.000000	nan	61.000000
gattiev01	566.000000	nan	139.000000	27.000000	66.000000
gomezca01	149.000000	nan	36.000000	4.000000	19.000000

- 黄色セル箇所は「インデックスに要素がない or カラムの要素がない」
- よつて値を埋められない！

```
In [65]: (df_14.index == 'congeha01').sum()

Out[65]: 0
```

```
In [72]: (df_14.columns == 'G').sum()

Out[72]: 1
```

```
In [66]: (df_15.index == 'congeha01').sum()

Out[66]: 1
```

```
In [71]: (df_15.columns == 'G').sum()

Out[71]: 0
```