

レシピ36 インデックス選択によるBooleanインデックス法の代用（P95）

- Boolean選択の特別な場合は、インデックスを活用して済ませることが出来る
 - インデックスによる選択は、直感的で読みやすい
-
- このレシピでは、collegeデータセットを使い、Booleanインデックス法とインデックス選択の両方で指定した州の大学を選び、それぞれの性能を比較する

(1) データセットを読み込み、Booleanインデックス法でテキサス州（TX）の全大学を選ぶ

```
In [1]: import os
os.getcwd()
```

Out[1]: '/Users/mbp441/Desktop/github/PYTHON/jupyter_notebook/cookbook_data'

```
In [3]: import pandas as pd
college = pd.read_csv('college.csv')
college[college['STABBR'] == 'TX'].head()
```

Out[3]:

	INSTNM	CITY	STABBR	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEON
3610	Abilene Christian University	Abilene	TX	0.0	0.0	0.0	1	530.0	545.0	
3611	Alvin Community College	Alvin	TX	0.0	0.0	0.0	0	NaN	NaN	
3612	Amarillo College	Amarillo	TX	0.0	0.0	0.0	0	NaN	NaN	
3613	Angelina College	Lufkin	TX	0.0	0.0	0.0	0	NaN	NaN	
3614	Angelo State University	San Angelo	TX	0.0	0.0	0.0	0	475.0	490.0	

5 rows × 27 columns

(2) 同じことをインデックス選択で行うためには、**STABBRカラムをインデックスにする必要がある**。その後、.locインデクサでラベルを使って選択。

```
In [6]: college2 = college.set_index('STABBR')
college2.loc['TX'].head()
```

Out[6]:

	INSTNM	CITY	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	L
STABBR										
TX	Abilene Christian University	Abilene	0.0	0.0	0.0	1	530.0	545.0	0.0	31
TX	Alvin Community College	Alvin	0.0	0.0	0.0	0	NaN	NaN	0.0	46
TX	Amarillo College	Amarillo	0.0	0.0	0.0	0	NaN	NaN	0.0	93

	INSTNM	CITY	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	L
STABBR										
TX	Angelina College	Lufkin	0.0	0.0	0.0	0	NaN	NaN	0.0	38
TX	Angelo State University	San Angelo	0.0	0.0	0.0	0	475.0	490.0	0.0	52

5 rows × 26 columns



(3) 1と2の速度を比較

- 単独でみた場合、インデックス法がBoolean選択よりも速い。

In [16]:

```
# 1.Boolean選択
%timeit college[college['STABBR'] == 'TX']
```

734 μs ± 17.7 μs per loop (mean ± std. dev. of 7 runs, 1000 loops each)

In [15]:

```
# 2.インデックス選択
%timeit college2.loc['TX']
```

366 μs ± 22.1 μs per loop (mean ± std. dev. of 7 runs, 1000 loops each)

(4) Booleanインデックス法>インデックス選択よりも2倍以上時間がかかっている。インデックスの手間を考慮する必要があるのもはかる、意外にかかる。

In [14]:

```
%timeit college2 = college.set_index('STABBR')
```

1.05 ms ± 190 μs per loop (mean ± std. dev. of 7 runs, 100 loops each)

(補足)

- Boolean選択でも、インデックス選択でも、複数の州を指定することができる
- ex.州「TX、CA、NY」を使う
- Boolean選択では、複数を指定する場合、isinを使う
 - college[college['STABBR'] == states]だとエラーになる

In [20]:

```
states = ['TX', 'CA', 'NY']

# 1.Boolean選択
college[college['STABBR'].isin(states)].head()

# 2.インデックス選択
college2.loc[states].head()
```

Out[20]:

	INSTNM	CITY	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	L
STABBR										
TX	Abilene Christian University	Abilene	0.0	0.0	0.0	1	530.0	545.0	0.0	38
TX	Alvin Community College	Alvin	0.0	0.0	0.0	0	NaN	NaN	0.0	46
TX	Amarillo College	Amarillo	0.0	0.0	0.0	0	NaN	NaN	0.0	93

	INSTNM	CITY	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	U
STABBR										
TX	Angelina College	Lufkin	0.0	0.0	0.0	0	NaN	NaN	0.0	38
	Angelo State University	San Angelo	0.0	0.0	0.0	0	475.0	490.0	0.0	52

5 rows × 26 columns



- pandasでは、インデックスに重複がないか、ソートされているか、によりインデックスの実装が異なる
- 詳細は次のレシピを参照

In []: