

レシピ58 groupbyオブジェクトの検討（P160）

- DataFrameでgroupbyメソッドを使うと、groupbyオブジェクトが戻る
- このレシピでは、groupbyオブジェクトそのものをメソッド呼び出しや、そのグループでイテレーションして調べる

(1) collegeデータセットを州と宗教系のカラムでグループ分けして変数に格納しておく。一応、型を調べてみる

In [82]:

```
grouped = college.groupby(['STABBR', 'RELAFFIL'])
type(grouped)
```

Out[82]: pandas.core.groupby.generic.DataFrameGroupBy

(2) dir関数で、利用可能な機能を統べてみる

In [83]:

```
print([attr for attr in dir(grouped) if not attr.startswith('_')])
```

```
['CITY', 'CURROPER', 'DISTANCEONLY', 'GRAD_DEBT_MDN_SUPP', 'HBCU', 'INSTNM', 'MD_EARN_WNE_P10', 'MENONLY', 'PCTFLOAN', 'PCTPELL', 'PPTUG_EF', 'RELAFFIL', 'SATMTMID', 'SATVRMID', 'STABBR', 'UG25ABV', 'UGDS', 'UGDS_2MOR', 'UGDS_AI
TAN', 'UGDS_ASIAN', 'UGDS_BLACK', 'UGDS_HISP', 'UGDS_NHPI', 'UGDS_NRA', 'UGDS_UNKN', 'UGDS_WHITE', 'WOMENONLY', 'agg', 'aggregate', 'all', 'any', 'apply', 'backfill', 'bfill', 'boxplot', 'corr', 'corrwith', 'count', 'cov',
'cumcount', 'cummax', 'cummin', 'cumprod', 'cumsum', 'describe', 'diff', 'dtypes', 'ewm', 'expanding', 'ffill', 'fillna', 'filter', 'first', 'get_group', 'groups', 'head', 'hist', 'idxmax', 'idxmin', 'indices', 'last', 'mad',
'max', 'mean', 'median', 'min', 'ndim', 'ngroup', 'ngroups', 'nth', 'nunique', 'ohlc', 'pad', 'pct_change', 'pipe', 'plot', 'prod', 'quantile', 'rank', 'resample', 'rolling', 'sample', 'sem', 'shift', 'size', 'skew', 'std', '
sum', 'tail', 'take', 'transform', 'tshift', 'var']
```

(3) ngroups 属性でグループの個数を確認

In [84]:

```
grouped.ngroups
```

Out[84]: 112

(4)

- 各グループで識別可能なラベルを調べるため
- 各グループが、そのグループのすべてに対応するインデックスラベルにマップした辞書をもつgroup属性を中を調べる
- grouped = college.groupby(['STABBR', 'RELAFFIL']) の「STABBR」と「RELAFFIL」のキーの組み合わせ

In [85]:

```
groups = list(grouped.groups.keys())
groups[:10]
```

Out[85]: [('AK', 0),
('AK', 1),
('AL', 0),
('AL', 1),
('AR', 0),
('AR', 1),
('AS', 0),
('AZ', 0),
('AZ', 1),
('CA', 0)]

(5)

- get_group メソッドに正確なグループラベルのタプルを渡してグループを取り出す。
- たとえば、AK州の宗教系大学（1）のすべてを取得する

In [86]:

```
grouped.get_group(('AK', 1)).head()
```

	INSTNM	CITY	STABBR	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	...	UGDS_2MOR	UGDS_NRA	UGDS_UNKN	PPTUG_EF	CURROPER	PCTPELL	PCTFLOAN	UG25ABV	MD_EARN_WNE_P10	GRAD_DEBT_MDN_SUPP
61	Alaska Bible College	Palmer	AK	0.0	0.0	0.0	1	NaN	NaN	0.0	...	0.0370	0.0	0.0000	0.1481	1	0.3571	0.2857	0.4286	NaN	PrivacySuppressed
64	Alaska Pacific University	Anchorage	AK	0.0	0.0	0.0	1	555.0	503.0	0.0	...	0.0945	0.0	0.0873	0.3745	1	0.3152	0.5297	0.4910	47000	23250
5417	Alaska Christian College	Soldotna	AK	0.0	0.0	0.0	1	NaN	NaN	0.0	...	0.0147	0.0	0.1324	0.0735	1	0.8868	0.6792	0.2264	NaN	PrivacySuppressed

3 rows × 27 columns

(6) 各グループがどうなっているかをざっと確認したい場合、groupbyオブジェクトでイテレートすればよい

```
: 1 for name, group in grouped:
2     print(name)
3     display(group.head(3))
```

- 結果は縦に長いので割愛

```
('AK', 0)
```

	INSTNM	CITY	STABBR	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	...	UGDS_2MOR	UGDS_NRA	UGDS_UNKN	PPTUG_EF	CURROPER	PCTPELL	PCTFLOAN	UG25ABV	MD_EARN_WNE_P10	GRAD_DEBT_MDN_SUPP
60	University of Alaska Anchorage	Anchorage	AK	0.0	0.0	0.0	0	NaN	NaN	0.0	...	0.0980	0.0181	0.0873	0.3745	1	0.3152	0.5297	0.4910	47000	23250
62	University of Alaska Fairbanks	Fairbanks	AK	0.0	0.0	0.0	0	NaN	NaN	0.0	...	0.0401	0.0110	0.0873	0.3745	1	0.3152	0.5297	0.4910	47000	23250
63	University of Alaska Southeast	Juneau	AK	0.0	0.0	0.0	0	NaN	NaN	0.0	...	0.0686	0.0049	0.0873	0.3745	1	0.3152	0.5297	0.4910	47000	23250