レシピ56 集約関数のカスタマイズ（P154）

- groupbyで使われる集約関数を自作したい場合がある
- ここでは、collegeデータセットを使い、州ごとの学部学生数の平均と標準偏差を求める
- それから、学生数の最大偏差値を州ごとに求める

（1）collegeデータセットを読み込み、州ごとに学部学生数の平均と標準偏差を求める

- STABBR：州の略語
- UGDS：学部への入学者数

In [68]:
```
college = pd.read_csv('college.csv')
display(college.head(2))
print(college.shape)
```

	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
0	Alabama A & M University	Normal	AL	1.0	0.0	0.0	0	424.0	420.0	0.0	...	0.0000	0.0059	0.0138	0.0656	1	0.7356	0.8284	0.1049	30300	33888
1	University of Alabama at Birmingham	Birmingham	AL	0.0	0.0	0.0	0	570.0	565.0	0.0	...	0.0368	0.0179	0.0100	0.2607	1	0.3460	0.5214	0.2422	39700	21941.5

2 rows × 27 columns

```
(7535, 27)
```

In [69]:
```
college.groupby('STABBR')['UGDS'].agg(['mean', 'std']).round(0).head()
#               グループ分けカラム↑              集約カラム              集約関数
```

Out[69]:

	mean	std
STABBR		
AK	2493.0	4052.0
AL	2790.0	4658.0
AR	1644.0	3143.0
AS	1276.0	NaN
AZ	4130.0	14894.0

- 上記のやり方は、母数が州全体の平均や標準偏差になっている
- 例）AK州の1大学あたりの平均入学者数と標準偏差

（2）

- （1）のような州全体の平均や標準偏差ではなく、大学の偏差値（平均から標準偏差で何個分離れているか）の最大値が欲しい
- 計算式 = {各大学の学部学生数 - 州平均の学生数} / 標準偏差
- これによりグループごとに学部学生数が標準化できる
- この点数の絶対値の最大値によって、平均から最も離れている大学を見つける
- Pandasにはこれらを求める関数がないため、自分でつくる

In [70]:
```
def max_deviation(s):
    std_score = (s - s.mean()) / s.std()
    return std_score.abs().max()
```

（3）関数定義後、それをaggメソッドに渡して集約する

- max_deviation(s)のパラメーターsは必須だが、s自体は読み出されない
- **カスタム関数はSeriesとして渡される**

In [71]:
```
college.groupby('STABBR')['UGDS'].agg(max_deviation).round(1).head()
#                                                  カスタムした集約関数
```

Out[71]:
```
STABBR
AK    2.6
AL    5.8
AR    6.3
AS    NaN
AZ    9.9
Name: UGDS, dtype: float64
```

（補足）

- カスタム関数は複数の集約カラムに適用することも可能

In [72]:
```
college.groupby('STABBR')[['UGDS', 'SATVRMID', 'SATMTMID']]\
       .agg(max_deviation).round(1).head()
```

Out[72]:

	UGDS	SATVRMID	SATMTMID
STABBR			
AK	2.6	NaN	NaN
AL	5.8	1.6	1.8
AR	6.3	2.2	2.3
AS	NaN	NaN	NaN
AZ	9.9	1.9	1.4

- カスタム関数をすでに組み込まれた関数と同時に使うこともできる。
- 集約関数とは異なり、クォートで囲む必要はない

In [73]:
```
college.groupby(['STABBR', 'RELAFFIL'])[['UGDS', 'SATVRMID']]\
       .agg([max_deviation, 'sum', 'count', 'size', 'mean', 'std']).round(1).head()
```

Out[73]:

		UGDS						SATVRMID					
		max_deviation	sum	count	size	mean	std	max_deviation	sum	count	size	mean	std
STABBR	**RELAFFIL**												
AK	**0**	2.1	24562.0	7	7	3508.9	4539.5	NaN	0.0	0	7	NaN	NaN
	1	1.1	370.0	3	3	123.3	132.9	NaN	555.0	1	3	555.0	NaN
AL	**0**	5.2	230663.0	71	72	3248.8	5102.4	1.6	6694.0	13	72	514.9	56.5
	1	2.4	17635.0	18	24	979.7	870.8	1.5	3984.0	8	24	498.0	53.0
AR	**0**	5.8	121971.0	68	68	1793.7	3401.6	1.9	4330.0	9	68	481.1	37.9

＜グループ分けカラムとは？＞

- 'STABBR'、'RELAFFIL'でグループ分けされている状態、**このイメージがつきにくい**
- 'STABBR'、'RELAFFIL'でグループ分けされている状態とは？
- **「グループ分けカラム」でまとめられた行単位で計算が行われる**
- データで例えるならば（下表がわかりやすい）
- 「AK-0」に属する複数の行における「集約カラム」の個々のセル値や件数 が**このグループの「計算ベース」になり「分母」になる**
- 例えば、AK-0の行数（count）、AK-0のUGDSの数値の合計値（sum）、など

＜集約された表の見方・解釈方法＞

上記の１行目（AK-0）をみた場合（下の表も参照）

- ①グループ分けカラム：'STABBR'=「AK」かつ'RELAFFIL'=「0」 でまとめる（**その行しかない**と考えるとわかりやすい、下表の色あり行）
- ②集約カラム：'UGDS'の列（件数や数値）を使って計算を行う
- ③集約関数：
  - その `sum` =24562.0
  - その `count,size` =7 ※**件数を指す**
  - その `mean` =3508.9
  - その `std` =4539.5、を意味する
- もしカスタム関数を使った場合は、集約カラム（'UGDS')の対象セルをSeriesとして1つずつ関数に渡される（？）

	INSTNM	CITY	STABBR	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	UGDS
1											
62	University of Alaska Anchorage	Anchorage	AK	0.0	0.0	0.0	0			0.0	12865.0
63	Alaska Bible College	Palmer	AK	0.0	0.0	0.0	1			0.0	27.0
64	University of Alaska Fairbanks	Fairbanks	AK	0.0	0.0	0.0	0			0.0	5536.0
65	University of Alaska Southeast	Juneau	AK	0.0	0.0	0.0	0			0.0	1428.0
66	Alaska Pacific University	Anchorage	AK	0.0	0.0	0.0	1	555.0	503.0	0.0	275.0
67	AVTEC-Alaska's Institute of Technology	Seward	AK	0.0	0.0	0.0	0			0.0	889.0
68	Charter College-Anchorage	Anchorage	AK	0.0	0.0	0.0	0			0.0	3256.0
69	Alaska Career College	Anchorage	AK	0.0	0.0	0.0	0			0.0	479.0
5173	Ilisagvik College	Barrow	AK	0.0	0.0	0.0	0			0.0	109.0
5419	Alaska Christian College	Soldotna	AK	0.0	0.0	0.0	1			0.0	68.0
7537											

SUM 24,562.0　AVERAGE 3,508.9　MIN 109.0　MAX 12,865.0　COUNTA 7　STDEV 4,539.5
