レシピ59 マイノリティが多数派の州をフィルタリング（P160）

- ４章では、行に対してTrueかFalseの印を付けて、Falseの行をフィルタリングして取り除く等おこなった
- 同様に、グループ分けしたデータに「TrueかFalse」の印を付けて、いずれかのグループ分けをフィルタリングできる
- そのためには、まずgroupbyメソッドでグループ分けをし、それからfilterメソッドを適用する

`filter`メソッドを使うため、TrueかFalseのいずれかを残すための関数をつくる

- 注）このfilterメソッドは、2章のレシピ12であったDataFrameのfilterメソッドとは全く違うので注意

- このレシピでは、collegeデータセットを使い、学部学生で非白人の方が白人より多い州をすべて求める

（1）州ごとにグループ分けし、グループの総数を表示。

In [87]:
```
college = pd.read_csv('college.csv', index_col='INSTNM')
grouped = college.groupby('STABBR')
grouped.ngroups
```

Out[87]: 59

In [88]:
```
college['STABBR'].nunique() # と同じで問題なし
```

Out[88]: 59

（2）

- 変数groupedにはfilterメソッドがあり、どのグループを保持するか決定するユーザー定義関数を渡すことができる
- 白人以外のマイノリティ率を返す関数を作成する（しきい値の設定も可能）

- 'UGDS_WHITE'：白人率、threshold：しきい値、pct:%

In [89]:
```
def check_minority(df, threshold):
    # 白人以外（マイノリティ）の率
    minority_pct = 1 - df['UGDS_WHITE']
    total_minority = (df['UGDS'] * minority_pct).sum()
    total_udgs = df['UGDS'].sum()
    total_minority_pct = total_minority / total_udgs
    # 戻り値は Boolean にする
    return total_minority_pct > threshold
```

（3）しきい値を50%にして、マイノリティが多数派の州をすべて求める

- つまり、check_minority関数で「True」となった行を`filter`している
- まず、州ごとにグループ化しているため、**分母は州単位で計算**されていることを理解する
- 各州に属する各行のUDGS_WHITEのセル値を使って、マイノリティの%合計を出している
- 例えば、**州AZであれば、AZだけの行があると考える**
- それらAZだけのデータを使い、total_minority　/ total_udgs を集計しているということ
- 結果、AZ州はマイノリティ率50%以上としてTrueとなり、**そこに属する各行にTrueが付与され、フィルタされた結果**がcollege_filteredに格納される
- これは、州の全大学生のうち「マイノリティが50%以上を占める州」にある大学のリスト、を出しているということか（？）
- 分母は州の大学生数、分子は州の大学生のマイノリティ数

In [90]:
```
college_filtered = grouped.filter(check_minority, threshold=.5)
college_filtered.head()
```

Out[90]:

INSTNM	CITY	STABBR	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATMTMID	DISTANCEONLY	UGDS	...	UGDS_2MOR	UGDS_NRA	UGDS_UNKN	PPTUG_EF	CURROPER	PCTPELL	PCTFLOAN	UG25ABV	MD_EARN_WNE_P10	GRAD_DEBT_MDN_SUPP
Everest College-Phoenix	Phoenix	AZ	0.0	0.0	0.0	1	NaN	NaN	0.0	4102.0	...	0.0373	0.0	0.1026	0.4749	0	0.8291	0.7151	0.6700	28600	9500
Collins College	Phoenix	AZ	0.0	0.0	0.0	0	NaN	NaN	0.0	83.0	...	0.0241	0.0	0.3855	0.3373	0	0.7205	0.8228	0.4764	25700	47000
Empire Beauty School-Paradise Valley	Phoenix	AZ	0.0	0.0	0.0	1	NaN	NaN	0.0	25.0	...	0.0400	0.0	0.0000	0.1600	0	0.6349	0.5873	0.4651	17800	9588
Empire Beauty School-Tucson	Tucson	AZ	0.0	0.0	0.0	0	NaN	NaN	0.0	126.0	...	0.0000	0.0	0.0079	0.2222	1	0.7962	0.6615	0.4229	18200	9833
Thunderbird School of Global Management	Glendale	AZ	0.0	0.0	0.0	0	NaN	NaN	0.0	1.0	...	0.0000	0.0	0.0000	1.0000	0	0.0000	0.0000	0.0000	118900	PrivacySuppressed

5 rows × 26 columns

（4）フィルタ結果を比較する

In [91]:
```
college.shape
```

Out[91]: (7535, 26)

In [92]:
```
college_filtered.shape
```

Out[92]: (3028, 26)

In [93]:
```
college_filtered['STABBR'].nunique()
```

Out[93]: 20

（補足）

- 処理がきちんとなされているかをチェックする
- しきいち値を変えて、shpaeや州の数がどうなるか調べる

In [94]:
```
college_filtered_20 = grouped.filter(check_minority, threshold=.2)
college_filtered_20.shape
```

Out[94]: (7461, 26)

In [95]:
```
college_filtered_20['STABBR'].nunique()
```

Out[95]: 57

In [96]:
```
college_filtered_70 = grouped.filter(check_minority, threshold=.7)
college_filtered_70.shape
```

Out[96]: (957, 26)

In [97]:
```
college_filtered_70['STABBR'].nunique()
```

Out[97]: 10
