

レシピ42 whereメソッドによるSeriesの保持（P110）

- Booleanインデックス法により、基準に合致しない行を取り除くことのできる
- whereメソッドを使って同じことができる
- このレシピでは whereメソッドにBoolean条件を渡して、movieデータセットのactor1のFacebookの「いいね！」の数の最小値と最大値の切り下げ値と切り上げ値を決める

(1) movieデータセットを読み込み、facebook_likesカラムの欠損値以外をすべて選ぶ

- ここで欠損値を削除したのは、後続でwhereメソッドが欠損値を条件に合った値で置き換えてしまうことを防止するため

In [10]:

```
import pathlib
import pandas as pd
import os

p = pathlib.Path(os.getcwd())

movie = pd.read_csv(p / 'movie.csv', index_col='movie_title')
movie.head()
```

Out[10]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes
movie_title						
Avatar	Color	James Cameron	723.0	178.0	0.0	855.0
Pirates of the Caribbean: At World's End	Color	Gore Verbinski	302.0	169.0	563.0	1000.0
Spectre	Color	Sam Mendes	602.0	148.0	0.0	161.0
The Dark Knight Rises	Color	Christopher Nolan	813.0	164.0	22000.0	23000.0
Star Wars: Episode VII - The Force Awakens	NaN	Doug Walker	NaN	NaN	131.0	NaN

5 rows × 27 columns

In [21]:

```
fb_likes = movie['actor_1_facebook_likes'].dropna()
fb_likes.head()
```

Out[21]:

movie_title	
Avatar	1000.0
Pirates of the Caribbean: At World's End	40000.0
Spectre	11000.0
The Dark Knight Rises	27000.0
Star Wars: Episode VII - The Force Awakens	131.0
Name: actor_1_facebook_likes, dtype: float64	

(2) 分布をみる

In [22]:

```
fb_likes.describe(percentiles=[.1, .25, .5, .75, .9]).astype(int)
```

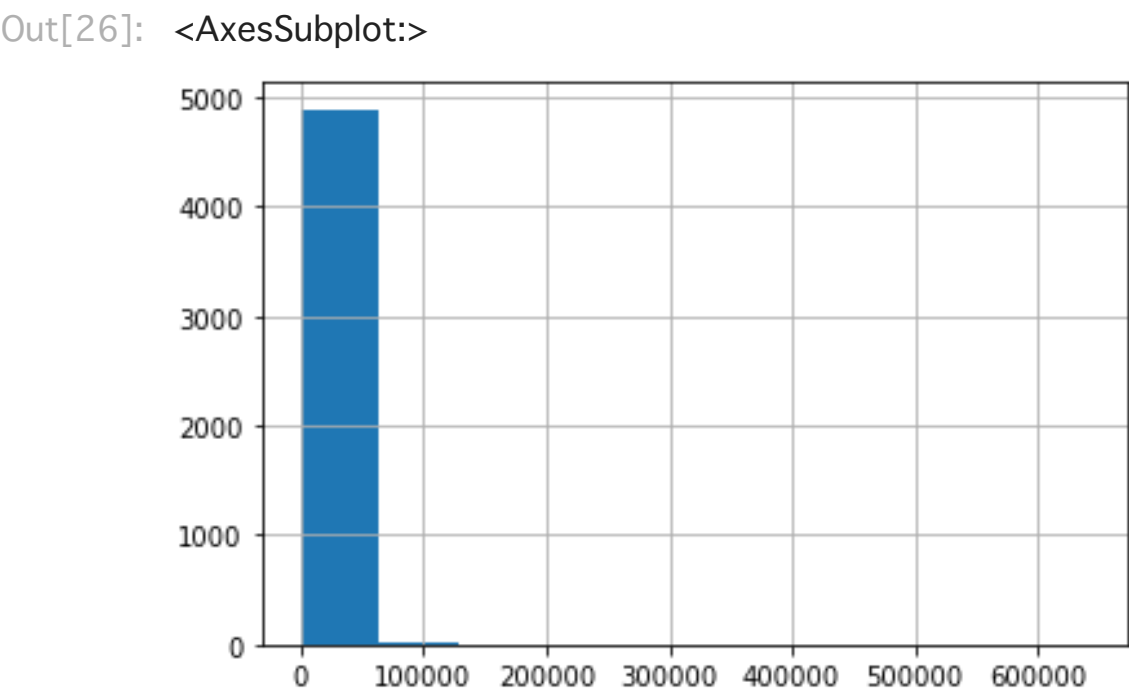
```
Out[22]: count    4909
         mean     6494
         std    15106
         min       0
         10%     240
         25%     607
         50%     982
         75%    11000
         90%    18000
         max   640000
         Name: actor_1_facebook_likes, dtype: int64
```

```
In [25]: # cf.デフォルト
         fb_likes.describe()
```

```
Out[25]: count    4909.000000
         mean    6494.488491
         std    15106.986884
         min      0.000000
         25%     607.000000
         50%     982.000000
         75%    11000.000000
         max   640000.000000
         Name: actor_1_facebook_likes, dtype: float64
```

(3) さらにグラフで見る

```
In [26]: fb_likes.hist()
```



(4)

- 上記はわかりにくい。また (2) の要約統計量から右に大きく歪んでいる。中央値より1桁多い観察値が多いことがわかる <---なぜ？
- 「いいね！」の個数が2万より少ないかどうかをチェックする基準を作る

```
In [27]: criteria_high = fb_likes < 20000
         criteria_high.mean().round(2) # <- mean:条件を満たすもの(True) ÷ 全体数
```

```
Out[27]: 0.91
```

```
In [29]: # cf.
         criteria_high
```

```
Out[29]: movie_title
```

```
Avatar                                True
Pirates of the Caribbean: At World's End    False
Spectre                                True
The Dark Knight Rises                    False
Star Wars: Episode VII - The Force Awakens    True
...
Signed Sealed Delivered                True
The Following                        True
A Plague So Pleasant                  True
Shanghai Calling                     True
My Date with Drew                     True
Name: actor_1_facebook_likes, Length: 4909, dtype: bool
```

```
In [62]: # cf.Trueの数 ÷ 全体数
int(criteria_high.sum()) / len(criteria_high)
```

Out[62]: 0.9085353432470972

(5)

- 約91%の映画でactor1が2万未満の「いいね！」をもらっている
- Booleanを渡せる where メソッドを使う

where

- デフォルトの挙動は同じサイズのSeriesを返す
- 条件に合致すれば「True」、合致しなければ「False」を返す
- そして、****Falseの箇所を書き換える****（Trueではない）
- つまり、**書き換えたい、置換したい箇所が「False」となるような条件をつくること**
- 引数なしの場合は、**NaNで置き換える**
- `where(criteria, other=*)` で対象を*に置換する
- 便利

```
In [70]: # cf.デフォルト
fb_likes.head()
```

Out[70]: movie_title
Avatar 1000.0
Pirates of the Caribbean: At World's End 40000.0
Spectre 11000.0
The Dark Knight Rises 27000.0
Star Wars: Episode VII - The Force Awakens 131.0
Name: actor_1_facebook_likes, dtype: float64

```
In [67]: # 「いいね!」が2万より少なければ True、多ければ False
fb_likes.where(criteria_high).head()
```

Out[67]: movie_title
Avatar 1000.0
Pirates of the Caribbean: At World's End NaN
Spectre 11000.0
The Dark Knight Rises NaN
Star Wars: Episode VII - The Force Awakens 131.0
Name: actor_1_facebook_likes, dtype: float64

(6) whereメソッドの第2パラメーターotherで置換する値を制御できる。欠損値をすべて2万に修正する

```
In [69]: fb_likes.where(criteria_high, other=20000).head()
```

Out[69]: movie_title
Avatar 1000.0
Pirates of the Caribbean: At World's End 20000.0
Spectre 11000.0
The Dark Knight Rises 20000.0
Star Wars: Episode VII - The Force Awakens 131.0
Name: actor_1_facebook_likes, dtype: float64

(7)

- 同様に「いいね！」の最小個数の下限設定基準を設けることができる
- 別のwhereメソッドをチェイニングして基準に合致しない値を300に変える
- ****メソッドチェインで複数条件を適用できる****

In [71]: criteria_low = fb_likes > 300
fb_likes_cap = fb_likes.where(criteria_high, other=20000) \
 .where(criteria_low, 300)
fb_likes_cap.head()

Out[71]: movie_title
Avatar 1000.0
Pirates of the Caribbean: At World's End 20000.0
Spectre 11000.0
The Dark Knight Rises 20000.0
Star Wars: Episode VII - The Force Awakens 300.0
Name: actor_1_facebook_likes, dtype: float64

(8) 元のSeriesの長さを変更したSeriesの長さは同じ

In [72]: len(fb_likes), len(fb_likes_cap)

Out[72]: (4909, 4909)

(9) 修正Seriesのヒストグラムを作成。データの範囲を絞ったので見やすくなった

In [73]: fb_likes_cap.hist()

Out[73]: <AxesSubplot:>

