Chpapter.5 Booleanインデックス法

レシピ33 Boolean統計量の計算（P88）

- Boolean Seriesを学ぶときは基本的な要約統計量を計算するとよい
- 各値が0か1になるので数値に関するSeriesメソッドすべてがBooleanに使える

- このレシピではデータのカラムに対して条件を当てはめ、その結果の要約統計量を計算する

（1）movieデータセットを読み込み、インデックスを映画の題名に

In [4]:
```
import pandas as pd
import pathlib
movie = pd.read_csv('movie.csv', index_col='movie_title')
movie.head(2)
```

Out[4]:

movie_title	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes
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

2 rows × 27 columns

（2）映画の上演時間「duration」が2h以上かを調べる

In [6]:
```
movie_2_hours = movie['duration'] > 120
movie_2_hours.head()
```

Out[6]:
```
movie_title
Avatar                                         True
Pirates of the Caribbean: At World's End       True
Spectre                                        True
The Dark Knight Rises                          True
Star Wars: Episode VII - The Force Awakens    False
Name: duration, dtype: bool
```

（3）このSeriesから2h以上の映画の本数がわかる

In [7]:
```
movie_2_hours.sum()
```

Out[7]: 1039

（4）2h超の映画の%を求めるには、 `mean` メソッドを使う

In [8]:
```
movie_2_hours.mean()
```

Out[8]: 0.2113506916192026

（5）**注意：**（4）の結果は間違っている。カラムdurationにはNaNが含まれている。

- Boolean条件は、欠損値に対し「False」を返すため計算に含まれてしまう
- だからdropして計算しないと結果が違ってくる

In [13]:
```
movie['duration'].isna().sum()
```

Out[13]: 15

In [14]:
```
movie['duration'].dropna().gt(120).mean()
```

Out[14]: 0.21199755152009794

（6） `describe` で、Boolean Seriesの要約統計量を出力する

In [15]:
```
movie_2_hours.describe()
```

Out[15]:
```
count     4916
unique       2
top      False
freq      3877
Name: duration, dtype: object
```

（補足）割愛