

レシピ35 Booleanインデックス法によるフィルタリング (P93)

- このレシピでは、映画の別々のデータセットにそれぞれ独立のBoolean基準を2つ作る
- データセット1：①imdb_score>8, ②content_rating==PG13, ③title_yearが2000年より前または2009より後
- データセット2：①imdb_score<5, ②content_rating==R, ③title_yearが2000年～2010年

(1) 1つめの基準を作成

```
In [3]: import pandas as pd
movie = pd.read_csv('movie.csv', index_col='movie_title')

crit_a1 = movie.imdb_score > 8
crit_a2 = movie.content_rating == 'PG-13'
crit_a3 = (movie.title_year < 2000) | (movie.title_year > 2009)

final_crit_a = crit_a1 & crit_a2 & crit_a3
```

(2) 2つめの基準を作成

```
In [6]: crit_b1 = movie.imdb_score > 5
crit_b2 = movie.content_rating == 'R'
crit_b3 = (movie.title_year >= 2000) | (movie.title_year <= 2010)
final_crit_b = crit_b1 & crit_b2 & crit_b3
```

(3) 2つの基準集合の和をとる。これによりどちらかの基準に適合する映画のBoolean Seriesが得られる

```
In [8]: final_crit_all = final_crit_a | final_crit_b
final_crit_all.head()
```

Out[8]: movie_title
Avatar False
Pirates of the Caribbean: At World's End False
Spectre False
The Dark Knight Rises True
Star Wars: Episode VII - The Force Awakens False
dtype: bool

(4) Boolean Seriesができれば、データをフィルタリングするためにインデックス演算子に渡す

```
In [28]: movie[final_crit_all].head()
```

Out[28]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes
movie_title						
The Dark Knight Rises	Color	Christopher Nolan	813.0	164.0	22000.0	23000.0
The Avengers	Color	Joss Whedon	703.0	173.0	0.0	19000.0
Captain America: Civil War	Color	Anthony Russo	516.0	147.0	94.0	11000.0

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes
movie_title						
Terminator 3: Rise of the Machines	Color	Jonathan Mostow	280.0	109.0	84.0	191.0
Guardians of the Galaxy	Color	James Gunn	653.0	121.0	571.0	3000.0

5 rows × 27 columns

```
In [32]: # 列を整理すると
movie[final_crit_all][['imdb_score', 'content_rating', 'title_year']]
```

	imdb_score	content_rating	title_year
movie_title			
The Dark Knight Rises	8.5	PG-13	2012.0
The Avengers	8.1	PG-13	2012.0
Captain America: Civil War	8.2	PG-13	2016.0
Terminator 3: Rise of the Machines	6.4	R	2003.0
Guardians of the Galaxy	8.1	PG-13	2014.0
...	...	...	...
Rampage	6.3	R	2009.0
Slacker	7.1	R	1991.0
The Puffy Chair	6.6	R	2005.0
Clean	6.9	R	2004.0
El Mariachi	6.9	R	1992.0

1933 rows × 3 columns

(5) フィルタリングが正しく行われたか？確認する（→4と同じ結果）

- **df.loc[行、列]の形でまとめている**
- **Booleanインデックス法は、.ilocインデクサではうまくいかない**（<-index番号を使うから。列名じゃないからでは？）
- Boolean Seriesを渡すと例外が起きる。しかし、Boolean ndarayを渡せば、うまくいく

```
In [25]: # cf.これはBooleanのSeriesになっている
print(type(final_crit_all))
final_crit_all
```

<class 'pandas.core.series.Series'>

movie_title		
Avatar	False	
Pirates of the Caribbean: At World's End	False	
Spectre	False	
The Dark Knight Rises	True	
Star Wars: Episode VII - The Force Awakens	False	
...		
Signed Sealed Delivered	False	
The Following	False	
A Plague So Pleasant	False	
Shanghai Calling	False	

My Date with DrewFalse

Length: 4916, dtype: bool

```
In [15]: cols = ['imdb_score', 'content_rating', 'title_year']

movie_filtered = movie.loc[final_crit_all, cols]
movie_filtered
```

Out[15]:

	imdb_score	content_rating	title_year
movie_title			
The Dark Knight Rises	8.5	PG-13	2012.0
The Avengers	8.1	PG-13	2012.0
Captain America: Civil War	8.2	PG-13	2016.0
Terminator 3: Rise of the Machines	6.4	R	2003.0
Guardians of the Galaxy	8.1	PG-13	2014.0
...	...	...	...
Rampage	6.3	R	2009.0
Slacker	7.1	R	1991.0
The Puffy Chair	6.6	R	2005.0
Clean	6.9	R	2004.0
El Mariachi	6.9	R	1992.0

1933 rows × 3 columns

