

# 빅데이터 실습

7주차 2차시

데이터 실전 분석 - 영화 평점 분석 [1]

## 데이터 실전 분석 영화 평점 분석 [ 1 ]



### 학습개요

- 1/ 영화 평점 데이터 소개
- 2/ 영화 평점 분석 실습



01

# 영화 평점 데이터 소개





영화 평점 데이터 소개

## ‘영화 평점 데이터’

사용자  
데이터

영화를  
리뷰한  
데이터

영화  
데이터



## 영화 평점 데이터 소개

## ➤ 사용자 데이터 (약 6,000명)

	사용자아이디	성별	연령	직업	지역
0	1	F	1	10	48067
1	2	M	56	16	70072
2	3	M	25	15	55117
3	4	M	45	7	02460
4	5	M	25	20	55455





## 영화 평점 데이터 소개

## ➤ 영화 평점 데이터 (약 1,000,000건)

	사용자아이디	영화아이디	평점	타임스탬프
0	1	1193	5	978300760
1	1	661	3	978302109
2	1	914	3	978301968
3	1	3408	4	978300275
4	1	2355	5	978824291

- ✓ 어떤 영화가 평점을 높게 받았는지 분석
- ✓ 특정 영화의 평점을 알기 어려움



영화 평점 데이터 소개

## ➤ 영화 데이터 (약 3,880건)

	영화아이디	영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
1	2	Jumanji (1995)	Adventure Children's Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama
4	5	Father of the Bride Part II (1995)	Comedy



02

## 영화 평점 분석 실습







## 영화 평점 분석 실습

# ➤ 평점의 평균과 개수를 활용하여 보고 싶은 영화 찾기

영화제목	평균	개수
American Beauty (1999)	4.317386	3428
Godfather, The (1972)	4.524966	2223
Matrix, The (1999)	4.315830	2590
Princess Bride, The (1987)	4.303710	2318
Raiders of the Lost Ark (1981)	4.477725	2514
Saving Private Ryan (1998)	4.337354	2653
Schindler's List (1993)	4.510417	2304
Shawshank Redemption, The (1994)	4.554558	2227
Silence of the Lambs, The (1991)	4.351823	2578
Sixth Sense, The (1999)	4.406263	2459
Star Wars: Episode IV - A New Hope (1977)	4.453694	2991

✓ 평점이 높고 평점의 개수도 많은 영화



## 영화 평점 분석 실습

## ➤ 여자들이 좋아하는 영화 찾기

성별	mean		count	
	F	M	F	M
영화제목				
American Beauty (1999)	4.238901	4.347301	946.0	2482.0
Being John Malkovich (1999)	4.159930	4.113636	569.0	1672.0
Braveheart (1995)	4.016484	4.297839	546.0	1897.0
Casablanca (1942)	4.300990	4.461340	505.0	1164.0
E.T. the Extra-Terrestrial (1982)	4.089850	3.920264	601.0	1668.0

✓ 남자와 여자의 평점 계산





## 영화 평점 분석 실습

# ➤ 여자들이 좋아하는 영화들의 장르 분석하기

Drama	12.0
Action	7.0
War	6.0
Comedy	6.0
Thriller	5.0
Adventure	5.0
Sci-Fi	4.0
Romance	4.0
Crime	3.0
Children's	3.0
Fantasy	2.0
Mystery	1.0
Musical	1.0
Film-Noir	1.0
Animation	1.0



## 영화 평점 분석 실습

# 남녀의 호불호가 크게 갈리는 영화 찾기

성별	영화제목	mean		count		평점차이
		F	M	F	M	
	Dirty Dancing (1987)	3.790378	2.959596	291.0	396.0	0.830782
	Good, The Bad and The Ugly, The (1966)	3.494949	4.221300	99.0	723.0	0.726351
	Dumb & Dumber (1994)	2.697987	3.336595	149.0	511.0	0.638608
	Evil Dead II (Dead By Dawn) (1987)	3.297297	3.909283	74.0	474.0	0.611985
	Grease (1978)	3.975265	3.367041	283.0	534.0	0.608224
	Caddyshack (1980)	3.396135	3.969737	207.0	760.0	0.573602
	Animal House (1978)	3.628906	4.167192	256.0	951.0	0.538286
	Exorcist, The (1973)	3.537634	4.067239	186.0	699.0	0.529605
	Rocky Horror Picture Show, The (1975)	3.673016	3.160131	315.0	918.0	0.512885
	Big Trouble in Little China (1986)	2.987952	3.485030	83.0	501.0	0.497078

✓ 남자와 여자의 평점이 차이가 큰 영화





## 영화 평점 분석 실습

## ➤ 연령대별 영화 평점 구하기

연령대	10대 미만	10대	20대	30대	40대	50대 이상
영화제목						
\$1,000,000 Duck (1971)	-	3	3.09091	3.13333	2	2.75
'Night Mother (1986)	2	4.66667	3.42308	2.90476	3.83333	3.75
'Til There Was You (1997)	3.5	2.5	2.66667	2.9	2.33333	2.6
'burbs, The (1989)	4.5	3.24444	2.65217	2.81818	2.54545	3.1
...And Justice for All (1979)	3	3.42857	3.72414	3.65714	4.1	3.67442

- ✓ 연령 정보를 통해 연령대 계산
- ✓ 각 연령대의 평점 계산

## 영화 평점 분석 실습

```
In [1]: import pandas as pd
        from pandas import Series, DataFrame
        import numpy as np
```

### 1. 영화 평점 데이터 적재 및 전처리

```
In [ ]: # 사용자 데이터 읽어오기
        users = pd.read_csv('data/movielens/users.dat', sep = '::', engine = 'python',
                             names = ['사용자아이디', '성별', '연령', '직업', '지역'])
        users.head()
```

```
In [ ]: # 평점 데이터 읽어오기
        ratings = pd.read_csv('data/movielens/ratings.dat', sep = '::', engine = 'python',
                               names = ['사용자아이디', '영화아이디', '평점', '타임스탬프'])
        ratings.head()
```

```
In [ ]: # 영화데이터 읽어오기
```



# 실습 준비하기

## ◎ 사용자 데이터(users), 평점 데이터(ratings), 영화 데이터(movies) 읽어오기

### ✓ 영화 데이터 인자값 살펴보기

In [3]: # 평점 데이터 읽어오기

```
ratings = pd.read_csv('data/movielens/ratings.dat', sep = '::', engine = 'python',  
                      names = ['사용자아이디', '영화아이디', '평점', '타임스탬프'])
```

Out [3]:

	사용자아이디	영화아이디	평점	타임스탬프
0	1	1193	5	978300760
1	1	661	3	978302109
2	1	914	3	978301968
3	1	3408	4	978300275
4	1	2355	5	978824291

In [4]: # 영화데이터 읽어오기

```
movies = pd.read_csv('data/movielens/movies.dat', sep = '::', engine = 'python',  
                     names = ['영화아이디', '영화제목', '장르'], encoding = 'latin-1')  
movies.head()
```

Out [4]:

	영화아이디	영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
1	2	Jumanji (1995)	Adventure Children's Fantasy

# 실습 준비하기

## ◎ 사용자 데이터(users), 평점 데이터(ratings), 영화 데이터(movies) 읽어오기

```
3      1      3408      4  978300275
4      1      2355      5  978824291
```

```
In [4]: # 영화데이터 읽어오기
movies = pd.read_csv('data/movielens/movies.dat', sep = '::', engine = 'python',
                    names = ['영화아이디', '영화제목', '장르'], encoding = 'latin-1')
movies.head()
```

Out[4]:

	영화아이디	영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
1	2	Jumanji (1995)	Adventure Children's Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama
4	5	Father of the Bride Part II (1995)	Comedy

```
In [ ]: #3개의 데이터프레임을 하나로 합치기
```

## 2. 보고 싶은 영화 찾기

여하튼이 편저 편규은 구하여 사라드에게 이전반느 (편저이 노으) 영화 찾기



# 실습 준비하기

## ◎ 3개의 데이터프레임을 하나로 합치기

- ✓ ratings 데이터를 기반으로 영화아이디, 영화제목을 같이 봐야 할 경우
- ✓ 사용자아이디를 users의 성별, 나이 정보를 같이 평점을 분석할 경우

```
3      1      3408      4      978300275
```

```
names = ['영화아이디', '영화제목', '장르'], encoding = 'latin-1')  
movies.head()
```

Out[4]:

	영화아이디	영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
1	2	Jumanji (1995)	Adventure Children's Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama
4	5	Father of the Bride Part II (1995)	Comedy

```
In [ ]: #3개의 데이터프레임을 하나로 합치기
```

## 2. 보고 싶은 영화 찾기

# 실습 준비하기

## ◎ 3개의 데이터프레임을 하나로 합치기

- ✓ merge : 사용자 아이디를 기준으로 합침
- ✓ data에 users, ratings 2개의 dataframe을 합침

```
, sep = '::', engine = 'python',  
제목', '장르'], encoding = 'latin-1')
```

Out[4]:

	영화아이디	영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
1	2	Jumanji (1995)	Adventure Children's Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama
4	5	Father of the Bride Part II (1995)	Comedy

```
In [5]: #3개의 데이터프레임을 하나로 합치기  
data = pd.merge(users, ratings)  
data
```

Out[5]:

	사용자아이디	성별	연령	직업	지역	영화아이디	평점	타임스탬프
0	1	F	1	10	48067	1193	5	978300760
1	1	F	1	10	48067	661	3	978302109



# 실습 준비하기

## ◎ 3개의 데이터프레임을 하나로 합치기

- ✓ data에 추가로 movies의 data를 합침
- ✓ on 인자는 사용하지 않아도 영화아이디 공통 columns이 있으므로 상관 없음

	영화아이디	영화제목	장르
3	4	Waiting to Exhale (1995)	Comedy Drama
4	5	Father of the Bride Part II (1995)	Comedy

```
In [5]: #3개의 데이터프레임을 하나로 합치기
data = pd.merge(users, ratings)
data
```

...

```
In [6]: data = pd.merge(data, movies)
```

```
In [7]: data
```

Out[7]:

	사용자아이디	영화아이디	영화제목	장르
--	--------	-------	------	----

One Flew Over

# 실습 준비하기

## ◎ 3개의 데이터프레임을 하나로 합치기

✓ 합쳐진 데이터 확인

17	M	50	1	95350	1193	5	978158471	One Flew Over the Cuckoo's Nest (1975)	Drama
----	---	----	---	-------	------	---	-----------	--	-------

## 2. 보고 싶은 영화 찾기

영화들의 평점 평균을 구하여, 사람들에게 인정받는 (평점이 높은) 영화 찾기

```
In [ ]: # 영화들의 평점 평균을 구하여, 평점이 높은 영화 찾기 |
```

```
In [ ]:
```

평균 평점이 만점인 영화들이 최상위에 위치함. 일반적으로 평점이 만점인 경우는 대부분 평점의 개수가 매우 적은 경우이므로, 이를 확인하기 위해 평점의 개수도 함께 구해본다.

```
In [ ]:
```

## [실습 #1] 여자들이 좋아하는 영화 찾기

- 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화



## ◎ 영화들의 평점 평균을 구하여, 평점이 높은 영화 찾기

## 2. 보고 싶은 영화 찾기

- ✓ 영화별 평점은 그룹 집계(pivot\_table) 사용
- ✓ 동일한 영화제목끼리 평점의 평균(mean)을 구함

평점이 높은) 영화 찾기

찾기

```
data.pivot_table(index = '영화제목', aggregate = 'mean', values = '평점')
```

Out[11]:

영화제목		평점
\$1,000,000 Duck (1971)		3.027027
'Night Mother (1986)		3.371429
'Til There Was You (1997)		2.692308
'burbs, The (1989)		2.910891
...And Justice for All (1979)		3.713568
...		...
Zed & Two Noughts, A (1985)		3.413793

영화별로 평점이 매겨지는 것을 확인해 볼 수 있습니다.

Zeus and Roxanne (1997) 2.521739

## ◎ 영화들의 평점 평균을 구하여, 평점이 높은 영화 찾기

```
In [12]: # 영화들의 평점 평균을 구하여, 평점이 높은 영화 찾기  
영화평점 = data.pivot_table(index = '영화제목', aggfunc = 'mean', values = '평점')
```

```
In [14]: 영화평점.sort_values(by = '평점', ascending = False)
```

Out[14]:

	평점
영화제목	
Ulysses (Ulisse) (1954)	5.0
Lured (1947)	5.0
How the Bitch (1998)	5.0
Sweet Motel (2000)	5.0
Fantasia (1975)	1.0
Torso (Corpi Presentano Tracce di Violenza Carnale) (1973)	1.0
Mutters Courage (1995)	1.0

**주의하기**

- ✓ 평점이 5.0인 경우는 합리적인 의심을 할 수 있으며,  
리뷰어가 1명 등 적은 사람이 점수를 주었다고 볼 수 있음  
→ 평점의 개수로 보완 가능



## ◎ 평점에 평점의 개수 확인하기

✓ aggfunc에 평균(mean)과 개수(count)를 리스트로 전달

```
mean', values = '평점')
```

```
In [14]: 영화평점.sort_values(by = '평점', ascending = False)
```

...

평균 평점이 만점인 영화들이 최상위에 위치함. 일반적으로 평점이 만점인 경우는 대부분 평점의 개수가 매우 적은 경우이므로, 이를 확인하기 위해 평점의 개수도 함께 구해본다.

```
In [ ]: 영화평점 = data.pivot_table(index = '영화제목', aggfunc = ['mean', 'count'], values =
```

### [실습 #1] 여자들이 좋아하는 영화 찾기

- 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

```
In [ ]:
```

```
In [ ]:
```

## ◎ 평점에 평점의 개수 확인하기

## ✓ 평점 정렬

평균 평점이 만점인 영화들이 최상위에 위치함. 일반적으로 평점이 만점인 경우는 대부분 평점의 개수가 매우 적은 경우이므로, 이를 확인하기 위해 평점의 개수도 함께 구해본다.

```
In [15]: 영화평점 = data.pivot_table(index = '영화제목',  
                                     aggfunc = ['mean', 'count'], values = '평점')
```

```
In [16]: 영화평점.sort_values(by = ('mean', '평점'), ascending = False)
```

Out [16]:

	mean	count
	평점	평점
영화제목		
\$1,000,000 Duck (1971)	3.027027	37
'Night Mother (1986)	3.371429	70
'Til There Was You (1997)	2.692308	52

## 주의하기

- ✓ 정렬 시 복사하여 사용하면 에러 발생  
: 컬럼명이 2줄로 되어 있는 경우에는 계층 색인이 되므로 튜플의 형태를 사용해야 함



## ◎ 기준 세우기

- ✓ 평점이 높은 영화 or 평점의 개수가 많은 영화 or 하이브리드
- ✓ 예 : 평점 4.3이상이며, 평점의 개수가 2,000개 이상인 영화

3706 rows × 2 columns

```
In [19]: # 평점평균이 4.3 이상이고, 평점의 개수가 2000개 이상인 영화를 찾기  
영화평점
```

Out[19]:

	mean	count
	평점	평점
영화제목		
\$1,000,000 Duck (1971)	3.027027	37
'Night Mother (1986)	3.371429	70
'Til There Was You (1997)	2.692308	52
'burbs, The (1989)	2.910891	303
...And Justice for All (1979)	3.713568	199
...	...	...

## ◎ 평점 4.3이상이며, 평점의 개수가 2,000개 이상인 영화 찾기

✓ 평점평균, 평점개수 2개의 columns으로 보기 좋게 변경

```
In [20]: # 평점평균이 4.3 이상이고, 평점의 개수가 2000개 이상인 영화를 찾기
영화평점.columns = ['평점평균', '평점개수']
영화평점
```

Out[20]:

	평점평균	평점개수
영화제목		
\$1,000,000 Duck (1971)	3.027027	37
'Night Mother (1986)	3.371429	70
'Til There Was You (1997)	2.692308	52
'burbs, The (1989)	2.910891	303
...And Justice for All (1979)	3.713568	199
...	...	...
Zed & Two Noughts, A (1985)	3.413793	29
Zero Effect (1998)	3.750831	301
Zero Kelvin (Kjærlighetens kjøtere) (1995)	3.500000	2



## ◎ 평점 4.3이상이며, 평점의 개수가 2,000개 이상인 영화 찾기

✓ 조건 색인 : 평점평균  $\geq 4.3$  & 평점개수  $\geq 2000$

(1980) 1.0 1

```
In [21]: # 평점평균이 4.3 이상이고, 평점의 개수가 2000개 이상인 영화를 찾기
영화평점.columns = ['평점평균', '평점개수']
(영화평점.평점평균 >= 4.3) & (영화평점.평점개수 >= 2000)
```

```
Out[21]: 영화제목
$1,000,000 Duck (1971)      False
'Night Mother (1986)      False
'Til There Was You (1997)  False
'burbs, The (1989)         False
...And Justice for All (1979) False
...
Zed & Two Noughts, A (1985) False
Zero Effect (1998)         False
Zero Kelvin (Kjærlighetens kjøtere) (1995) False
Zeus and Roxanne (1997)    False
eXistenZ (1999)             False
Length: 3706, dtype: bool
```

◎ 평점 4.3이상이며, 평점의 개수가 2,000개 이상인 영화 찾기

✓ 조건 색인 : 평점평균 $\geq$ 4.3 & 평점개수 $\geq$ 2000

평점.평점개수  $\geq$  2000)]

영화제목	평점평균	평점개수
American Beauty (1999)	4.317386	3428
Godfather, The (1972)	4.524966	2223
Matrix, The (1999)	4.315830	2590
Princess Bride, The (1987)	4.303710	2318
Raiders of the Lost Ark (1981)	4.477725	2514
Saving Private Ryan (1998)	4.337354	2653
Schindler's List (1993)	4.510417	2304
Shawshank Redemption, The (1994)	4.554558	2227
Silence of the Lambs, The (1991)	4.351823	2578
Sixth Sense, The (1999)	4.406263	2459

이 조건을 충족하는 11개의 영화만 볼 수가 있습니다.



◎ 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

✓ 여자들이 좋아하는 영화를 찾는 방법

① 모든 영화 평점 데이터에서 여성 평점 데이터만 선택하여 영화별로

Star Wars: Episode IV - A New Hope (1977) 4.453694 2991

## [실습 #1] 여자들이 좋아하는 영화 찾기

- 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

In [23]: data

Out[23]:

장르	영화제목	타임스탬프	평점	영화 아이 디	지역	직업	연령	성별	사용 자아 이디
Drama	One Flew Over the Cuckoo's Nest (1975)	978300760	5	1193	48067	10	1	F	1

# 보고 싶은 영화 찾기    여자들이 좋아하는 영화 찾기

◎ 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

✓ 여자들이 좋아하는 영화를 찾는 방법

② 성별로 평점을 각각 구해서 여성 평점만 취함

## [실습 #1] 여자들이 좋아하는 영화 찾기

- 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화 ¶

In [23]: data

Out [23]:

	사용자아이디	성별	연령	직업	지역	영화아이디	평점	타임스탬프	영화제목	장르
0	1	F	1	10	48067	1193	5	978300760	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2	M	56	16	70072	1193	5	978298413	One Flew Over the Cuckoo's Nest (1975)	Drama



# 보고 싶은 영화 찾기

# 여자들이 좋아하는 영화 찾기

◎ 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

✓ 영화별 성별 평점 구하기

Bank Redemption, The (1994) 4.554558 2227  
e of the Lambs, The (1991) 4.351823 2578  
Sixth Sense, The (1999) 4.406263 2459

✓ pivot\_table에 index='영화제목', columns='성별', 평점에 대한 평균과 개수 구함

## [실습 #1] 여자들이 좋아하는 영화 찾기

- 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

```
In [23]: data.pivot_table(index = '영화제목', columns = '성별',  
                           aggfunc = ['mean', 'count'], values = '평점')
```

Out[23]:

	사용자아이디	성별	연령	직업	지역	영화아이디	평점	타임스탬프	영화제목	장르
0	1	F	1	10	48067	1193	5	978300760	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2	M	56	16	70072	1193	5	978298413	One Flew Over the Cuckoo's	Drama

◎ 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

- 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

```
In [24]: = data.pivot_table(index = '영화제목', columns = '성별',
                        aggfunc = ['mean', 'count'], values = '평점')
```

Out[24]:

성별	영화제목	mean		count	
		F	M	F	M
	\$1,000,000 Duck (1971)	3.375000	2.761905	16.0	21.0
	'Night Mother (1986)	3.388889	3.352941	36.0	34.0
	'Til There Was You (1997)	2.675676	2.733333	37.0	15.0
	'burbs, The (1989)	2.793478	2.962085	92.0	211.0
	...And Justice for All (1979)	3.828571	3.689024	35.0	164.0
	...	...	...	...	...
	Zed & Two Noughts, A (1985)	3.500000	3.380952	8.0	21.0

평점의 평균과 개수가 각각 female과 male로 남성과 여성으로 각각 나뉘어서 데이터가 저장된 것을 확인해 볼 수 있습니다.

Zeus and Roxanne (1997) 2.777778 2.357143 9.0 14.0



# 보고 싶은 영화 찾기    여자들이 좋아하는 영화 찾기

◎ 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

- ✓ 조건 색인
- ✓ 튜플의 형태로, (mean, '여성'), 즉 여성평점의 평균을 선택
- ✓ 4.0이상인 경우만 True가 나옴

```
In [27]: (ex1[('mean', 'F')]) >= 4.0
```

```
Out[27]: 영화제목
$1,000,000 Duck (1971)      False
'Night Mother (1986)      False
'Til There Was You (1997)  False
'burbs, The (1989)         False
...And Justice for All (1979) False
...
Zed & Two Noughts, A (1985) False
Zero Effect (1998)         False
Zero Kelvin (Kjærlighetens kjøtere) (1995) False
Zeus and Roxanne (1997)    False
eXistenZ (1999)            False
Name: (mean, F), Length: 3706, dtype: bool
```

```
In [ ]:
```

# 보고 싶은 영화 찾기    여자들이 좋아하는 영화 찾기

◎ 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

[실습 #11] 여자들이 좋아하는 영화 찾기

- ✓ &로 조건 색인 추가
- ✓ 튜플의 형태로, ('count', '여성'), 즉 여성평점의 개수를 선택
- ✓ 2개를 모두 만족하는 경우만 True가 나옴

개 이상인 영화

```
In [28]: ex1[(ex1[('mean', 'F')])>= 4.0) & (ex1[('count', 'F')])>= 500)
```

```
Out[28]: 영화제목
$1,000,000 Duck (1971)      False
'Night Mother (1986)      False
'Til There Was You (1997)  False
'burbs, The (1989)         False
...And Justice for All (1979) False
...
Zed & Two Noughts, A (1985) False
Zero Effect (1998)         False
Zero Kelvin (Kjærlighetens kjøtere) (1995) False
Zeus and Roxanne (1997)    False
eXistenZ (1999)             False
Length: 3706, dtype: bool
```

```
In [ ]:
```



◎ 여성 평점이 4.0 이상이고 여성 평점의 개수가 500개 이상인 영화

```
In [30]: 여성인기영화 = ex1[(ex1['mean', 'F'] >= 4.0) & (ex1['count', 'F'] >= 500)]
```

✓ 여성인기영화로 저장

Out[31]:

성별	영화제목	mean		count	
		F	M	F	M
	American Beauty (1999)	4.238901	4.347301	946.0	2482.0
	Being John Malkovich (1999)	4.159930	4.113636	569.0	1672.0
	Braveheart (1995)	4.016484	4.297839	546.0	1897.0
	Casablanca (1942)	4.300990	4.461340	505.0	1164.0
	E.T. the Extra-Terrestrial (1982)	4.089850	3.920264	601.0	1668.0
	Fargo (1996)	4.217656	4.267780	657.0	1856.0
	Forrest Gump (1994)	4.045031	4.105806	644.0	1550.0
	L.A. Confidential (1997)	4.106007	4.256678	566.0	1722.0
	Matrix, The (1999)	4.128405	4.362235	514.0	2076.0
	Princess Bride, The (1987)	4.342767	4.288942	636.0	1682.0
	Pulp Fiction (1994)	4.071956	4.346839	542.0	1629.0

◎ 장르 정보는 movies에 존재

◎ 전체 영화에 대한 장르가 아닌 여성인기영화에 해당하는 장르만 필요함

식애 모사.

여성인기영화의 장르 통계 구하기

예를 들어, 여성인기영화 중 Drama 장르의 영화는 10개, Action 영화는 3개, ...

In [36]: movies

Out [36]:

	영화아이디	영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
1	2	Jumanji (1995)	Adventure Children's Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama
4	5	Father of the Bride Part II (1995)	Comedy
...	...	...	...
3878	3948	Meet the Parents (2000)	Comedy
3879	3949	Requiem for a Dream (2000)	Drama
3880	3950	Tigerland (2000)	Drama



© isin()을 사용하여 movies에서 여성인기영화의 장르만 선택

✓ data.isin() : data의 값이 인자로 받은 리스트에 포함되어 있는 경우에만 True

예를 들어, 여성인기영화 중 Drama 장르의 영화는 10개, Action 영화는 3개, ...

```
In [36]: # isin() 함수를 사용해서 movies에서 여성인기영화의 장르만 선택
movies.영화아이디.isin([1,3,5,9])
```

Out [36]:

영화아이디		영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
1	2	Jumanji (1995)	Adventure Children's Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama
4	5	Father of the Bride Part II (1995)	Comedy
...	...	...	...
3878	3948	Meet the Parents (2000)	Comedy
3879	3949	Requiem for a Dream (2000)	Drama

© isin( )을 사용하여 movies에서 여성인기영화의 장르만 선택

✓ 영화아이디를 영화제목으로 변경

극에 표시.

여성인기영화의 장르 통계 구하기

예를 들어, 여성인기영화 중 Drama 장르의 영화는 10개, Action 영화는 3개, ...

```
In [38]: # isin() 함수를 사용해서 movies에서 여성인기영화의 장르만 선택
movies[movies.영화아이디.isin([1,3,5,9])]
```

Out [38]:

	영화아이디	영화제목	장르
0	1	Toy Story (1995)	Animation Children's Comedy
2	3	Grumpier Old Men (1995)	Comedy Romance
4	5	Father of the Bride Part II (1995)	Comedy
8	9	Sudden Death (1995)	Action

In [ ]:

In [ ]:



© isin()을 사용하여 movies에서 여성인기영화의 장르만 선택

✓ 1,3,5,9대신 여성인기영화.index로 바꿈

예를 들어, 여성인기영화 중 Drama 장르의 영화는 10개, Action 영화는 3개, ...

In [40]: 여성인기영화.index

Out [40]: Index(['American Beauty (1999)', 'Being John Malkovich (1999)',  
'Braveheart (1995)', 'Casablanca (1942)',  
'E.T. the Extra-Terrestrial (1982)', 'Fargo (1996)',  
'Forrest Gump (1994)', 'L.A. Confidential (1997)', 'Matrix, The (1999)',  
'Princess Bride, The (1987)', 'Pulp Fiction (1994)',  
'Raiders of the Lost Ark (1981)', 'Saving Private Ryan (1998)',  
'Schindler's List (1993)', 'Shakespeare in Love (1998)',  
'Shawshank Redemption, The (1994)', 'Silence of the Lambs, The (1991)',  
'Sixth Sense, The (1999)', 'Star Wars: Episode IV - A New Hope (1977)',  
'Star Wars: Episode V - The Empire Strikes Back (1980)',  
'Toy Story (1995)', 'Wizard of Oz, The (1939)'],  
dtype='object', name='영화제목')

In [38]: # isin() 함수를 사용해서 movies에서 여성인기영화의 장르만 선택  
movies[movies.영화제목.isin([1,3,5,9])]

Out [38]:

영화아이디

영화제목

장르



© isin()을 사용하여 movies에서 여성인기영화의 장르만 선택

✓ 장르 정보만 선택

```
In [42]: # isin() 함수를 사용해서 movies에서 여성인기영화의 장르만 선택
movies[movies.영화제목.isin(여성인기영화.index)].장르
```

```
Out[42]: 0      Animation|Children's|Comedy
108      Action|Drama|War
257      Action|Adventure|Fantasy|Sci-Fi
293      Crime|Drama
315      Drama
352      Comedy|Romance|War
523      Drama|War
589      Drama|Thriller
604      Crime|Drama|Thriller
900      Drama|Romance|War
907      Adventure|Children's|Drama|Musical
1081     Children's|Drama|Fantasy|Sci-Fi
1178     Action|Adventure|Drama|Sci-Fi|War
1179     Action|Adventure|Comedy|Romance
1180     Action|Adventure
1575     Crime|Film-Noir|Mystery|Thriller
1959     Action|Drama|War
2327     Comedy|Romance
2502     Action|Sci-Fi|Thriller
2602     Thriller
```



© isin( )을 사용하여 movies에서 여성인기영화의 장르만 선택

- ✓ 여성인기영화장르로 저장
- ✓ 하나의 영화가 여러 장르에 속해 있는 경우, 세로줄로 구분되어 있음

```
Out[44]: 0      Animation|Children's|Comedy
          108      Action|Drama|War
          257      Action|Adventure|Fantasy|Sci-Fi
          293      Crime|Drama
          315      Drama
          352      Comedy|Romance|War
          523      Drama|War
          589      Drama|Thriller
          604      Crime|Drama|Thriller
          900      Drama|Romance|War
          907      Adventure|Children's|Drama|Musical
          1081     Children's|Drama|Fantasy|Sci-Fi
          1178     Action|Adventure|Drama|Sci-Fi|War
          1179     Action|Adventure|Comedy|Romance
          1180     Action|Adventure
          1575     Crime|Film-Noir|Mystery|Thriller
          1959     Action|Drama|War
          2327     Comedy|Romance
          2502     Action|Sci-Fi|Thriller
```

© str.split() 함수를 활용하여 장르 정보를 구분자로 구분



```
여성인기영화장르 = movies[movies.영화제목.isin(여성인기영화.index)].장르
```

```
In [46]: # str.split() 함수를 활용하여 장르 정보를 구분자로 구분
여성인기영화장르 = 여성인기영화장르.str.split('|', expand = True)
```

Out[46]:

	0	1	2	3	4
0	Animation	Children's	Comedy	None	None
108	Action	Drama	War	None	None
257	Action	Adventure	Fantasy	Sci-Fi	None
293	Crime	Drama	None	None	None
315	Drama	None	None	None	None
352	Comedy	Romance	War	None	None
523	Drama	War	None	None	None
589	Drama	Thriller	None	None	None
604	Crime	Drama	Thriller	None	None
900	Drama	Romance	War	None	None
907	Adventure	Children's	Drama	Musical	None



## ◎ 영화장르 카운트하기

✓ 컬럼별로 `value_count()`를 수행한 결과값들을 모두 더함

```
In [47]: # str.split() 함수를 활용하여 장르 정보를 구분자로 구분  
여성인기영화장르 = 여성인기영화장르.str.split('|', expand = True)
```

```
In [50]: 여성인기영화장르[0].value_counts() + 여성인기영화장르[1].value_counts()
```

```
Out[50]: Action      NaN  
Adventure    5.0  
Animation     NaN  
Children's    3.0  
Comedy        NaN  
Crime         NaN  
Drama        10.0  
Film-Noir     NaN  
Romance       NaN  
Sci-Fi        NaN  
Thriller      2.0  
War           NaN  
dtype: float64
```

**주의하기**

✓ 동일한 인덱스가 두 군데 모두 있는 경우에만  
정상적으로 산술연산이 수행되고, 둘 중 하나라도  
인덱스가 없다면 NaN값으로 처리됨

```
In [ ]:
```

## ◎ 영화장르 카운트하기

✓ +를 add로 변경

✓ fill\_value = 0

```
In [51]: 여성인기영화장르[0].value_counts().add(여성인기영화장르[1].value_counts(), fill_value =
```

```
Out[51]: Action          7.0  
Adventure       5.0  
Animation        1.0  
Children's       3.0  
Comedy           4.0  
Crime            3.0  
Drama           10.0  
Film-Noir        1.0  
Romance          3.0  
Sci-Fi           1.0  
Thriller         2.0  
War              1.0  
dtype: float64
```

```
In [ ]:
```



## ◎ 영화장르 카운트하기

✓ for문 활용

```
# str.split() 함수를 활용하여 장르 정보를 구분자로 구분  
여성인기영화장르 = 여성인기영화장르.str.split('|', expand = True)
```

✓ 여성인기장르를 series로 생성

```
counts().add(여성인기영화장르[1].value_counts(), fill_value =
```

```
Action      1.0  
Adventure    5.0  
Animation     1.0  
Children's   3.0  
Comedy        4.0  
Crime         3.0  
Drama        10.0  
Film-Noir     1.0  
Romance       3.0  
Sci-Fi        1.0  
Thriller      2.0  
War           1.0  
dtype: float64
```

```
In [ ]: 여성인기장르 = Series()  
for col |
```

```
In [ ]:
```

## ◎ 영화장르 카운트하기

✓ for문 활용

Children's	3.0
Comedy	4.0
Crime	3.0

- ✓ 5개의 columns 순회
- ✓ '여성인기장르'에 5번을 더함
- ✓ 더할 때 index로 col을 넣어줌

```
In [52]: 여성인기장르 = Series()  
for col in 여성인기영화장르.columns:  
    여성인기장르.add(여성인기영화장르[col].value_counts(), fill_value = 0)
```

<ipython-input-52-b3aff1e180a8>:1: DeprecationWarning: The default dtype for empty Series will be 'object' instead of 'float64' in a future version. Specify a dtype explicitly to silence this warning.

```
여성인기장르 = Series()
```

```
In [ ]:
```

```
In [ ]:
```



## ◎ 영화장르 카운트하기

- ✓ DeprecationWarning : series 생성 시 object로 기본 데이터 타입이 설정되므로 float로 하기 위해서는 float로 변경하라는 뜻

```
Film-Noir      1.0  
Romance        3.0  
Sci-Fi         1.0  
Thriller       2.0  
War            1.0  
dtype: float64
```

```
In [52]: 여성인기장르 = Series()  
for col in 여성인기영화장르.columns:  
    여성인기장르.add(여성인기영화장르[col].value_counts(), fill_value = 0)
```

```
<ipython-input-52-b3aff1e180a8>:1: DeprecationWarning: The default dtype for empty Series will be 'object' instead of 'float64' in a future version. Specify a dtype explicitly to silence this warning.  
    여성인기장르 = Series()
```

```
In [ ]:
```

```
In [ ]:
```

## ◎ 영화장르 카운트하기

✓ DeprecationWarning을 없애기 위해서는 dtype을 float64로 지정

```
0.0  
Drama      10.0  
Film-Noir   1.0  
Romance     3.0  
Sci-Fi      1.0  
Thriller    2.0  
War         1.0  
dtype: float64
```

```
In [52]: 여성인기장르 = Series(dtype = 'float64')  
for col in 여성인기영화장르.columns:  
    여성인기장르.add(여성인기영화장르[col].value_counts(), fill_value = 0)
```

```
<ipython-input-52-b3aff1e180a8>:1: DeprecationWarning: The default dtype for empty Series will be 'object' instead of 'float64' in a future version. Specify a dtype explicitly to silence this warning.  
여성인기장르 = Series()
```

In [ ]:

In [ ]:



## ◎ 영화장르 카운트하기

✓ 여성인기장르를 더한 값을 여성인기장르에 할당하여 누적

```
Out[56]: Action      7.0  
         Adventure   5.0  
         Animation    1.0  
         Children's   3.0  
         Comedy       6.0  
         Crime        3.0  
         Drama        12.0  
         Fantasy      2.0  
         Film-Noir    1.0  
         Musical      1.0  
         Mystery      1.0  
         Romance      4.0  
         Sci-Fi       4.0  
         Thriller     5.0  
         War          6.0  
         dtype: float64
```

In [ ]:

In [ ]: