

전처리 |

**DF Group 4**

**Netflix**

**N**

# Table of Contents

---

- > 분석 데이터 및 목표
- > 가설1: '제목의 길이'와 '관객 수'의 상관관계
- > 가설2: '생산 국가'와 '관객 수'의 상관관계
- > 가설3: '시즌 수'와 '관객 수'의 상관관계
- > 가설4: '관객 등급'과 '관객 수'의 상관관계



## 분석 데이터 및 목표

---

- > 데이터 셋: 캐글 Netflix-movies-and-tv-shows
- > 1차 목표: 관객 수(imdb\_votes)에 영향을 미치는 변수 회귀분석
- > 2차 목표: 머신러닝을 통한 '관객수' 예측 모델링
- > 3차 목표(잠정): User 데이터셋을 이용한 Recommendation System 개발

NETFLIX

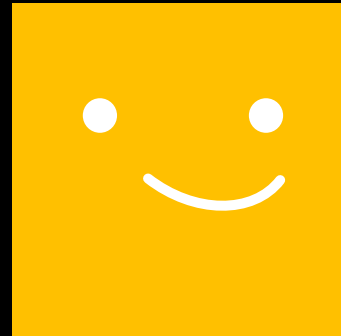
NETFLIX를 시청할 프로필을 선택하세요



변서현



김동호



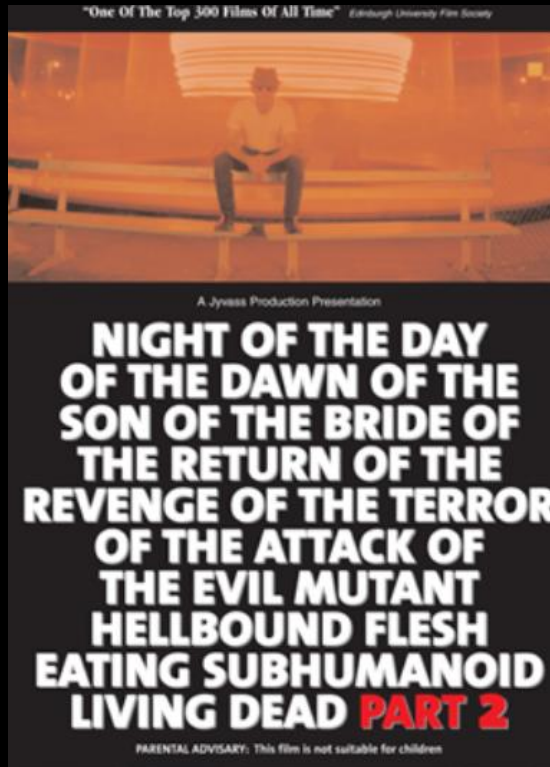
김수현



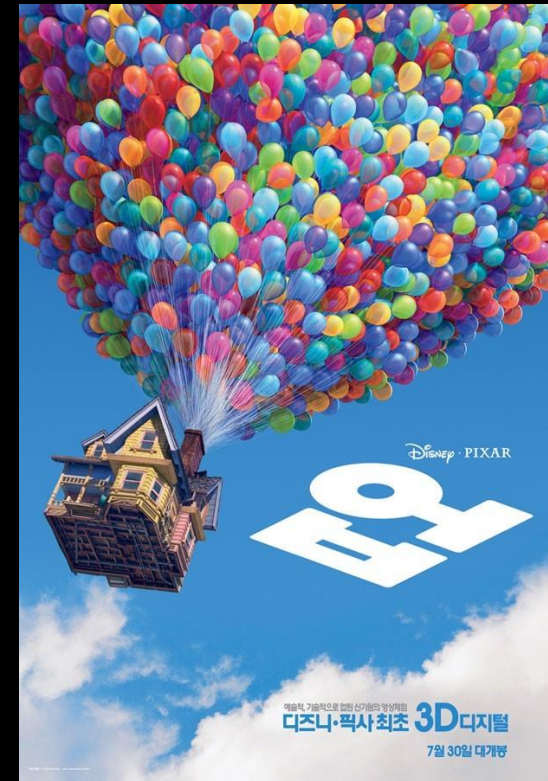
박예은

시작하기

# 가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?



VS





가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?

[전처리 과정]

Step1. 변수 선택

Step2. 결측치 제거

Step3. '제목의 길이' 변수 추가



## 가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?

### [전처리 과정]

#### Step1. 변수 선택

제목의 길이 → title 변수 -> word\_count 변수 추가

관객 수 → imdb\_votes 변수



## 가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?

### [전처리 과정]

#### Step2. 결측치 제거

```
▶ ## 결측치 제거  
▶ df.dropna(subset=['imdb_votes'], inplace=True)
```

imdb\_votes 칼럼의 결측치 제거

5850 rows × 15 columns -> 5352 rows × 15 columns





## 가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?

### [전처리 과정]

#### Step3. '제목의 길이' 데이터 추가

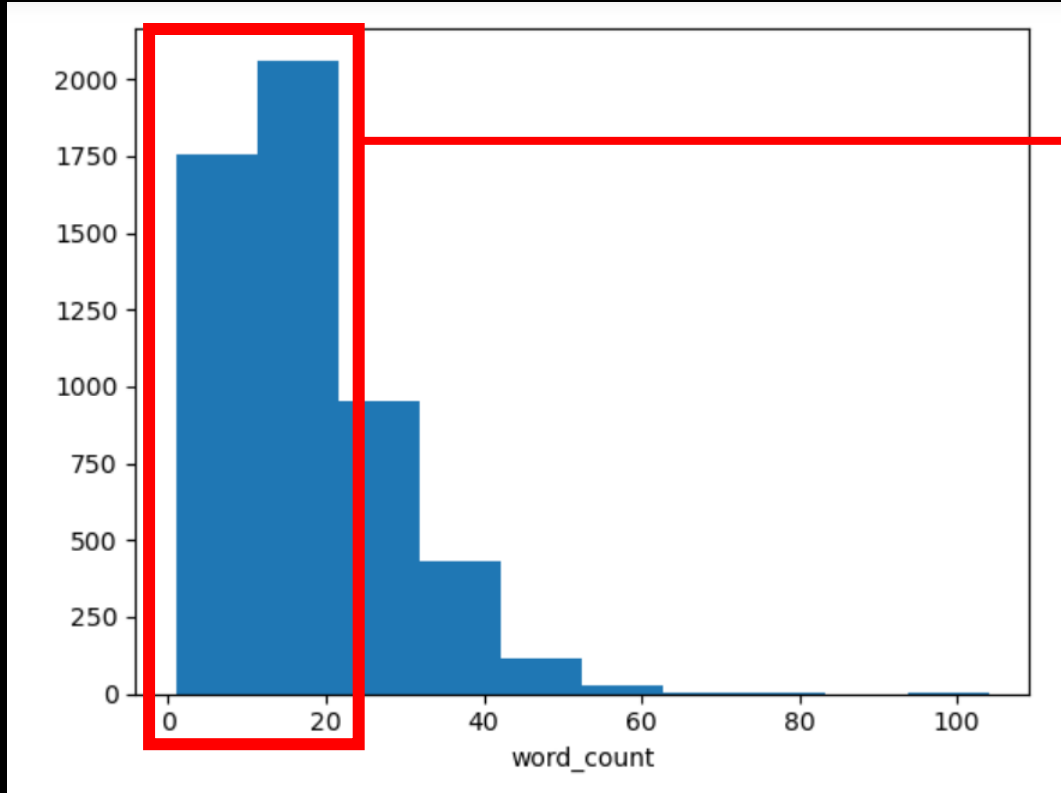
```
▶ ## word_count 칼럼 추가
```

```
▶ df[['title']]
```

```
▶ df['word_count']=df.title.str.count(' ')-1
```

## 가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?

[summary]



10~20글자 사이의 영화가 가장 많음

## 가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?

[summary]

```
In [50]: ## 상관계수
```

```
In [39]: from scipy import stats
```

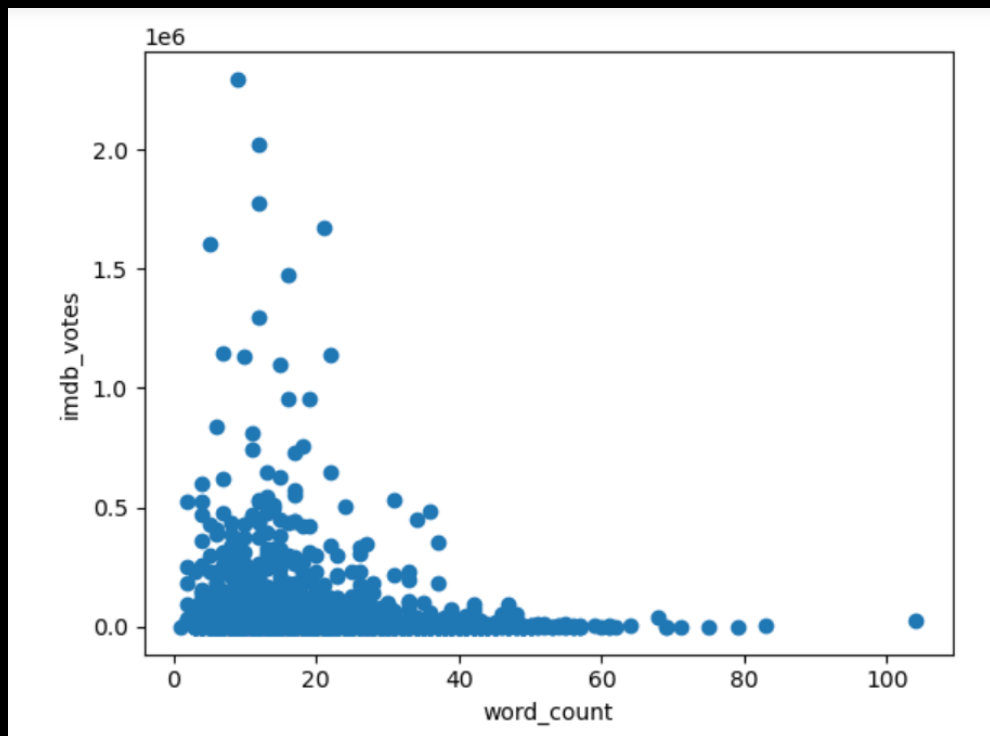
```
In [41]: stats.pearsonr(df['word_count'], df['imdb_votes'])
```

```
Out[41]: PearsonRResult(statistic=-0.08260074783067509, pvalue=1.4324961371944998e-09)
```

상관계수는 -0.08로, 두 변수 사이의 직접적인 상관성은 없는 것으로 보임

## 가설1: '제목의 길이'와 '관객 수'간의 상관관계가 있을까?

[summary]



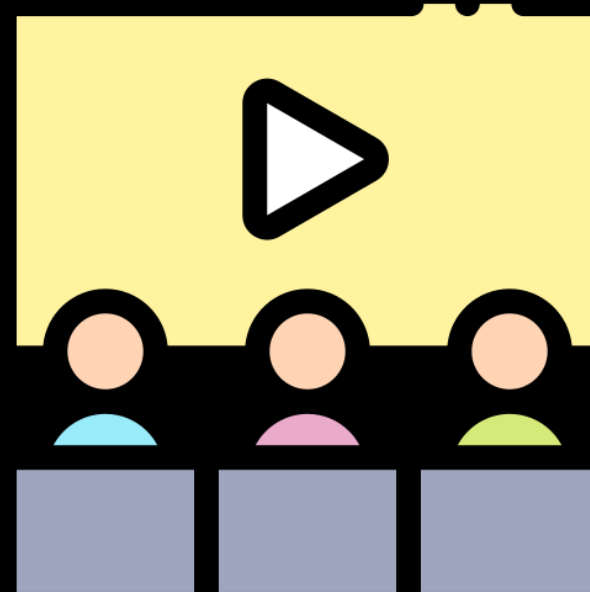
산점도 역시 상관관계를 발견하긴 어려움

하지만, 높은 평점을 받은 영화들은 대체로  
10~20 글자 구간에 분포되어 있는 것을 알  
수 있음

## 가설2: 작품 생산 '국가'와 '관객 수' 간에 상관관계가 있을까?



VS





## 가설2: 작품 생산 '국가'와 '관객 수' 간에 상관관계가 있을까?

### [전처리 과정]

Step1. 결측치 제거

Step2. 국가별 평균 관객수 분석

Step3. Categorical > Numerical

## 가설2: 작품 생산 '국가'와 '관객 수' 간에 상관관계가 있을까?

## [전처리 과정]

## Step1. 결측치 제거

1. 국가명 앞뒤로 붙어있는 [ ]와 ' ' 등을 제거
2. imdb\_votes를 가지고 있지 않은 row 제거

production_countries
['US']
['US']
['US']
['GB']



:	0	US
	1	US
	2	US
	3	GB
	4	GB
		..
	5845	NG
	5846	
	5847	CO
	5848	US
	5849	

## 가설2: 작품 생산 '국가'와 '관객 수' 간에 상관관계가 있을까?

### [전처리 과정]

#### Step2. 국가별 평균 관객수 분석

1. Country: 작품이 많이 등록된 국가 순위
2. Average votes: 해당 국가의 평균 관객수 (imdb\_votes)

>> 생산한 영화가 많다고 꼭 관객수가 많은 건 아님

	Country	Average votes
0	US	43983.779503
1	IN	11274.654577
2	GB	42212.551095
3	JP	10783.419355
4	KR	7057.963918
...	...	...
90	BY	97.000000
91	SN	135.000000
92	GR	58882.000000
93	MU	1029.000000
94	BT	3257.000000





## 가설2: 작품 생산 '국가'와 '관객 수' 간에 상관관계가 있을까?

0]: 86 527447.000000  
92 58882.000000  
0 43983.779503  
2 42212.551095  
66 39272.500000  
...  
51 97.250000  
90 97.000000  
70 82.000000  
78 64.000000  
50 46.000000

### Top cast >



Brad Pitt  
Achilles



Eric Bana  
Hector



Orlando Bloom  
Paris



Julian Glover  
Triopas



Brian Cox  
Agamemnon



Nathan Jones  
Boagrius

### Top cast >



Olivia Colman  
Leda



Jessie Buckley  
Young Leda



Dakota Johnson  
Nina



Ed Harris  
Lyle



Peter Sarsgaard  
Professor Hardy



Paul Mescal  
Will

1. 86번 국가: 몰타

2. 92번 국가: 그리스

>> 감독과 배우 모두 사실상 미국 영화인이므로 이상치 교체  
(미국으로 편입)



## 가설2: 작품 생산 '국가'와 '관객 수' 간에 상관관계가 있을까?

### [전처리 과정]

### Step3. Categorical > Numerical

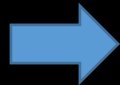
1. Outlier 제거 후 다시 평균 관객수 계산
2. 1순위부터 94순위까지 가중치 부여

	Country	Average votes	Weight
0	US	44241.463806	94
1	GB	42212.551095	93
2	HU	39272.500000	92
3	DE	36636.046512	91
4	CH	32977.333333	90
...	...	...	...
90	KE	82.000000	4
91	HR	64.000000	3
92	TZ	46.000000	2
93	MT	NaN	1
94	GR	NaN	0



## 가설2: 작품 생산 '국가'와 '관객 수' 간에 상관관계가 있을까?

age_certification	runtime	genres	production_countries	seasons
R	114	['drama', 'crime']	US	NaN
R	109	['drama', 'action', 'thriller', 'european']	US	NaN
PG	91	['fantasy', 'action', 'comedy']	GB	NaN
NaN	150	['war', 'action']	GB	NaN
TV-14	30	['comedy', 'european']	GB	4.0
...	...	...	...	...



age_certification	runtime	genres	production_countries	seasons
R	114	['drama', 'crime']	94	NaN
R	109	['drama', 'action', 'thriller', 'european']	94	NaN
PG	91	['fantasy', 'action', 'comedy']	93	NaN
NaN	150	['war', 'action']	93	NaN
TV-14	30	['comedy', 'european']	93	4.0



## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

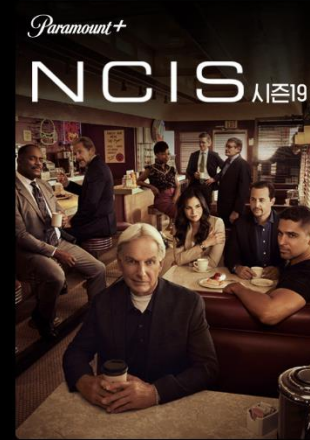


슬기로운 의사생활 시즌2



슬기로운 의사생활

VS



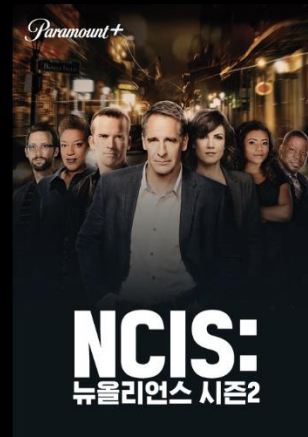
NCIS 시즌19



NCIS : 뉴올리언스 시즌6



NCIS 로스앤젤레스 시즌13



NCIS : 뉴올리언스 시즌2



NCIS : 뉴올리언스 시즌7



NCIS 로스앤젤레스 시즌12



## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

### [전처리 과정]

Step1. 변수 선택

Step2. 결측치 대체

Step3. 필요한 데이터 추출



## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

### [전처리 과정]

#### Step1. 변수 선택

시즌 수 → seasons 변수

관객 수 → imdb\_votes 변수

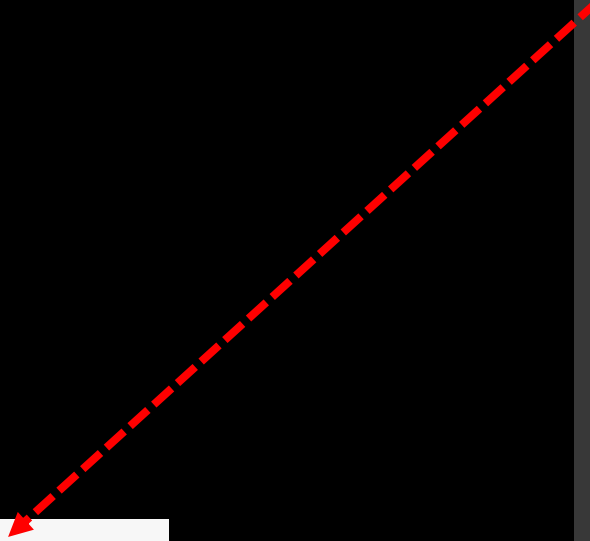
## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

### [전처리 과정]

#### Step2. 결측치 대체

```
# seasons의 값 확인  
df['seasons'].value_counts()  
  
# seasons 결측치를 최빈값인 1.0으로 대체  
df['seasons'].fillna(1.0, inplace=True)
```

seasons 칼럼의 최빈값인 1.0으로 결측치 대체



1.0	1221
2.0	389
3.0	187
4.0	120
5.0	79
6.0	36
7.0	18
8.0	11
9.0	9
11.0	8
10.0	6
12.0	4
15.0	3
24.0	2
13.0	2
20.0	1
19.0	1
32.0	1
29.0	1
14.0	1
37.0	1
21.0	1
25.0	1
42.0	1
39.0	1
16.0	1

Name: seasons, dtype: int64



## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

### [전처리 과정]

#### Step3. 필요한 데이터 추출

```
# type이 SHOW인 데이터를 tvshow로 저장  
tvshow = df[df['type'] == 'SHOW']  
tvshow
```

	id	title	type	description	release_year	age_certification	runtime	genre
7	ts22164	Monty Python's Flying Circus	SHOW	A British sketch comedy series with the shows ...	1969	TV-14	30	com
17	ts45948	Monty Python's Fliegender Zirkus	SHOW	Monty Python's Fliegender Zirkus consisted of ...	1972	TV-MA	43	com
35	ts20681	Seinfeld	SHOW	A stand-up comedian and his three offbeat frie...	1989	TV-PG	24	com

SHOW 작품의 seasons과 imdb\_votes간의 관계를 살펴보기 위해 해당하는 데이터만 추출

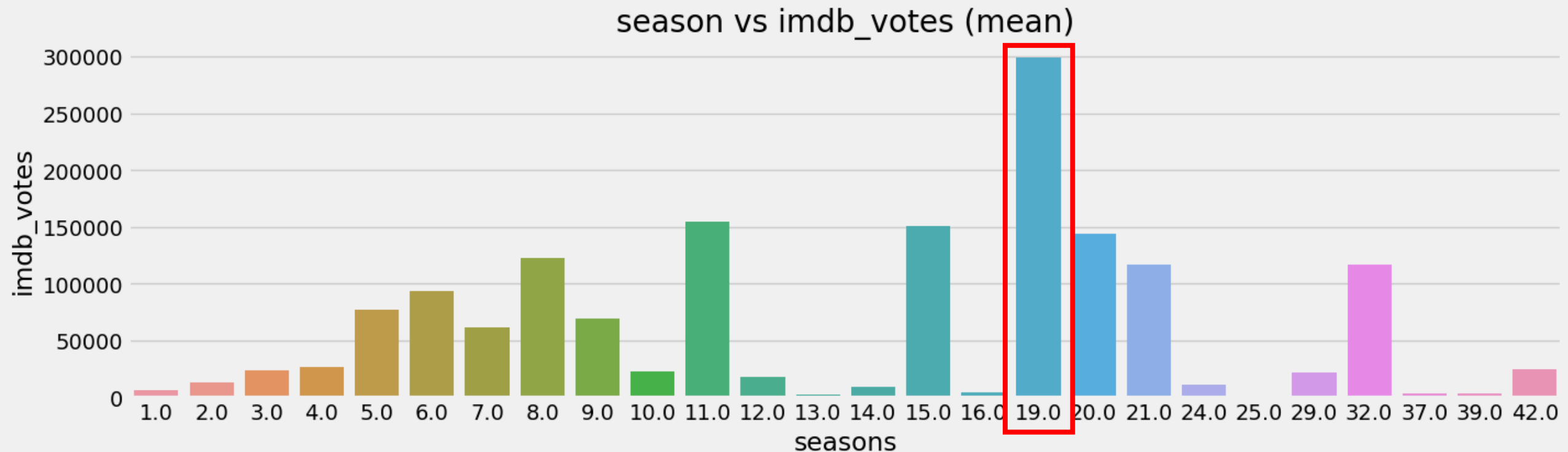




## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

[summary]

평균적으로 작품의 시즌 수가 19일 때 가장 높은 평점을 받음

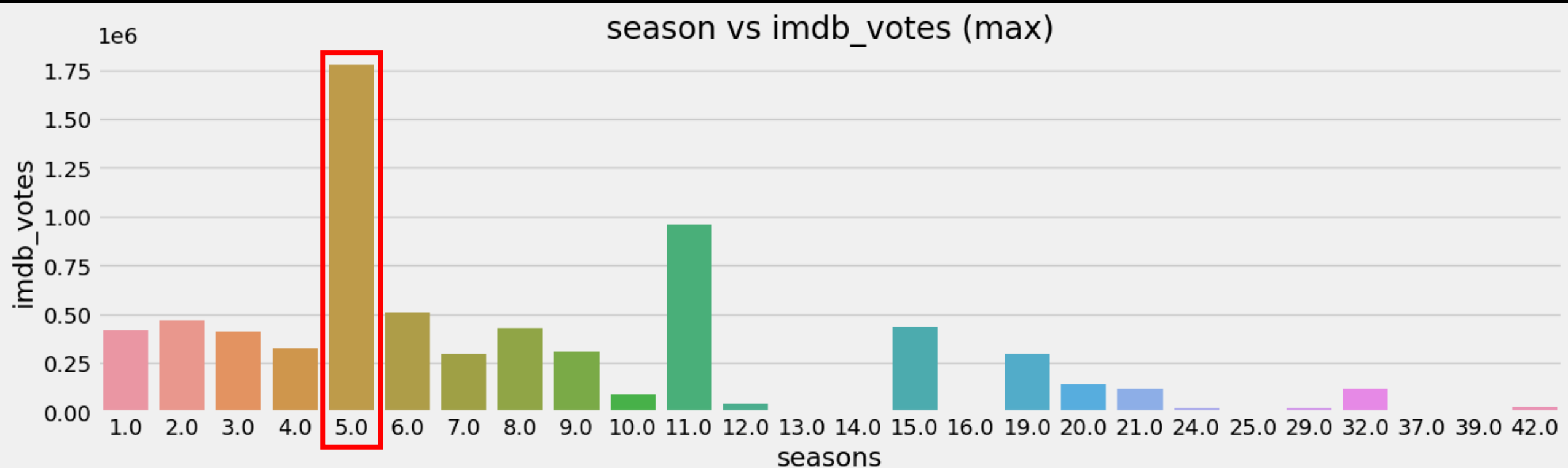




## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

[summary]

하지만, 시즌 수가 5인 작품이 최대 평점을 받음



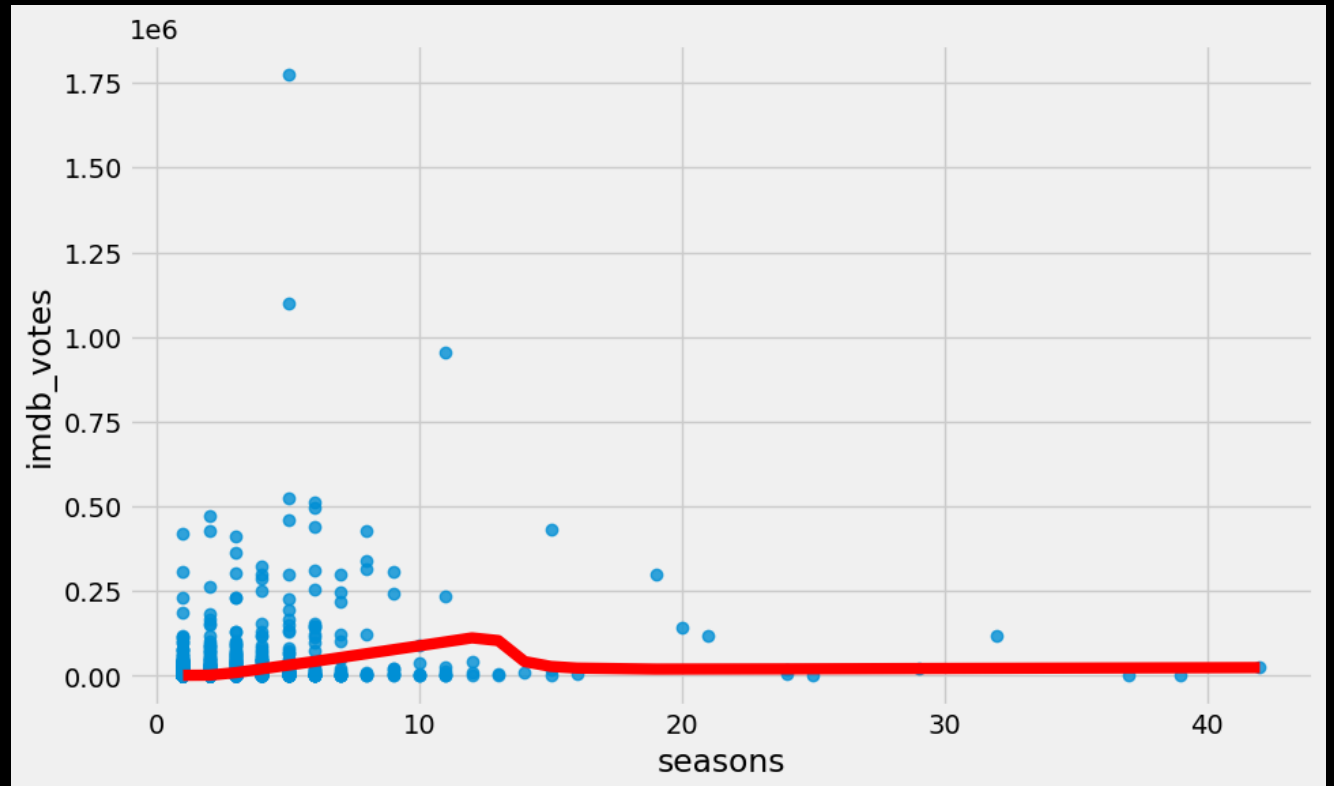


## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

[summary]

전세계 작품의 평점 추세를 살펴보면,

- 시즌 10 초반까지는 작품 시즌 개수가 증가할 수록 imdb\_votes도 증가함
- 10 중반부터는 급격하게 imdb\_votes가 하락하는 것을 알 수 있음



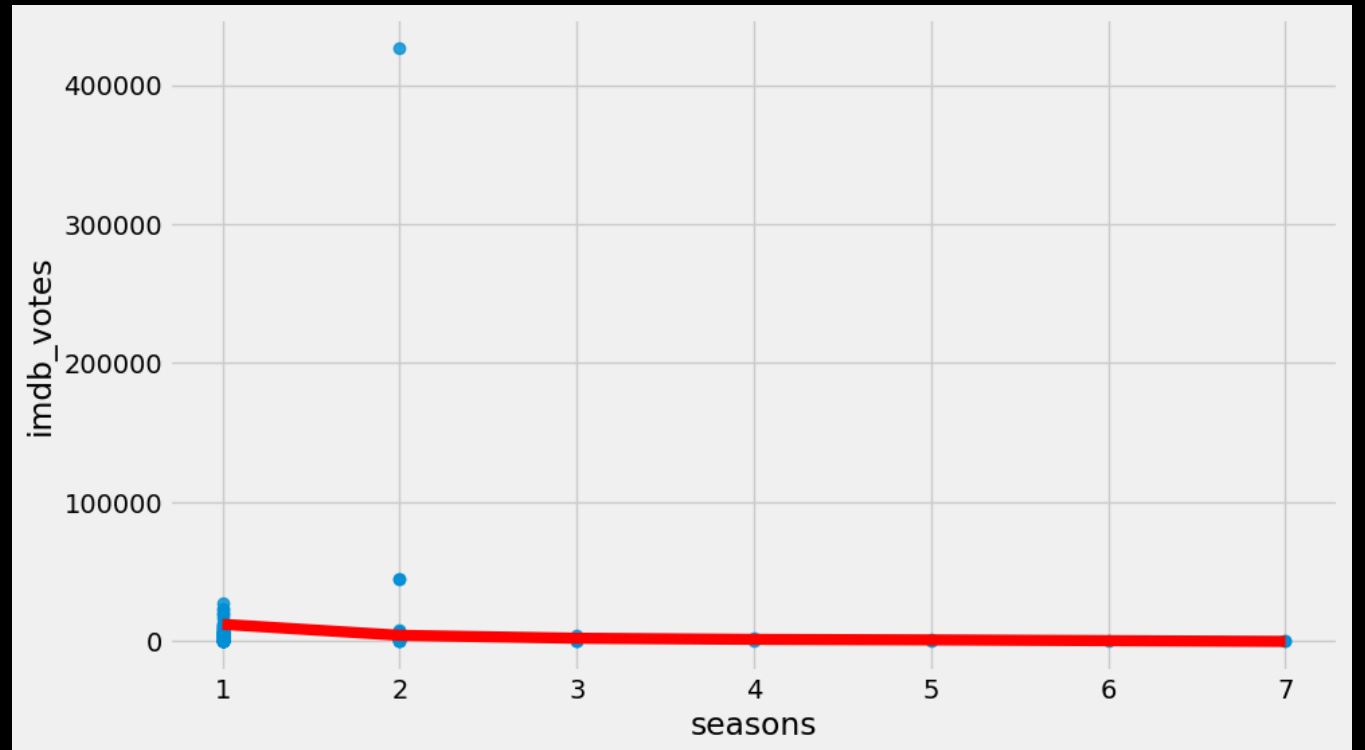


## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

[summary]

한국 작품의 평점 추세를 살펴보면,

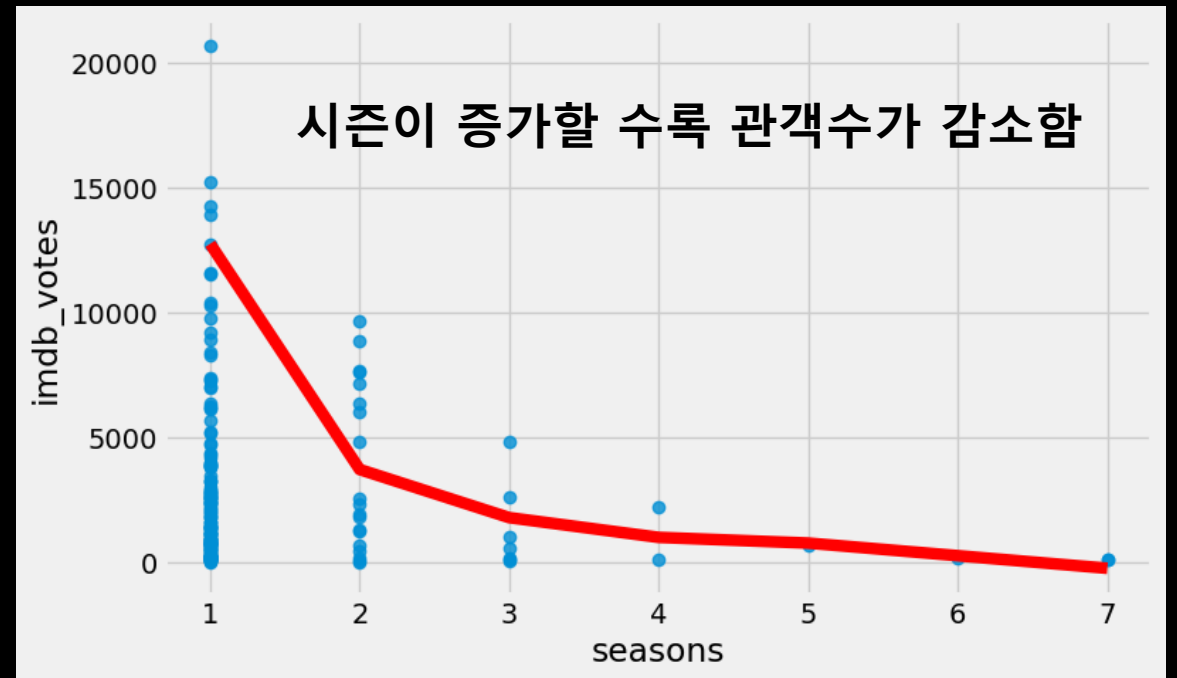
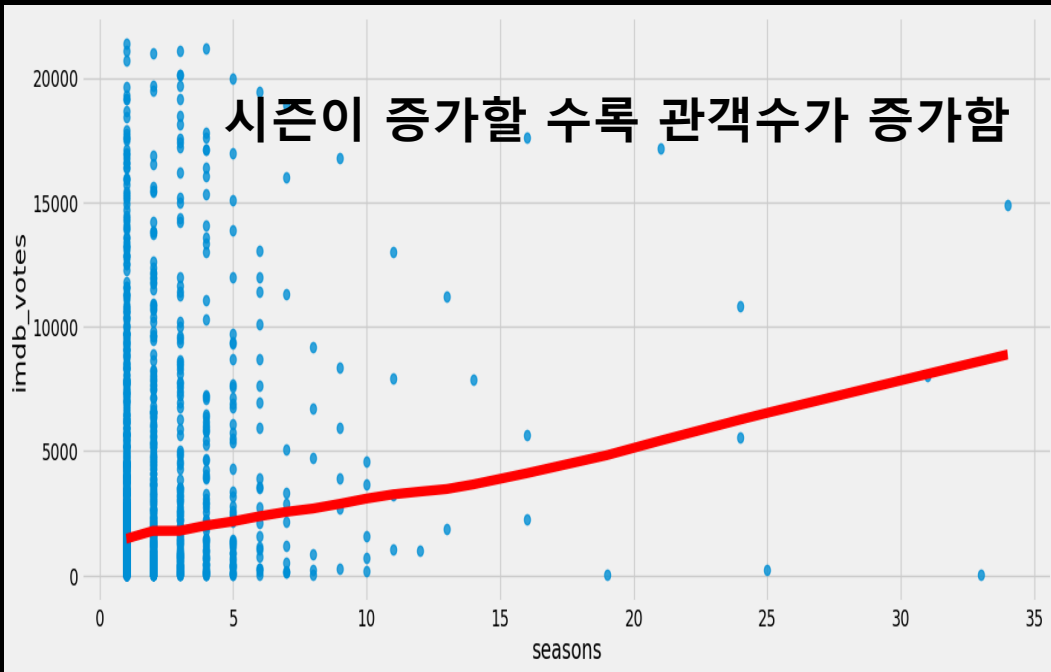
- 시즌이 증가할 수록 imdb\_votes가 감소하나, 큰 차이가 없는 것으로 보임
- 이는 시즌 2에 존재하는 이상치가 영향을 미치는 것으로 파악됨



## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

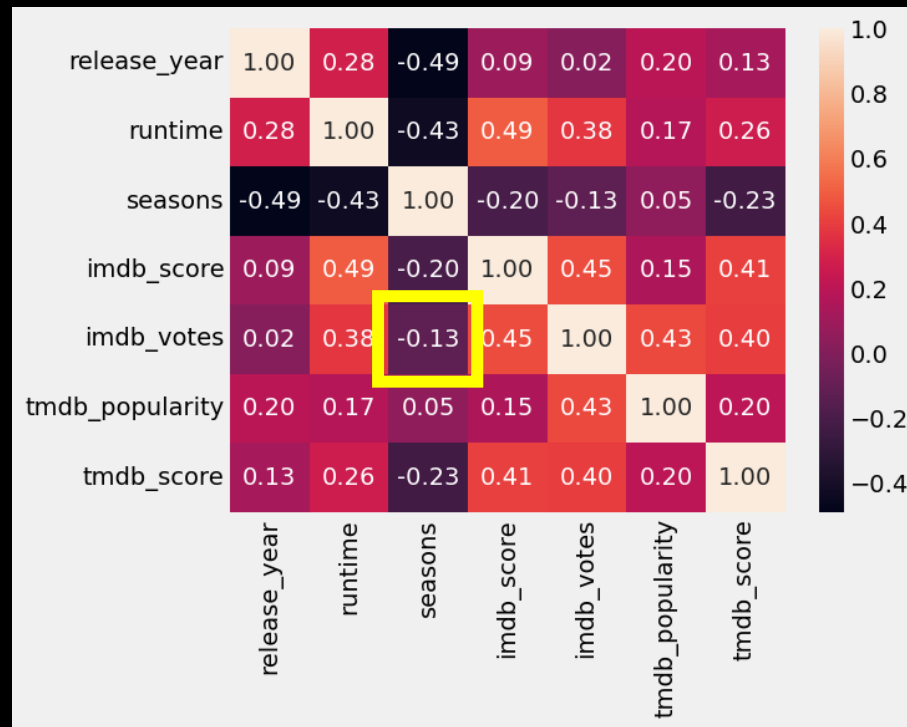
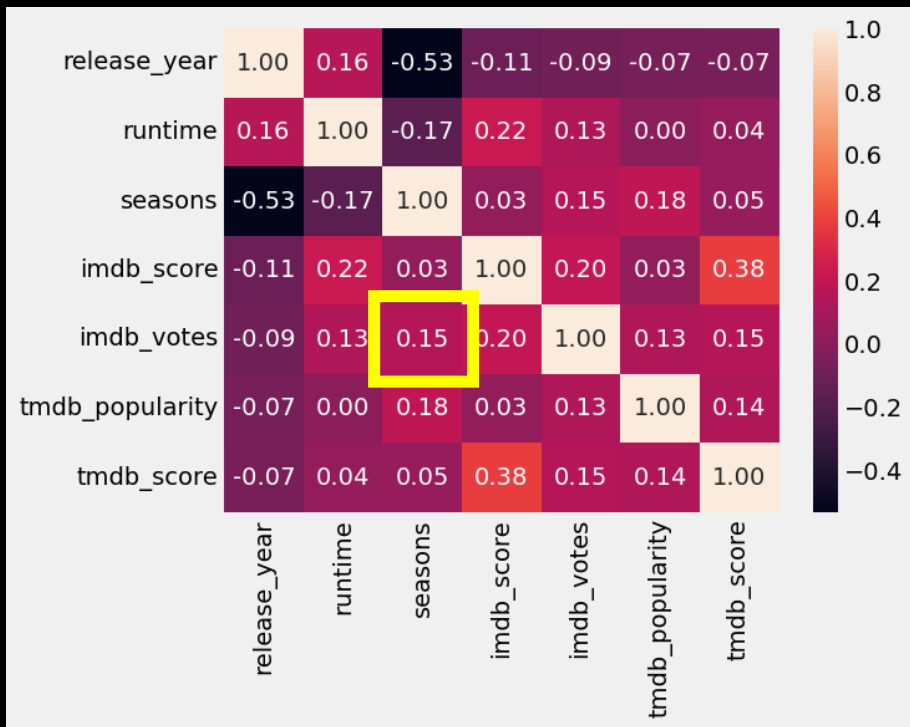
[summary]

이상치를 제거한 후 다시 전세계 작품과 한국 작품의 추세를 살펴보면,



## 가설3: '시즌 수'와 '관객 수' 간의 상관관계가 있을까?

[summary] seasons과 imdb\_votes 사이의 상관계수는 0.15, -0.13이지만 산점도와 추세선을 통해, 시즌 수에 따라 관객수가 변함을 볼 수 있음





## 가설4: '관객 등급'과 '관객 수' 간의 상관관계가 있을까?



VS





## 가설4: '관객 등급'과 '관객 수' 간의 상관관계가 있을까?

### [전처리 과정]

Step1. 변수 선택

Step2. 결측치 제거

Step3. 데이터 추출





## 가설4: '관객 등급'과 '관객 수' 간의 상관관계가 있을까?

### [전처리 과정]

#### Step1. 변수 선택

관객 등급 -> age\_certification 변수

영화 평점 -> imdb\_votes 변수



## 가설4: '관객 등급'과 '관객 수' 간의 상관관계가 있을까?

[전처리 과정]

Step2. 결측치 제거

```
## age_certification 결측값 제거  
~~~~~  
df = df[df[AGE].notna()]  
## imdb_votes 결측값 제거  
~~~~~  
df = df[df[VOTE] > 0]
```



## 가설4: '관객 등급'과 '관객 수' 간의 상관관계가 있을까?

### [전처리 과정]

#### Step3. 데이터 추출

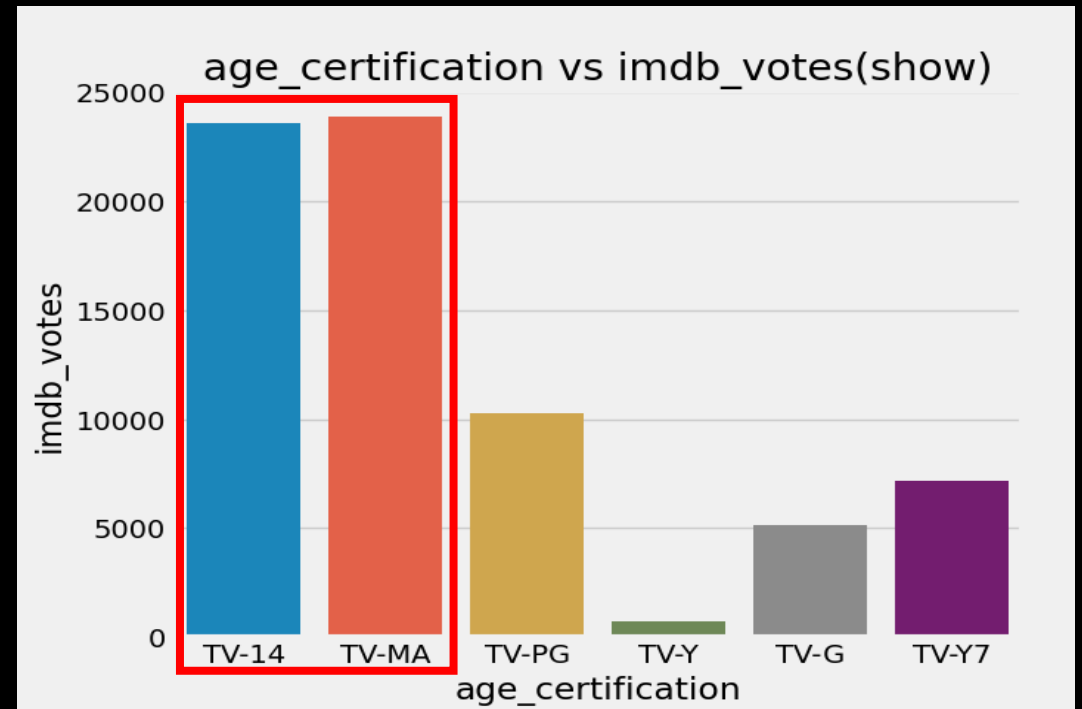
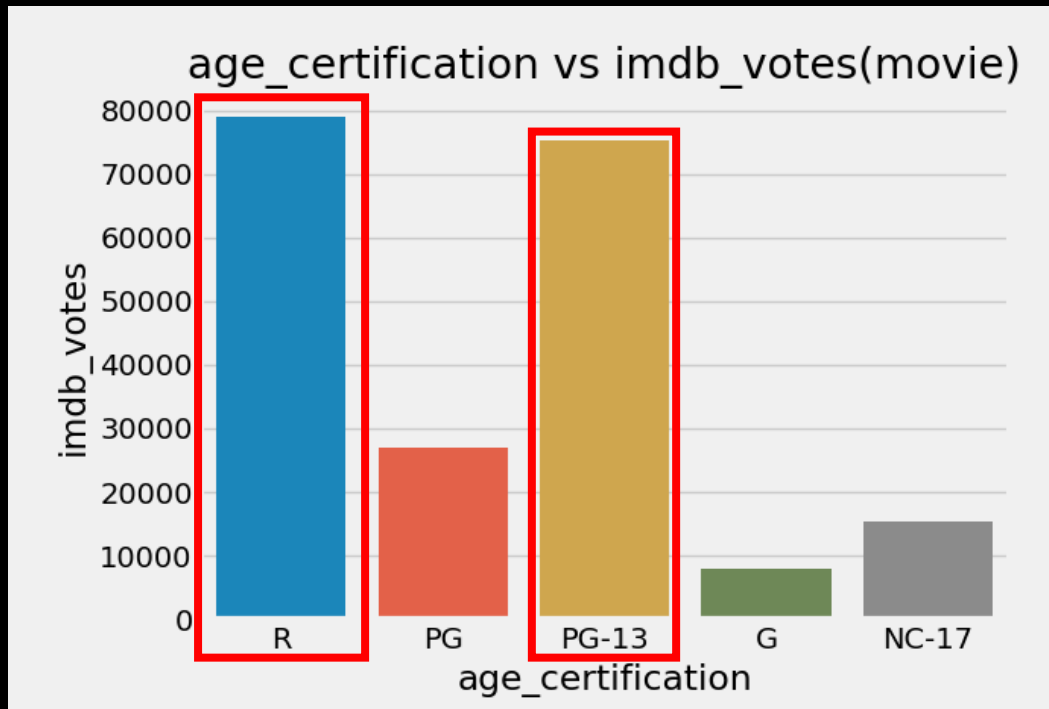
```
movie = df[df['type'] == 'MOVIE']  
show = df[df['type'] == 'SHOW']
```

TV-MA	883
R	556
TV-14	474
PG-13	451
PG	233
TV-PG	188
G	124
TV-Y7	120
TV-Y	107
TV-G	79
NC-17	16



## 가설4: '관객 등급'과 '관객 수' 간의 상관관계가 있을까?

[summary] movie와 show 모두 청소년(만 14세 이상)과 성인(만 18세 이상) 관객 등급에서 높은 관객 수를 보이지만, 성인 관객 등급이 제일 높은 관객 수를 보임

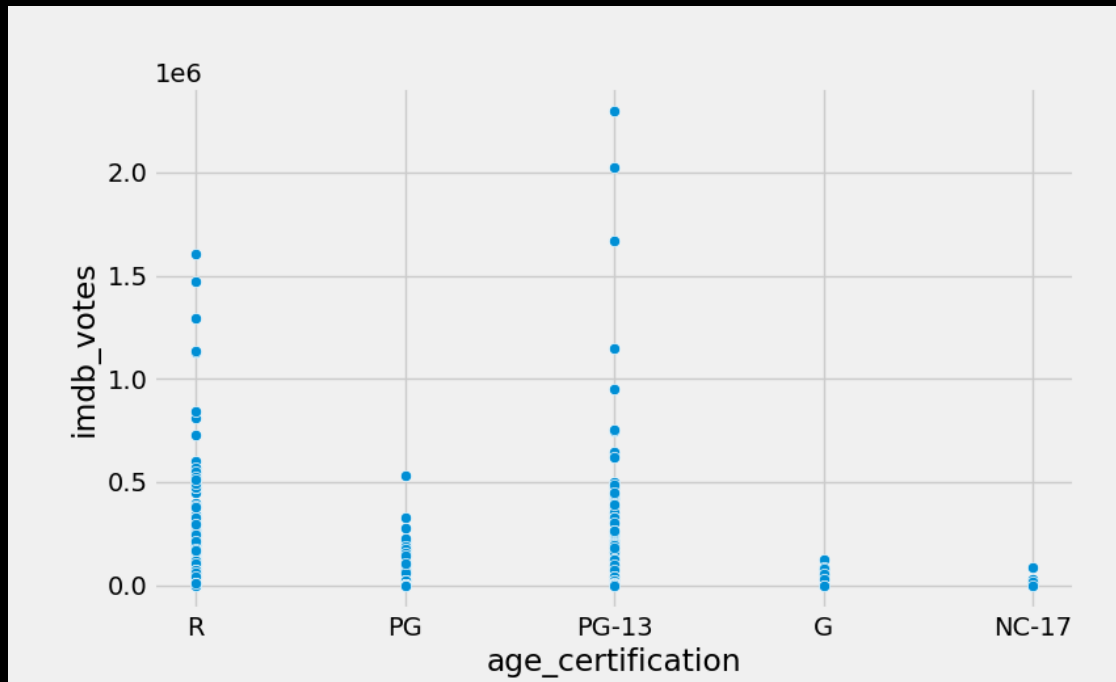




## 가설4: '관객 등급'과 '관객 수' 간의 상관관계가 있을까?

[summary]

movie



show

