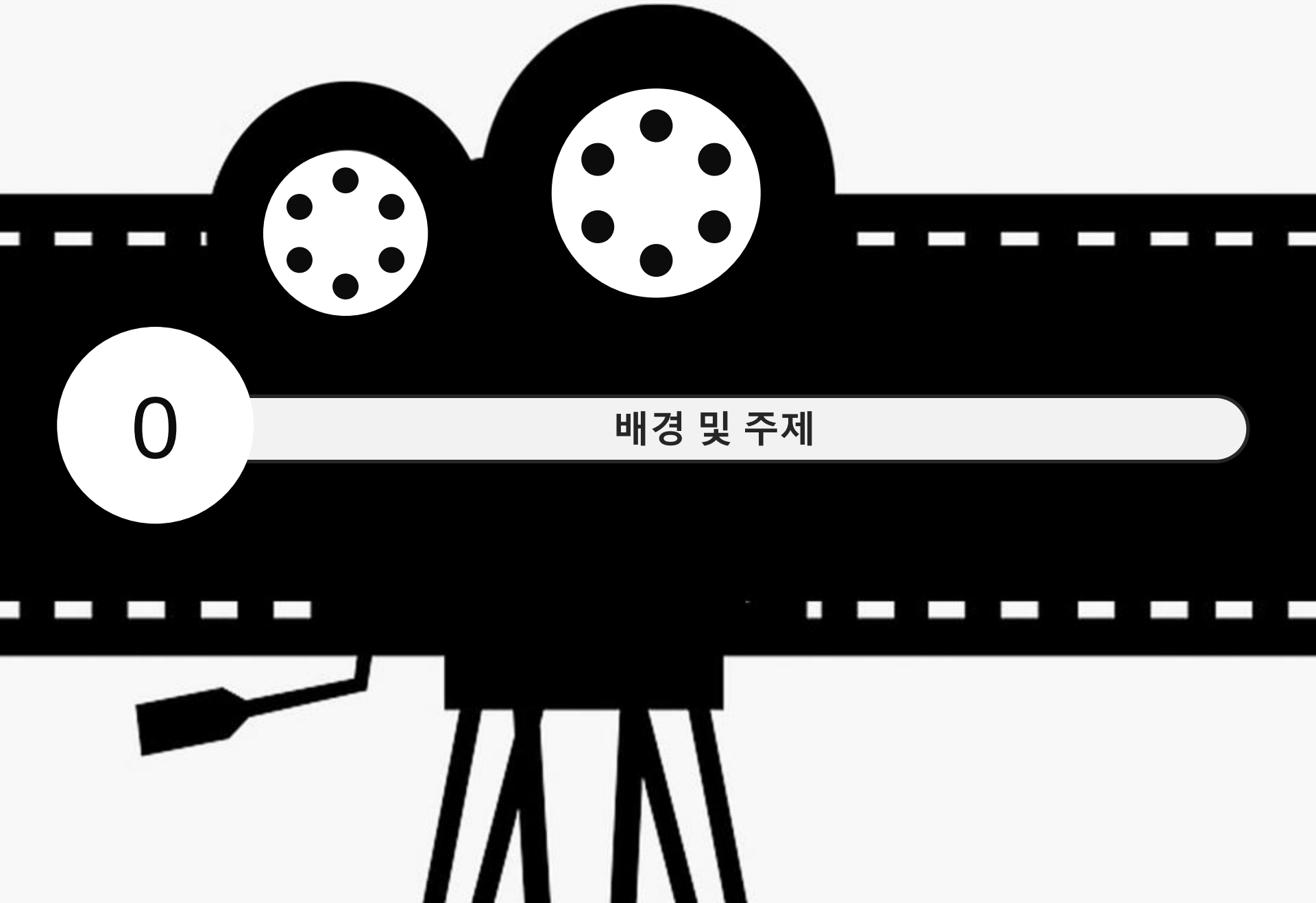
S T A R T

머신러닝 프로젝트 1조

movie

김남규 / 노연우 / 이은지 / 이태기

CONTENTS

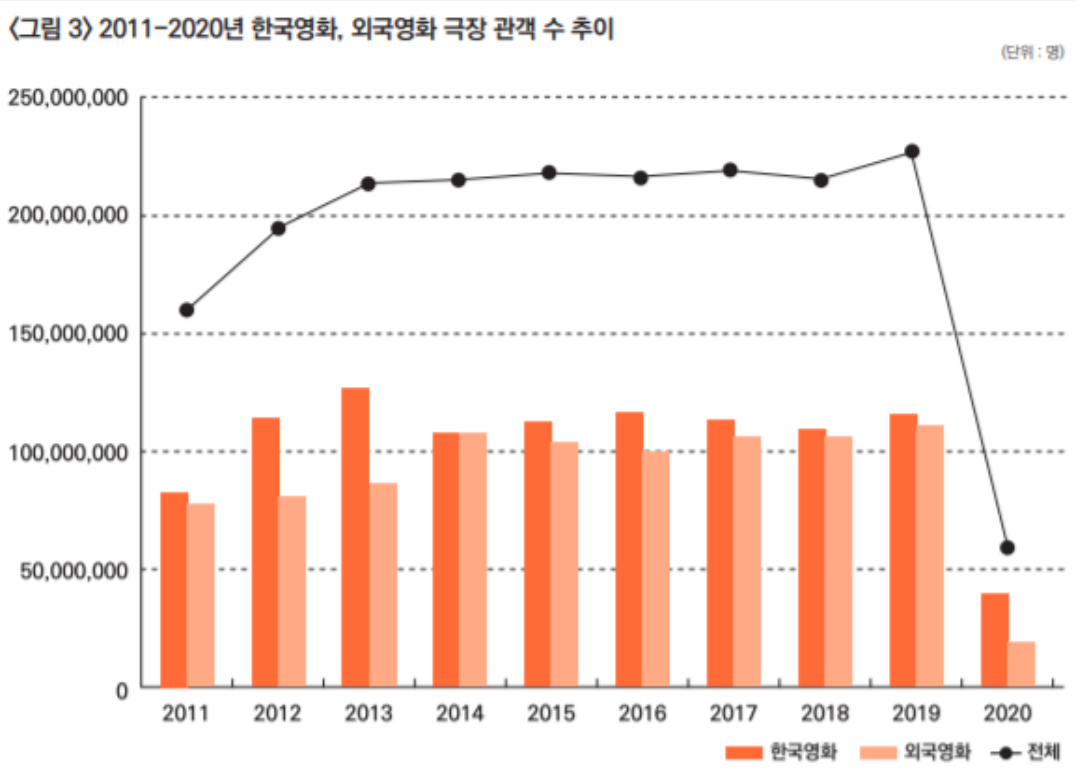


0

배경 및 주제

0 배경

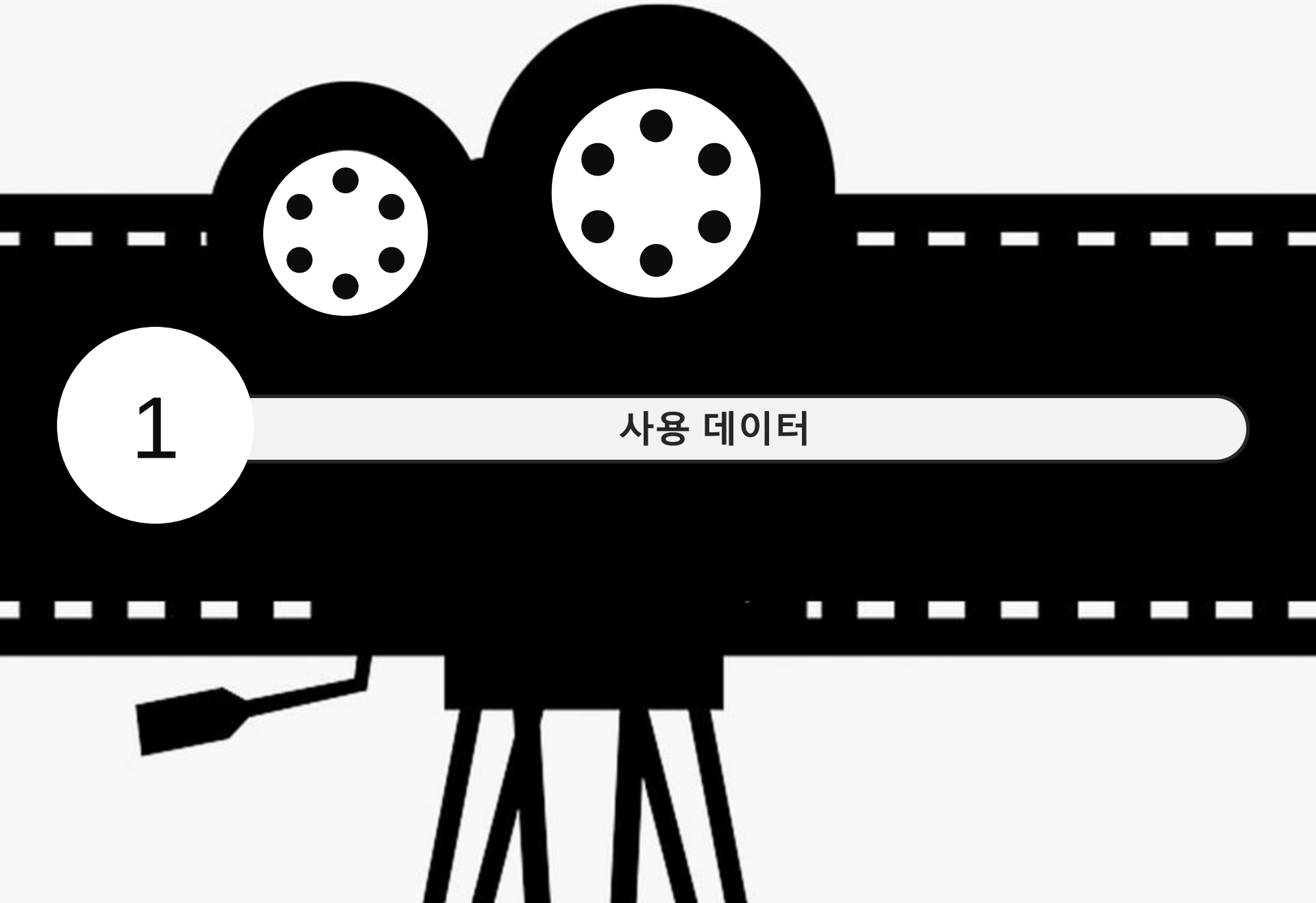
2020년 영화 관객수 전년대비 70% 가량 감소



출처 : 영화진흥위원회

0 주제

코로나19 발생 이전의 영화 데이터로 관객수에 영향을 미치는 요소를 분석하고 **2020년에 개봉한 영화들이 코로나가 없었다면 영화 관객수가 얼마나 되었을지를 예측**해본다.



1 사용 데이터

1 사용 데이터

1. 영화 데이터(영화의 전반적인 정보) –> KOBIS공식 통계

- 2011년 ~ 2020년 데이터(13308개)
- 컬럼(17개) : 영화명, 감독, 제작사, 수입사, 배급사, 개봉일, 영화유형, 영화형태, 국적, 전국스크린수, 전국매출액, 전국관객수, 서울매출액, 서울관객수, 장르, 등급, 영화구분

2. KOBIS에 없는 데이터
-> 네이버 영화 페이지 크롤링 + 네이버 영화 API

- 영화 데이터를 기준으로 크롤링(2414개)
- 컬럼(5개) : 영화명, 주연배우, 평점, 평가자수, 상영시간

*KOBIS : 영화관 입장권 통합 전산망

1 사용 데이터

3. 역대 박스 오피스 Top 300 - KOBIS 공식 통계 파일

- 감독, 배급사, 배우의 흥행 실적을 수치화하기 위해 사용
- 컬럼(8개) : 영화명, 감독, 국적, 전국관객수, 개봉년도, 개봉일, 배우, 배급사

4. 영화 데이터, 네이버 영화 데이터 결합 (2379개)

- 1,2번에서 얻은 데이터를 결합한 형태로 최종 사용 데이터이다.
- 컬럼(15개) : 영화명, 감독(감독_흥행), 배급사(배급사_흥행), 개봉일, 영화형태, 국적, 전국스크린수, 전국관객수, 장르, 등급, 영화구분, 주연배우(주연배우_흥행), 평점, 평가자수, 상영시간

2 탐색적 데이터 분석(EDA)

② 탐색적 데이터 분석(EDA)

> *데이터 과학의 80%는 데이터 클리닝에 소비되고,*
> *나머지 20%는 데이터 클리닝하는 시간을 불평하는데 쓰인다*
>
> Kaggle 창림자 Anthony Goldbloom

② 탐색적 데이터 분석(EDA)

- 데이터 구성

```
Int64Index: 13308 entries, 1 to 13308
Data columns (total 17 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   영화명     13308 non-null  object
 1   감독      12619 non-null  object
 2   제작사     3941 non-null   object
 3   수입사     9095 non-null   object
 4   배급사     13292 non-null  object
 5   개봉일     13308 non-null  datetime64[ns]
 6   영화유형    13308 non-null  object
 7   영화형태    13307 non-null  object
 8   국적      13308 non-null  object
 9   전국스크린수  13308 non-null  int64
 10  전국매출액   13308 non-null  int64
 11  전국관객수   13308 non-null  int64
 12  서울매출액   13302 non-null  float64
 13  서울관객수   13308 non-null  int64
 14  장르      13238 non-null  object
 15  등급      13308 non-null  object
 16  영화구분    13308 non-null  object
```

② 탐색적 데이터 분석(EDA)

- 제작사(9367개), 수입사(4213개)의 경우 결측치가 많기 때문에 삭제
- 다른 컬럼의 결측치는 '기타'로 대체

```
movie.isnull().sum()
```

```
영화명          0
감독          689
제작사        9367
수입사        4213
배급사          16
개봉일           0
영화유형         0
영화형태         1
국적           0
전국스크린수       0
전국매출액        0
전국관객수        0
서울매출액        6
서울관객수        0
장르           70
등급           0
영화구분         0
dtype: int64
```

```
# 결측치 처리
movie = movie.drop(['제작사', '수입사'], axis=1)
movie = movie.drop(['서울매출액', '서울관객수'], axis=1)
movie['감독'].fillna('기타', inplace=True)
movie['장르'].fillna('기타', inplace=True)
movie['배급사'].fillna('기타', inplace=True)
movie['영화형태'].fillna('기타', inplace=True)
```

```
movie.isnull().sum()
```

```
영화명         0
감독          0
배급사         0
개봉일         0
영화유형        0
영화형태        0
국적          0
전국스크린수      0
전국매출액       0
전국관객수       0
장르          0
등급          0
영화구분        0
dtype: int64
```

② 탐색적 데이터 분석(EDA)

- 영화 유형 컬럼은 '개봉영화' 1개의 값이므로 삭제

```
movie['영화유형'].unique()
```

```
array(['개봉영화'], dtype=object)
```

```
movie.drop('영화유형', axis=1, inplace=True)
```

② 탐색적 데이터 분석(EDA)

- 감독, 배급사, 등급에서 값이 복수로 주어진 경우 대표값으로 조정

```
컬럼 [감독] UNIQUE : 6131
컬럼 [배급사] UNIQUE : 905
컬럼 [등급] UNIQUE : 10
```

```
컬럼 [감독] UNIQUE : 5987
컬럼 [배급사] UNIQUE : 606
컬럼 [등급] UNIQUE : 4
```

② 탐색적 데이터 분석(EDA)

- 이상치 제외 : 남은 데이터 2933개

- '전국스크린수', '전국관객수' 기준으로 이상치 제외
- 1,000 < 전국관객수 < 11,000,000
- 전국스크린수 > 50

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2933 entries, 18 to 4697
Data columns (total 12 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   영화명     2933 non-null   object
 1   감독      2933 non-null   object
 2   배급사     2933 non-null   object
 3   개봉일     2933 non-null   datetime64[ns]
 4   영화형태    2933 non-null   object
 5   국적      2933 non-null   object
 6   전국스크린수  2933 non-null   int64
 7   전국매출액   2933 non-null   int64
 8   전국관객수   2933 non-null   int64
 9   장르      2933 non-null   object
 10  등급      2933 non-null   object
 11  영화구분    2933 non-null   object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 297.9+ KB
```

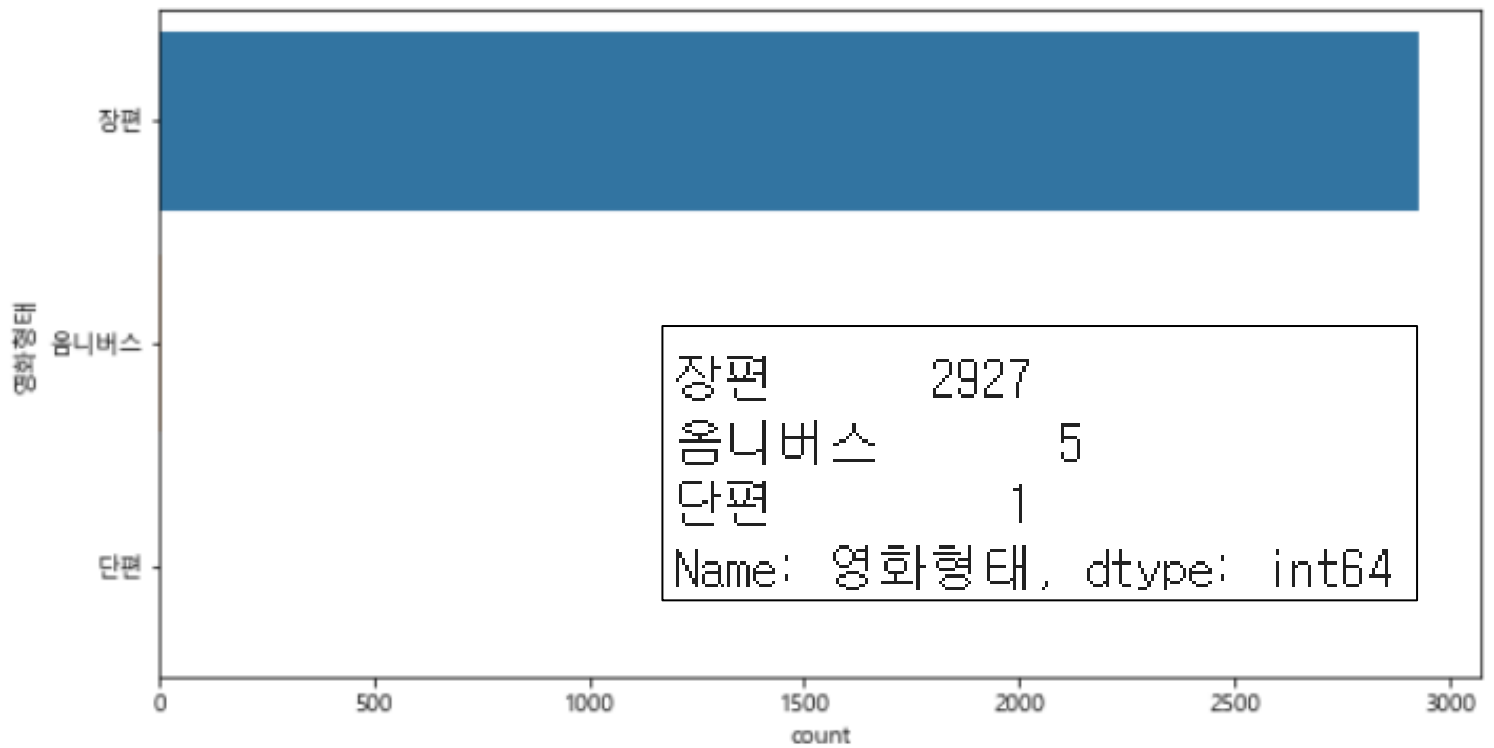
② 탐색적 데이터 분석(EDA)

- 수치형 독립변수간 상관관계 확인


```
movie_resize.drop('전국매출액', axis=1, inplace=True)
```

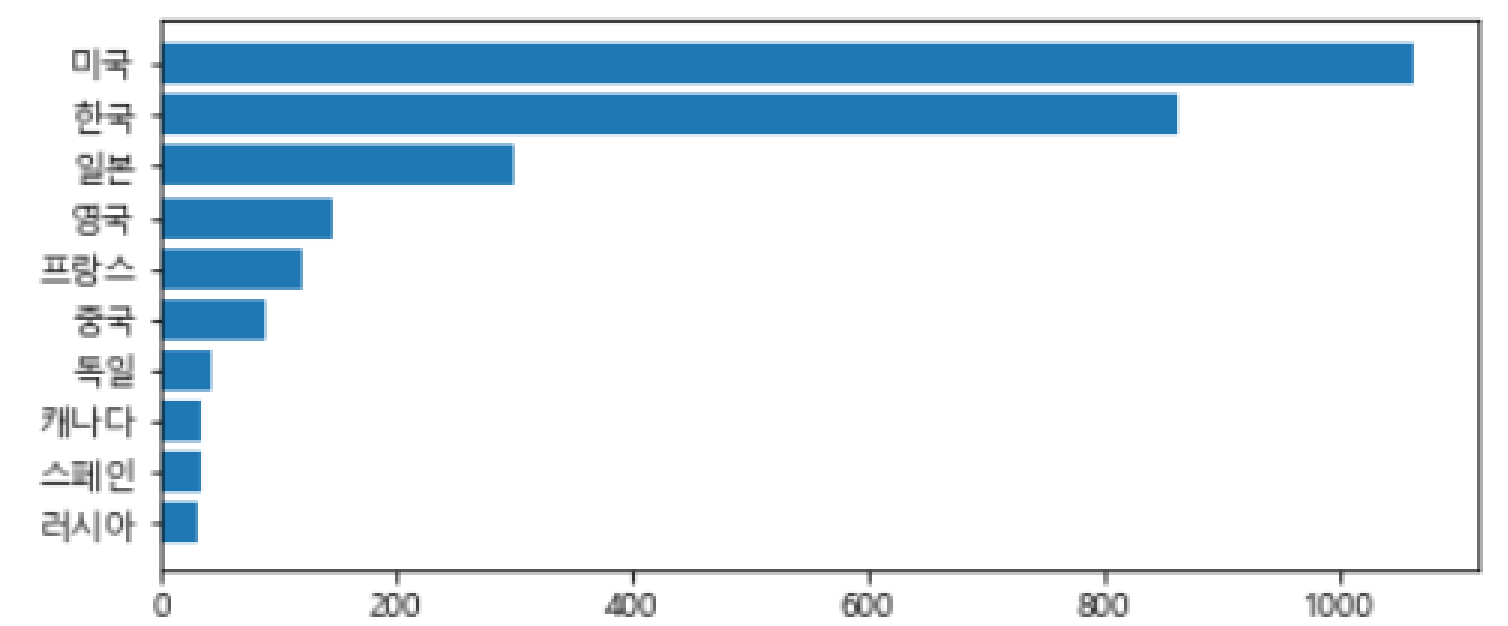
② 탐색적 데이터 분석(EDA)

- 영화 형태



② 탐색적 데이터 분석(EDA)

- 국적 상위 10개



② 탐색적 데이터 분석(EDA)

- 국적의 경우 총 68개의 국가 => Top5 국가 + '기타'로 대체

```
미국      1064
한국       862
일본       300
영국       145
프랑스      119
중국        89
독일        43
스페인       34
캐나다       34
러시아       31
Name: 국적, dtype: int64
```

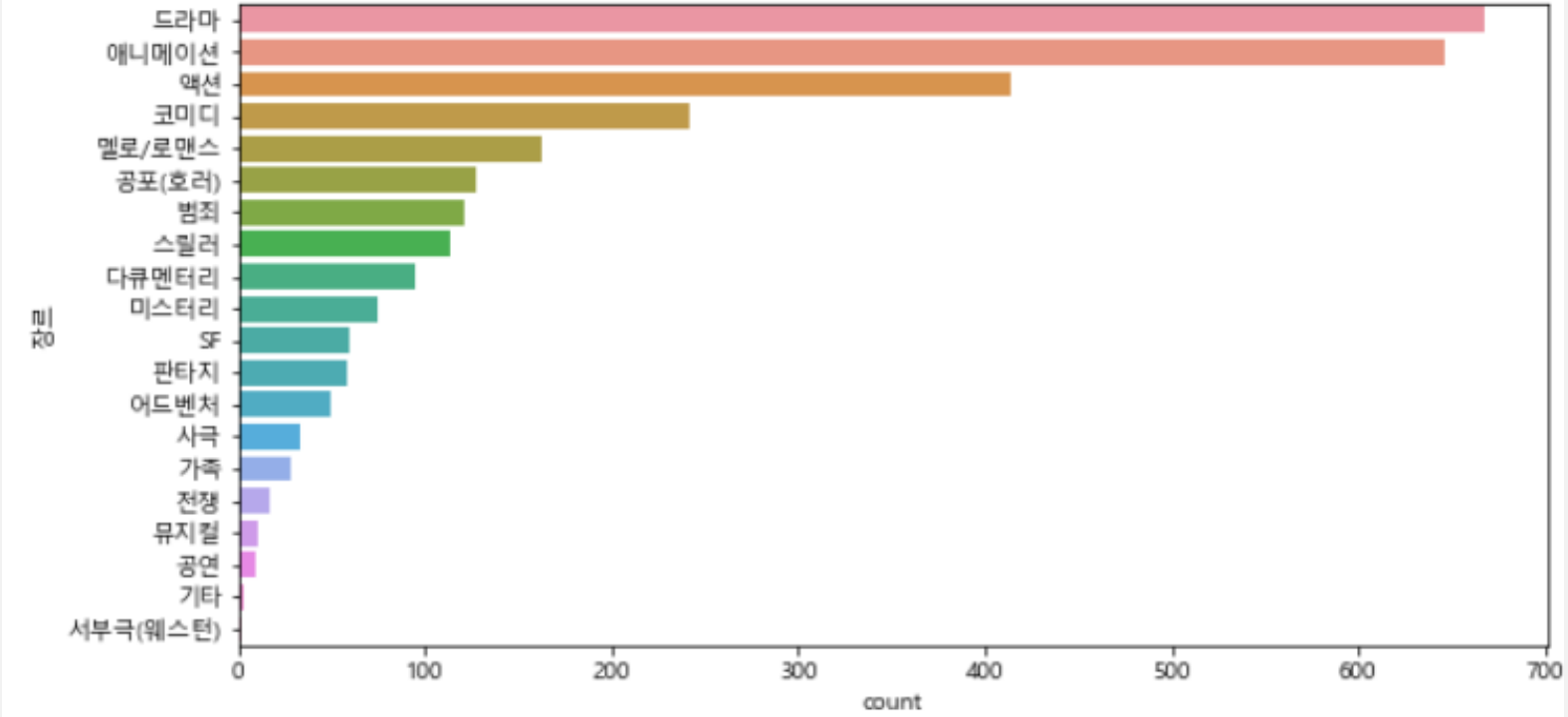
변환전 상위 10개

미국	1064
한국	862
기타	443
일본	300
영국	145
프랑스	119

변환후

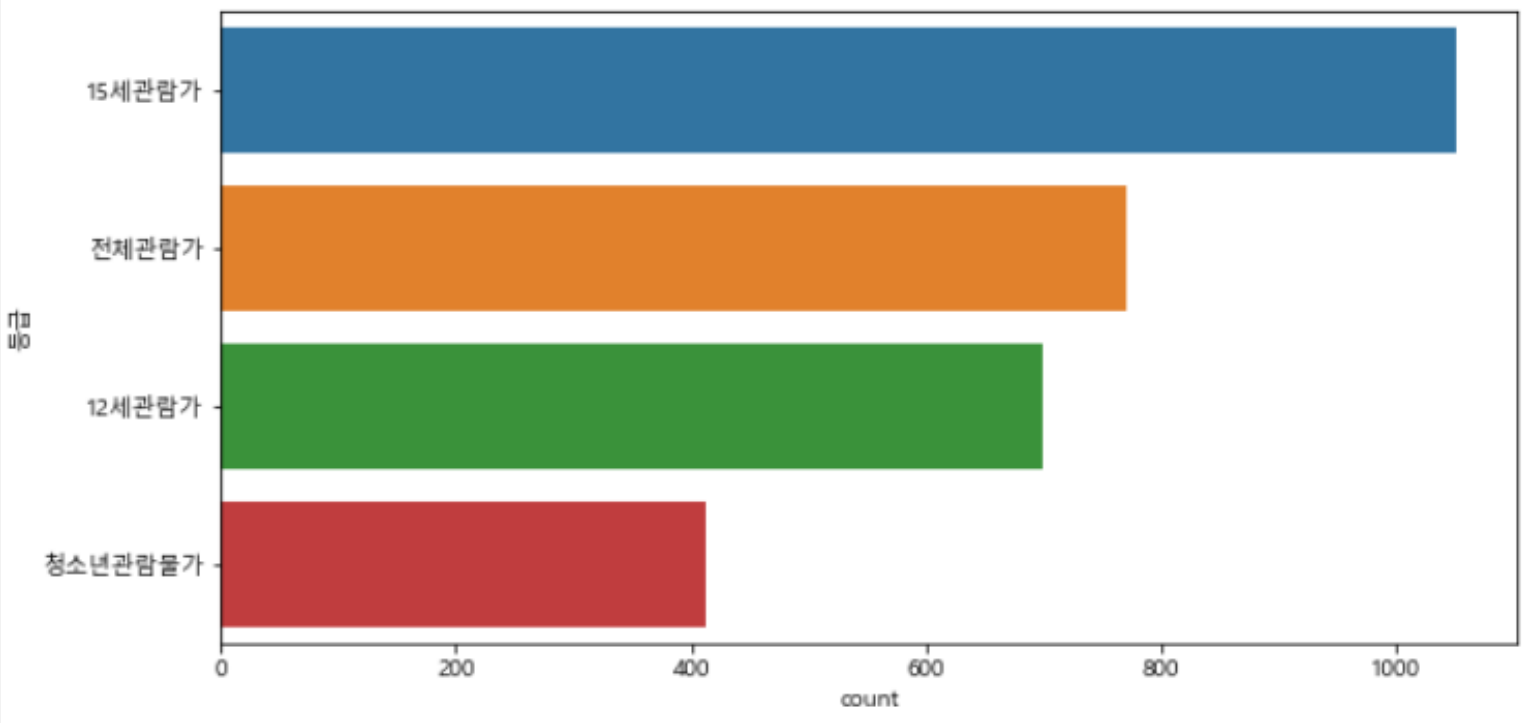
2 탐색적 데이터 분석(EDA)

- 장르



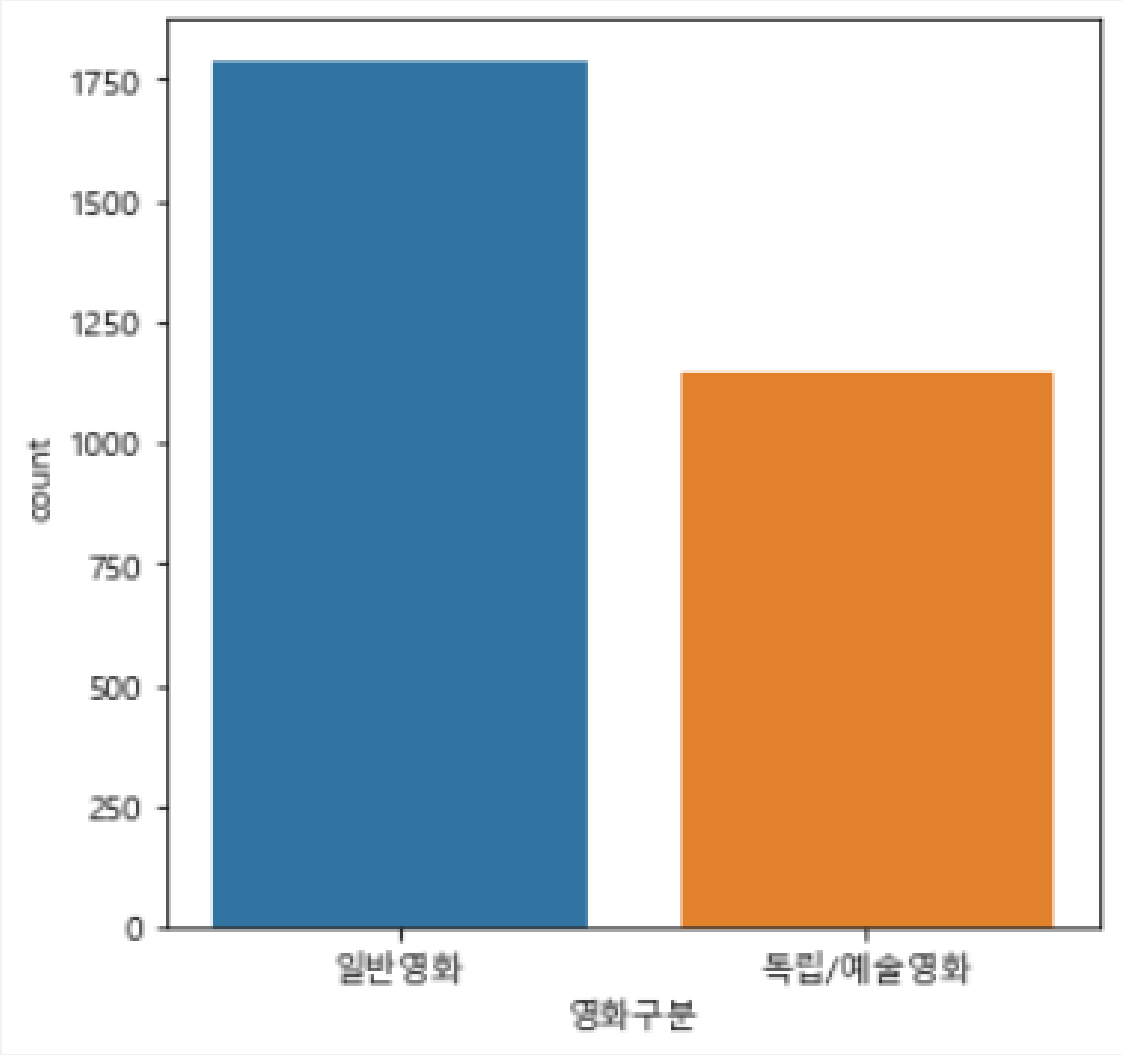
2 탐색적 데이터 분석(EDA)

- 등급



② 탐색적 데이터 분석(EDA)

- 영화 구분



② 탐색적 데이터 분석(EDA)

- 영화형태, 국적, 장르, 등급, 영화구분 -> One-Hot Encoding

```
# One-Hot Encoding
movie_dummy = movie_resize.copy()
movie_labels = pd.get_dummies(movie_dummy,columns = ['영화형태','국적','장르','등급','영화구분'])
movie_labels.head(3)
```

순번	영화명	감독	배급사	개봉일	전국스크린수	전국관객수	영화형태_단편	영화형태_옴니버스	영화형태_장편	국적_기타	...	장르_어드벤처	장르_전쟁	장르_코미디	장르_판타지	등급_12세관람가	등급_15세관람가	등급_전체관람가	등급_청소년관람불가	영화구분_독립/예술영화	영화구분_일반영화
18	어벤져스: 에이지 오브 울트론	조스 웨던	월트디즈니컴퍼니코리아 유한책임회사	2015-04-23	1843	10494499	0	0	1	0	...	0	0	0	0	1	0	0	0	0	1
19	기생충	봉준호	(주)씨제이이엔엠	2019-05-30	1948	10313086	0	0	1	0	...	0	0	0	0	0	1	0	0	0	1
20	겨울왕국	제니퍼 리	소니픽쳐스릴리징월트디즈니스튜디오스코리아(주)	2014-01-16	1010	10296101	0	0	1	0	...	0	0	0	0	0	0	1	0	0	1

3 rows × 41 columns

② 탐색적 데이터 분석(EDA)

- 영화 데이터, 네이버 영화 데이터 결합 : 컬럼 45개

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2379 entries, 0 to 2378
Data columns (total 45 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   영화명            2379 non-null   object
 1   감독              2379 non-null   object
 2   배급사             2379 non-null   object
 3   개봉일             2379 non-null   datetime64[ns]
 4   전국스크린수         2379 non-null   int64
 5   전국관객수          2379 non-null   int64
 6   영화형태_단편        2379 non-null   uint8
 7   영화형태_옴니버스      2379 non-null   uint8
 8   영화형태_장편        2379 non-null   uint8
 9   국적_기타          2379 non-null   uint8
 10  국적_미국          2379 non-null   uint8
 11  국적_영국          2379 non-null   uint8
 12  국적_일본          2379 non-null   uint8
 13  국적_프랑스         2379 non-null   uint8
```

•
•
•

```
 30  장르_액션          2379 non-null   uint8
 31  장르_어드벤처        2379 non-null   uint8
 32  장르_전쟁          2379 non-null   uint8
 33  장르_코미디         2379 non-null   uint8
 34  장르_판타지         2379 non-null   uint8
 35  등급_12세관람가      2379 non-null   uint8
 36  등급_15세관람가      2379 non-null   uint8
 37  등급_전체관람가       2379 non-null   uint8
 38  등급_청소년관람불가     2379 non-null   uint8
 39  영화구분_독립/예술영화  2379 non-null   uint8
 40  영화구분_일반영화      2379 non-null   uint8
 41  주연배우           2379 non-null   object
 42  평점              2379 non-null   float64
 43  평가자수           2379 non-null   int64
 44  상영시간           2379 non-null   int64
dtypes: datetime64[ns](1), float64(1), int64(4), object(4), uint8(35)
memory usage: 285.8+ KB
```

```
41  주연배우          2379 non-null   object
42  평점             2379 non-null   float64
43  평가자수          2379 non-null   int64
44  상영시간          2379 non-null   int64
```

② 탐색적 데이터 분석(EDA)

- 감독_흥행 : 개봉일 이전 영화중 박스 오피스 Top 300 에 있는 가장 높은 순위의 영화의 등수를 이용해 점수화(0~10)

개봉일 이전 영화 기준											
최고순위	미포함	300 ~ 271	270 ~ 241	240 ~ 211	210 ~ 181	180 ~ 151	150 ~ 121	120 ~ 91	90 ~ 61	60 ~ 31	30 ~ 1
점수	0	1	2	3	4	5	6	7	8	9	10

```
[ ]  (movie_final['감독_흥행'] == 0).value_counts()

True     2137
False     242
Name: 감독_흥행, dtype: int64
```

② 탐색적 데이터 분석(EDA)

- 배급사_흥행 : 개봉일 이전 영화중 박스오피스 Top300 에 있는 영화의 수

```
[ ]  (movie_final['배급사_흥행'] == 0).value_counts()
```

```
True     1366
False    1013
Name: 배급사_흥행, dtype: int64
```

② 탐색적 데이터 분석(EDA)

- 배우_흥행 : 각 배우의 개봉일 이전 영화중 박스 오피스 Top 300 에 있는 가장 높은 순위의 영화 등수를 이용해 점수화한 값의 합

개봉일 이전 영화 기준											
최고순위	미포함	300 ~ 271	270 ~ 241	240 ~ 211	210 ~ 181	180 ~ 151	150 ~ 121	120 ~ 91	90 ~ 61	60 ~ 31	30 ~ 1
점수	0	1	2	3	4	5	6	7	8	9	10

```
[ ]  (movie_final['주연배우_흥행'] == 0).value_counts()

     True     1618
     False     761
     Name: 주연배우_흥행, dtype: int64
```

② 탐색적 데이터 분석(EDA)

- 2011년~2019년(2145개) : 데이터 분리

```
movie_2011_2019 = movie_final.query('개봉일 < "2020-01-01"')
print(len(movie_2011_2019))
movie_2011_2019.head(3)
```

2145

	영화명	감독	배급사	개봉일	전국스크린수	전국관객수	영화형태_단편	영화형태_옴니버스	영화형태_장편	국적_기타	...	등급_청소년관람불가	영화구분_독립/예술영화	영화구분_일반영화	주연배우	평점	평가자수	상영시간	감독_흥행	배급사_흥행	주연배우_흥행
0	어벤져스: 에이지 오브 울트론	조스 웨던	월트디즈니컴퍼니코리아 유한책임회사	2015-04-23	1843	10494499	0	0	1	0	...	0	0	1	['로버트 다우니 주니어', '크리스 헴스워스', '마크 러팔로', '크리스 ...	8.32	31015	141	9	0	9
1	기생충	봉준호	(주)씨제이이엔엠	2019-05-30	1948	10313086	0	0	1	0	...	0	0	1	['송강호', '이선균', '조여정', '최우식', '박소담', '이정은'...	8.48	37395	131	10	39	10
2	인터스텔라	크리스토퍼 놀란	워너브러더스 코리아(주)	2014-11-06	1342	10273803	0	0	1	0	...	0	0	1	['마를린 맥코헨', '브라이언 랠리']	9.11	51135	83	8	13	0

3 rows × 48 columns

② 탐색적 데이터 분석(EDA)

- 2020년(234개) : 데이터 분리

```
movie_2020 = movie_final.query('개봉일 >= "20200101"')
print(len(movie_2020))
movie_2020.head(3)
```

```
234
```

	영화명	감독	배급사	개봉일	전국스크린수	전국관객수	영화형태_단편	영화형태_옴니버스	영화형태_장편	국적_기타	...	등급_청소년관람불가	영화구분_독립/예술영화	영화구분_일반영화	주연배우	평점	평가자수	상영시간	감독_흥행	배급사_흥행	주연배우_흥행
70	남산의 부장들	우민호	(주)쇼박스	2020-01-22	1659	4750104	0	0	1	0	...	0	0	1	['이병헌', '이성민', '곽도원', '이희준', '김소진']	7.47	18168	114	9	40	10
83	다만 악에서 구하소서	홍원찬	(주)씨제이이엔엠	2020-08-05	1998	4352669	0	0	1	0	...	0	0	1	['황정민', '이정재', '박정민', '박소이']	7.61	21397	108	0	42	10
99	반도	연상호	(주)넥스트엔터테인먼트월드(NEW)	2020-07-15	2575	3812080	0	0	1	0	...	0	0	1	['강동원', '이정현']	5.64	31323	116	10	19	10

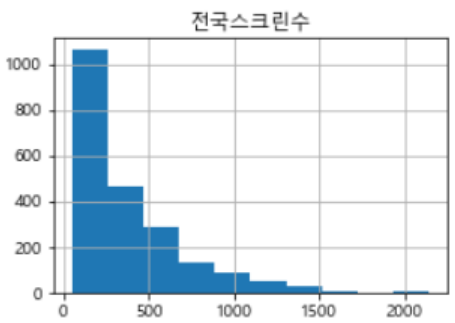
3 rows × 48 columns

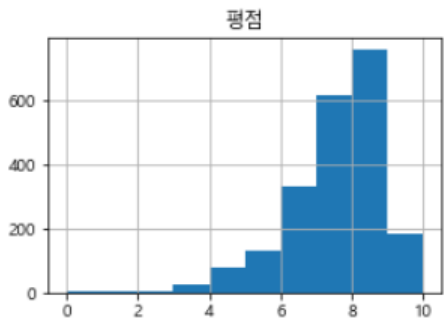
② 탐색적 데이터 분석(EDA)

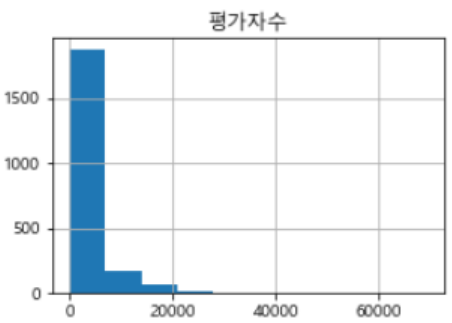
1. 2011년~2019년의 수치형 컬럼의 상관관계 확인

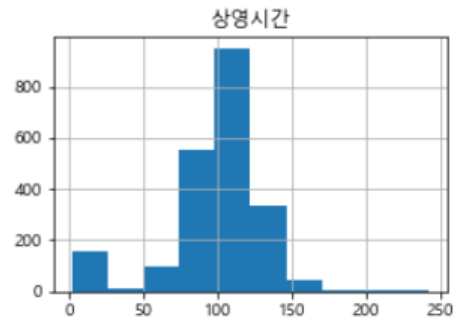
② 탐색적 데이터 분석(EDA)

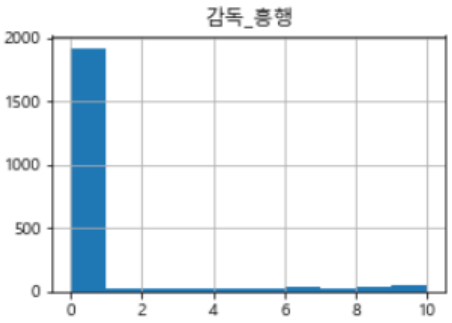
2. 2011년~2019년의 수치형 데이터 히스토그램

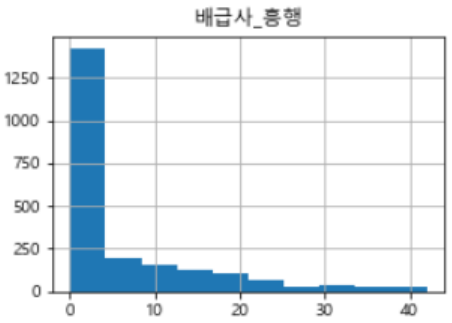


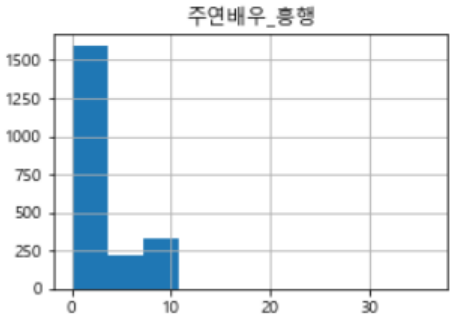












② 탐색적 데이터 분석(EDA)

3. 강한 상관관계를 지닌 “전국스크린수, 평가자수” 2020년과 비교

2011~2019년

	전국스크린수
10	2142
30	2027
38	1972
18	1965
31	1957

	전국스크린수
count	2145.000000
mean	377.000000
std	339.324693
min	51.000000
25%	124.000000
50%	265.000000
75%	524.000000
max	2142.000000

2020년

	전국스크린수
99	2575
229	2228
254	2137
83	1998
237	1882

	전국스크린수
count	234.000000
mean	384.273504
std	444.270683
min	51.000000
25%	105.000000
50%	196.000000
75%	464.500000
max	2575.000000

② 탐색적 데이터 분석(EDA)

3. 강한 상관관계를 지닌 “전국스크린수, 평가자수” 2020년과 비교

2011~2019년

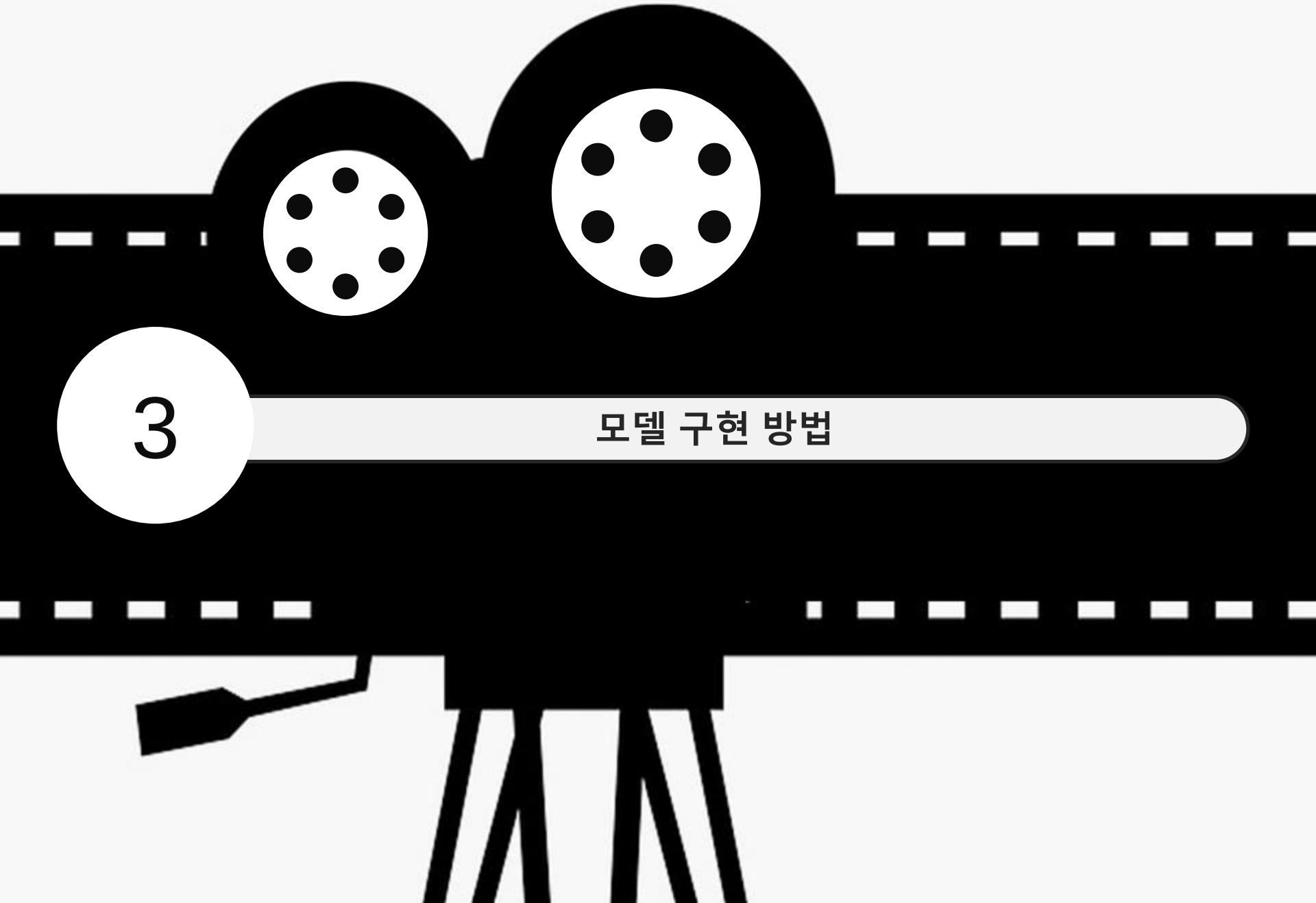
	평가자수
22	69234
30	53590
143	51611
2	51135
103	50563

	평가자수
count	2145.000000
mean	3076.002331
std	5773.756307
min	0.000000
25%	238.000000
50%	907.000000
75%	3277.000000
max	69234.000000

2020년

	평가자수
99	31323
83	21397
237	18648
70	18168
254	18049

	평가자수
count	234.000000
mean	1407.927350
std	3582.208541
min	4.000000
25%	96.500000
50%	208.500000
75%	748.750000
max	31323.000000



3 모델 구현 방법

3 모델 구현 방법

- 데이터 구분

1. 2011년 ~ 2019년 : Train(1716개) 0.8, Test(429개) 0.2
2. 2020년: 234개

- 독립변수(대표 컬럼 11개, 전체 컬럼 41개) :
감독_흥행, 배급사_흥행, 주연배우_흥행, 전국스크린수, 평점, 상영시간, 장르(20개), 국적(6개), 등급(4개), 영화형태(3개), 영화구분(2개)

- 종속변수(1개) : 전국관객수(단위 1000명)

3 모델 구현 방법

- 스케일러 : StandardScaler

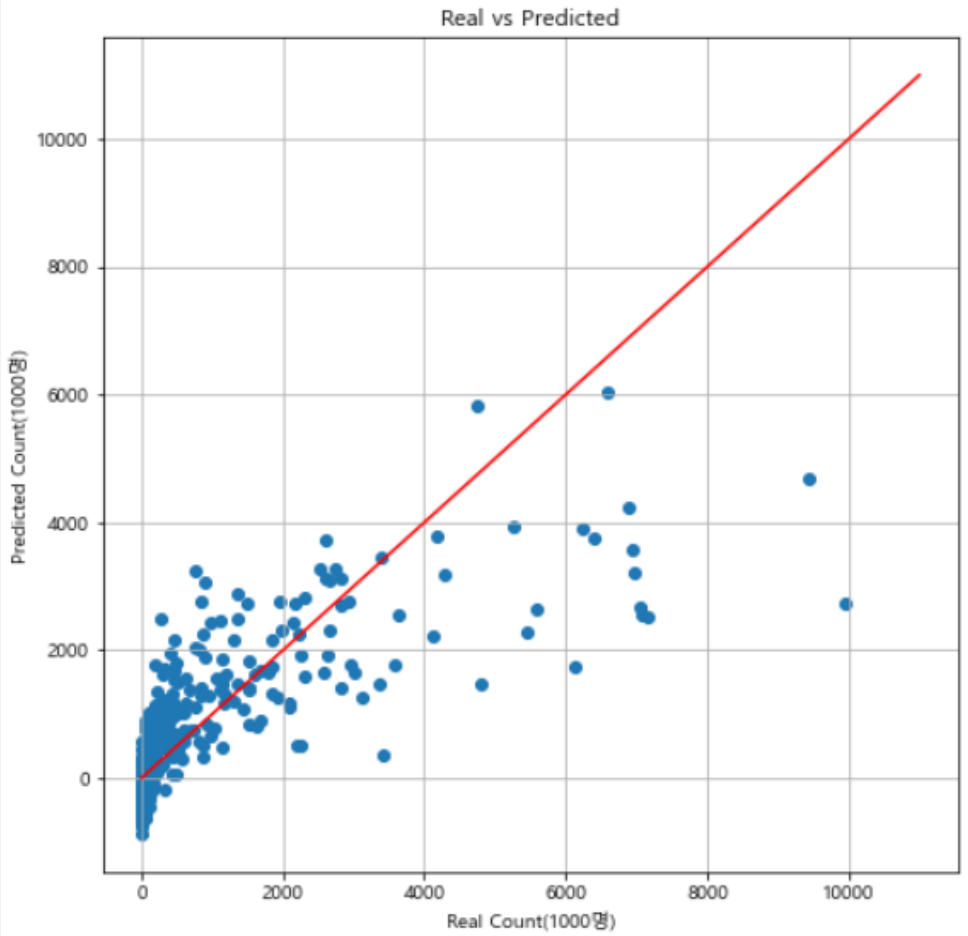
- 회귀 모델 :
  1. LinearRegression
  2. RandomForestRegressor
  3. GradientBoostingRegressor

- 평가 : RMSE(Root Mean Square Error)

4 모델 구현 결과

4 모델 구현 결과

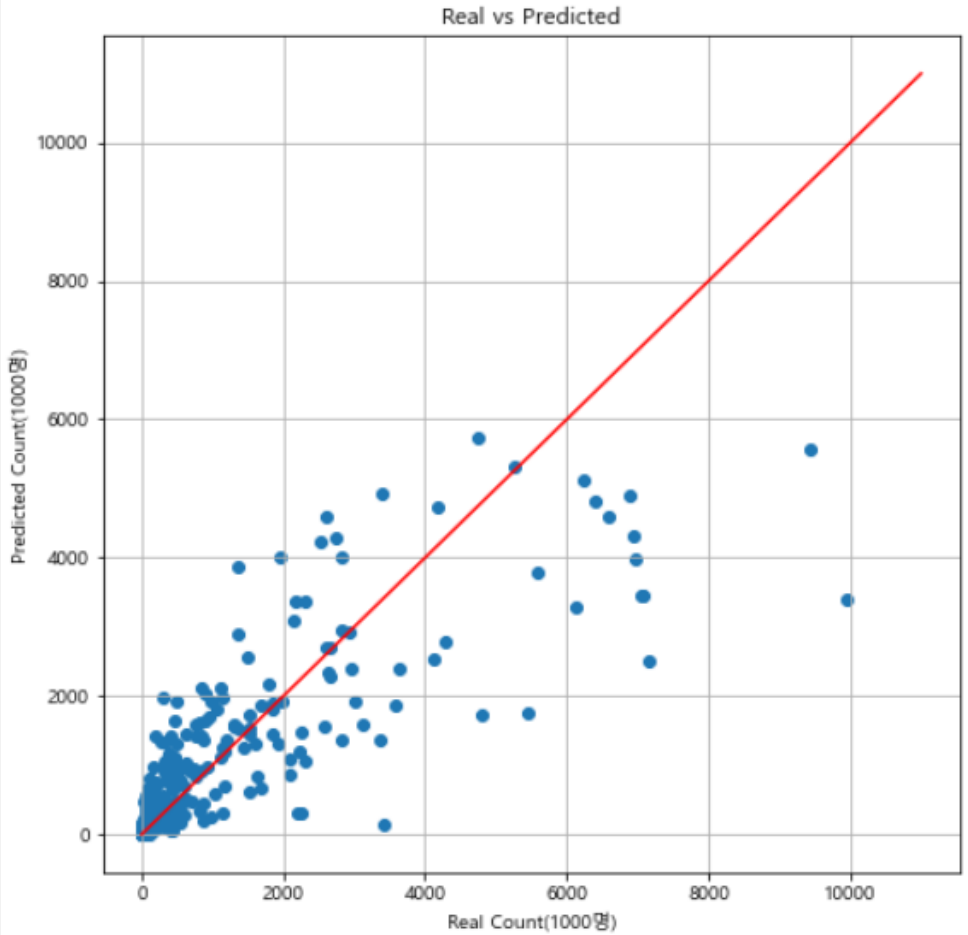
선형 회귀 모델



RMSE of Train Data : 854.4325900648541
RMSE of Test Data : 946.8134316072397

4 모델 구현 결과

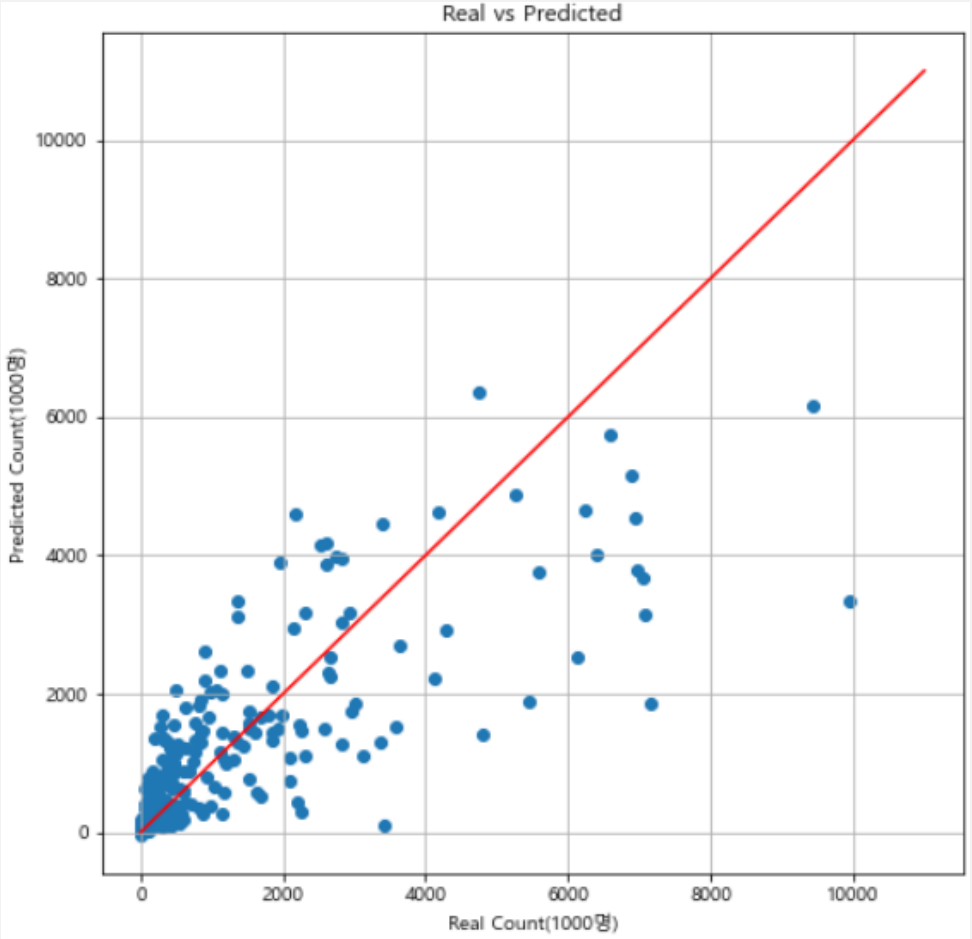
랜덤 포레스트 회귀 모델



RMSE of Train Data : 651.2508142530651
RMSE of Test Data : 812.5885202895471

4 모델 구현 결과

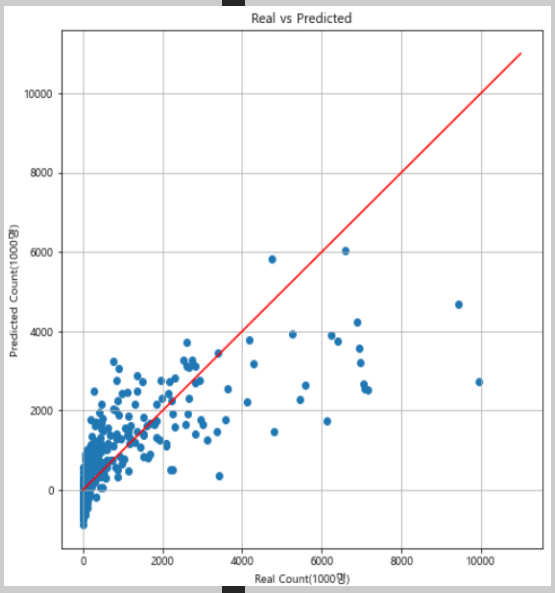
GBM 회귀 모델



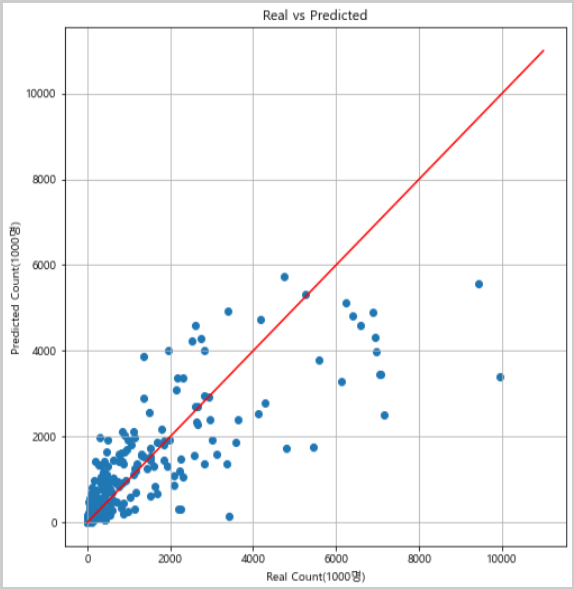
RMSE of Train Data : 643.3518769983194
RMSE of Test Data : 842.8098430892501

4 모델 구현 결과

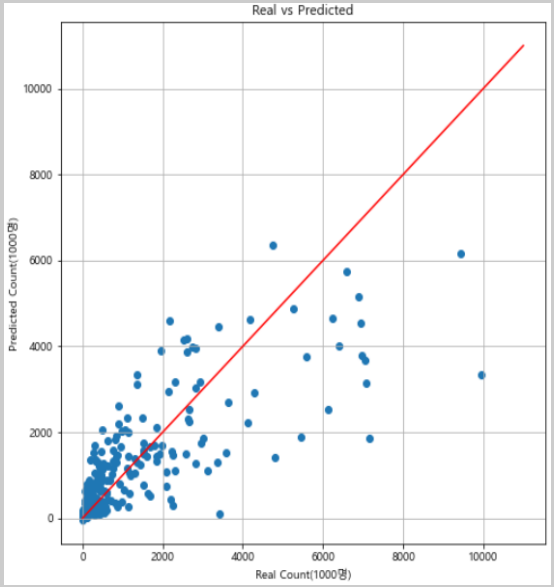
선형 회귀 모델



랜덤포레스트 회귀 모델



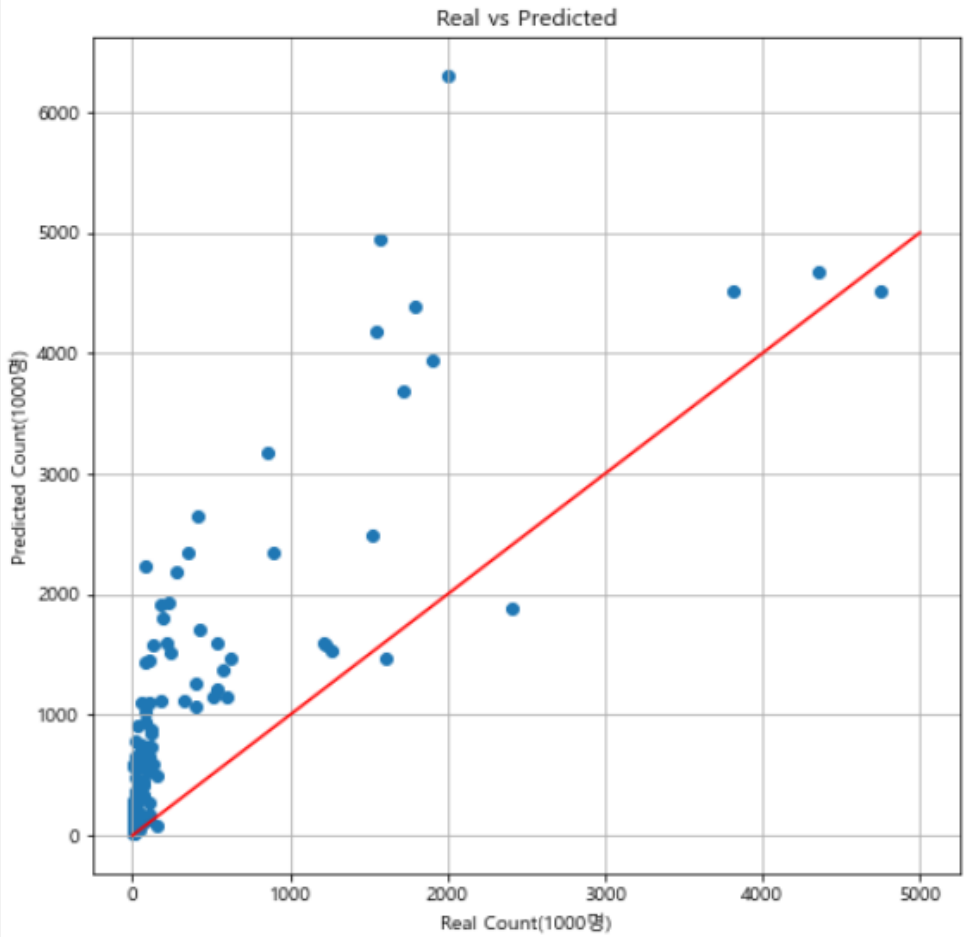
GBM 회귀 모델



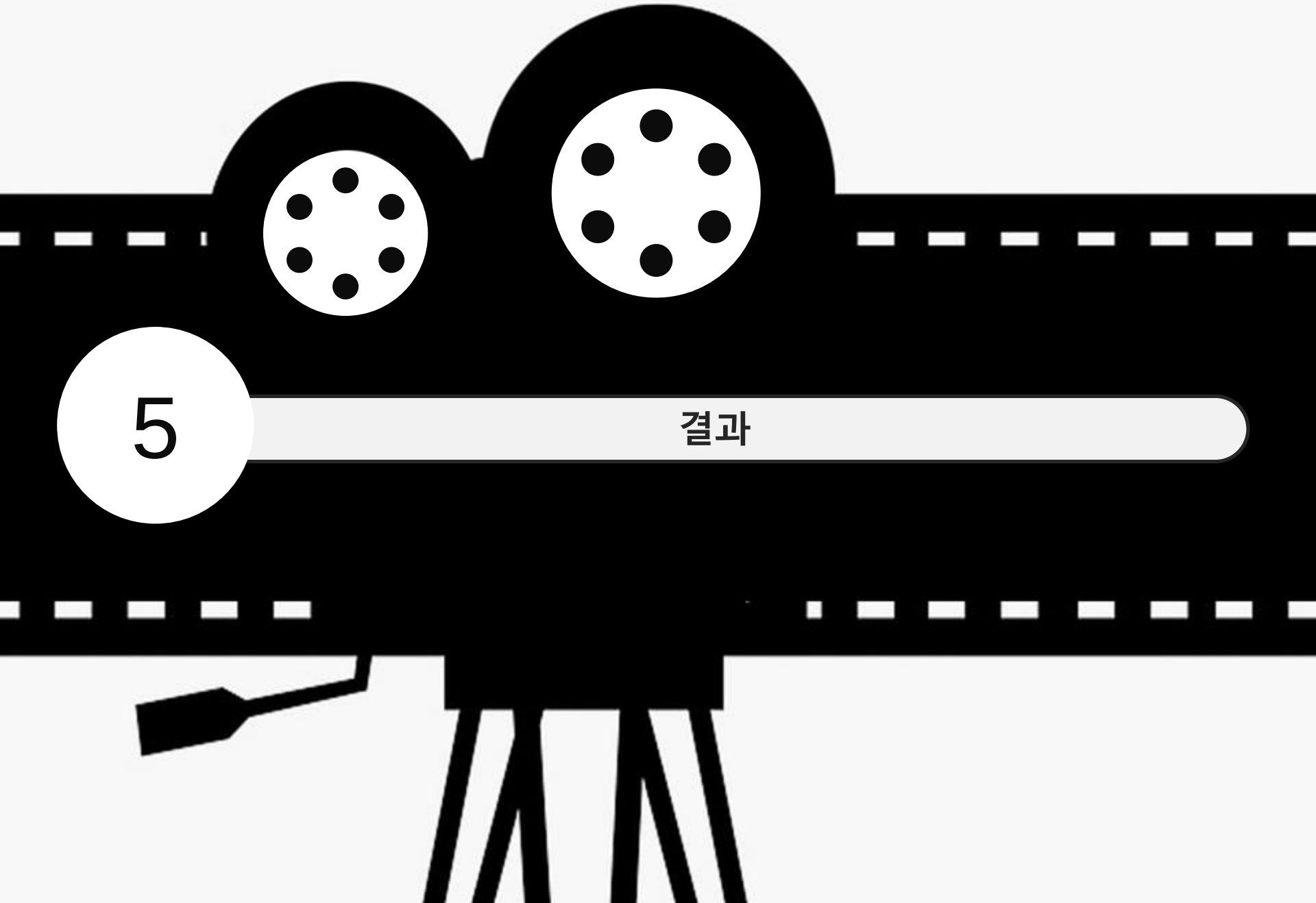
	Train RMSE	Test RMSE
Linear Regression	854.4	946.8
Random Forest	651.3	812.6
Gradient Boosting	643.4	842.8

4 모델 구현 결과

2020년 예측



RMSE of 2020 Data : 716.4146786077532



5 결과

5 결과

관객수 예측 TOP 10

	영화명	감독	개봉일	전국관객수	예측관객수	diff
229	테넷	크리스토퍼 놀란	2020-08-26	1998987	6307343	4308356
295	삼진그룹 영어토익반	이종필	2020-10-21	1571774	4945090	3373316
83	다만 악에서 구하소서	홍원찬	2020-08-05	4352669	4682867	330198
99	반도	연상호	2020-07-15	3812080	4516461	704381
70	남산의 부장들	우민호	2020-01-22	4750104	4509524	-240580
254	강철비2: 정상회담	양우석	2020-07-29	1790797	4384579	2593782
300	도굴	박정배	2020-11-04	1543813	4177462	2633649
237	#살아있다	조일형	2020-06-24	1903703	3934443	2030740
267	담보	강대규	2020-09-29	1719592	3687853	1968261
454	작은 아씨들	그레타 거윅	2020-02-12	859072	3172867	2313795

5 결과

코로나의 영향을 받은 2020년 영화 관객 수의 평균은 **19.9만명** 이었는데

코로나가 없었던 10년의 데이터로 학습해 2020년 영화들의 평균 관객을 예측한결과

56.9만명으로 3배 정도 많았을 것이라고 예측했다.

	전국관객수
count	2145
mean	672788
std	1395798
min	1303
25%	19846
50%	89358
75%	560620
max	10494499

2011~2019년

	전국관객수	예측관객수	diff
count	234	234	234
mean	199249	569805	370556
std	606490	1002373	614452
min	1140	23243	-520430
25%	4817	50617	40762
50%	15220	135770	113482
75%	72769	589181	461010
max	4750104	6307343	4308356

2020년

QUESTION & ANSWER

질문이 있으시면, 자유롭게 말씀해주세요!


THANK YOU FOR LISTENING