머신러닝을 활용한 영화 관객수 예측

2팀

팀장 : 최서*

수행 역할

- 데이터 크롤링
- 통계 검증
- 머신러닝 모델링

팀원 : 박종*

수행 역할

- 데이터 수집
- 데이터 전처리
- 머신러닝 모델링

팀원 : 오승*

수행 역할

- 데이터 정제
- 탐색적 데이터 분석
- 데이터 시각화

팀원 : 한대*

수행 역할

- 데이터 전처리
- 데이터 크롤링
- 탐색적 데이터 분석

프로젝트 주제

국내 개봉 영화 Box_Office 데이터를 활용한

Data_Analysis와 흥행 예측 ML 모델 개발

목적(비즈니스 기획 의도)

- 국내 영화 시장의 규모와 수요 및 소비 경향을 파악할 수 있음
- 영화 흥행 예측 모델이 개봉 전/후 의사결정에 긍정적 역할을 기대
- 국내 영화 산업의 경쟁력과 신성장 동력의 탐색 기회 제공

프로젝트 주제

- 영화 제작사, 배급사, 매니지먼트 회사, OTT 서비스 회사 등
  영화 산업에 일조하고 흥행할 것을 기대하는 목적으로 투자하는 기업체
- 영화 데이터와 영화 흥행 분석에 관심 있는 일반인, 소비자 등

목적(비즈니스 기획 의도)

- 기업의 입장에서 예측 결과에 따른 사용자의 영화 제작 전반에 걸친
  투자 규모 결정이나 제작 비용과 자원의 효율적 사용 기대
- 소비자 입장에서 흥행성 짙은 작품을 확인함으로써 관람 후 만족도 상승할 것으로 기대

index

1. Describe Dataset
2. Data Crawling & Scraping
3. EDA & Preprocessing
4. Machine Learning Algorithm
5. Results & Project Evaluation

1. Describe Dataset

1) 데이터 출처

영화진흥위원회의 영화관 입장권 통합 전산망에 등록된
2003년부터 2021년까지 총 19년간의
박스오피스 데이터 사용

1. Describe Dataset

2) 데이터 형식

Raw Data

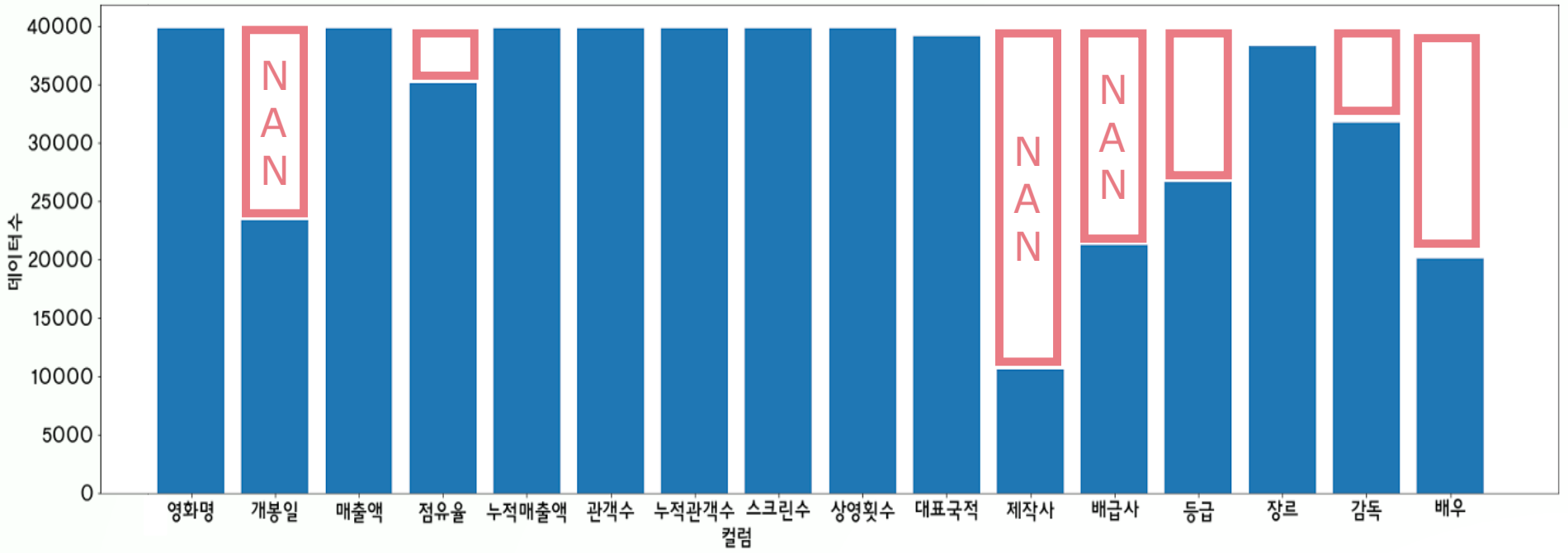
영화명	개봉일	매출액	점유율	누적매출액	관객수	누적관객수	스크린수
도둑들	2012-07-25	93,664,808,500	3.50%	93,664,808,500	12,983,178	12,983,178	1,091
광해, 왕이	2012-09-13	88,899,448,769	3.30%	88,899,448,769	12,319,390	12,319,390	1,001
트랜스포머	2011-06-29	74,840,681,500	2.80%	74,840,681,500	7,784,743	7,784,743	1,409

상영횟수	대표국적	제작사	배급사	등급	장르	감독	배우
155,392	한국	(주)케이퍼	(주)쇼박스	15세이상관	액션	최동훈	김윤석,김혜수,이정재,전지현
203,400	한국	리얼라이즈	(주)씨제이	15세이상관	사극	추창민	이병헌,류승룡,한효주,장광,김
124,450	미국		(주)씨제이	12세이상관	액션	마이클 베	조쉬 더하멜,존 말코비치,로지

1. Describe Dataset

2) 데이터 형식

Raw Data



Box Office Data 4만 개

총 변수 16 개

결측 값 보정 데이터 타입 변환 필요

2. Data Crawling & Scraping

출처 : 네이버
영화

블랙위도우 상영중

Black Widow, 2021

관람객 9.07 기자·평론가 6.70

네티즌 8.49 내 평점 등록

개요 액션, 모험, SF | 미국 | 134분 | 2021.07.07 개봉

감독 케이트 쇼트랜드

출연 스칼릿 조핸슨(나타샤 로마노프 / 블랙 위도우), 플로렌스 퓨(옐… 더보기

등급 [국내] 12세 관람가

흥행 누적관객 2,952,003명(08.20 기준)

성별·나이별 관람추이

```
<dd>
  <p>
    <a href="/movie/bi/pi/basic.naver?code=3253">케이트 쇼트랜드</a> == $0
    <!-- N=a:ifo.director -->
  </p>
</dd>
<dt class="step3">…</dt>
<dd>
  <p>
    <a href="/movie/bi/pi/basic.naver?code=5027">스칼릿 조핸슨</a>
    <!-- N=a:ifo.actor -->
    "(나타샤 로마노프 / 블랙 위도우), "
    <a href="/movie/bi/pi/basic.naver?code=368146">플로렌스 퓨</a>
    <!-- N=a:ifo.actor -->
    "(옐레나 벨로바), "
    <a href="/movie/bi/pi/basic.naver?code=2076">레이첼 와이즈</a>
    <!-- N=a:ifo.actor -->
    "(멜리나 보스토코프)"
  </p>
  <a href="detail.naver?code=184318" title="더보기" class="more">더보기</a>
  <!-- N=a:ifo.actmore -->
</dd>
<dt class="step4">…</dt>
<dd>
  <p>
    " [국내] "
    <a href="/movie/sdb/browsing/bmovie.naver?grade=1001002">12세 관람가</a>
    <!-- N=a:ifo.filmrate -->
```

… .basic div#container div#content div.article div.mv_info_area div.mv_info dl.info_spec dd p a

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

2. Data Crawling & Scraping

블랙위도우 상영중

Black Widow, 2021

관람객 9.07 기자·평론가 6.70

네티즌 8.49 내 평점 등록

개요 액션, 모험, SF | 미국 | 134분 | 2021.07.07 개봉

감독 케이트 쇼트랜드

출연 스칼릿 조핸슨(나타샤 로마노프 / 블랙 위도우), 플로렌스 퓨(옐… 더보기

등급 [국내] 12세 관람가

흥행 누적관객 2,952,003명(08.20 기준)

성별·나이별 관람추이

```
<dd>
  <p>
    <a href="/movie/bi/pi/basic.naver?code=3253">케이트 쇼트랜드</a> == $0
    <!-- N=a:ifo.director -->
  </p>
</dd>
<dt class="step3">…</dt>
<dd>
  <p>
    <a href="/movie/bi/pi/basic.naver?code=5027">스칼릿 조핸슨</a>
    <!-- N=a:ifo.actor -->
    "(나타샤 로마노프 / 블랙 위도우), "
    <a href="/movie/bi/pi/basic.naver?code=368146">플로렌스 퓨</a>
    <!-- N=a:ifo.actor -->
    "(옐레나 벨로바), "
    <a href="/movie/bi/pi/basic.naver?code=2076">레이첼 와이즈</a>
    <!-- N=a:ifo.actor -->
    "(멜리나 보스토코프)"
  </p>
  <a href="detail.naver?code=184318" title="더보기" class="more">더보기</a>
  <!-- N=a:ifo.actmore -->
</dd>
<dt class="step4">…</dt>
<dd>
  <p>
    " [국내] "
    <a href="/movie/sdb/browsing/bmovie.naver?grade=1001002">12세 관람가</a>
    <!-- N=a:ifo.filmrate -->
```

… .basic div#container div#content div.article div.mv_info_area div.mv_info dl.info_spec dd p a

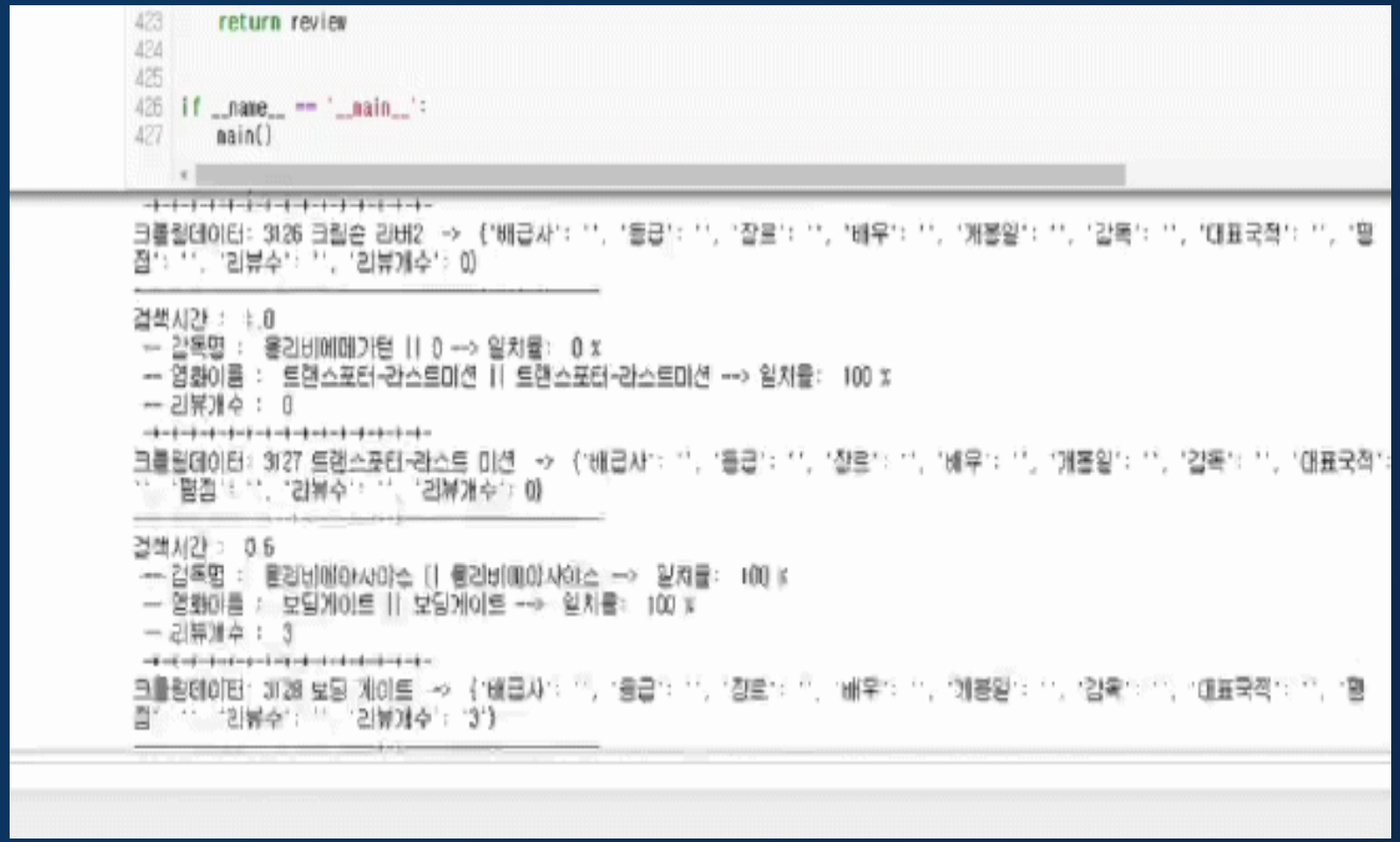
Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

2. Data Crawling & Scraping

Ver 1.
개봉일 / 감독 / 장르 /등급
배우 / 배급사 / 대표국적

Ver 2.
평점 / 평가자 수

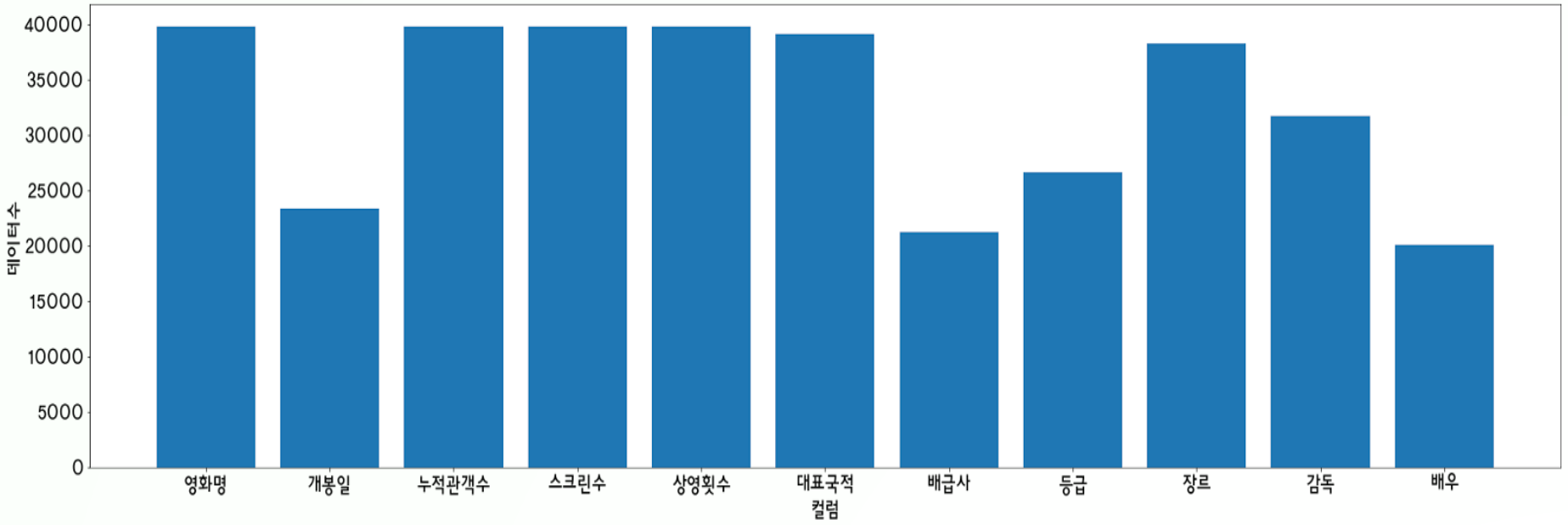
Ver 3.
리뷰 평가자 수



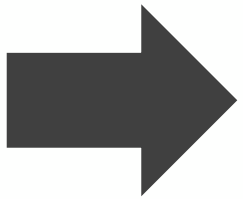
3. EDA & Preprocessing

1) 데이터 정제

Raw Data

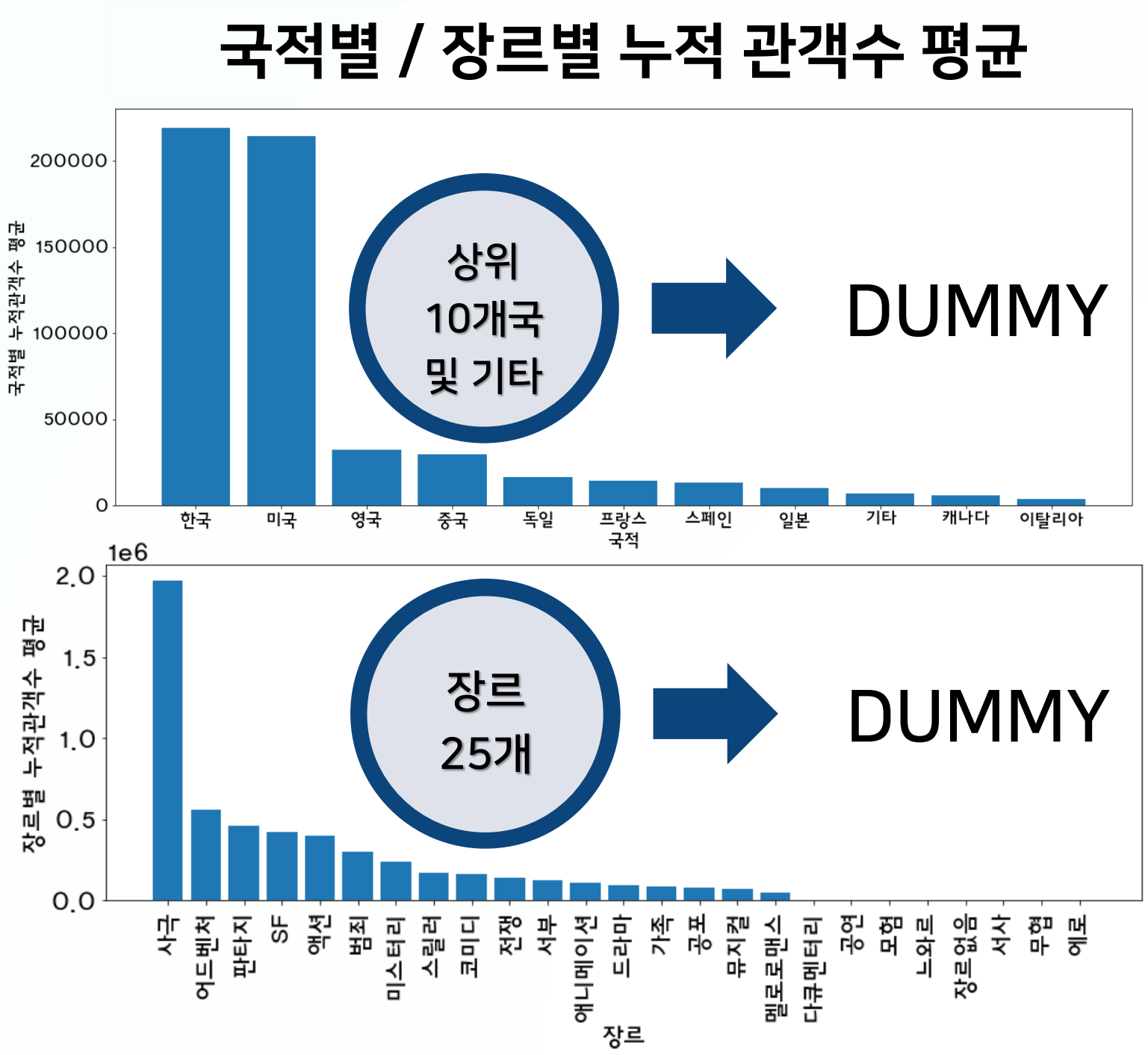


점유율 매출액
제작사 관객수



3. EDA & Preprocessing

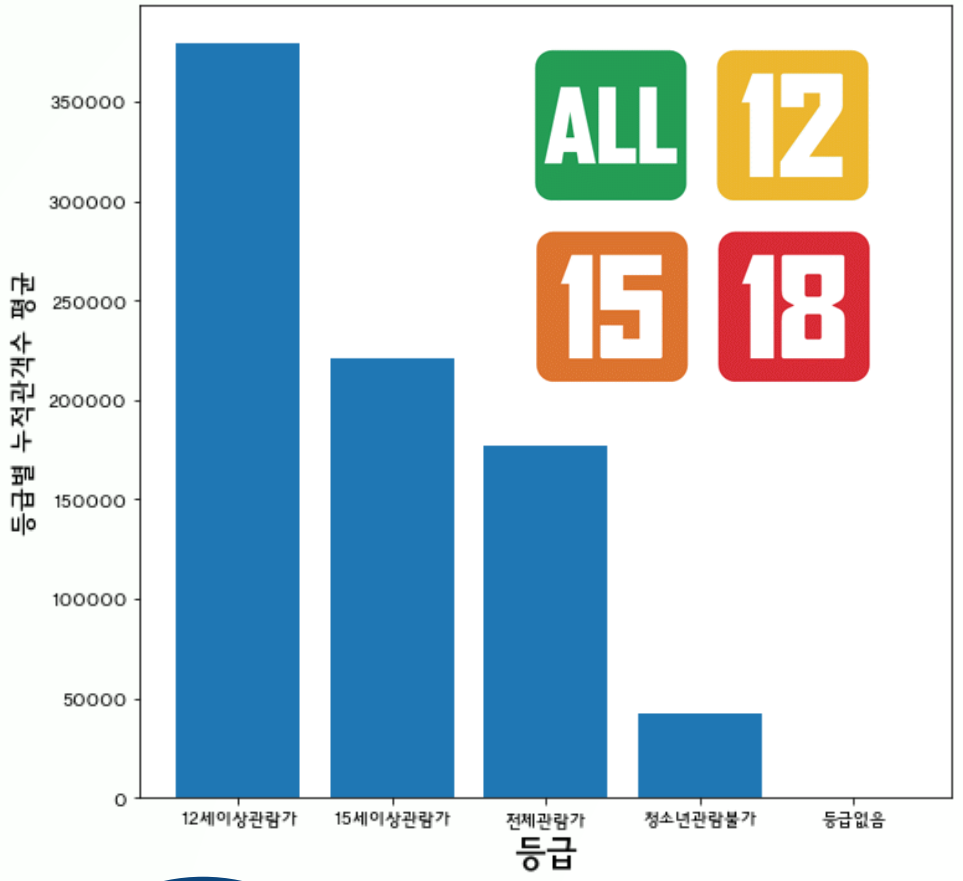
2) 더미 데이터 - 대표국적, 장르



3. EDA & Preprocessing

3) 더미 데이터 - 등급

등급별 누적 관객수 평균



등급
5개로 구분

→ DUMMY

3. EDA & Preprocessing

3) 더미 데이터

	기타	독일	미국	스페인	영국	이탈리아	일본	중국	캐나다	프랑스	한국
0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0

대표국적

	12세이상관람가	15세이상관람가	등급없음	전체관람가	청소년관람불가
0	0	0	1	0	0
1	0	1	0	0	0
2	0	1	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0

등급

	SF	가족	공연	공포	느와르	다큐멘터리	드라마	멜로로맨스	모험	무협	...	서사	스릴러	애니메이션	액션	어드벤처	에로	장르없음	전쟁	코미디	판타지
0	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0

장르

3. EDA & Preprocessing

4) 기타 전처리

개봉일 형식 변경

- datetime 형식으로 변환, 결측값에 대해서는 1900-01-01 형식으로 변환

누적관객수값 보정

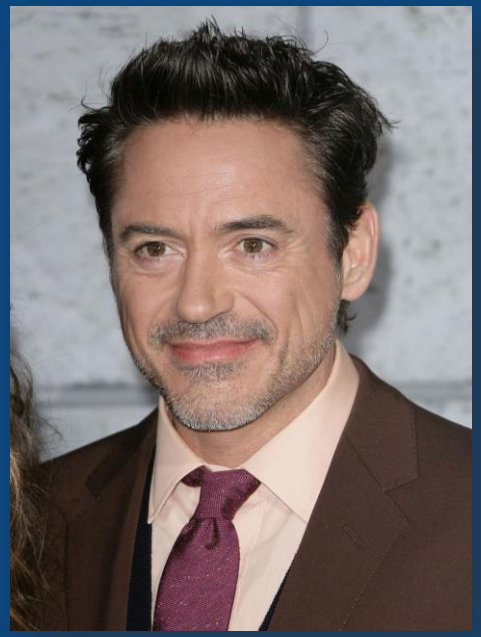
- 누적관객수 값 음수인 것 확인 후 양수값으로 변경

중복값 제거

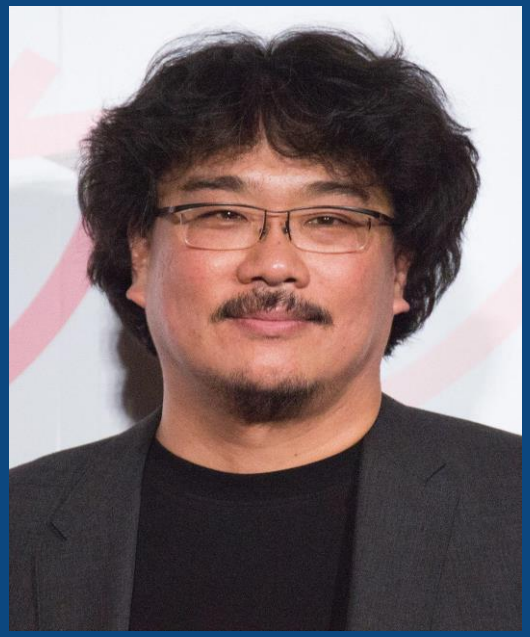
- 중복 영화데이터 누적관객수 높은 값 기준으로 제거

3. EDA & Preprocessing

5) 배우, 감독, 배급사 수치화



배우

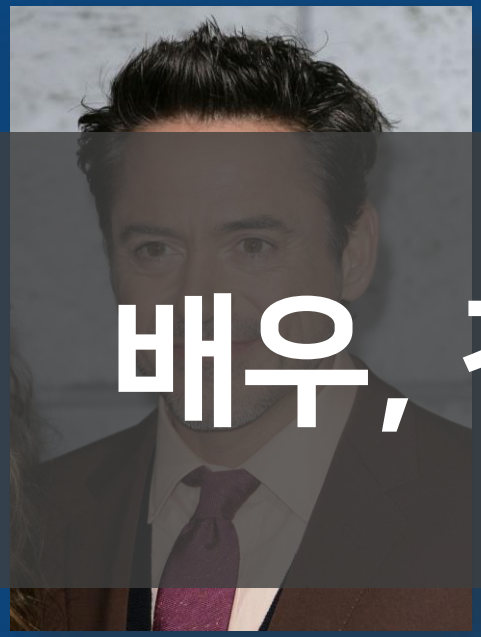


감독

배급사

3. EDA & Preprocessing

5) 배우, 감독, 배급사 수치화



배우, 감독, 배급사 전부 동일 처리

주연배우(최대 3인)

개봉일 기준
이전 6년 출연 영화 동원 관객수 ÷ 영화 수

3. EDA & Preprocessing

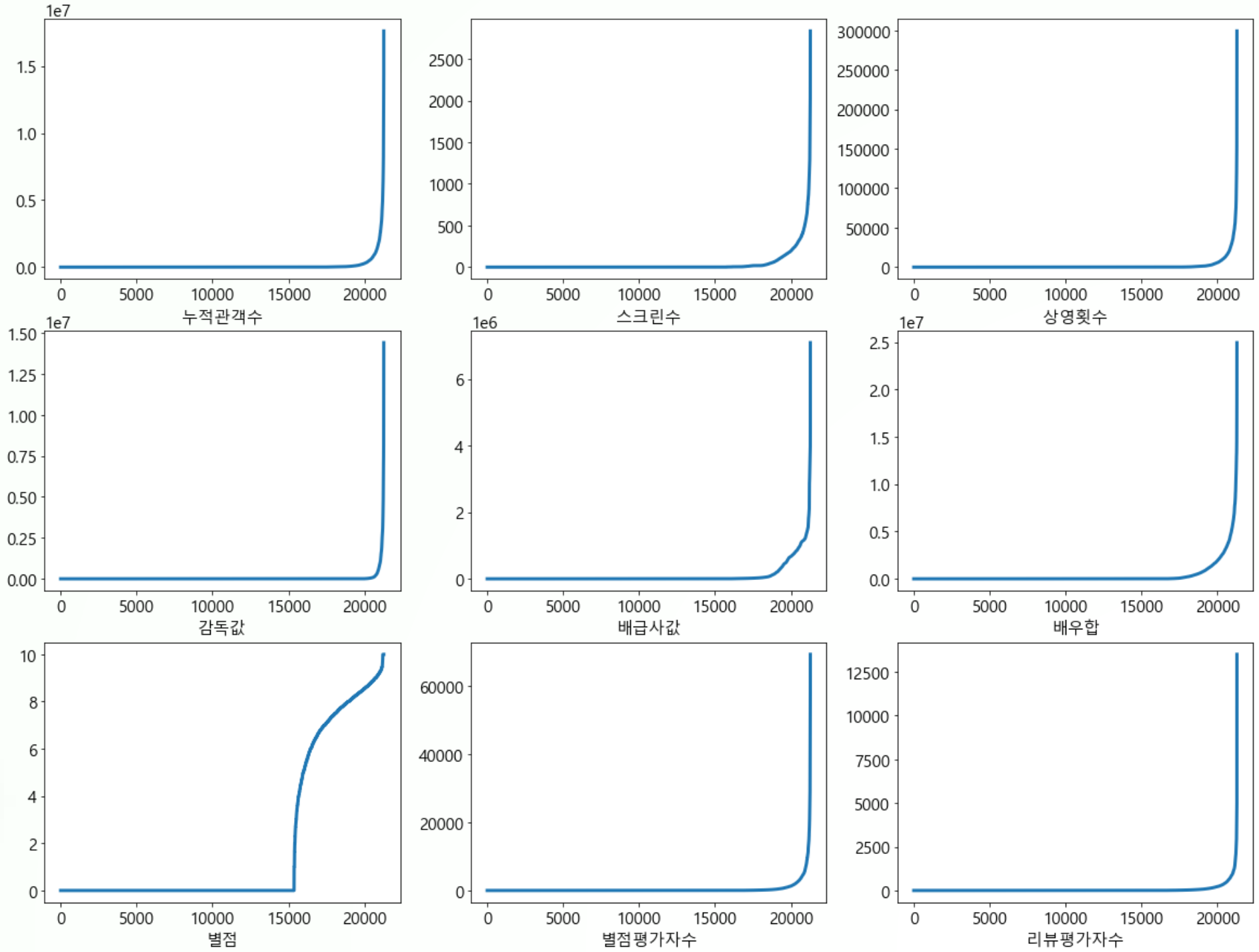
5) 배우, 감독, 배급사 수치화

15445	아메리칸 셰프	2015-01-07	155807	(주)영화사진진	존파브로	존파브로,엠제이안소니,소피아베르가라,스칼렛요한슨,더스틴호프만,로버트다우니주니어,존레귀자모
14932	어벤져스: 에이지 오브 울트론	2015-04-23	10504007	월트디즈니컴퍼니코리아유한책임회사	조스웨던	로버트다우니주니어,크리스헴스워스,마크러팔로,크리스에반스,스칼렛요한슨,제레미레너,돈치...
11787	캡틴 아메리카: 시빌 워	2016-04-27	8678117	월트디즈니컴퍼니코리아유한책임회사	안소니루소,조루소	크리스에반스,로버트다우니주니어,스칼렛요한슨,세바스찬스탠,안소니마키,돈치들,제레미레너...
15369	스파이더맨: 홈 커밍	2017-07-05	7258678	한국소니픽쳐스릴리징브에나비스타영화㈜	존왓츠	제이콥배덜런,톰홀랜드,로버트다우니주니어,마이클키튼,마리사토메이,젠데이아콜먼,존파브로...
11788	어벤져스: 인피니티 워	2018-04-25	11233176	월트디즈니컴퍼니코리아유한책임회사	안소니루소,조루소	로버트다우니주니어,조슈브롤린,마크러팔로,톰히들스턴,크리스에반스,크리스헴스워스,제레미...
11789	어벤져스: 엔드게임	2019-04-24	13977602	월트디즈니컴퍼니코리아유한책임회사	안소니루소,조루소	로버트다우니주니어,크리스에반스,마크러팔로,크리스헴스워스,스칼렛요한슨,제레미레너,돈치...
10648	닥터 두리틀	2020-01-08	1607082	유니버설픽쳐스인터내셔널코리아(유)	스티븐개건	로버트다우니주니어,톰홀랜드,라미말렉,안토니오반데라스,마리옹꼬띠아르,마이클쉰,옥타비아...

	누적관객수	스크린수	상영횟수	대표국적	배급사	등급	장르	감독	감독값	배급사값	배우합
10648	1607082	1290	68634	미국	유니버설픽쳐스인터내셔널코리아(유)	전체관람가	판타지	스티븐개건	25659	990746	21160120

3. EDA & Preprocessing

6) 변수 별 데이터분포



3. EDA & Preprocessing

7) VIF & Correlation

Person Correlation of Features

	누적관객수	스크린수	상영횟수	감독값	배급사값	배우합	별점	별점평가자수	리뷰평가자수	VIF
누적관객수	1	0.46	0.58	0.43	0.49	0.42	0.22	0.75	0.72	
스크린수	0.46	1	0.85	0.27	0.58	0.44	0.23	0.41	0.29	4.43
상영횟수	0.58	0.85	1	0.26	0.5	0.39	0.19	0.44	0.34	3.91
감독값	0.43	0.27	0.26	1	0.29	0.32	0.15	0.41	0.32	1.26
배급사값	0.49	0.58	0.5	0.29	1	0.46	0.26	0.41	0.39	1.76
배우값	0.42	0.44	0.39	0.32	0.46	1	0.23	0.38	0.32	1.44
별점	0.22	0.23	0.19	0.15	0.26	0.23	1	0.32	0.35	1.19
별점평가자수	0.75	0.41	0.44	0.41	0.41	0.38	0.32	1	0.72	2.43
리뷰평가자수	0.72	0.29	0.34	0.32	0.39	0.32	0.35	0.72	1	2.22

3. EDA & Preprocessing

8) 변수선택법

전체모형 : 누적관객수 ~ 스크린수 + 상영횟수 + 장르 + ~~대표국적~~ + 등급 + 감독값 + 배급사값 + 배우합 + 별점 + 별점평가자수 + 리뷰평가자수

forward

Step: AIC=539672.1
누적관객수 ~ 별점평가자수 + 상영횟수 + 리뷰평가자수 + 감독값 + 스크린수 + 배급사값 + 별점 + 배우합 + 장르 + 등급

backward

Step: AIC=539672.1
누적관객수 ~ 스크린수 + 상영횟수 + 등급 + 장르 + 감독값 + 배급사값 + 배우합 + 별점 + 별점평가자수 + 리뷰평가자수

stepwise

Step: AIC=539672.1
누적관객수 ~ 스크린수 + 상영횟수 + 등급 + 장르 + 감독값 + 배급사값 + 배우합 + 별점 + 별점평가자수 + 리뷰평가자수

3. EDA & Preprocessing

9) R-Square

Dep. Variable:	누적관객수	R-squared:	0.507
Model:	OLS	Adj. R-squared:	0.506
Method:	Least Squares	F-statistic:	662.6
Date:	Sun, 22 Aug 2021	Prob (F-statistic):	0.00
Time:	21:56:52	Log-Likelihood:	-3.0692e+05
No. Observations:	21280	AIC:	6.139e+05
Df Residuals:	21246	BIC:	6.142e+05
Df Model:	33		
Covariance Type:	nonrobust		

변수
스크린수 / 상영횟수 / 배급사 / 감독 / 배우 등급 / 장르

Dep. Variable:	누적관객수	R-squared:	0.696
Model:	OLS	Adj. R-squared:	0.695
Method:	Least Squares	F-statistic:	1389.
Date:	Sun, 22 Aug 2021	Prob (F-statistic):	0.00
Time:	21:56:15	Log-Likelihood:	-3.0178e+05
No. Observations:	21280	AIC:	6.036e+05
Df Residuals:	21244	BIC:	6.039e+05
Df Model:	35		
Covariance Type:	nonrobust		

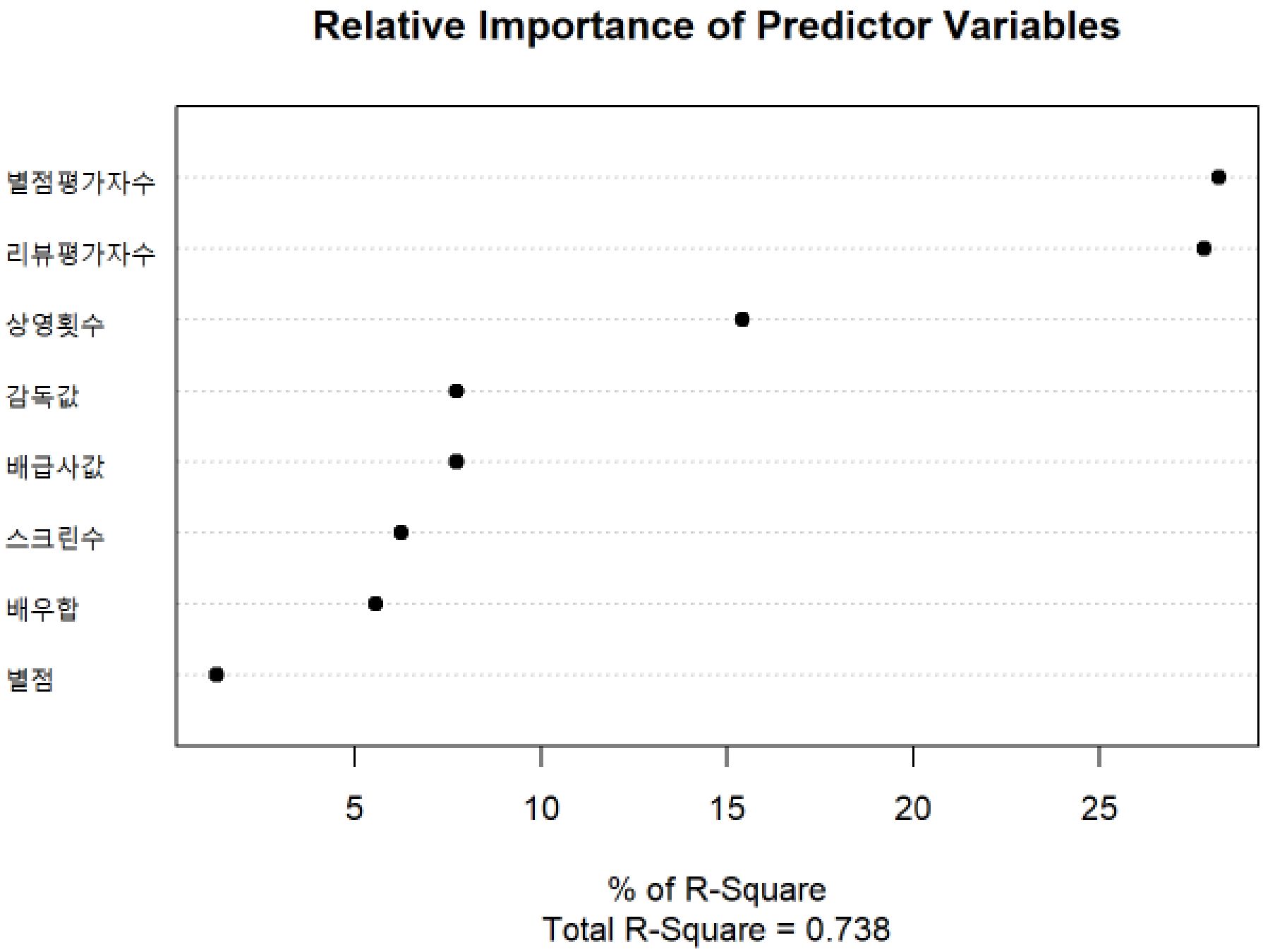
변수
스크린수 / 상영횟수 / 배급사 / 감독 / 배우 등급 / 장르 / 별점 / 별점 평가자 수

Dep. Variable:	누적관객수	R-squared:	0.743
Model:	OLS	Adj. R-squared:	0.742
Method:	Least Squares	F-statistic:	1705.
Date:	Sun, 22 Aug 2021	Prob (F-statistic):	0.00
Time:	21:58:23	Log-Likelihood:	-2.9999e+05
No. Observations:	21280	AIC:	6.001e+05
Df Residuals:	21243	BIC:	6.004e+05
Df Model:	36		
Covariance Type:	nonrobust		

변수
스크린수 / 상영횟수 / 배급사 / 감독 / 배우 등급 / 장르 / 별점 / 별점 평가자 수 / 리뷰 평가자수

3. EDA & Preprocessing

10) 변수 설명력



3. EDA & Preprocessing

11) 최종 데이터 형식

	누적관객수	스크린수	상영횟수	감독값	배급사값	배우합	SF	가족	공연	공포	다큐멘터리	드라마	멜로로맨스	뮤지컬	미스터리	범죄	사극	서부	스릴러	애니메이션
0	2947549	2528	153914	7011	3024611	5189961	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2442963	1688	119738	10003317	348679	9357161	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2292413	2297	131856	0	762252	5723478	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2151572	900	151874	0	826095	6951	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	2048137	2018	160725	4969735	3251960	15650	0	0	0	0	0	0	0	0	0	0	0	0	0	1

	액션	어드벤처	에로	전쟁	코미디	판타지	12세이상관람가	15세이상관람가	전체관람가	청소년관람불가	평점	평점평가자수	리뷰평가자수	모험	느와르	무협	서사	등급없음	장르없음
0	1	0	0	0	0	0	1	0	0	0	8.50	7061	194	0	0	0	0	0	0
1	1	0	0	0	0	0	0	1	0	0	8.22	11249	115	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0	0	0	7.46	5950	137	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	9.28	12774	94	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	1	0	9.31	9894	275	0	0	0	0	0	0

4. Machine Learning

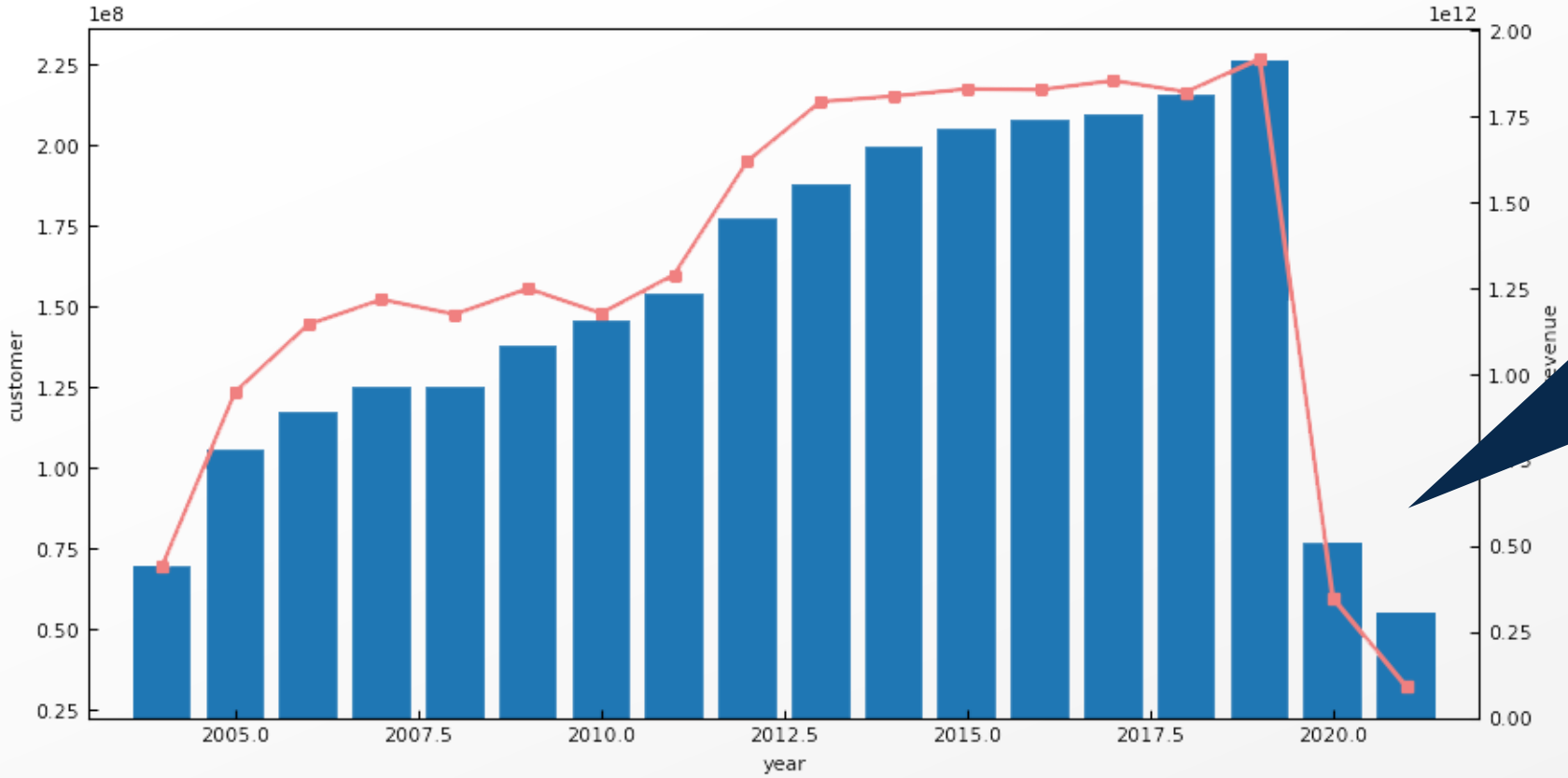
1) 데이터 구성

어벤져스 : 엔드게임	
누적관객수	13977602
스크린수	2835
상영횟수	246433
등급	12세 이상관람가
장르	액션
감독값	9955646
배우값	22102634
배급사값	2880132
별점	9.38
별점평가자수	69186
리뷰평가자수	1008

4. Machine Learning

2) 연도별 영화 데이터

연도별 관객수 증감 추이



19년
226,678,777명

COVID-19

20년
59,524,093명

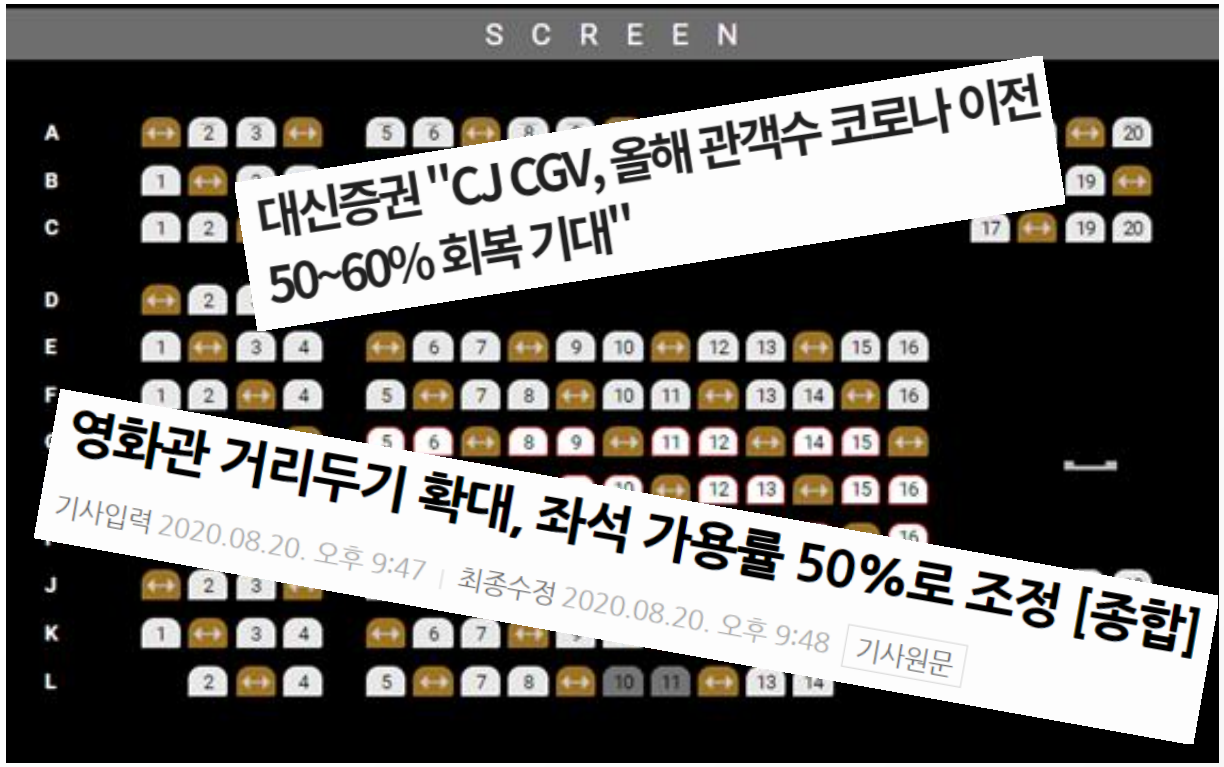
21년
32,163,118명

관객수 약 70% 감소

4. Machine Learning

3) 데이터 보정

기사 출처 : 연합뉴스(2020-08-20) / **이코노믹 리뷰** (노성인 기자 2021.02.2)



2020년 사회적 거리두기 추진(2.29~)

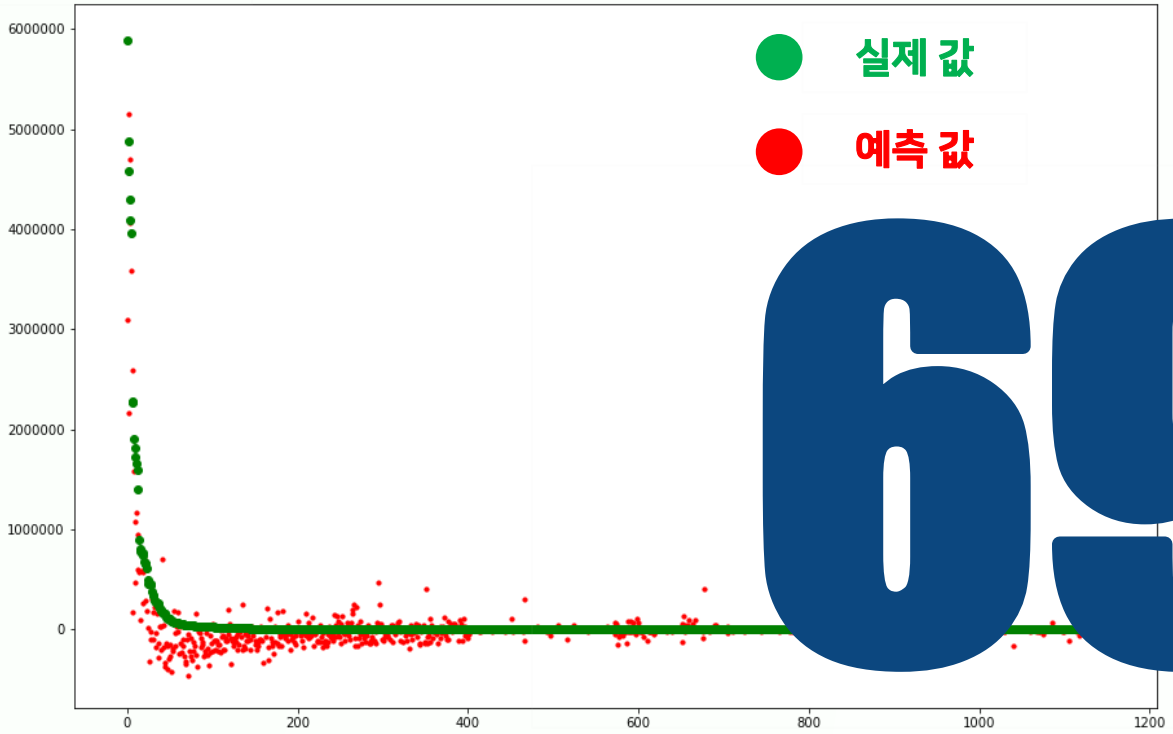
20년 3월 이 후 극장 내 좌석간 거리두기 시행

(좌석 가용률 70% -> 60% -> 50%)

Test data Pred(예측값) ÷ 2

Linear Regression

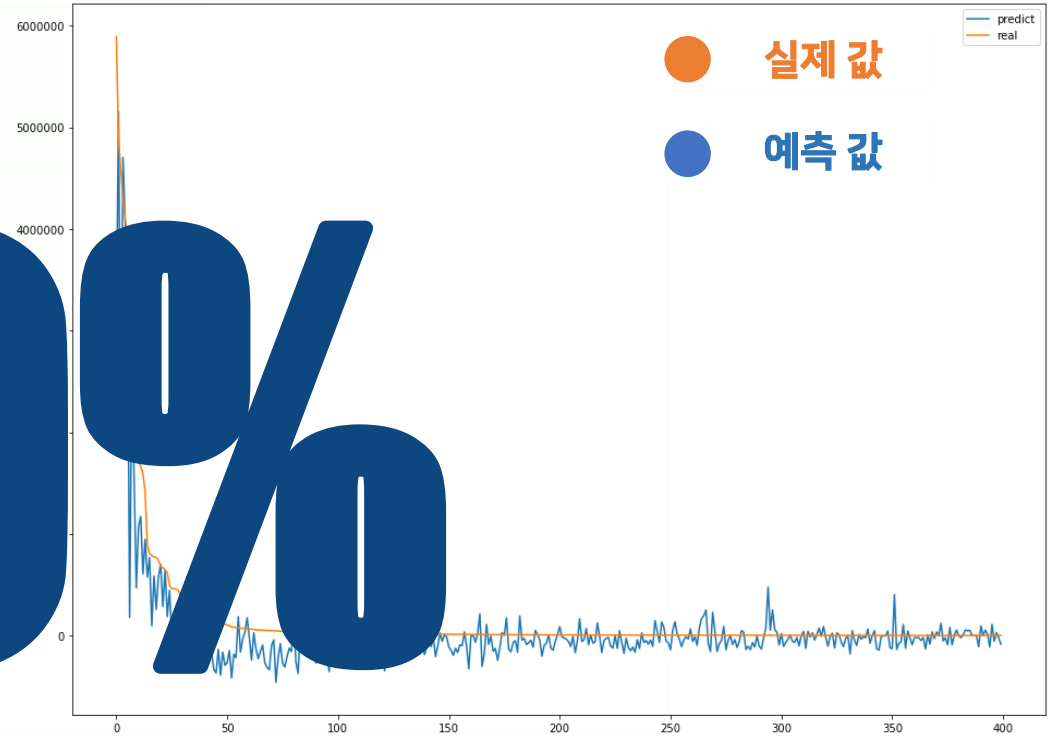
누적 관객수



SCATTER_GRAPH

69%

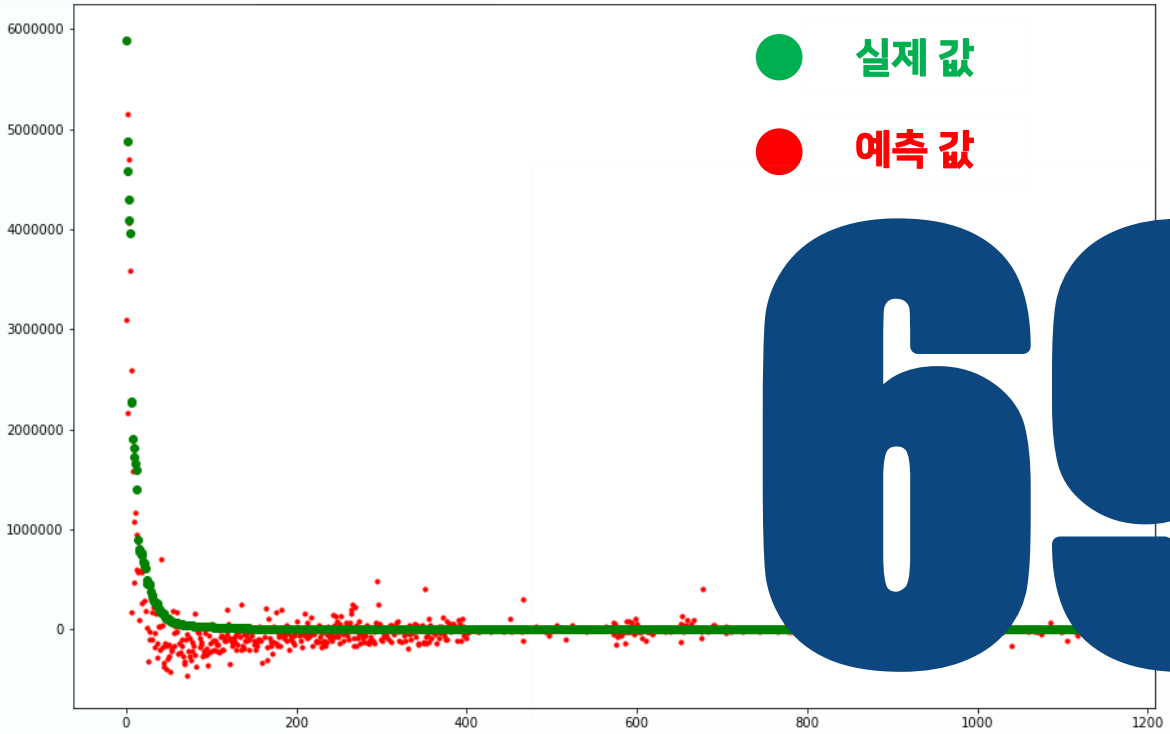
누적 관객수



LINE_GRAPH

Ridge Regression

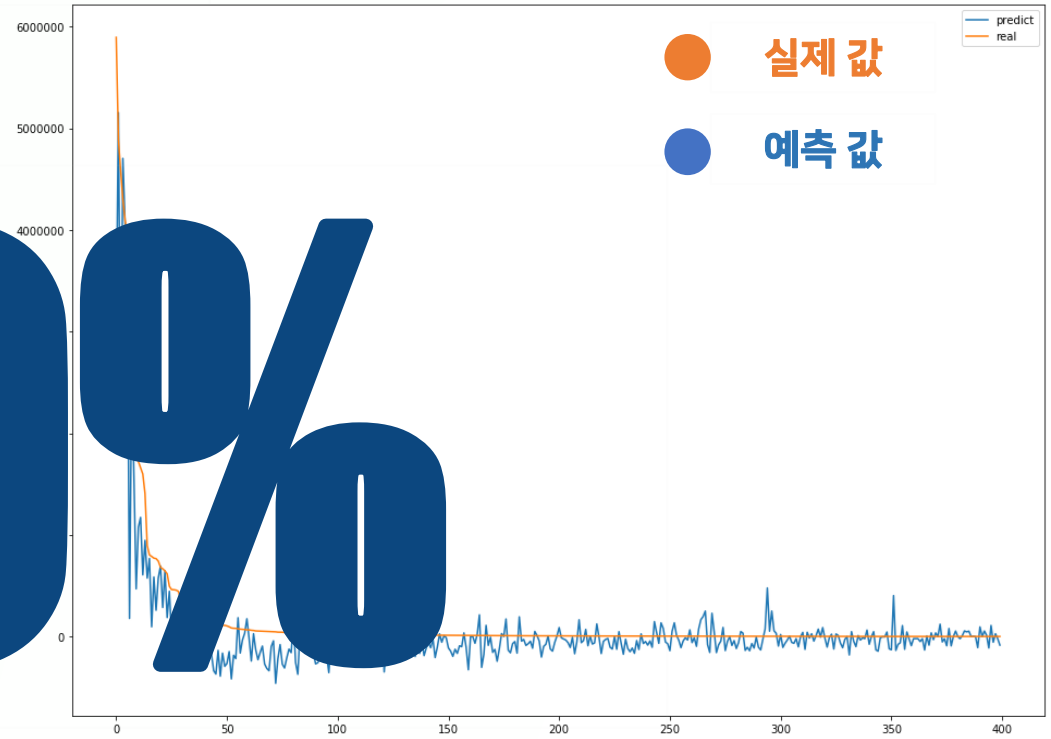
누적 관객수



SCATTER_GRAPH

69%

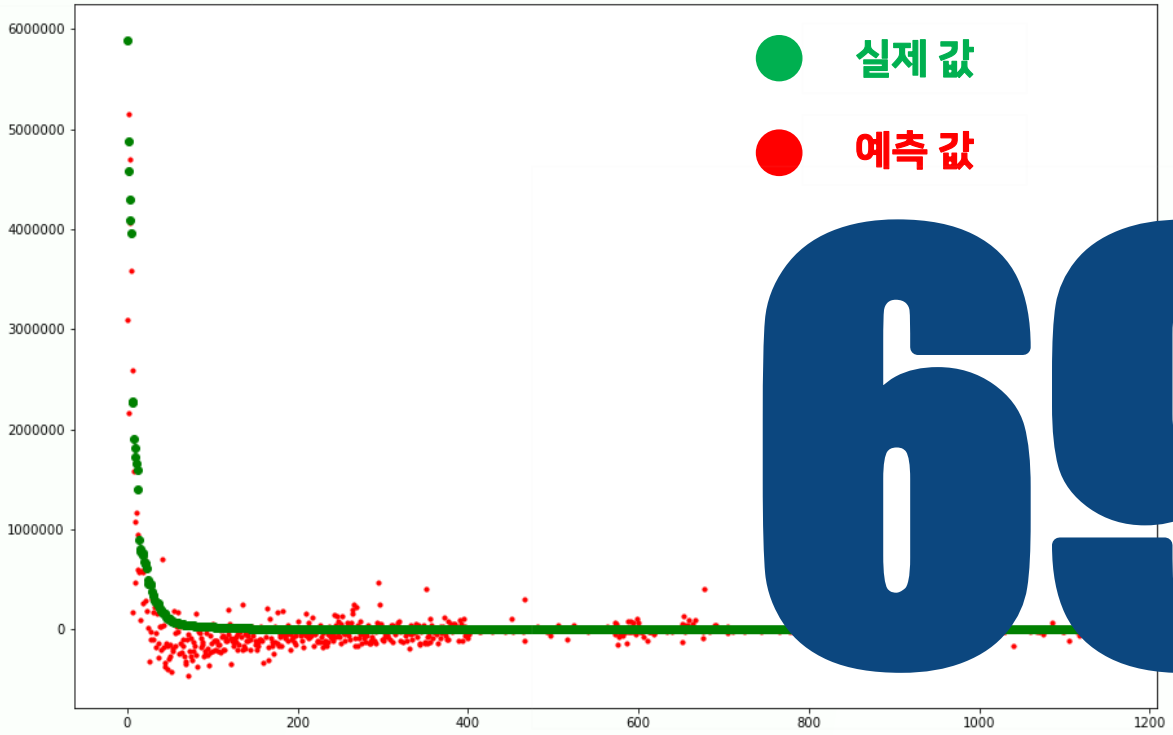
누적 관객수



LINE_GRAPH

Lasso Regression

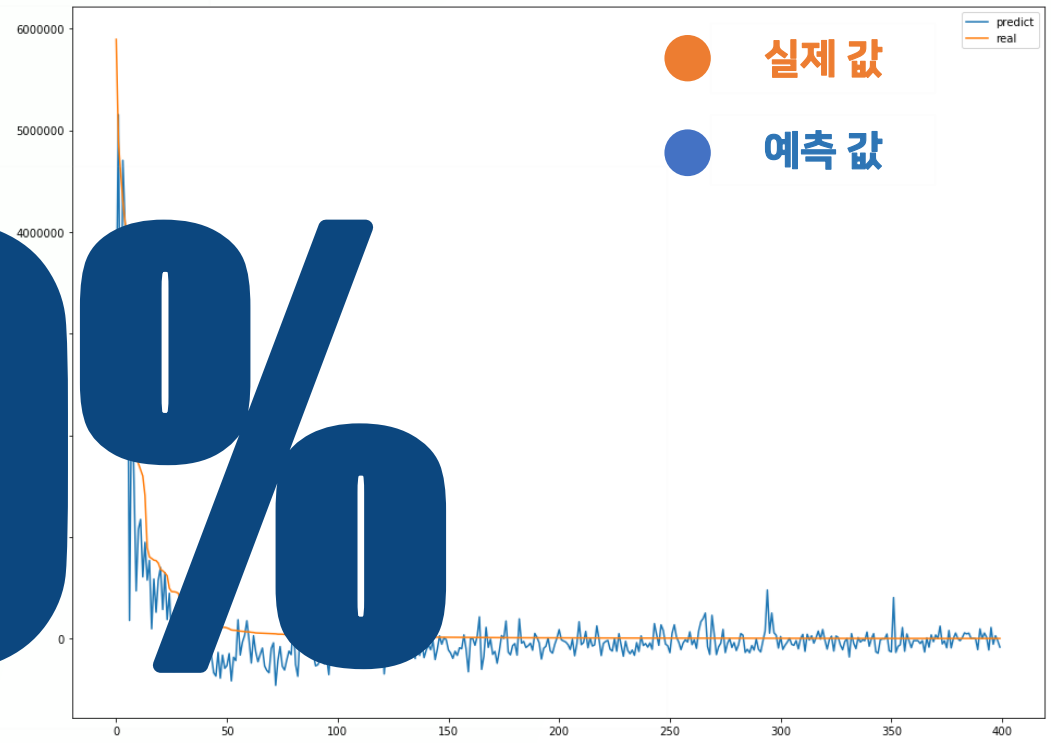
누적 관객수



SCATTER_GRAPH

69%

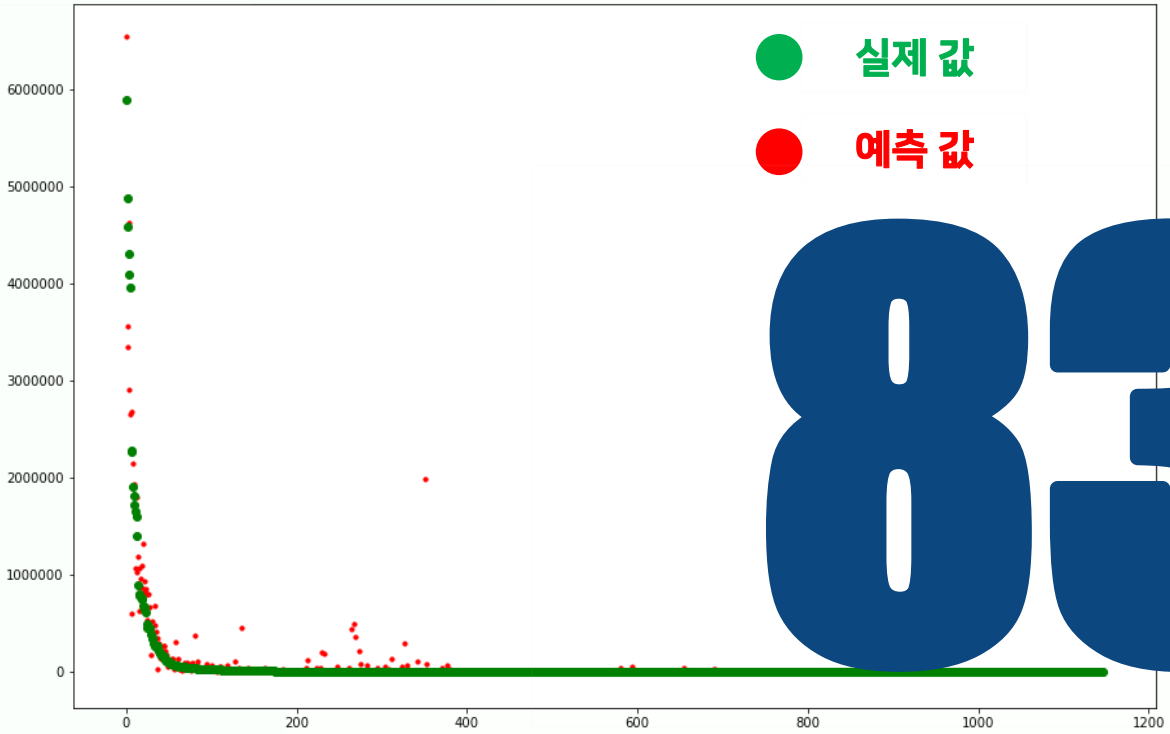
누적 관객수



LINE_GRAPH

XGB Regression

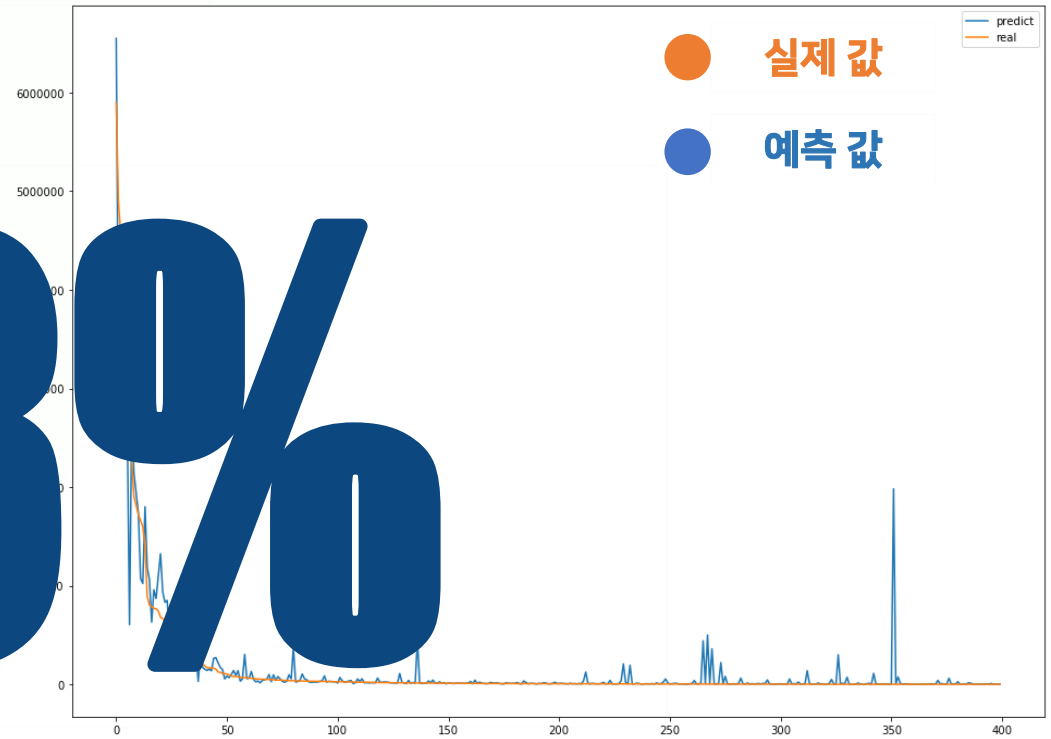
누적 관객수



SCATTER_GRAPH

83%

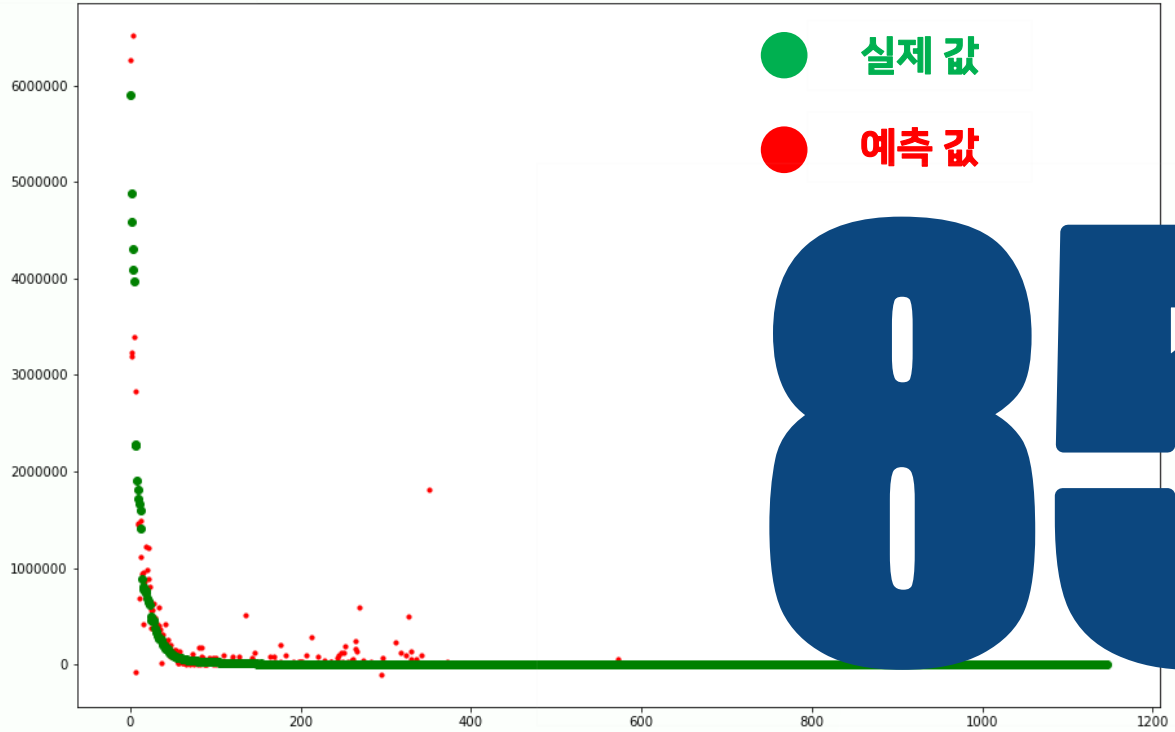
누적 관객수



LINE_GRAPH

Lgbm Regression

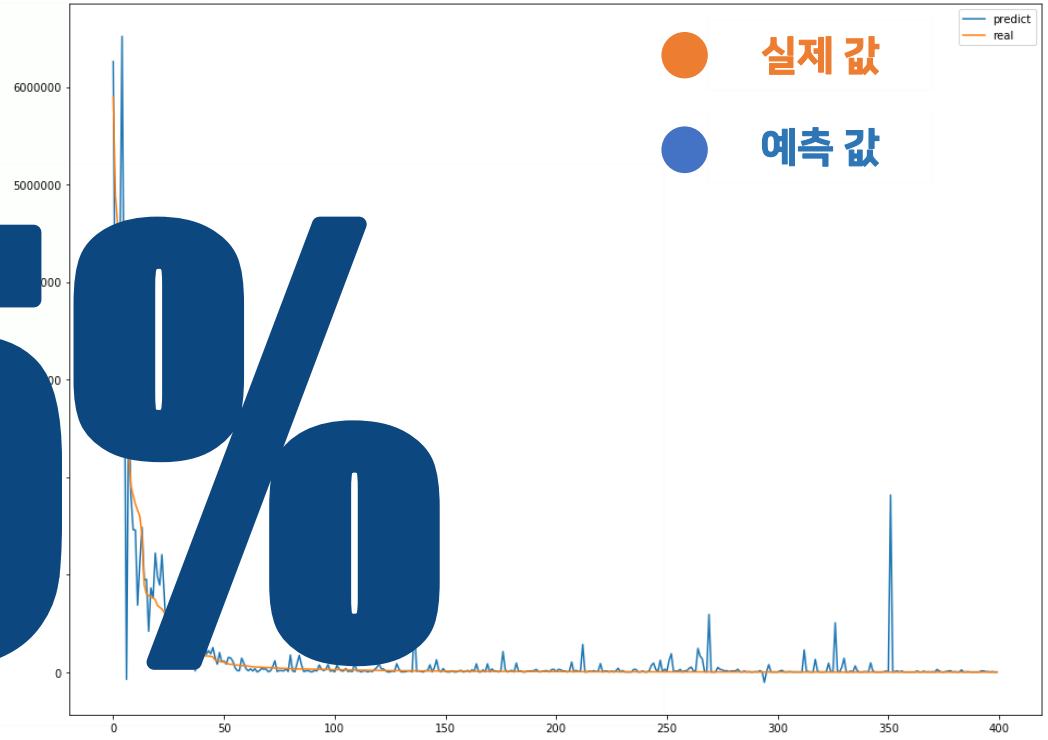
누적 관객수



SCATTER_GRAPH

85%

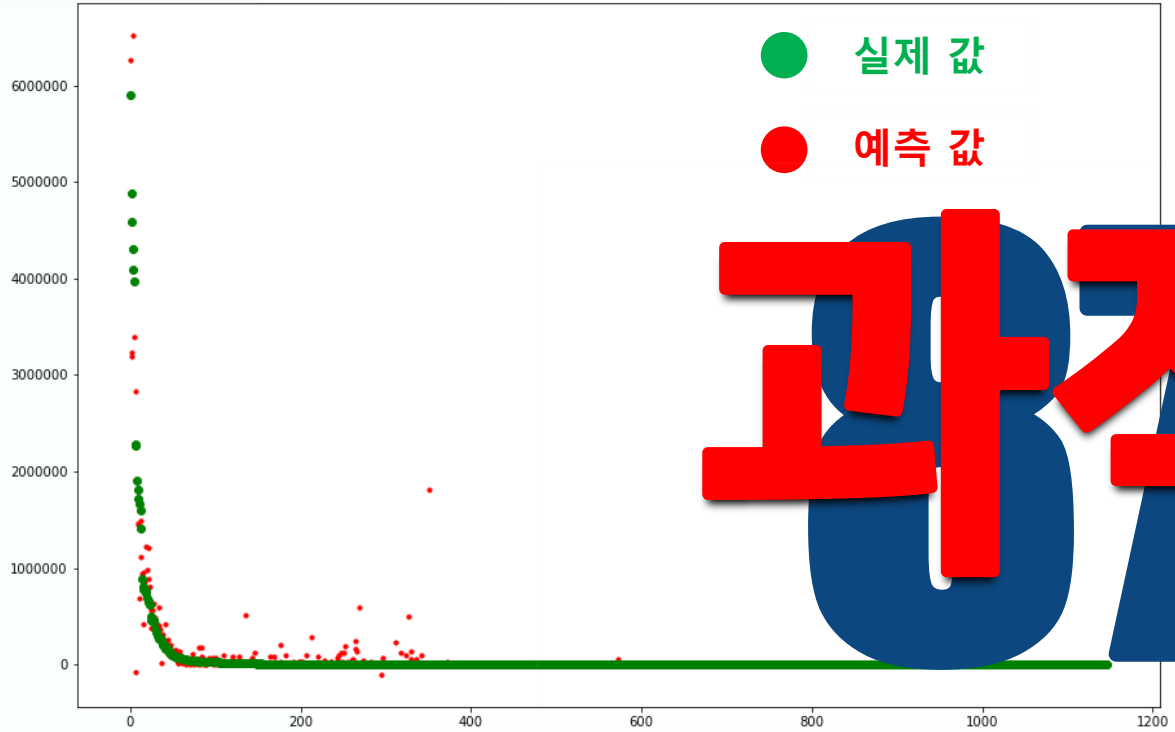
누적 관객수



LINE_GRAPH

Random Forest Regression

누적 관객수



SCATTER_GRAPH

누적 관객수

LINE_GRAPH

과적합

- ❑ XGB Regression
- ❑ LGBM Regression
- ❑ Linear Regression
- ❑ Ridge Regression
- ❑ Lasso Regression
- ❑ Random Forest

4. Machine Learning

5) LGBM Regression

```
# lgb 모델
model = lgb.LGBMRegressor(random_state = 1, n_estimators = 1000)

X_train, X_test, y_train, y_test = train_test_split(train_X, train_y, test_size = 0.2, random_state = 2)

params = {'learning_rate': [1, 0.1, 0.01, 0.001, 0.005],
          'min_child_samples': [5, 10, 20]}

grid_model = GridSearchCV(estimator = model, param_grid = params,
                          scoring = 'r2', cv = 5, refit = True, verbose = 100)

grid_model.fit(X_train, y_train)
```

영화명	예측값	실제값	차이	오차율
블랙 위도우	3129909.84	2947549	182361	6.186864
분노의 질주: 더 얼티메이트	1616799.04	2442963	826164	-33.8181
극장판 귀멸의 칼날: 무한열차편	1597022.98	2292413	695390	-30.3344
소울	2037840.78	2151572	113731.2	-5.28596
크루엘라	3259259.98	2048137	1211123	59.13291
싱크홀	1694898.95	1982221	287322.1	-14.495
미나리	1413544.04	1141829	271715	23.79647
보스베이비2	1413544.04	907883	505661	55.69672
콰이어트 플레이스 2	908770.13	860904	47866.1	5.559985
랑종	729935.7	830426	100490.3	-12.1011

5. Result & Project Evaluation

한계점

- 데이터의 변수가 대부분 결측 값

  Ex) 흥행하지 못한 영화, 독립 영화 등Ex)

- 데이터 특성 상 관객수가 없거나 관객수가 너무 높은 데이터는 이상 치로 해석 될 수 있음

  Ex) 관객수 5, 매출액 3000, 쪽박 영화, 영화 “명량” 등

- 코로나와 같은 급격한 외부환경의 변화가 패턴을 찾기 어렵게 함
- 분류 모델이 아닌 회귀 모델을 선택함으로, 모델 성능 향상에 어려움을 가졌음

개선사항으로 딥러닝 모델 추가 예정

5. Result & Project Evaluation

자체평가

- 일반적으로 배우, 배급사, 감독이 영화 흥행에 중요한 요소로 작용 할 것으로 예상 했으나, 변수의 설명력은 별점평가자수와 리뷰평가자수가 가장 크게 작용
- 회귀 알고리즘으로 예측 모델을 정밀하게 설계한다면 발전 가능성이 있음
- 현재 상영중인 영화에 대해서는 누적관객수 집계가 완벽하지 않음으로 인해 모델의 성능을 낮추는 요인으로 작용했을 것
- 단순히 데이터 양을 늘리는 것보다 영향력 있는 변수를 찾아 추가하는 것이 모델의 성능을 높일 수 있을 것

  Ex) 제작사, 제작비, 리뷰, 상영 시간, 마케팅 비용 등

수고하셨습니다람쥐

Q & A

THANK YOU