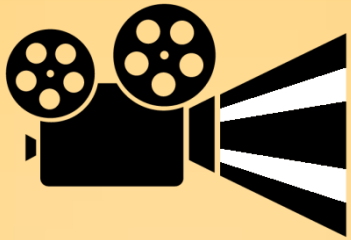
Fast Campus
Data Science School 12기
김민혜, 배빛나, 송이준



한국 개봉영화 관람객 예측

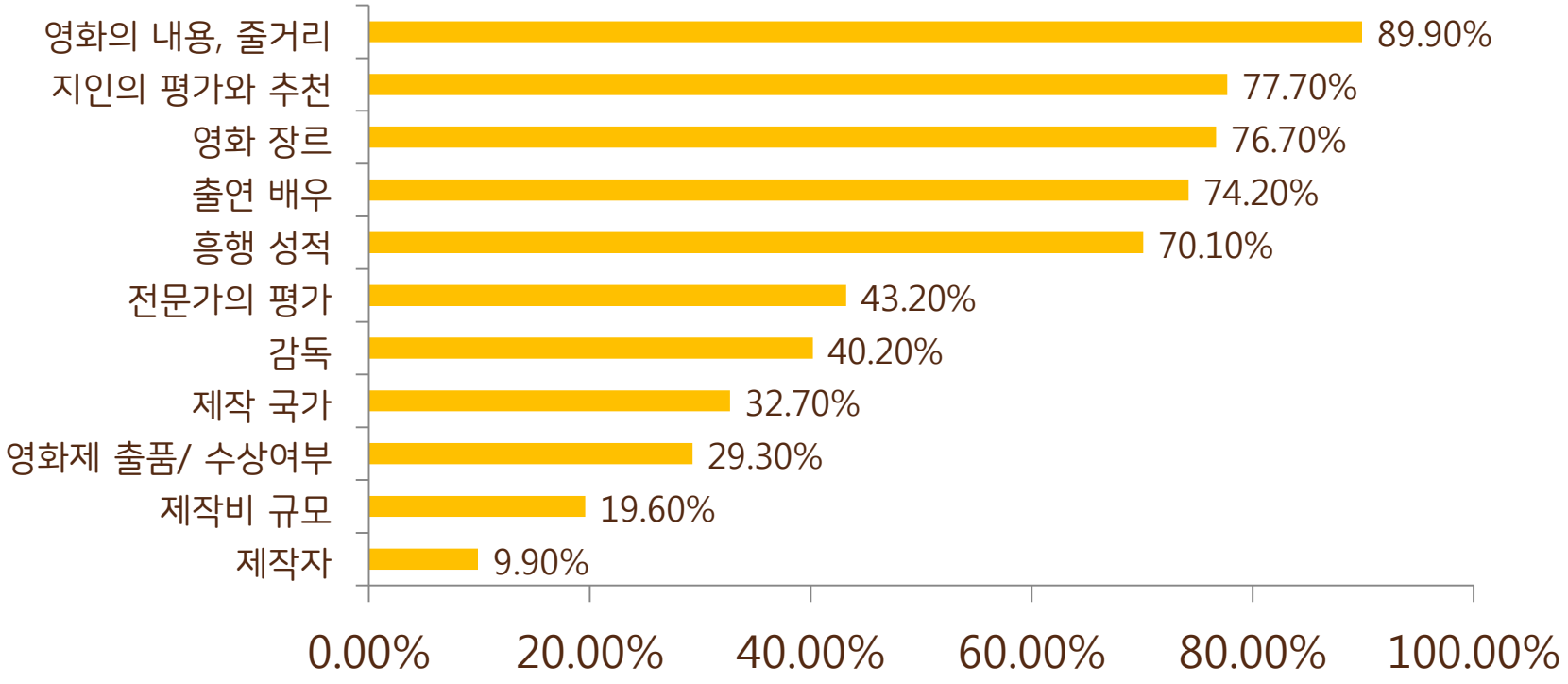
2015년 - 2019년 한국에서 개봉한 영화를 바탕으로 한 선형회귀분석

한국 개봉영화 관람객 예측

CONTENTS

01
주제 선정 동기

주제 선정 동기



전국 성인 남녀 1,217명 / 중복응답

자료출처 : 트렌드모니터

02
자료수집
및
특징

데이터 수집 시 참고한 논문

논문	저자	참고사항	응용
한국 영화의 흥행 결정 요인에 관한 연구(2003)	김은미	스타파워 산출 시 배우 수 (남주, 여주, 남조주, 여조주)	3명(네이버기사)
		마케팅 비용	홍보기사 개수(개봉 1주일 전후)
한국 영화시장의 흥행결정 요인에 관한 연구(2009)	박승현 외 1명	매년 흥행 순위 100위	6개월 단위 5년간 TOP100위
		배급사 파워 구분 - 헐리우드 배급사 자회사 - 한국 메이저 배급사 - 국내 미니 메이저 배급사 - 국내 독립 배급사	배급사 파워 구분 -헐리우드 배급사 -한국 메이저 배급사 -국내 미니 메이저 배급사
		개봉시점 - 7-8월 여름방학, -12-2월 겨울방학	시즌으로 구분 - 봄, 여름, 가을, 겨울
		원작, 실화, 시리즈물 영향	원작(영상, 실화)/ 시리즈 컬럼 추가
한국 영화시장의 흥행에 관한 통계분석 : 2006 ~ 2010년 흥행 작품을 중심으로	문병준 외 2명	비평가, 관람객 평가	네이버 전문가/관람객 평점 컬럼 추가
		개봉 첫주 상영관수	네이버 평점등록 관람객 수로 대체

데이터수집 방법

2015년~2019년에 개봉한 925개의 영화에 대한 데이터 수집

KOBIS
영화관입장권통합전산망

전국관객수
감독 & 배우
개봉일
국적
배급사
상영등급
상영시간
스크린 수
장르

NAVER 영화

전문가 평점
관객 평점
평점 매긴 관객 수

NAVER

Google

개봉 전/후 기사 건수
시리즈 여부
원작 여부

DATA 공공데이터포털 .GO.KR

공휴일

수집한 데이터의 특징

<실수형 데이터의 Describe>

	스크린개수	상영시간	전문가 평점	관객평점	전국관객수
count	925.000000	925.000000	925.000000	925.000000	9.250000e+02
mean	615.031351	109.778378	5.400735	8.166960	1.099422e+06
std	419.167442	19.932613	2.048343	1.492485	2.009962e+06
min	28.000000	48.000000	0.000000	0.000000	3.186600e+04
25%	322.000000	97.000000	5.000000	7.867150	9.701700e+04
50%	512.000000	110.000000	5.920000	8.457944	2.860420e+05
75%	827.000000	122.000000	6.700000	8.919948	1.124815e+06
max	2835.000000	222.000000	9.110000	10.000000	1.626336e+07

편차가 크다

48분짜리 영화는 무엇?

전문가와 관객평점은 차이가 있다

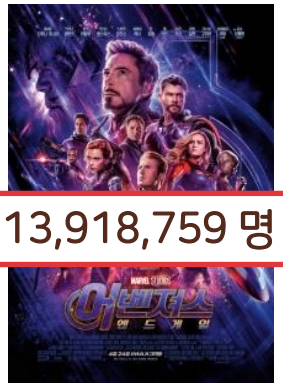
한국 개봉영화 관람객 예측

상위/하위 관객 동원 영화

<전국관객수 상위 5개 / 하위 5개 영화>

16,263,360 명

32,684 명



13,918,759 명

510배!

31,888명

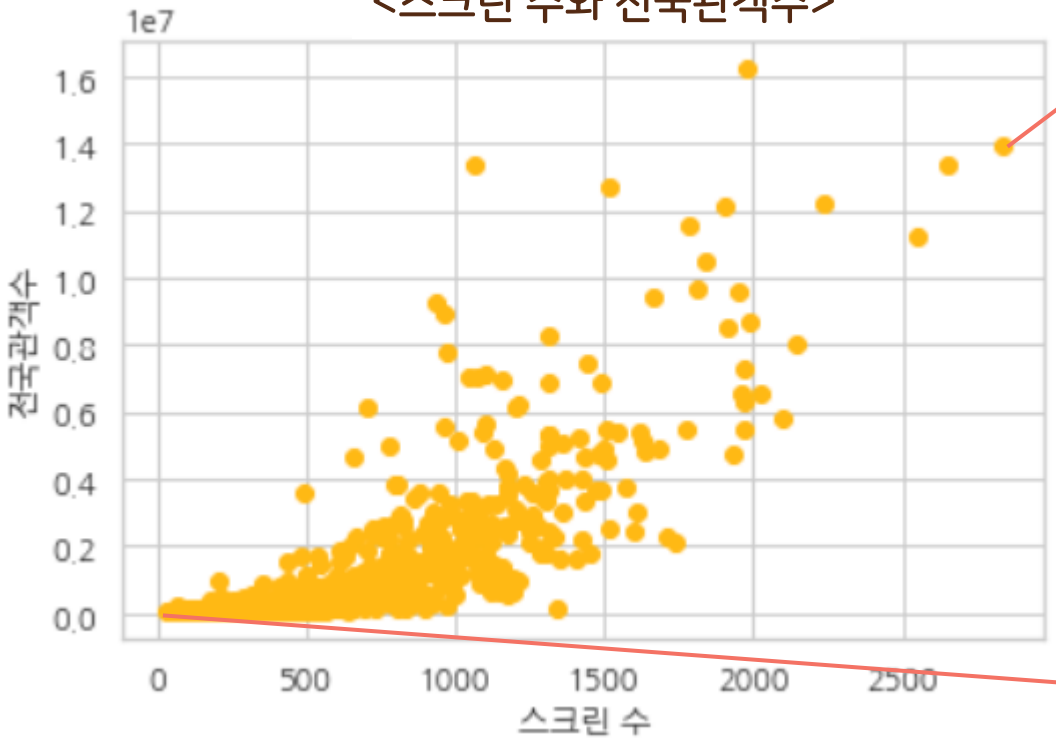
13,395,400 명

31,886명

한국 개봉영화 관람객 예측

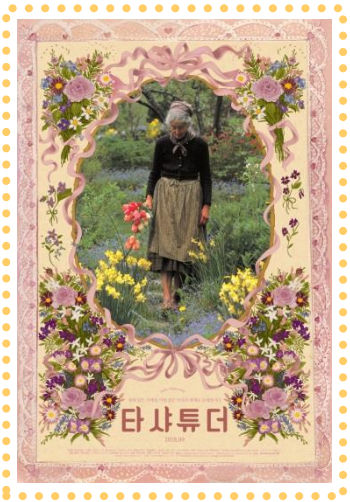
스크린 수 최대, 최소 영화

<스크린 수와 전국관객수>



어벤져스 : 엔드 게임

스크린 수 : 2835개

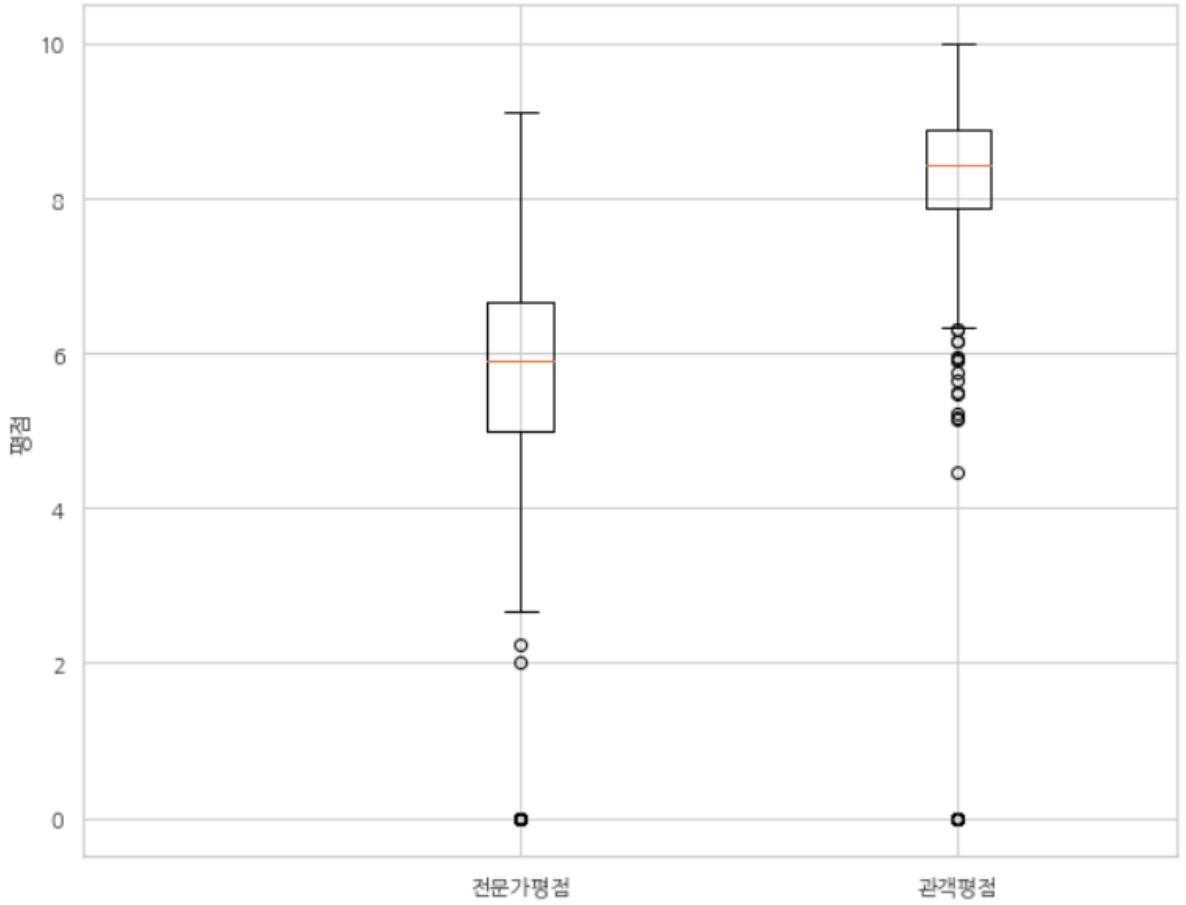


타샤 튜더

스크린 수 : 28개

전문가, 관객 평점과 전국관객수

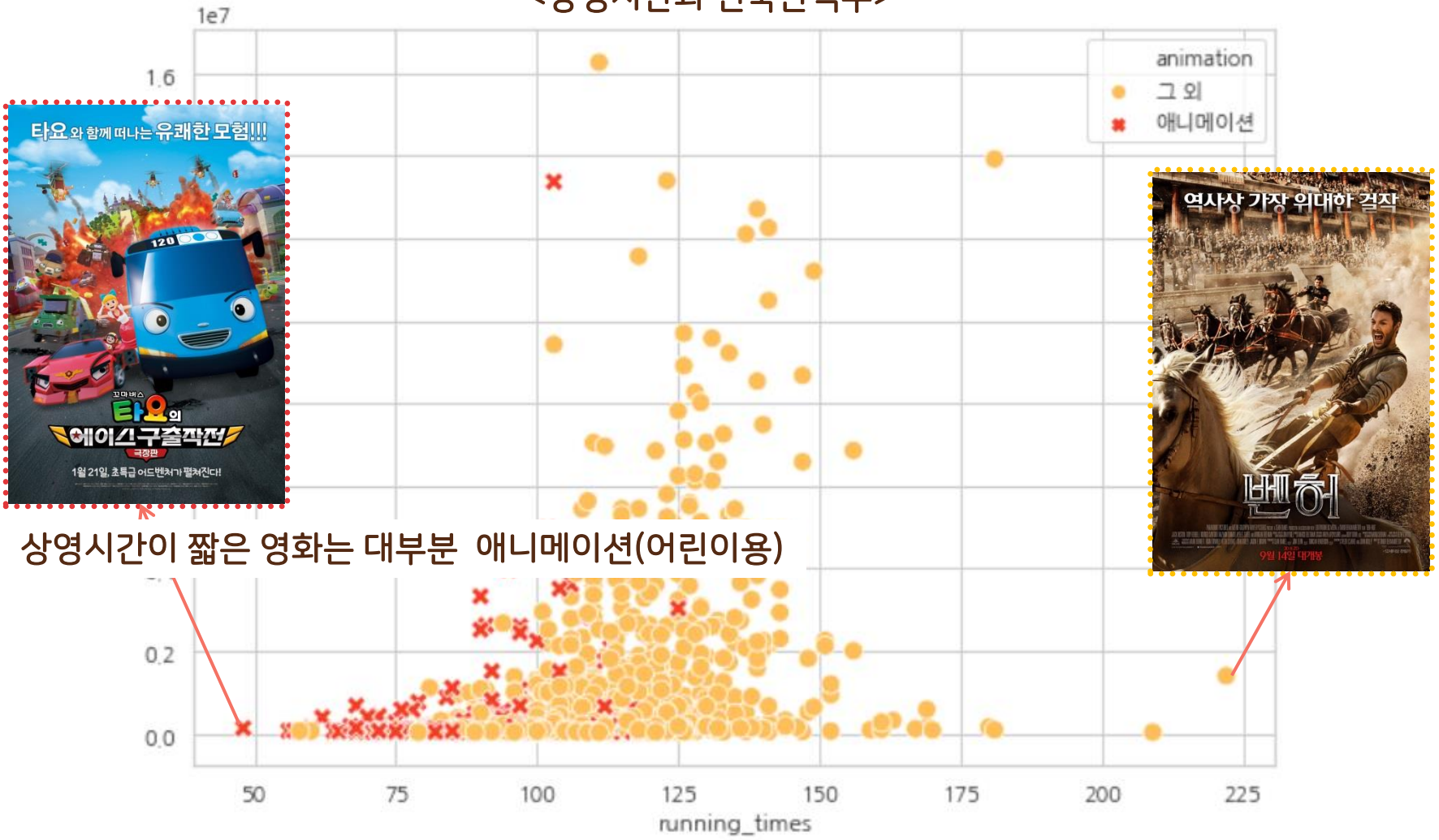
<전문가 / 관객 평점과 전국관객수>



한국 개봉영화 관람객 예측

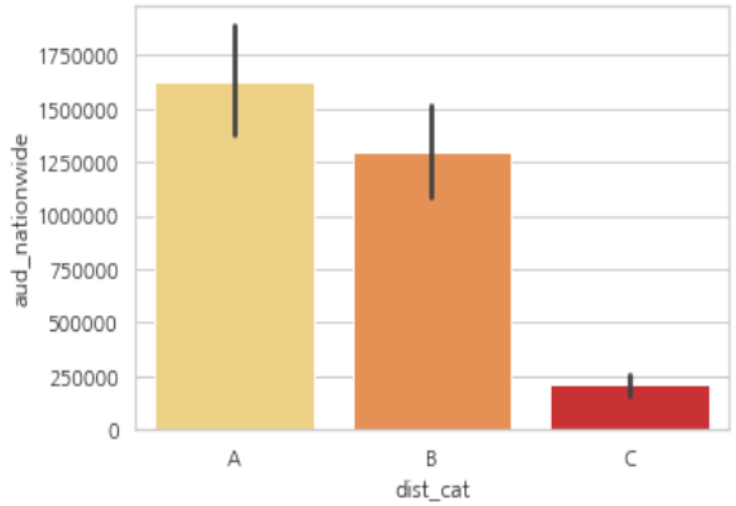
상영시간

<상영시간과 전국관객수>

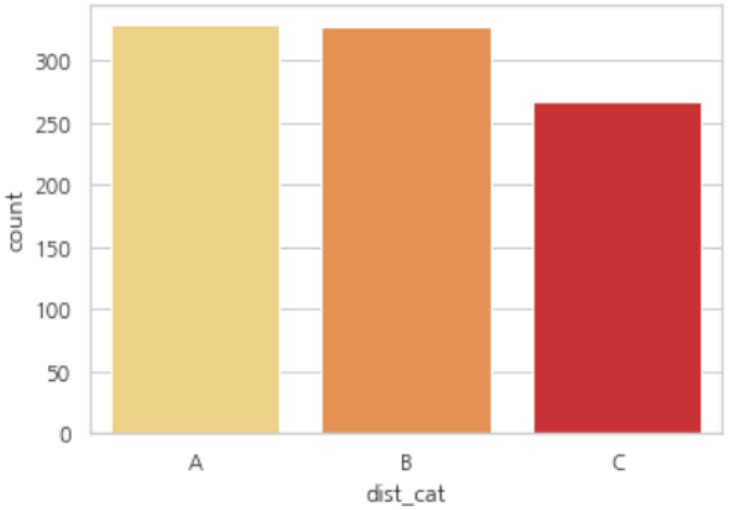


범주형 변수 정리 - 배급사

<배급사와 전국관객수>



<배급사별 배급영화 개수>

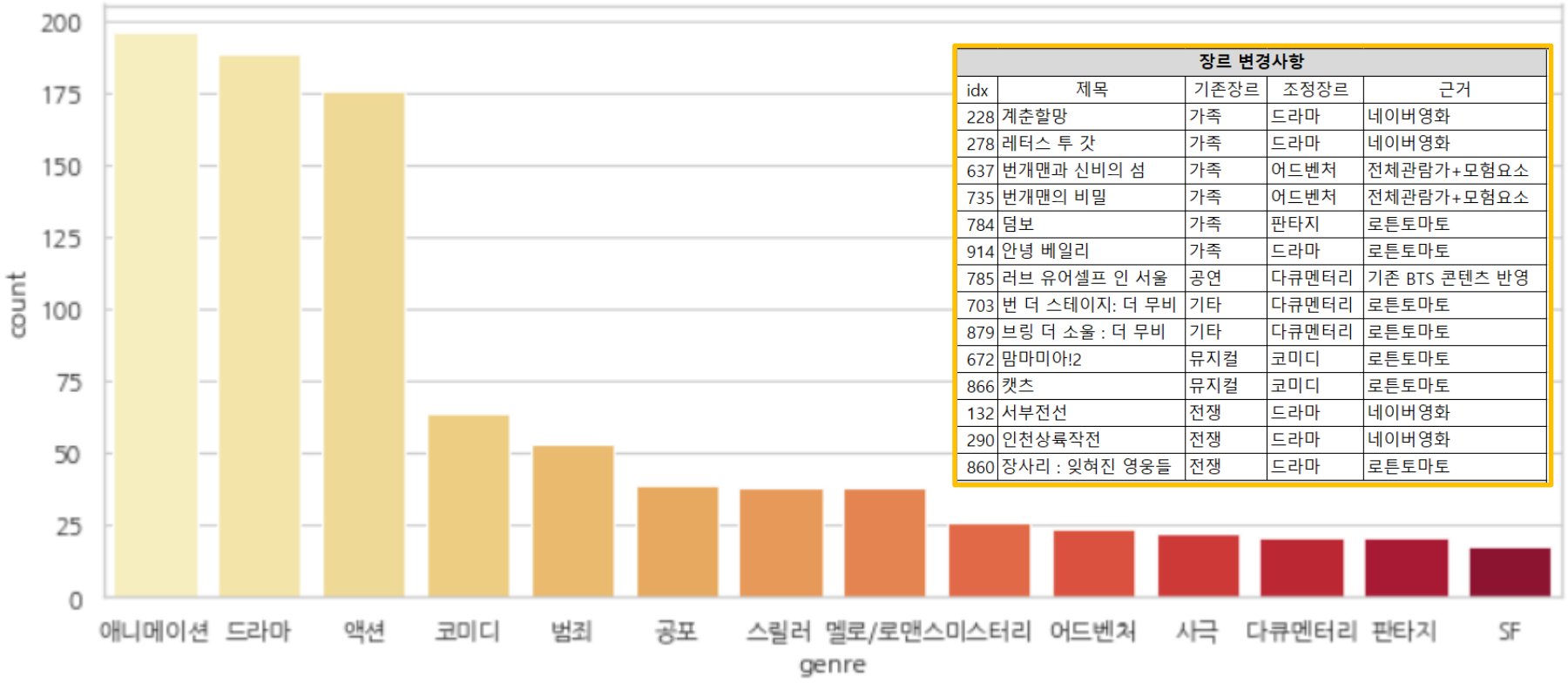


A	국내 메이저 배급사 (CJ엔터테인먼트, 롯데엔터테인먼트, 넥스트엔터테인먼트월드, 쇼박스 포함)
B	해외 메이저 배급사 (월트디즈니 픽처스, 워너브라더스 픽처스, 컬럼비아 픽처스, 유니버설 스튜디오, 파라마운트 픽처스 포함
C	기타

한국 개봉영화 관람객 예측

범주형 변수 정리 - 장르

<장르별 개봉영화 개수>

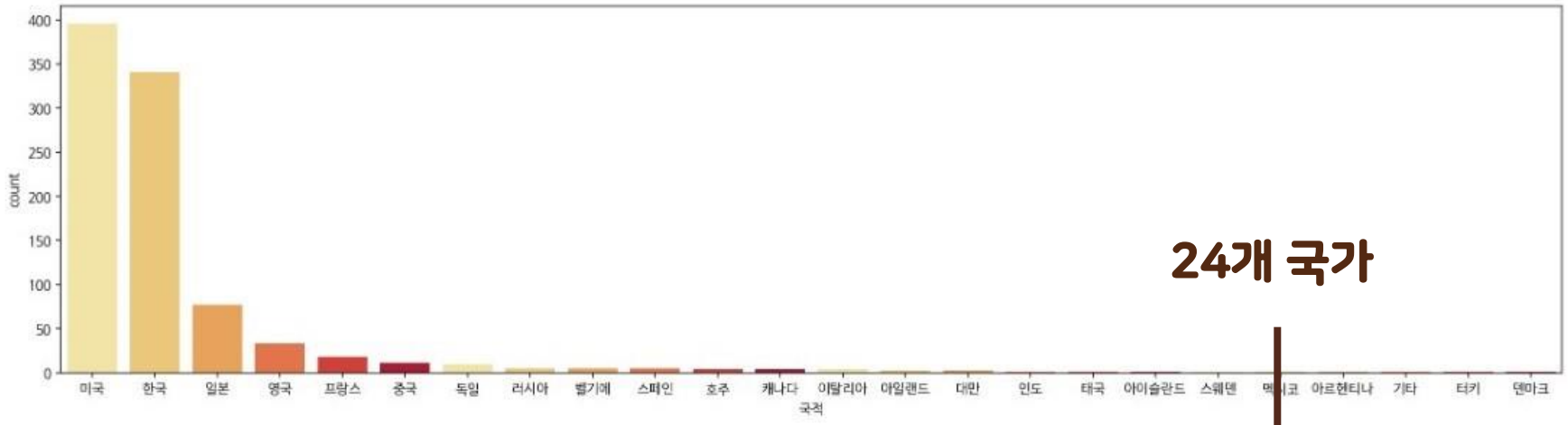


장르 변경사항				
idx	제목	기존장르	조정장르	근거
228	계춘할망	가족	드라마	네이버영화
278	레터스 투 갓	가족	드라마	네이버영화
637	번개맨과 신비의 섬	가족	어드벤처	전체관람가+모험요소
735	번개맨의 비밀	가족	어드벤처	전체관람가+모험요소
784	덤보	가족	판타지	로튼토마토
914	안녕 베일리	가족	드라마	로튼토마토
785	러브 유어셀프 인 서울	공연	다큐멘터리	기존 BTS 콘텐츠 반영
703	번 더 스테이지: 더 무비	기타	다큐멘터리	로튼토마토
879	브링 더 소울 : 더 무비	기타	다큐멘터리	로튼토마토
672	맘마미아!2	뮤지컬	코미디	로튼토마토
866	캣츠	뮤지컬	코미디	로튼토마토
132	서부전선	전쟁	드라마	네이버영화
290	인천상륙작전	전쟁	드라마	네이버영화
860	장사리 : 잊혀진 영웅들	전쟁	드라마	로튼토마토

19개 장르 ⟶ 14개 장르

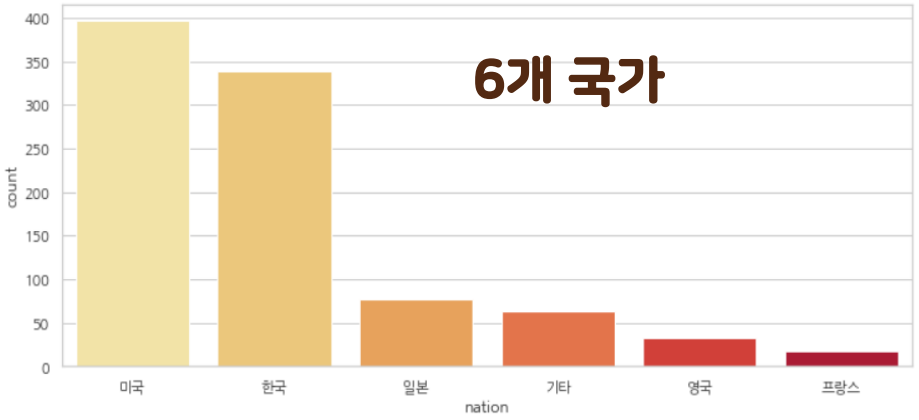
범주형 변수 정리 - 국가

<국가별 개봉영화 개수>



24개 국가

6개 국가

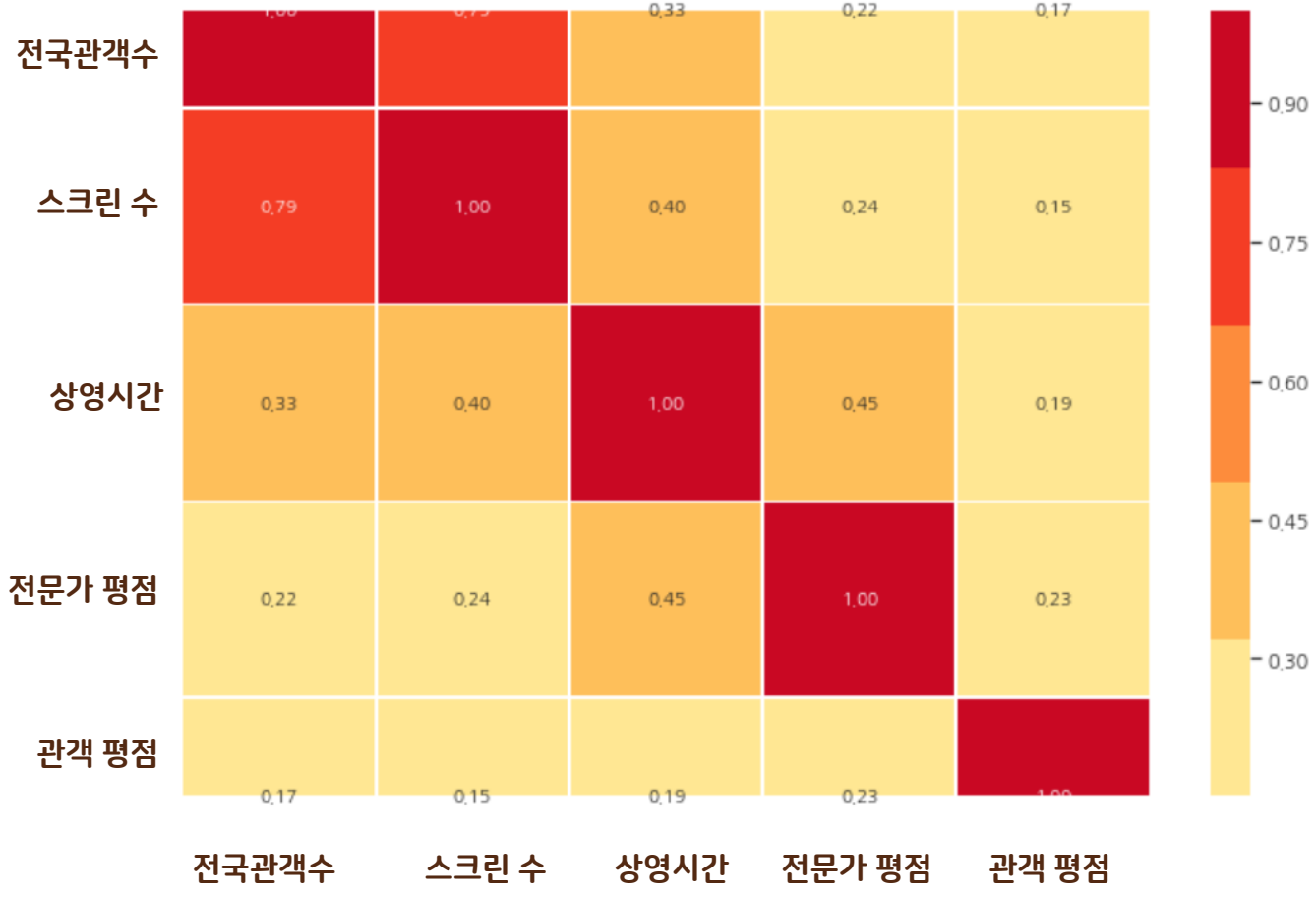


미국 > 한국 > 일본 > 기타 > 영국 > 프랑스

한국 개봉영화 관람객 예측

상관계수 히트맵

각 변수별 상관계수 확인



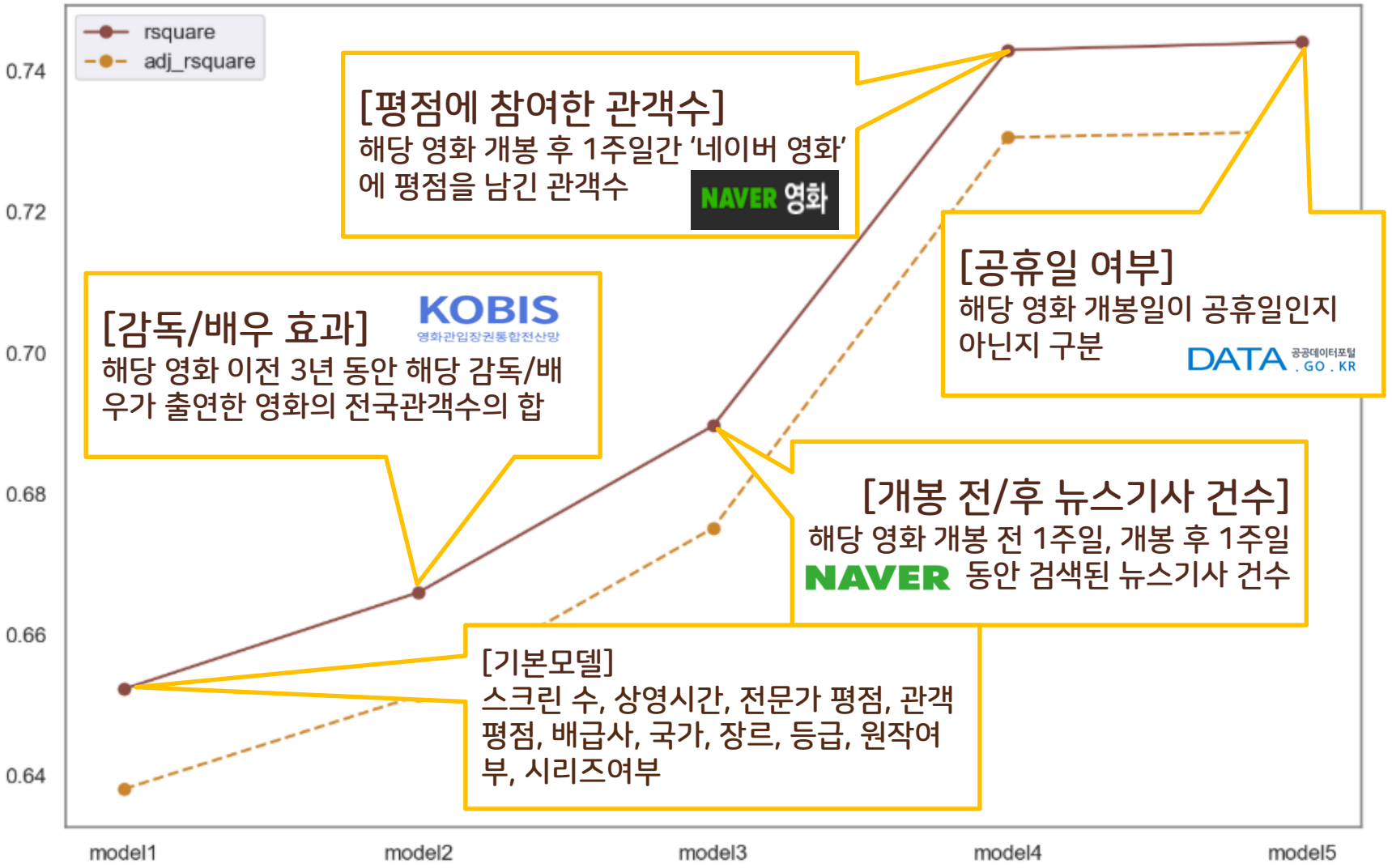
03
분석

Part 1

독립변수 추가

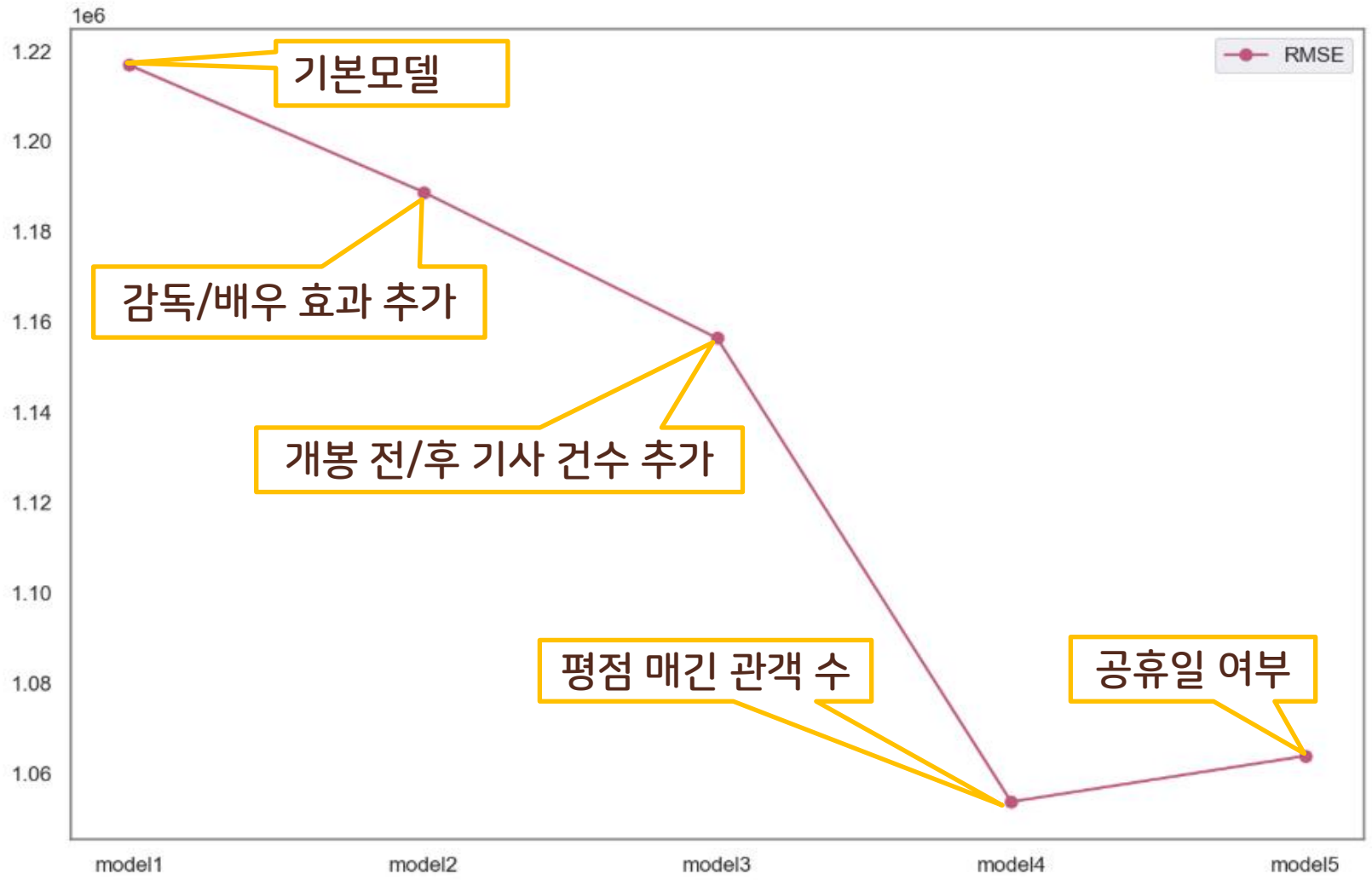
모델 01 – 모델 05 : 독립변수 추가

<각 모델의 결정계수(R^2) 비교>



모델 02 - 모델 05

<각 모델의 제곱근평균제곱오차 (RMSE) 비교>



Part 2

최종모델과 발전 과정

모델 09 - 최종

Formula

"log(전국관객수) ~ scale(스크린 수**(4/7)) + scale(상영시간) + scale(전문가 평점) + scale(관객 평점) + scale(감독파워) + scale(배우파워) + scale(log(개봉 전 뉴스)) + scale(log(평점 준 관객수)) + C(국가) + C(등급) + C(배급사) + C(장르) + C(시리즈여부) + C(원작여부) + C(공휴일)"

OLS Regression Results

Dep. Variable:	log_aud	R-squared:	0.811
Model:	OLS	Adj. R-squared:	0.801
Method:	Least Squares	F-statistic:	88.69
Date:	Sat, 18 Apr 2020	Prob (F-statistic):	3.11e-229
Time:	02:28:22	Log-Likelihood:	-732.10
No. Observations:	740	AIC:	1534.
Df Residuals:	705	BIC:	1695.
Df Model:	34		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.5648	0.217	57.987	0.000	12.139	12.990
dist_cat[T.B]	-0.0755	0.073	-1.030	0.303	-0.219	0.068
dist_cat[T.C]	-0.1714	0.077	-2.229	0.026	-0.322	-0.020
C(holiday)[T.1]	-0.0629	0.111	-0.568	0.570	-0.280	0.154

R-squared : 0.811
Adj. R-squared : 0.801

한국 개봉영화 관람객 예측

최종모델 분석 결과

OLS Regression Results

Dep. Variable:	log_aud	R-squared:	0.811
Model:	OLS	Adj. R-squared:	0.801
Method:	Least Squares	F-statistic:	88.69
Date:	Sat, 18 Apr 2020	Prob (F-statistic):	3.11e-229
Time:	02:28:22	Log-Likelihood:	-732.10
No. Observations:	740	AIC:	1534.
Df Residuals:	705	BIC:	1695.
Df Model:	34		
Covariance Type:	nonrobust		

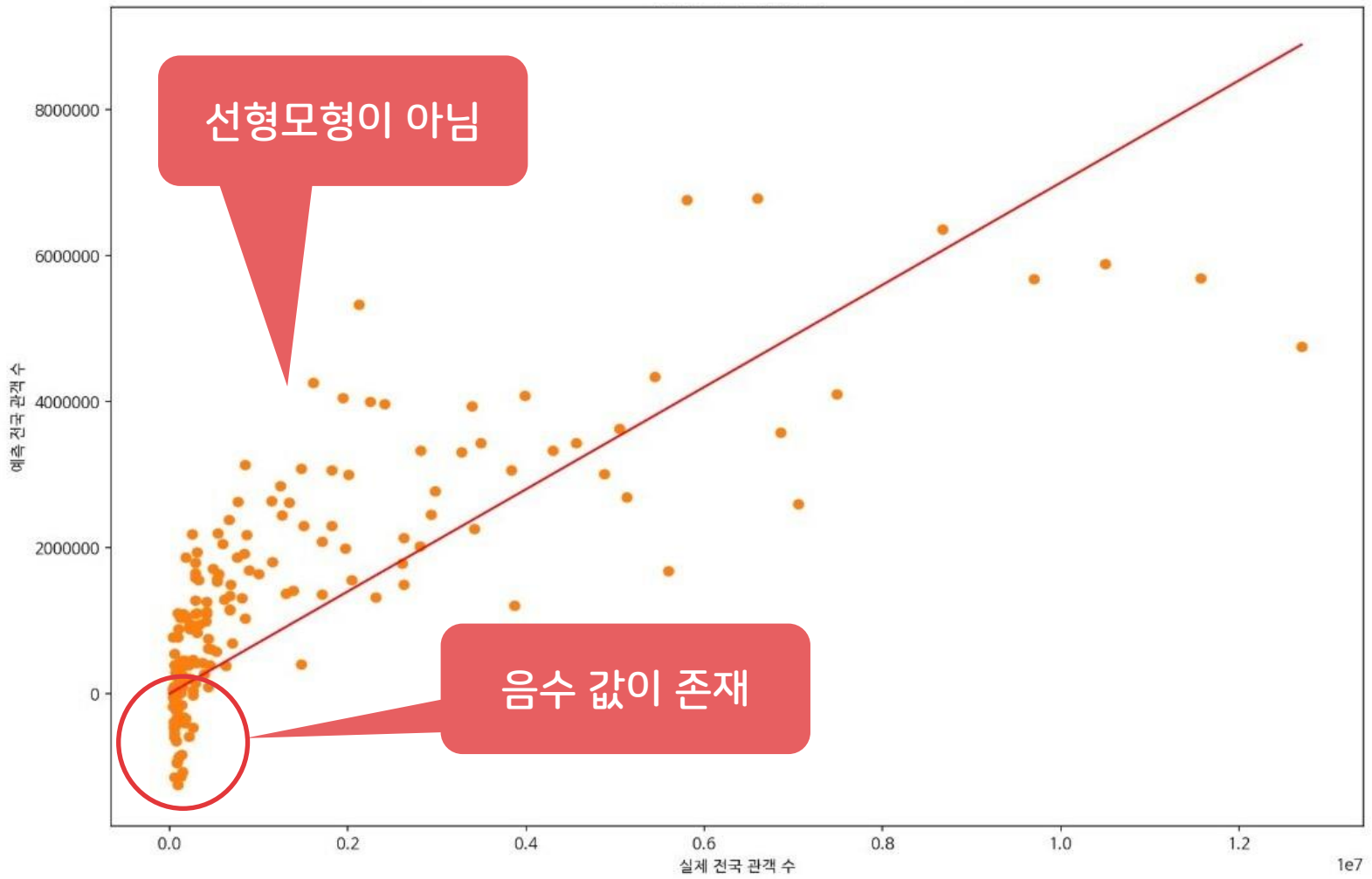
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.5648	0.217	57.987	0.000	12.139	12.990
dist_cat[T.B]	-0.0755	0.073	-1.030	0.303	-0.219	0.068
dist_cat[T.C]	-0.1714	0.077	-2.229	0.026	-0.322	-0.020
C(holiday)[T.1]	-0.0629	0.111	-0.568	0.570	-0.280	0.154
nation[T.미국]	-0.0241	0.118	-0.204	0.838	-0.256	0.208
nation[T.영국]	-0.0177	0.164	-0.108	0.914	-0.339	0.303
nation[T.일본]	0.1279	0.140	0.911	0.363	-0.148	0.404
nation[T.프랑스]	-0.2048	0.219	-0.937	0.349	-0.634	0.224
nation[T.한국]	-0.1070	0.132	-0.812	0.417	-0.366	0.152
genre[T.공포]	0.3926	0.210	1.870	0.062	-0.020	0.805
genre[T.다큐멘터리]	0.7861	0.255	3.083	0.002	0.285	1.287
genre[T.드라마]	0.2950	0.186	1.587	0.113	-0.070	0.660
genre[T.멜로/로맨스]	0.4442	0.221	2.012	0.045	0.011	0.878
genre[T.미스터리]	0.2164	0.226	0.958	0.338	-0.227	0.660
genre[T.범죄]	0.2900	0.205	1.414	0.158	-0.113	0.692
genre[T.사극]	0.5219	0.263	1.985	0.047	0.006	1.038
genre[T.스릴러]	0.3393	0.211	1.608	0.108	-0.075	0.754
genre[T.애니메이션]	0.4626	0.213	2.177	0.030	0.045	0.880
genre[T.액션]	0.4106	0.183	2.245	0.025	0.052	0.770
genre[T.어드벤처]	0.6405	0.233	2.745	0.006	0.182	1.099
genre[T.코미디]	0.4271	0.201	2.120	0.034	0.032	0.823
genre[T.판타지]	0.1353	0.241	0.562	0.575	-0.338	0.608
rate[T.15세이상관람가]	-0.0539	0.070	-0.771	0.441	-0.191	0.083
rate[T.전체관람가]	-0.1411	0.110	-1.279	0.201	-0.358	0.075
rate[T.청소년관람불가]	-0.1769	0.097	-1.830	0.068	-0.367	0.013
C(sequel)[T.1]	0.1573	0.058	2.712	0.007	0.043	0.271
C(original)[T.1]	-0.0227	0.056	-0.405	0.685	-0.133	0.087
scale(screen_sqrt)	0.8948	0.048	18.512	0.000	0.800	0.990
scale(expert_rating)	0.0703	0.030	2.307	0.021	0.010	0.130
scale(audience_rating)	-0.0263	0.031	-0.846	0.398	-0.087	0.035
scale(running_times)	0.0469	0.037	1.283	0.200	-0.025	0.119
scale(actor_power)	0.0362	0.032	1.144	0.253	-0.026	0.098
scale(director_power)	-0.0739	0.028	-2.676	0.008	-0.128	-0.020
scale(log_before_news)	0.1784	0.049	3.605	0.000	0.081	0.276
scale(log_rating_audiences)	0.3519	0.044	7.984	0.000	0.265	0.438

<유의미한 변수들>

	coef
Intercept	12.564784
scale(screen_sqrt)	0.894766
genre[T.다큐멘터리]	0.786093
genre[T.어드벤처]	0.640466
genre[T.사극]	0.521888
genre[T.애니메이션]	0.462609
genre[T.멜로/로맨스]	0.444206
genre[T.코미디]	0.427149
genre[T.액션]	0.410644
genre[T.공포]	0.392562
scale(log_rating_audiences)	0.351854
scale(log_before_news)	0.178407
C(sequel)[T.1]	0.157266
scale(expert_rating)	0.070317
scale(director_power)	-0.073943
dist_cat[T.C]	-0.171350
rate[T.청소년관람불가]	-0.176897

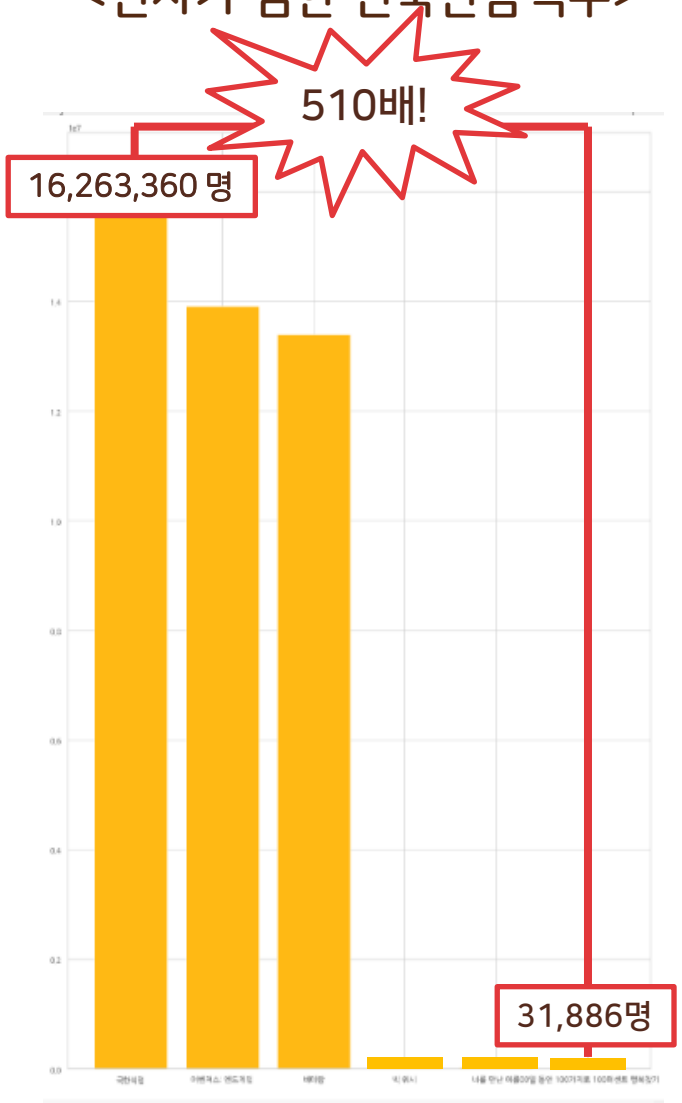
종속변수에 로그를 취한 이유

<실제 전국관객수와 예측값 >

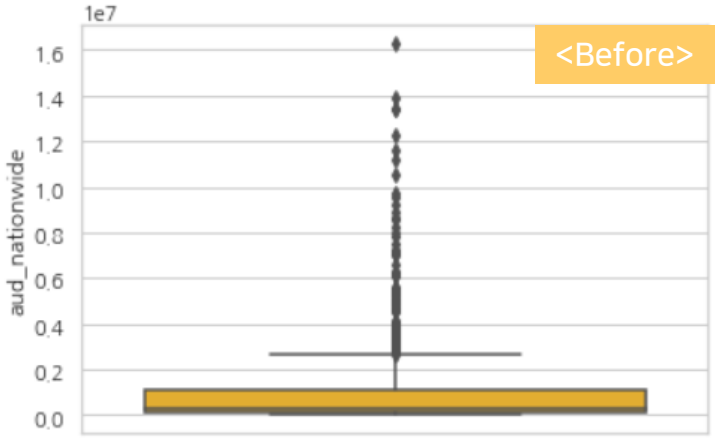


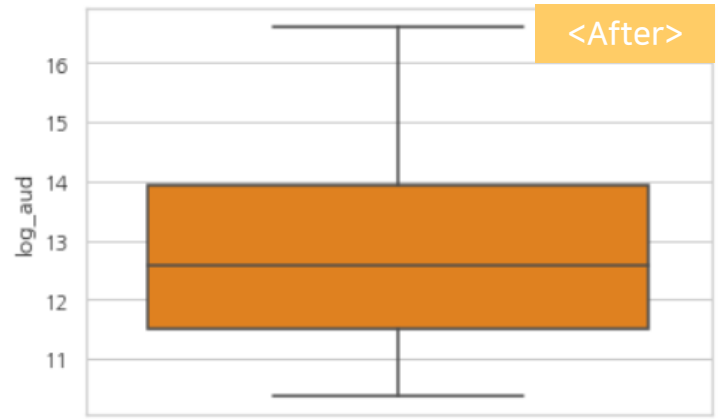
종속변수에 로그를 취한 이유

<편차가 심한 전국관람객수>



<로그 변환 전/후 전국관람객수 편차>

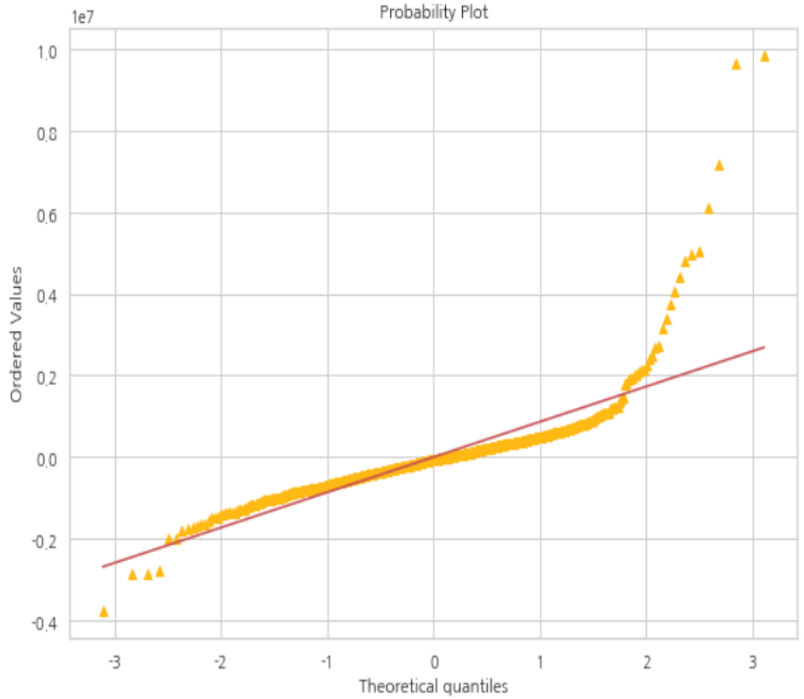




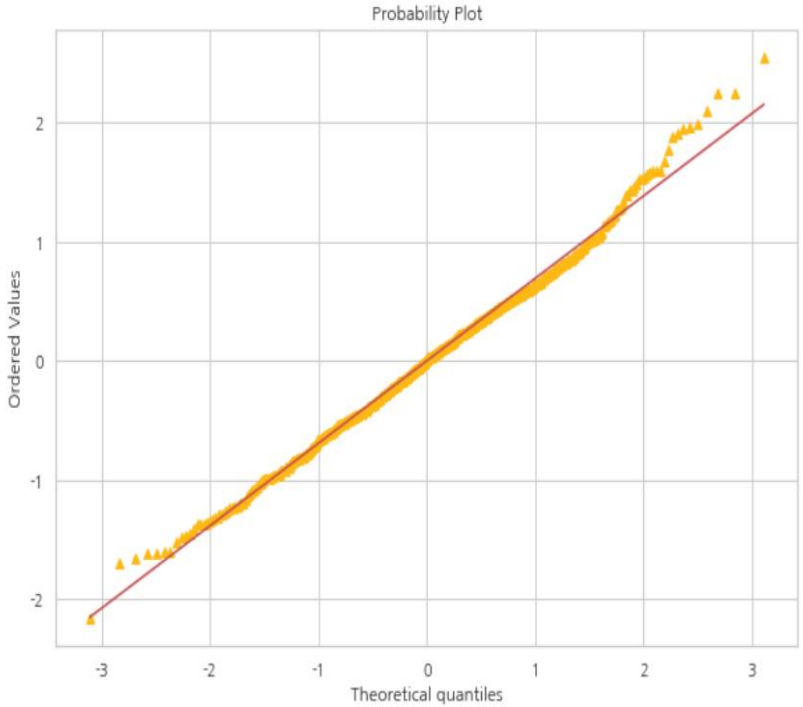
모델 06 – 종속변수 로그변환

Formula

"log(전국관객수) ~ scale(스크린 수) + scale(상영시간) + scale(전문가 평점) + scale(관객 평점) + scale(감독파워) + scale(배우파워) + scale(개봉 전 뉴스) + scale(개봉 후 뉴스) + scale(평점 준 관객수) + C(배급사) + C(국가) + C(장르) + C(등급) + C(원작여부) + C(시리즈여부) + C(공휴일)"



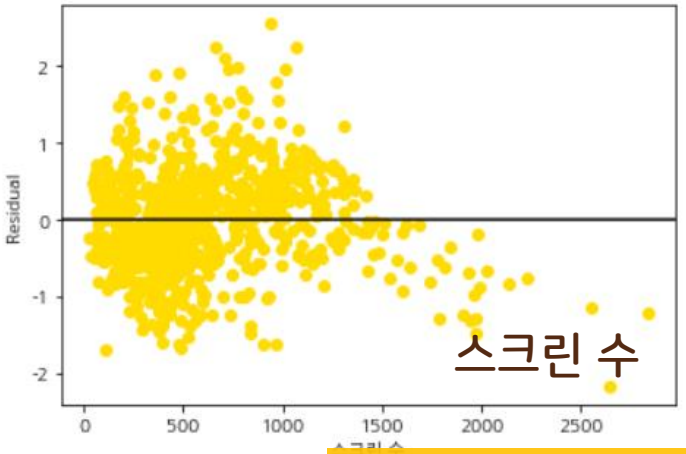
<종속변수 로그변환 전 QQ plot>

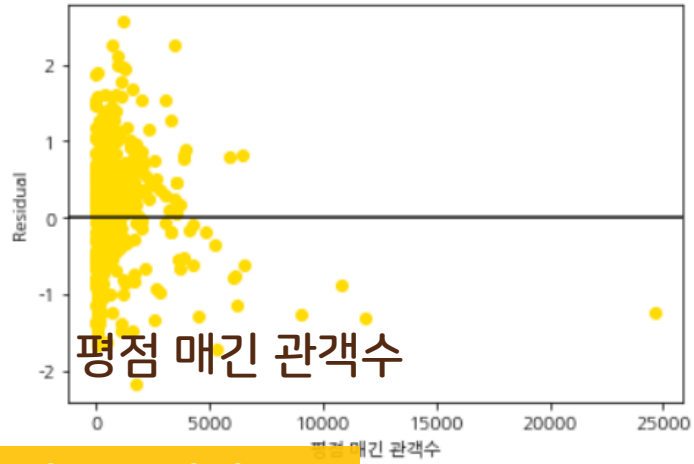


<종속변수 로그변환 후 QQ plot>

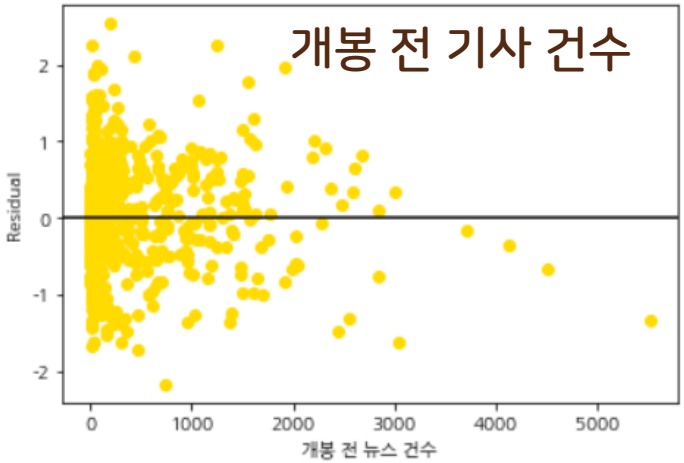
독립변수를 비선형 변환 한 이유

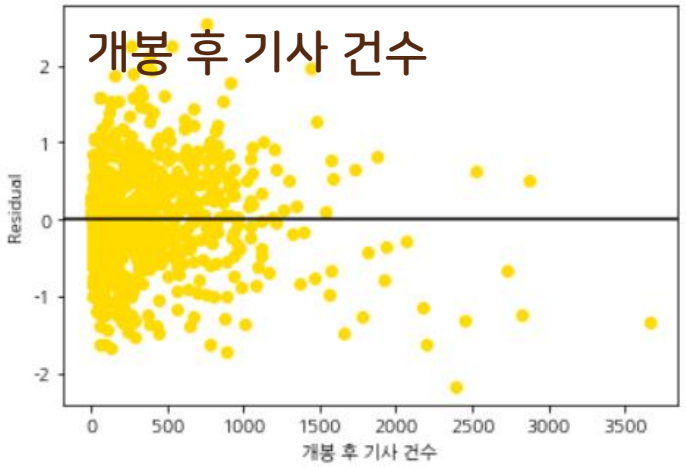
<모델06. 잔차와 독립변수의 관계>





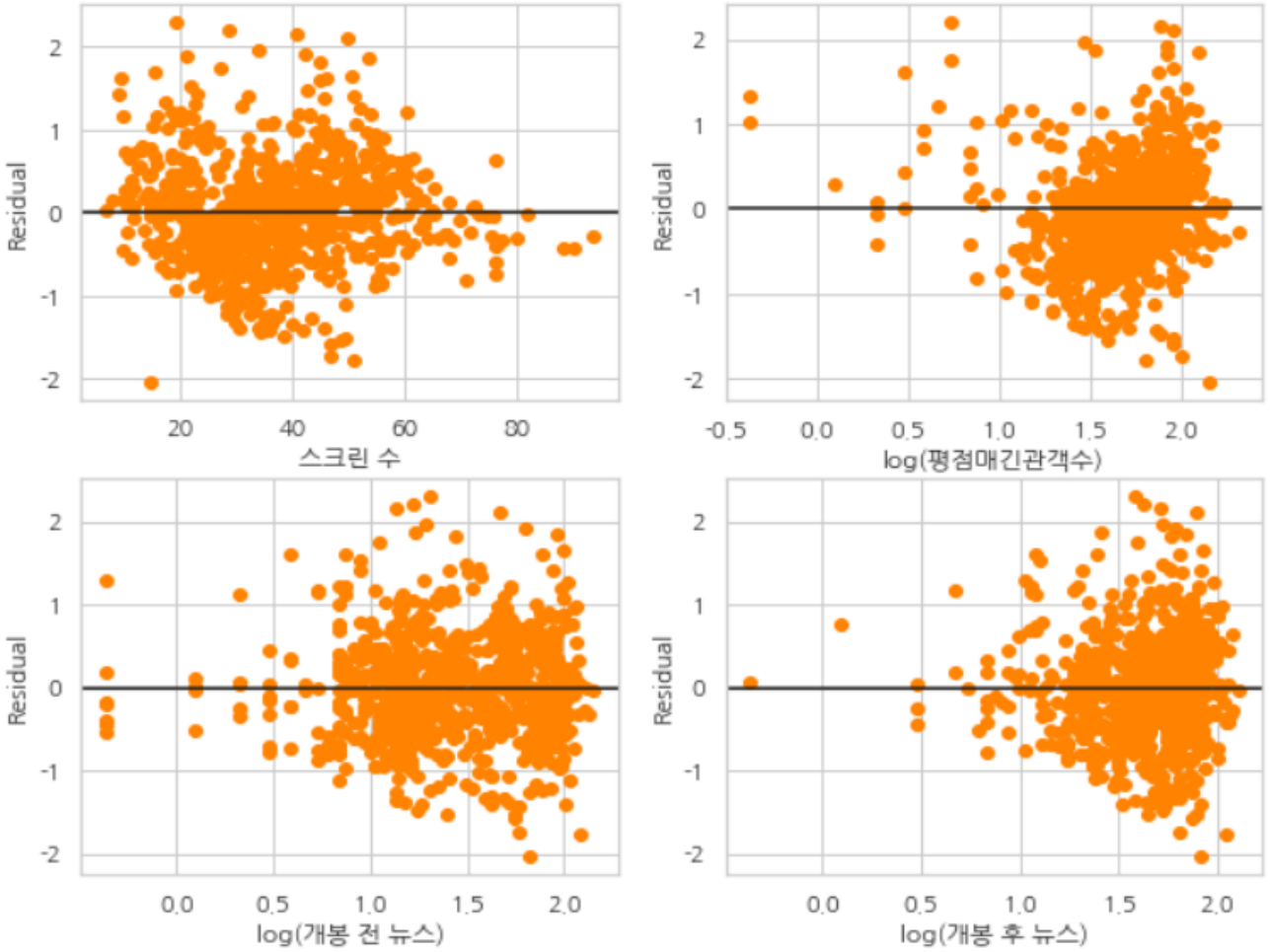
비선형 상관관계를 보이는 독립변수들





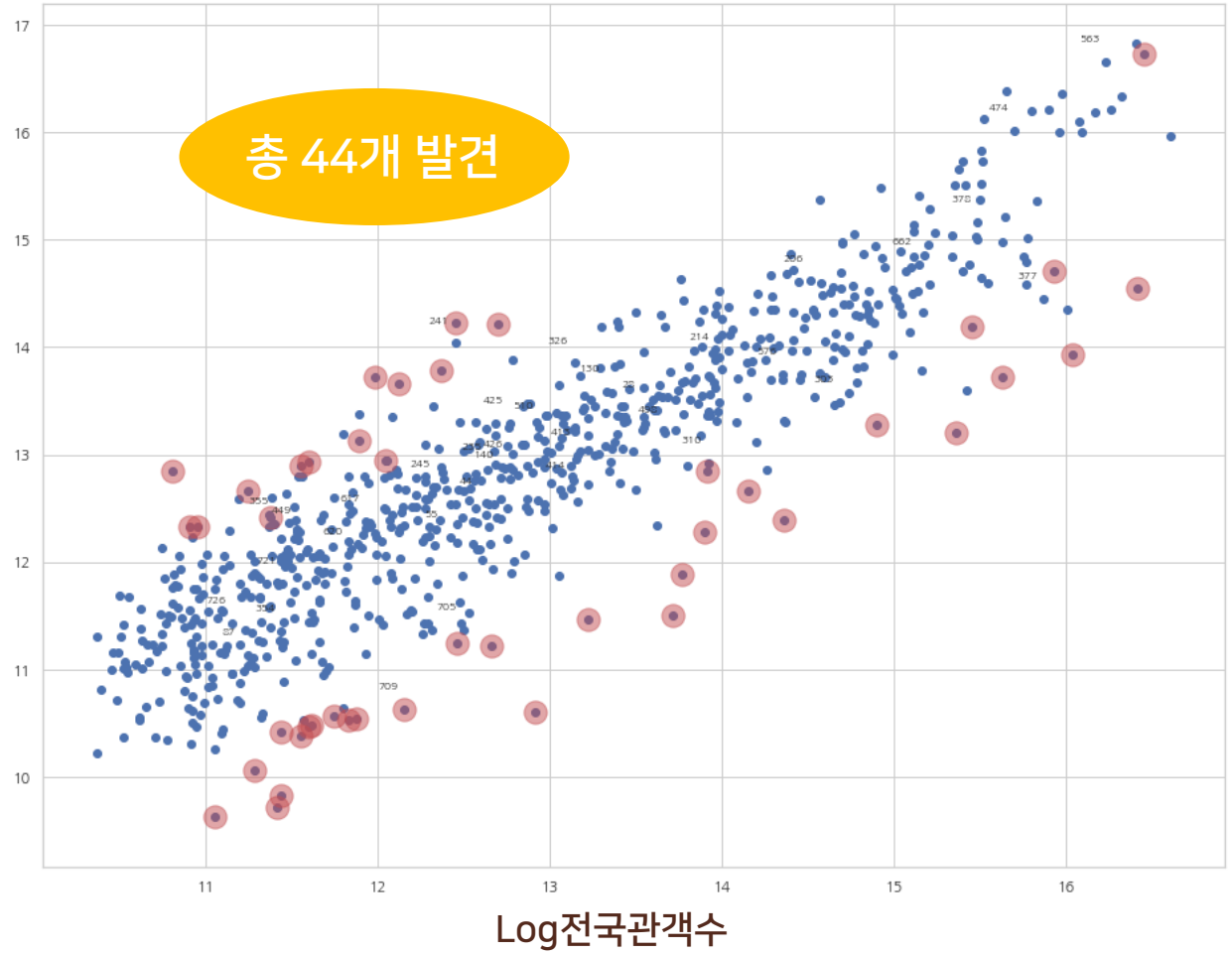
모델 07 - 스크린**(4/7), log(개봉전/후 뉴스, 평점 매긴 관객 수)

비선형 변형 후 잔차와 독립변수들과의 관계



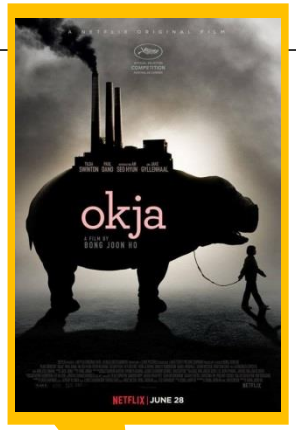
모델 07의 아웃라이어 확인

<Fox recommendation을 적용한 아웃라이어>



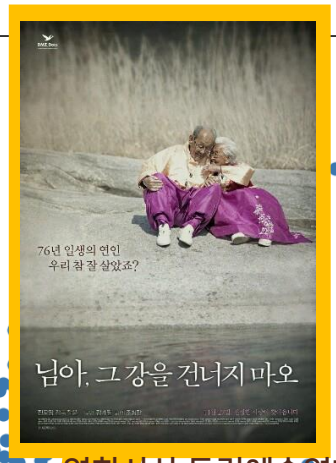
한국 개봉영화 관람객 예측

모델 08 - 7개 아웃라이어 제거



넷플릭스 동시개봉

옥자



영화사상 독립예술영화 전부문 흥행 1위

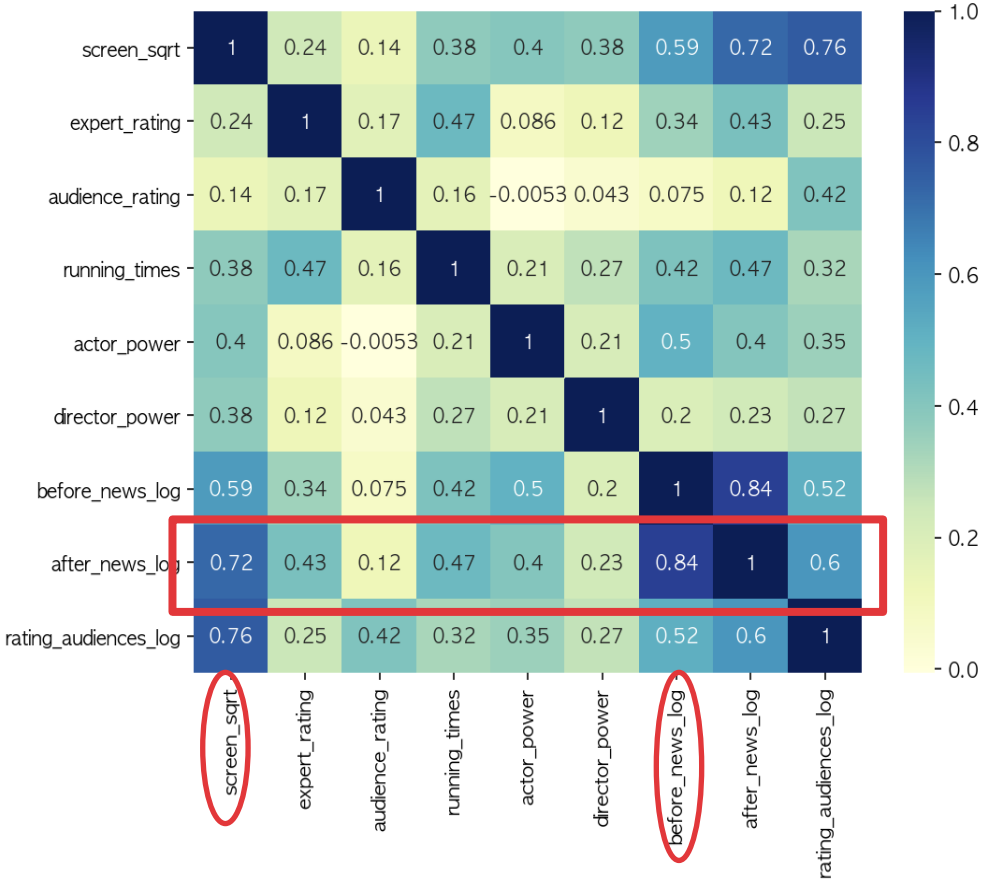
5월 12일 개봉한 '나의 소녀시대'는 지난 24일(영진위)까지 전국 40만 4,435명의 관객을 동원했다. 역대 대만 영화 최고 스코어다.

최근 침체된 아시아 영화로는 이례적으로 입소문 열풍과 배우들을 향한 팬들의 사랑으로 이뤄 낸 성과다.

출처 : 노컷뉴스, 영화 <나의 소녀시대> '40만 관객 돌파'가 전하는 의미

모델 09 – VIF로 다중공선성 제거

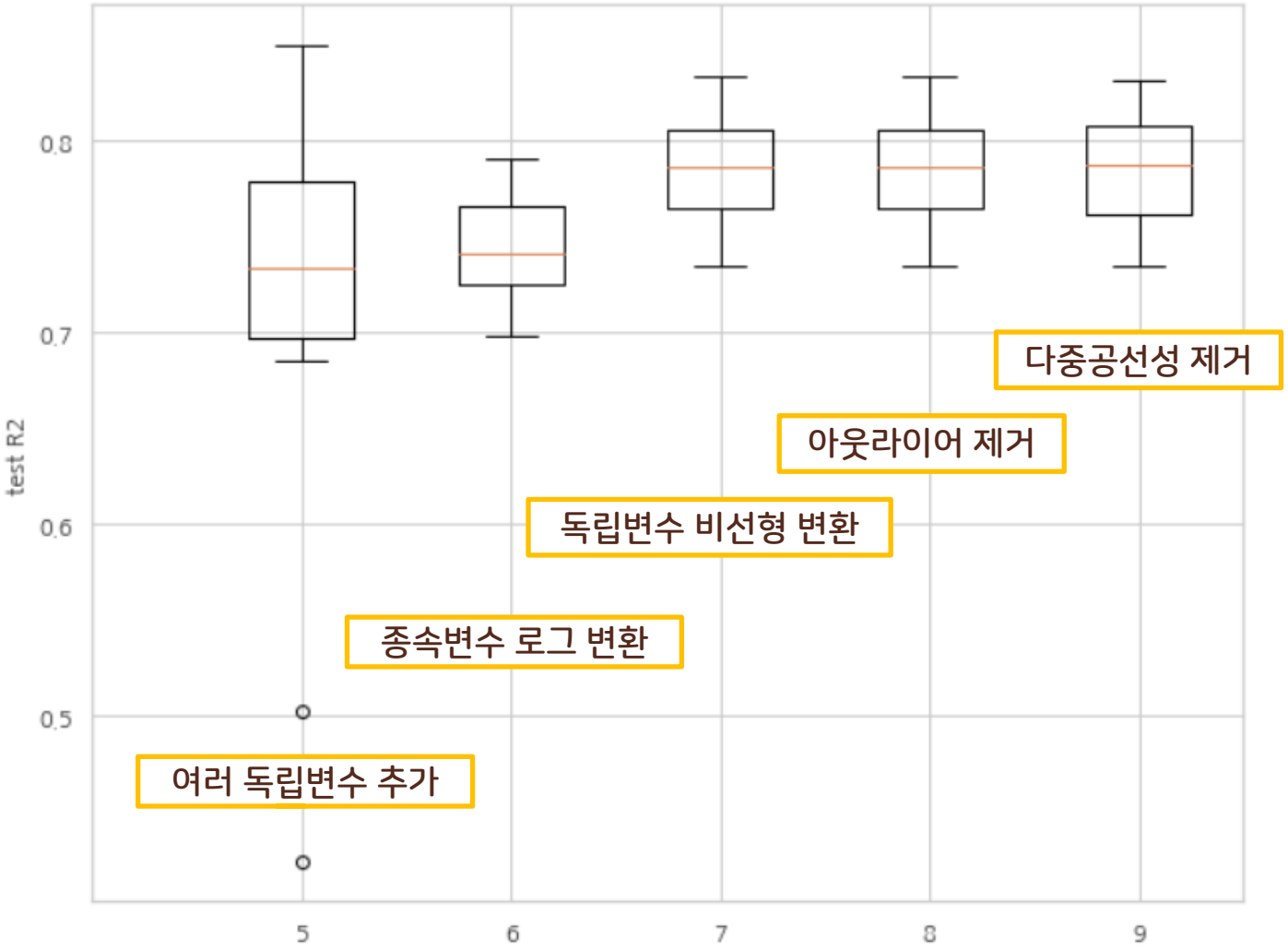
	VIF Factor	features
7	78.889837	log_after_news
3	36.728179	running_times
2	31.480788	audience_rating
8	30.700569	log_rating_audiences
6	26.720469	log_before_news
0	25.837795	screen_sqrt
1	11.551708	expert_rating
4	2.118846	actor_power
5	1.285970	director_power



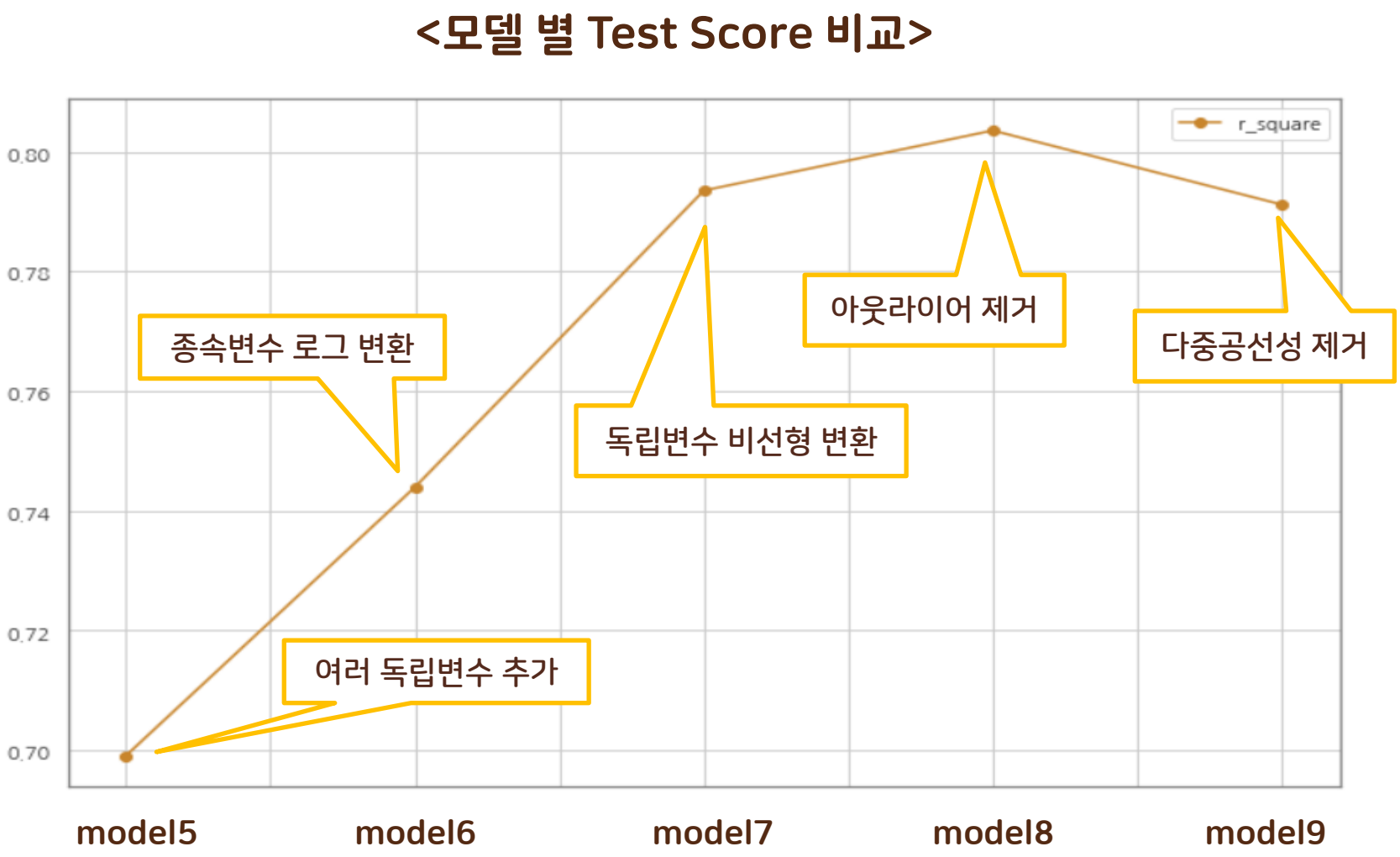
"개봉 후 기사 건수"의 VIF 수치 최고! "스크린 수", "개봉 전 기사 건수"와 높은 상관관계

모델05 - 모델10의 K-Fold

<모델 별 K-Fold 결과>



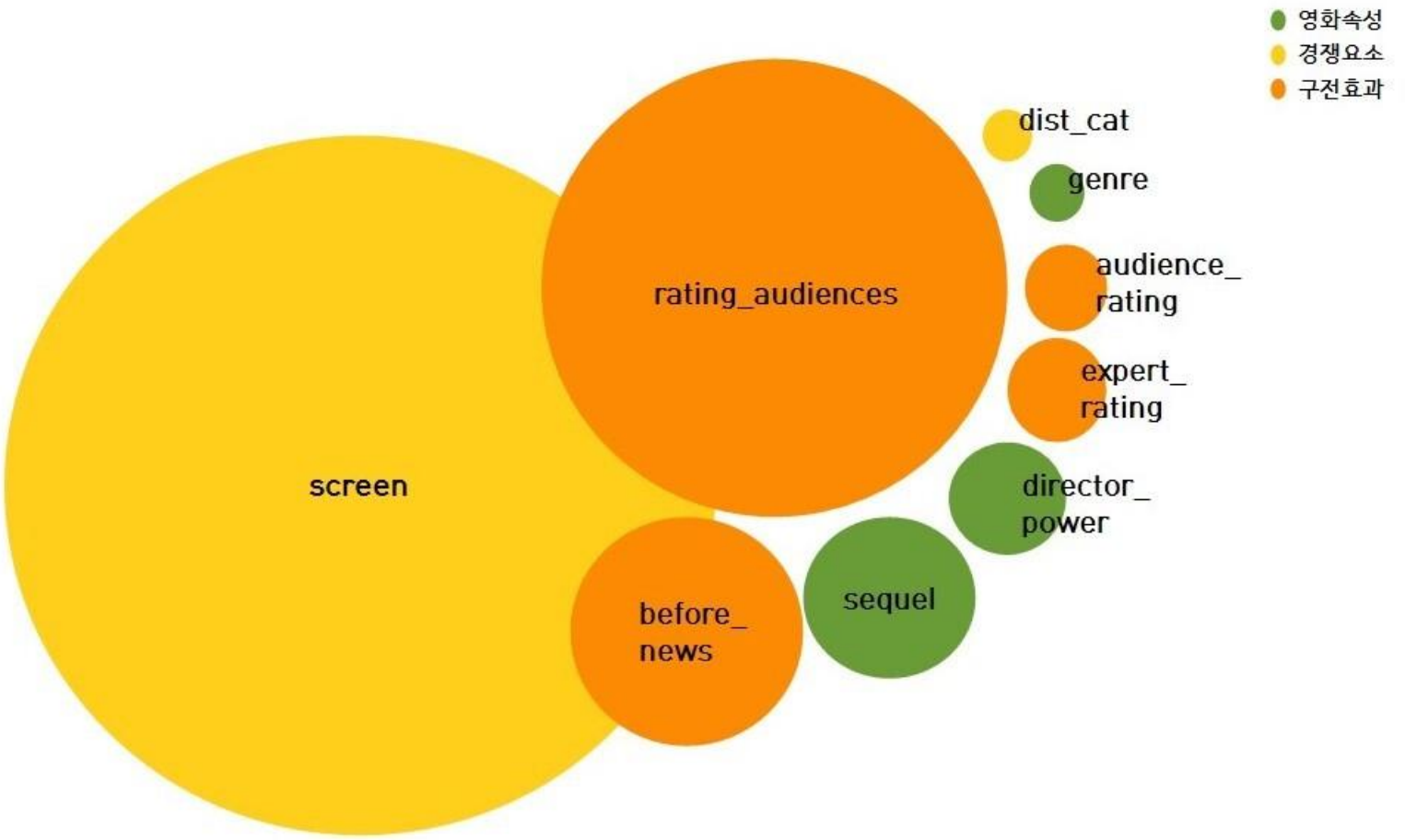
모델05 - 모델10의 Test Score



04
결과 요약 및 제언

분석 결과 요약

<영화관람객 예측에 영향을 미치는 요인>



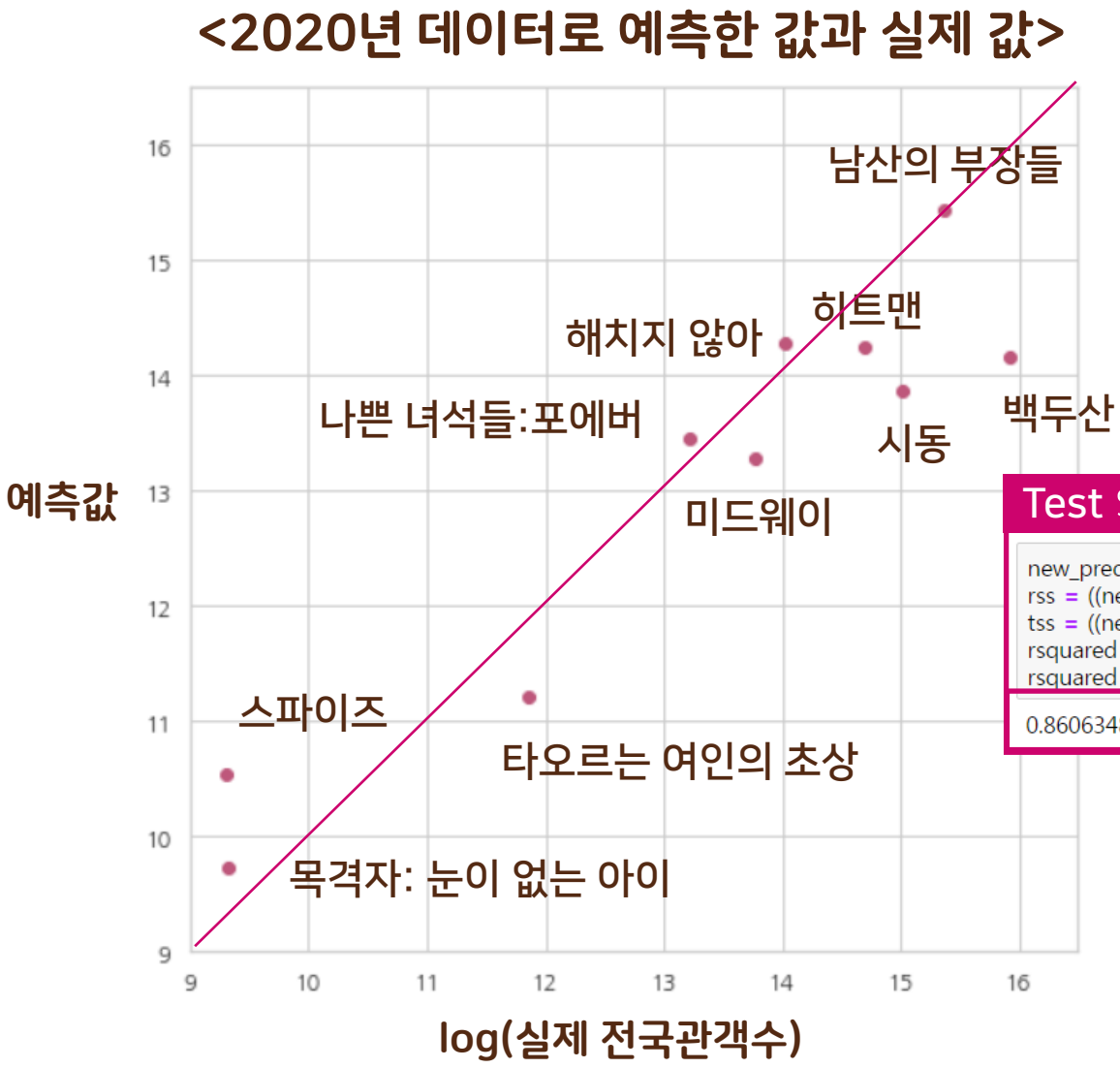
한국 개봉영화 관람객 예측

최종모델로 2020년 데이터 테스트

<2020년 데이터>

영화제목	전국관객수	스크린수	상영시간	배급사	국가	장르	등급	배우효과	감독효과	관객평점	전문가평점	평점매긴관객수	개봉전기사	공휴일	시리즈	원작
백두산	8,252,187	1241	128	A	한국	액션	12세이상관람가	117,627,166	29,747,237	7.844660194	5.67	1133	2820	0	0	0
미드웨이	955,294	842	136	C	미국	액션	15세이상관람가	21,830,583	1,500,744	8.766129032	5.5	248	33	0	0	1
시동	3,317,847	825	102	A	한국	드라마	15세이상관람가	61,023,236	-	8.919947507	6	762	960	0	0	1
남산의 부장들	4,750,104	1659	114	A	한국	드라마	15세이상관람가	51,729,579	1,864,077	8.514680484	6.9	1158	3203	0	0	1
히트맨	2,405,885	1122	110	A	한국	액션	15세이상관람가	11,184,384	-	8.238853503	5.67	314	830	0	0	0
해치지않아	1,224,726	1216	117	C	한국	코미디	12세이상관람가	9,159,551	-	7.846774194	6.5	248	787	0	0	1
스파이즈	11,000	224	98	C	프랑스	애니메이션	전체관람가	-	-	7.6	6	5	5	0	0	0
타오르는 여인의 초상	141,945	146	121	C	프랑스	드라마	12세이상관람가	39,658	10,591	9.396226415	9.22	106	61	0	0	0
목격자-눈이 없는 아이	11,235	43	87	C	기타	공포	15세이상관람가	30,116	-	10	5	1	74	0	0	1
나쁜 녀석들: 포에버	547,775	896	124	B	미국	액션	청소년관람불가	2,736,281	-	8.674846626	5.67	163	37	0	1	0

최종모델로 2020년 데이터 테스트



Test Score

```
new_pred = result9.predict(new_df)
rss = ((new_df['log_aud'] - new_pred) ** 2).sum()
tss = ((new_df['log_aud'] - new_df['log_aud'].mean()) ** 2).sum()
rsquared = 1 - rss/tss
rsquared
```

```
0.8606348493266779
```

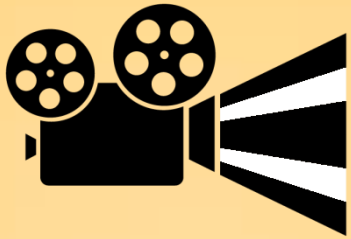
보완할 점과 추후 연구 방향 제안

아쉬운 점

- 직접적으로 수치화 할 수 없는 요소들(스토리, 연출의 질 등)의 영향력을 반영하지 못함
- 영화산업 흐름에 맞는 데이터를 수집하기 어려웠음
  - 넷플릭스 등 OTT 플랫폼을 통해 동시개봉하는 영화가 늘고 있는 추세
  - 이와 관련한 데이터(동시개봉여부 등) 얻기 어려움

추후 연구 방향 제안

- 자연어처리를 통한 스토리와 전국 관람객간의 관계 분석
- 매체(영화관, IPTV, OTT플랫폼 등)에 따른 인기도 분석



THANK YOU