

영화의 흥행 요인 분석 및 개봉 예정 영화의 관객수 예측

- 훈련 과정명 : 파이썬과 R을 활용한 빅데이터 분석(머신러닝, 딥러닝) 전문가 양성과정
- 훈련 기간 : 2019-01-11 ~ 2019-07-12
- 팀명 : NA
- 팀장 : 정호진 | 팀원 : 김지수 박건우 이형욱

NA팀 구성원

정호진 | 팀장

- ✓ 역할분담 및 의견수렴
- ✓ 공용저장소(GITHUB) 구축
- ✓ 데이터 수집(Naver, KMDB)
- ✓ 데이터 분석(분산 분석) 및 시각화
- ✓ 회귀

이형욱 | 팀원

- ✓ 데이터 수집(Naver)
- ✓ 데이터 분석(상관관계 분석, 텍스트 마이닝) 및 시각화
- ✓ 회귀
- ✓ 코드정리

김지수 | 팀원

- ✓ 데이터 수집(TMDB, Naver)
- ✓ 데이터 분석(텍스트 마이닝, 상관관계 분석) 및 시각화
- ✓ 회귀
- ✓ 보고서 작성

박건우 | 팀원

- ✓ 데이터 수집(KOBIS, KOFIC)
- ✓ 데이터 분석(분산 분석) 및 시각화
- ✓ 회귀 결과 분석
- ✓ 발표

목차



1. 서론

- 1.1 연구의 배경 및 필요성
- 1.2 연구 문제 및 가설

2. 연구 절차

3. 자료 수집

- 3.1 출처 및 변수 종류
- 3.2 데이터 전처리
- 3.3 변수 정의

4. 연구 결과

- 4.1 분석 결과
- 4.2 회귀 결과
- 4.3 예측 결과

5. 결론 및 한계점

- 5.1 연구의 결론
- 5.2 연구의 한계점

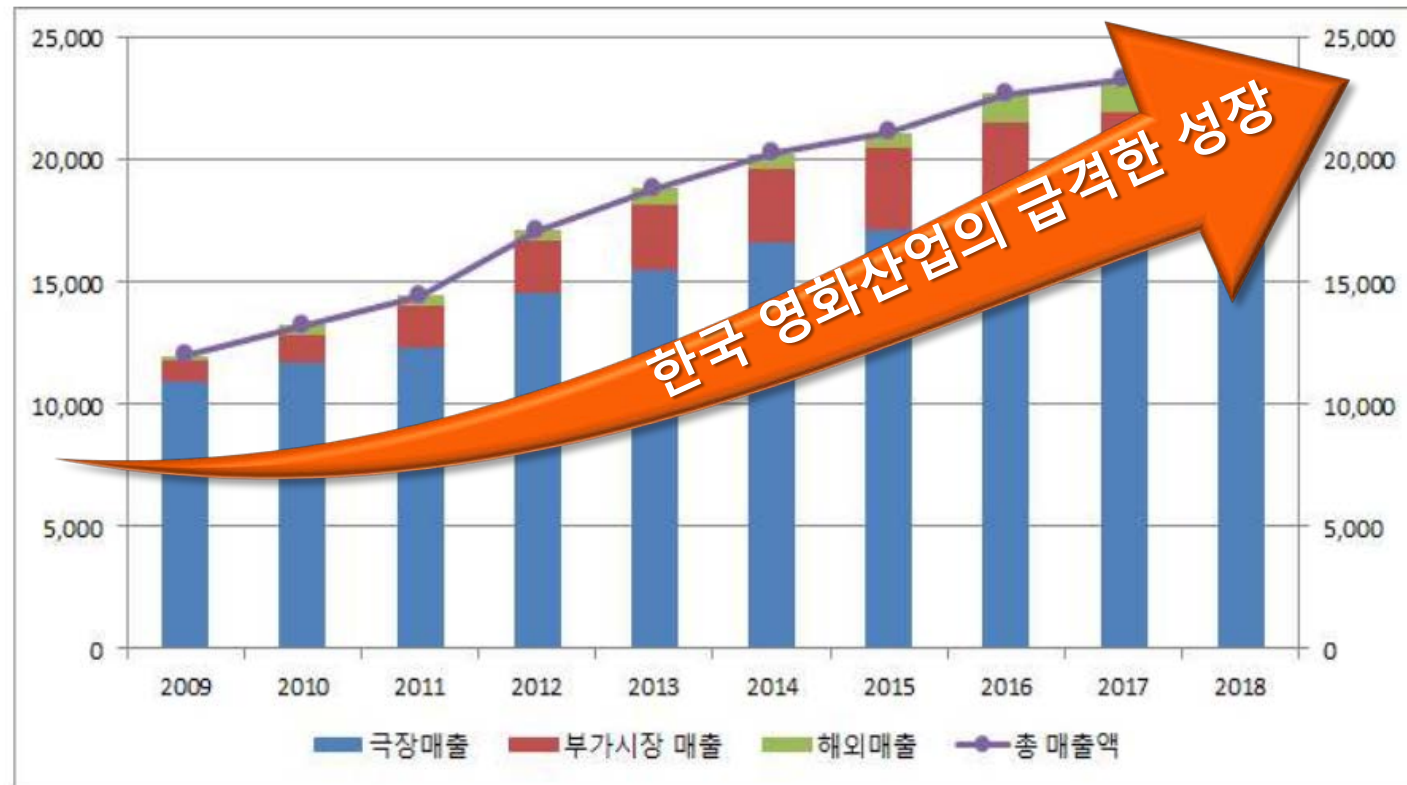
<참고문헌>

1. 서론

1.1 연구의 배경 및 필요성

<그림 1> 2009-2018년 한국 영화산업 매출 추이

(단위 : 억 원)



1. 서론

1.1 연구의 배경 및 필요성



기생충 칸영화제 첫 '황금종려상' 수상... '칸' 콧대 꺾은 봉준호 매직

한국경제 - 2019. 5. 26.

봉준호 감독의 '기생충'이 한국영화 사상 처음으로 세계 최고 권위의 국제영화제인 칸 영화제에서 최고상인 황금종려상을 받았다. 한국영화가 세계 3대 영화제(칸· ...

'칸' 선택은 봉준호... '기생충' 韓 최초 황금종려상

MBC뉴스 - 2019. 5. 26.

[빅픽처] 봉준호 '기생충' 황금종려상이 갖는 놀라운 의미

SBS연예뉴스 - 2019. 5. 26.

"황금종려상 수상 판타지같은 일... 위대한 한국감독·배우 덕분"

심층 뉴스 - 매일경제 - 2019. 5. 26.

[사설] 봉준호 감독 황금종려상, 한국 영화 100년의 쾌거

한겨레 - 2019. 5. 26.

봉준호, 세계 영화 지형도 바꿨다... 거장들 총출동한 칸서 韓 최초 황금 ...

심층 뉴스 - 동아일보 - 2019. 5. 26.

[모두 보기](#)

한국 영화의 국제적 위상 확대

1. 서론

1.1 연구의 배경 및 필요성

<표 2> 2009-2018년 한국 영화산업 주요 통계지표

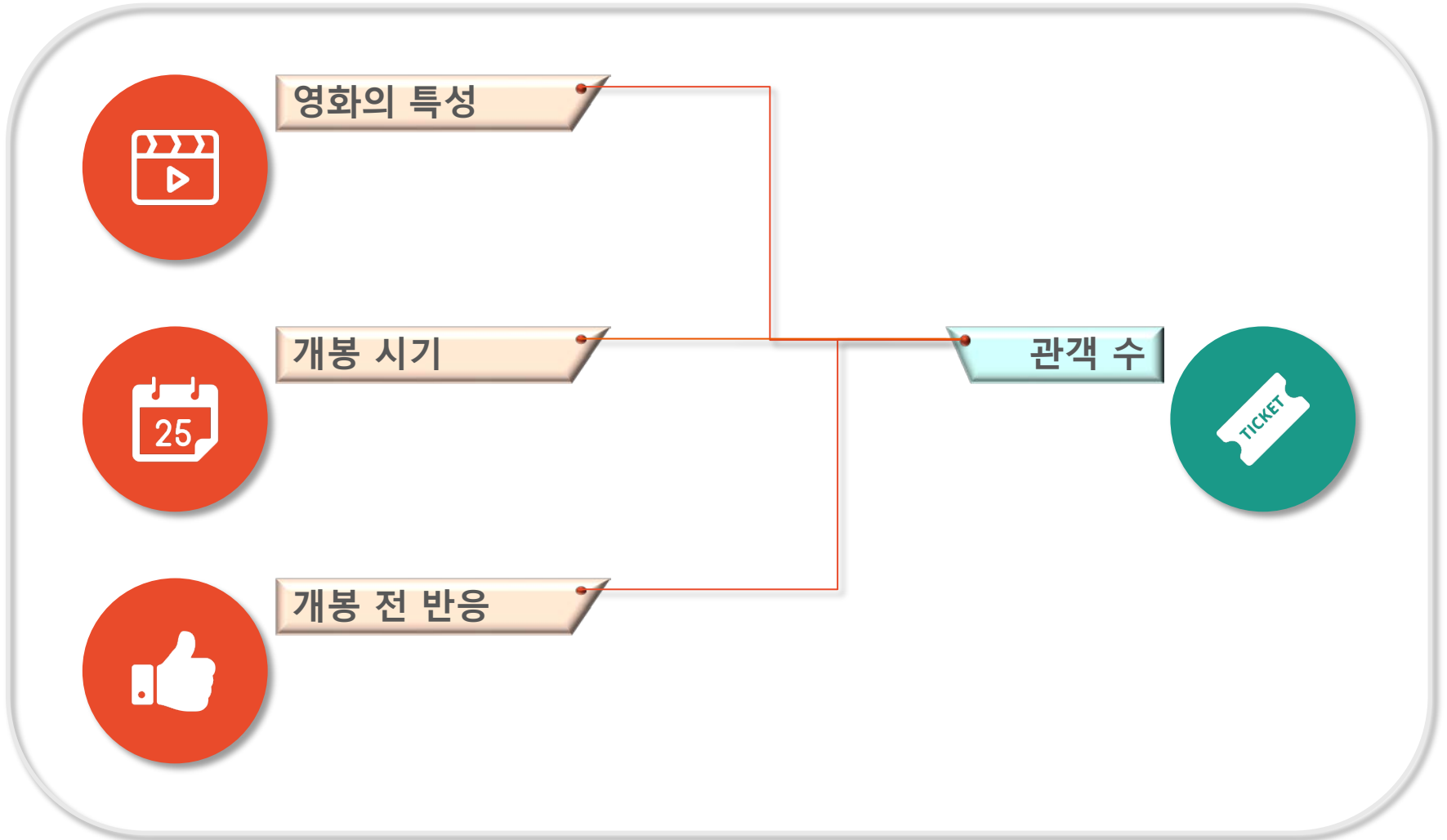
구분		2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
관객 수 (만 명)	총 관객 수	15,696	14,918	15,972	19,489	21,335	21,506	21,729	21,702	21,987	21,639
	한국영화	7,641	6,940	8,287	11,461	12,729	10,770	11,293	11,655	11,390	11,015
	점유율	48.7%	46.5%	51.9%	58.8%	59.7%	50.1%	52.0%	53.7%	51.8%	50.9%
	외국영화	8,055	7,978	7,685	8,028	8,606	10,736	10,436	10,047	10,597	10,624
	점유율	51.3%	53.5%	48.1%	41.2%	40.3%	49.9%	48.0%	46.3%	48.2%	49.1%
개봉 편수 (편) ⁶⁾	한국영화 (실질개봉)	118	140	150	175	183	217	232	302 (167)	376 (164)	454 (194)
	외국영화 (실질개봉)	243	286	289	456	722	878	944	1,218 (411)	1,245 (456)	1,192 (534)
전국 스크린 수 (개)		2,055	2,003	1,974	2,081	2,184	2,281	2,424	2,575	2,766	2,937
전국 극장 수 (개)		305	301	292	314	333	356	388	417	452	483
1인당 관람횟수 (회)		3.15	2.92	3.15	3.83	4.17	4.19	4.22	4.20	4.25	4.18
한국영화 투자수익률 ⁷⁾		-13.1%	-11.0%	-16.5%	15.9%	16.8%	7.6%	4.0%	17.6%	18%	-17.3%

2018년 한국 영화산업 결산 (영화진흥위원회 영화정책연구원, 2018)

최근 한국영화 투자수익률의 부진

1. 서론

1.2 연구 문제 및 가설



2. 연구 절차



3.1 출처 및 변수 종류

독립
변수

개봉 시기

영화
특성

개봉 전
반응

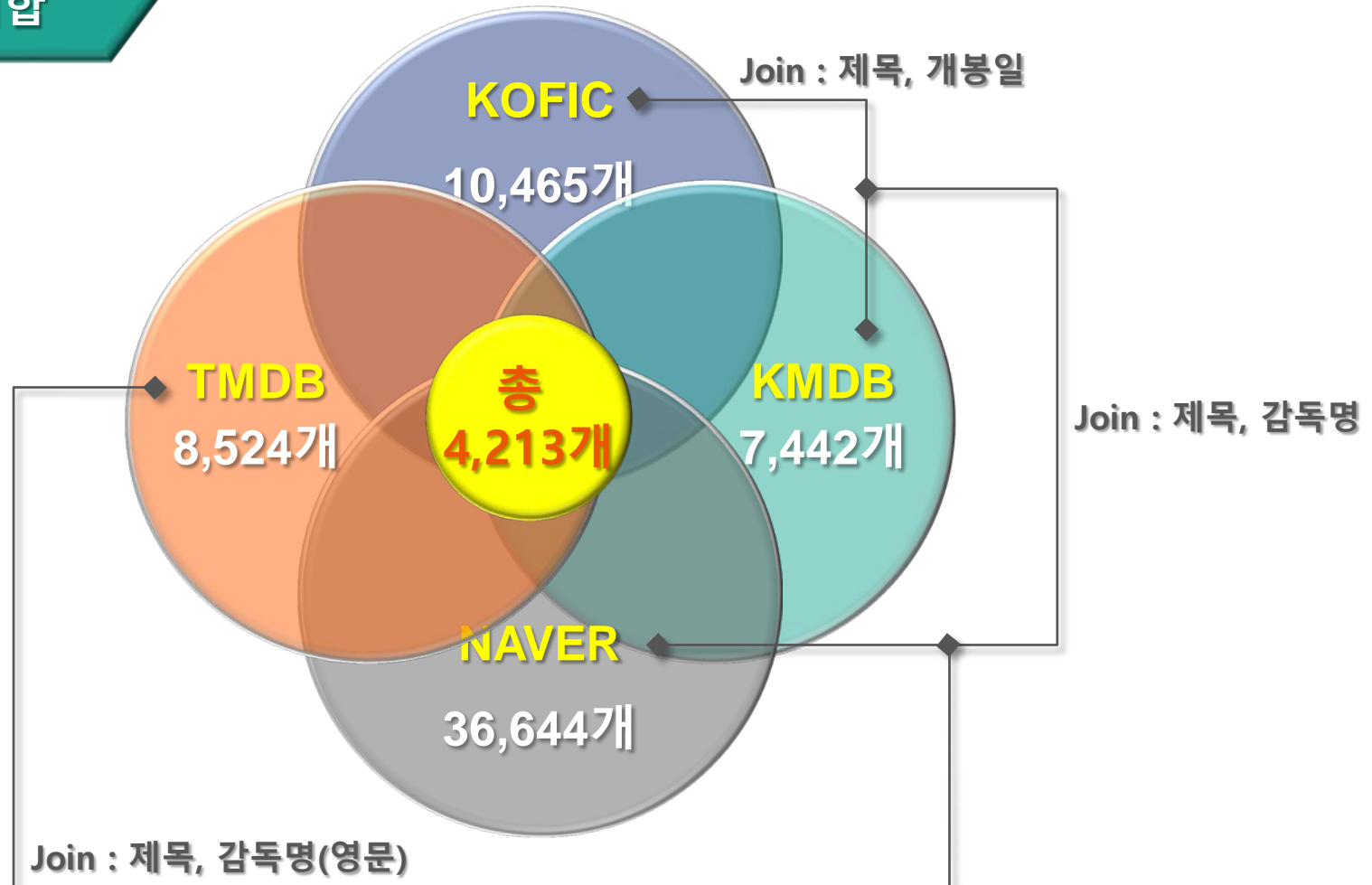
종속
변수

출처 \ 변수	제목	개봉일	감독	상영시간	제작국가	장르	관람등급	배급사	언어	예산	배우	시리즈	키워드	줄거리	수상내역	원작도서	평점	댓글	댓글개수	기대지수	누적관객수
KOFIC	●	●	●	●	●	●	●	●													●
KMDB	●	●	●										●	●	●						
TMDB	●	●	●						●	●	●	●	●								
NAVER	●		●													●	●	●	●	●	

3. 자료 수집

3.2 데이터 전처리

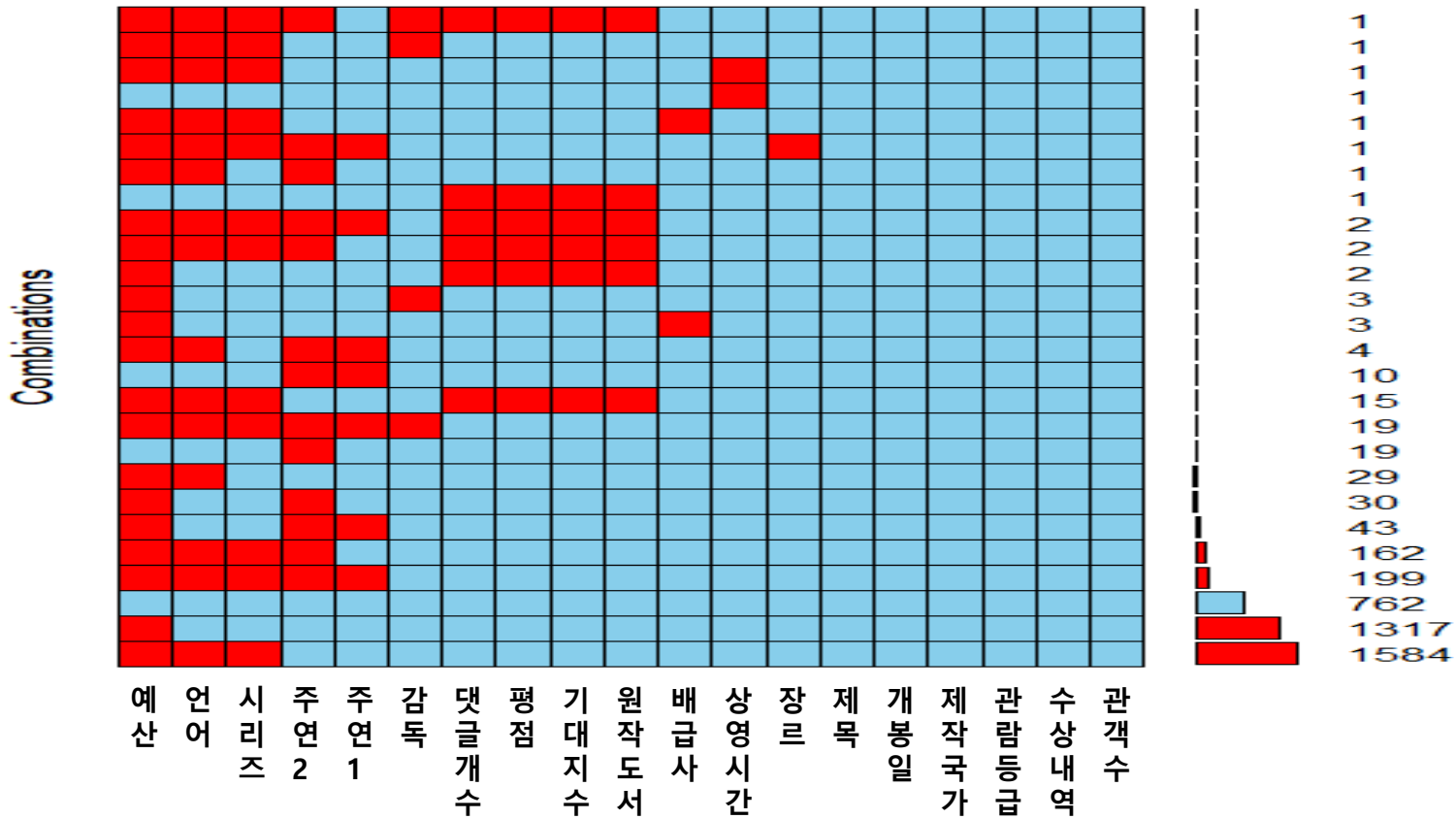
데이터 취합



3. 자료 수집

3.2 데이터 전처리

결측치



3. 자료 수집

3.2 데이터 전처리

데이터 전처리

변수	분석 전처리	회귀 전처리
감독		평균관객수 상위 457명 외 : '기타'
상영시간	결측치 2개 행 삭제	
제작국가		빈도수 상위 24개 외 : '기타'
장르	결측치 1개 행 삭제	
배급사		빈도수 상위 30개 외 : '기타'
언어	제작국가로 대체 (복수의 경우 주요 언어 추출 불가)	
예산		제외 (결측치 과다)
주인공(배우)		평균관객수 상위 179명 외 : '기타'
시리즈		결측치 : '비시리즈' ('비시리즈' 과다)
키워드	제외 (줄거리에서 추출한 키워드로 대체)	
줄거리	텍스트 마이닝	제외
수상내역	제외 (개봉 전/후 수상내역 분류 불가)	
원작도서		결측치 : '없음' ('없음' 과다)
평점		제외 (관객수와의 상관계수 저조)
댓글	텍스트 마이닝	제외
댓글개수	결측치 23개 행 삭제	
기대지수	결측치 23개 행 삭제	

3. 자료 수집

3.3 변수 정의

개봉일	2010년 ~ 2019년 3월까지 개봉한 영화의 개봉일
감독	감독명
상영시간	상영 시간 (단위 : 분)
제작국가	제작국가명
장르	장르명
관람등급	관람등급
배급사	배급사명
예산	제작비용 (단위 : 달러)
배우	배우명
시리즈	시리즈 영화 여부
줄거리	줄거리 내용
평점	개봉 전 평점 (1~10점)
댓글개수	개봉 전 댓글 총 개수
댓글	개봉 전 댓글 내용
원작도서	원작도서 유무
기대지수	개봉 전 '보고싶어요' 버튼 클릭 총 개수
누적관객수	개봉 기간 내 총 관객수

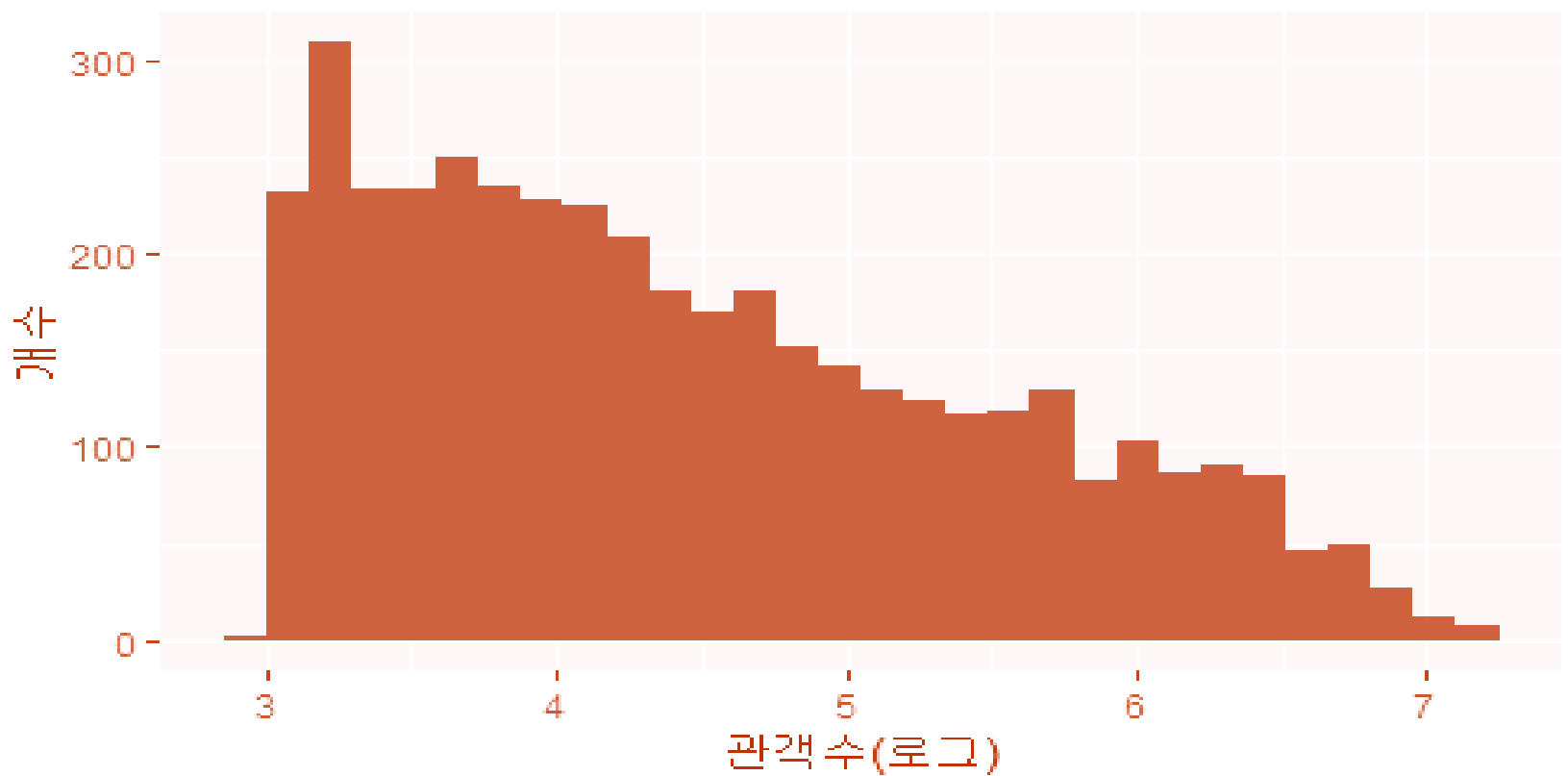


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

관객수

log(관객수) 빈도분포

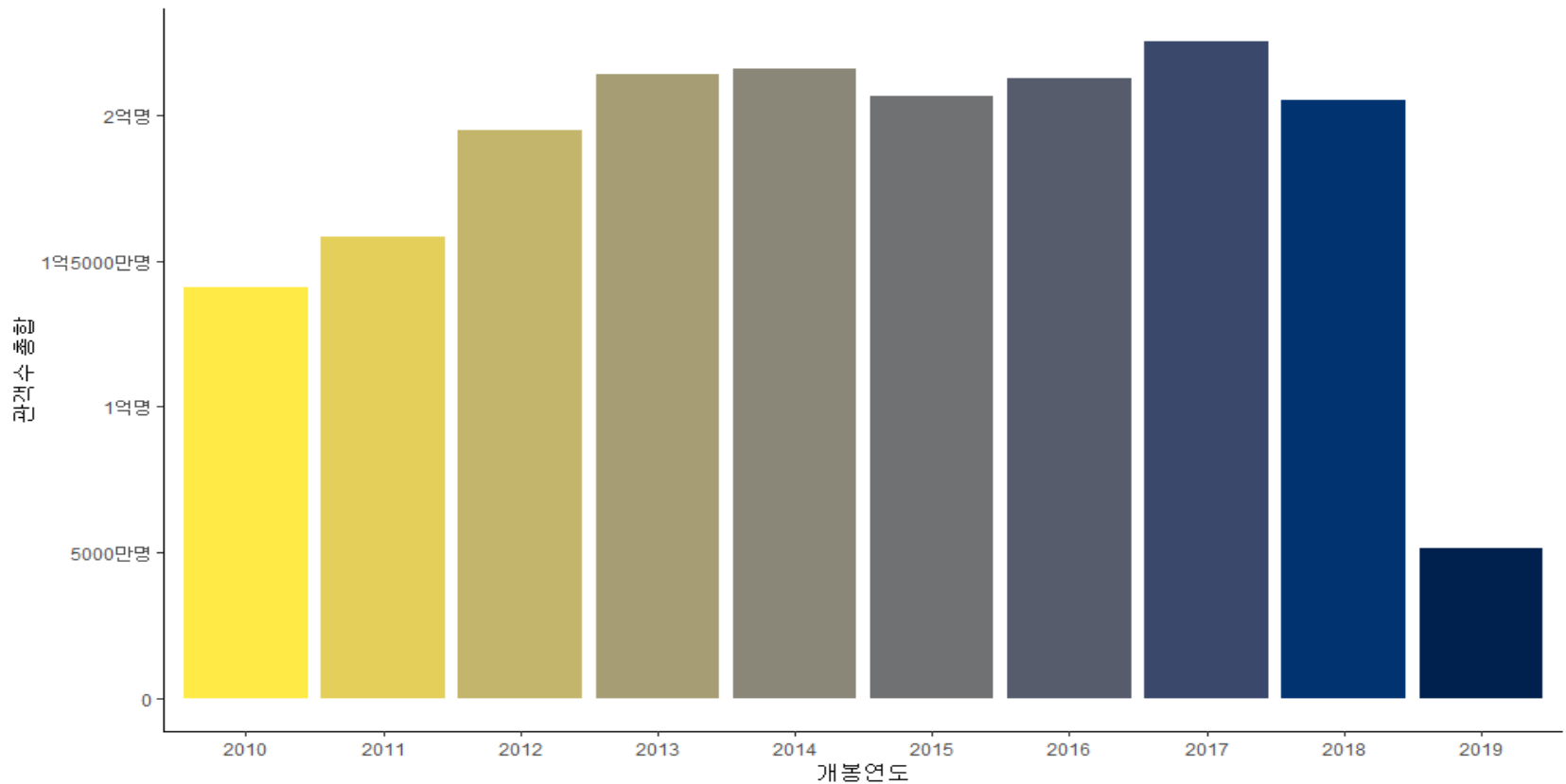


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉시기

개봉연도별 관객수 총합

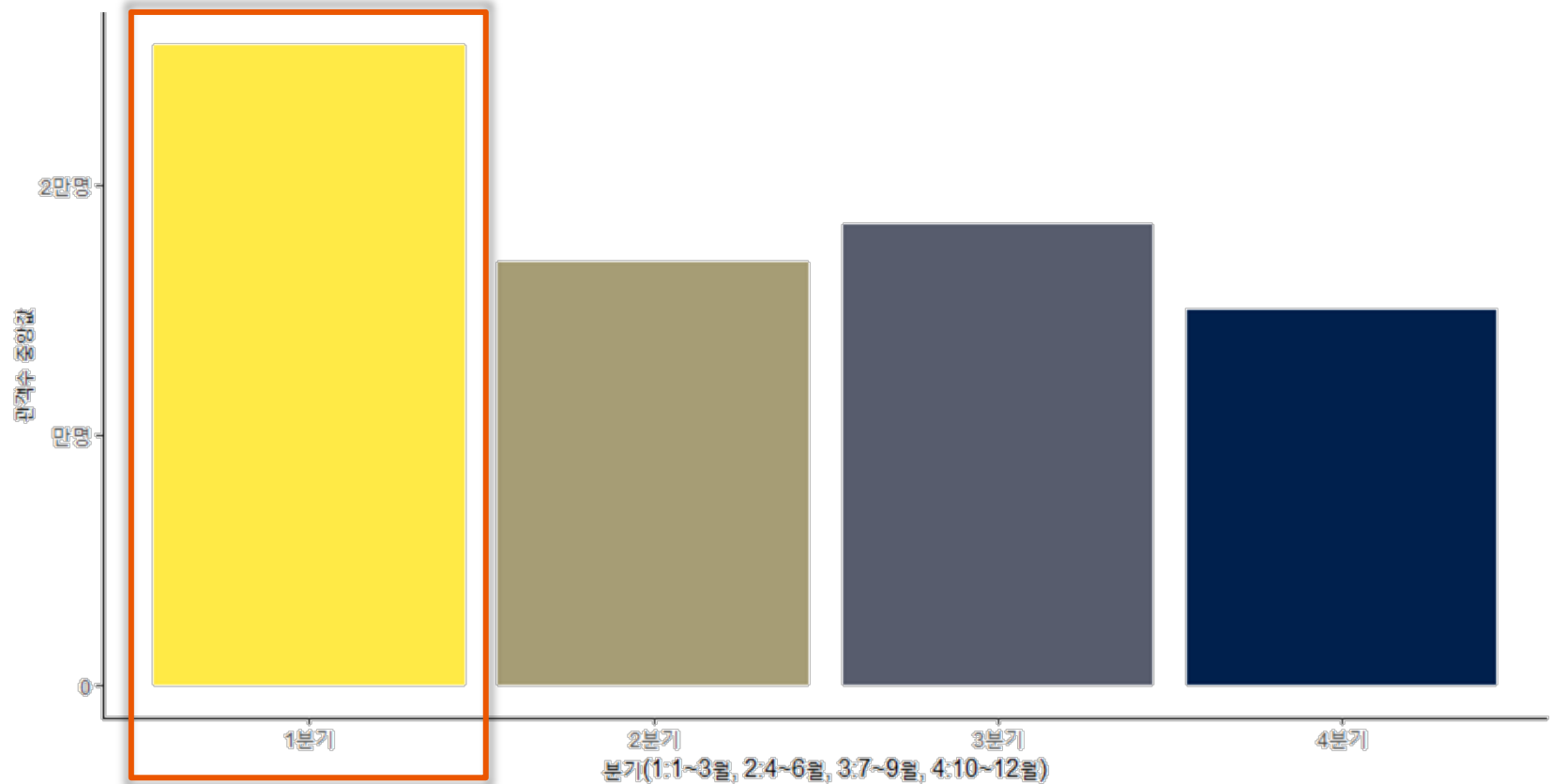


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉시기

개봉분기별 고객수 중앙값

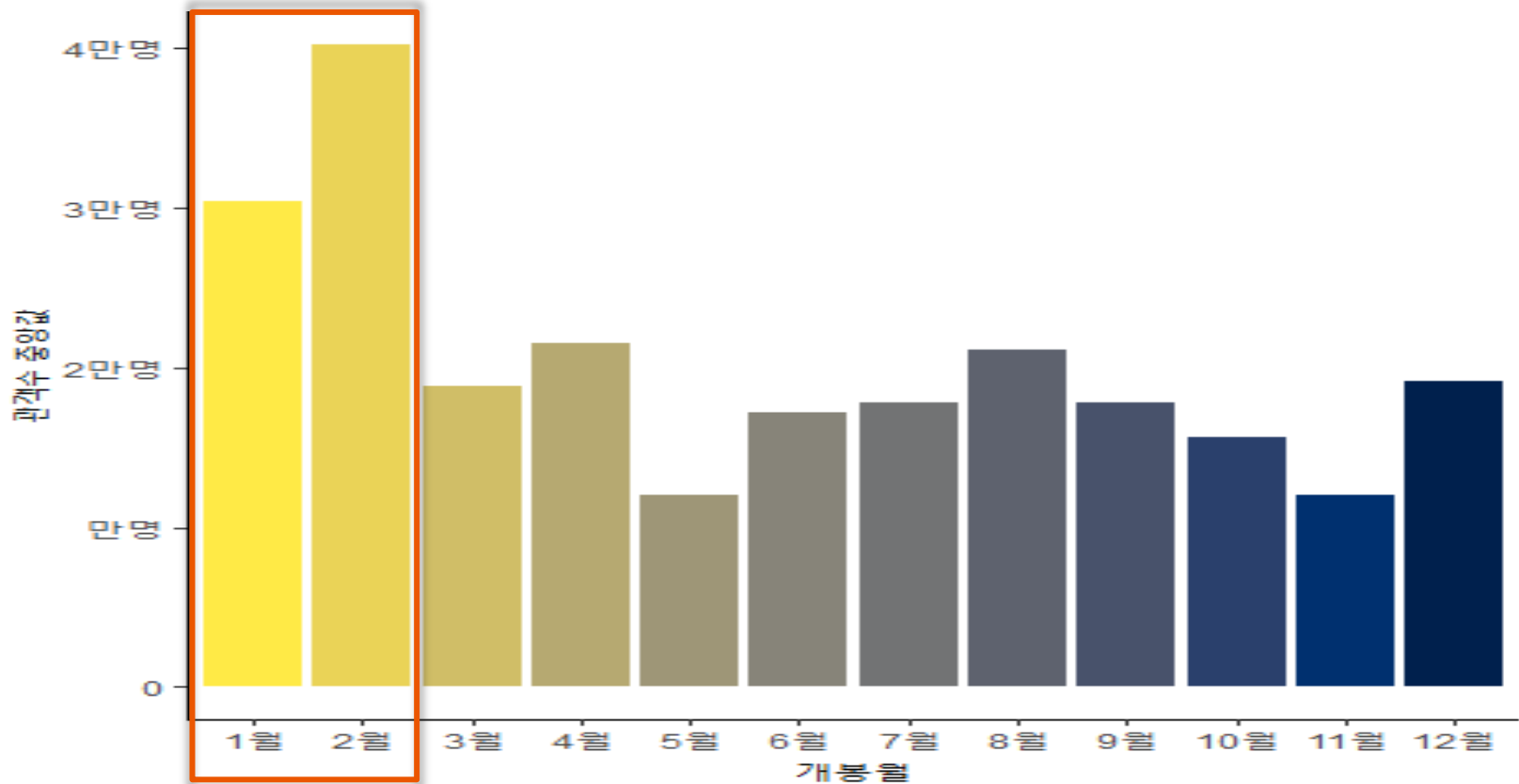


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉시기

개봉월별 관객수 중앙값

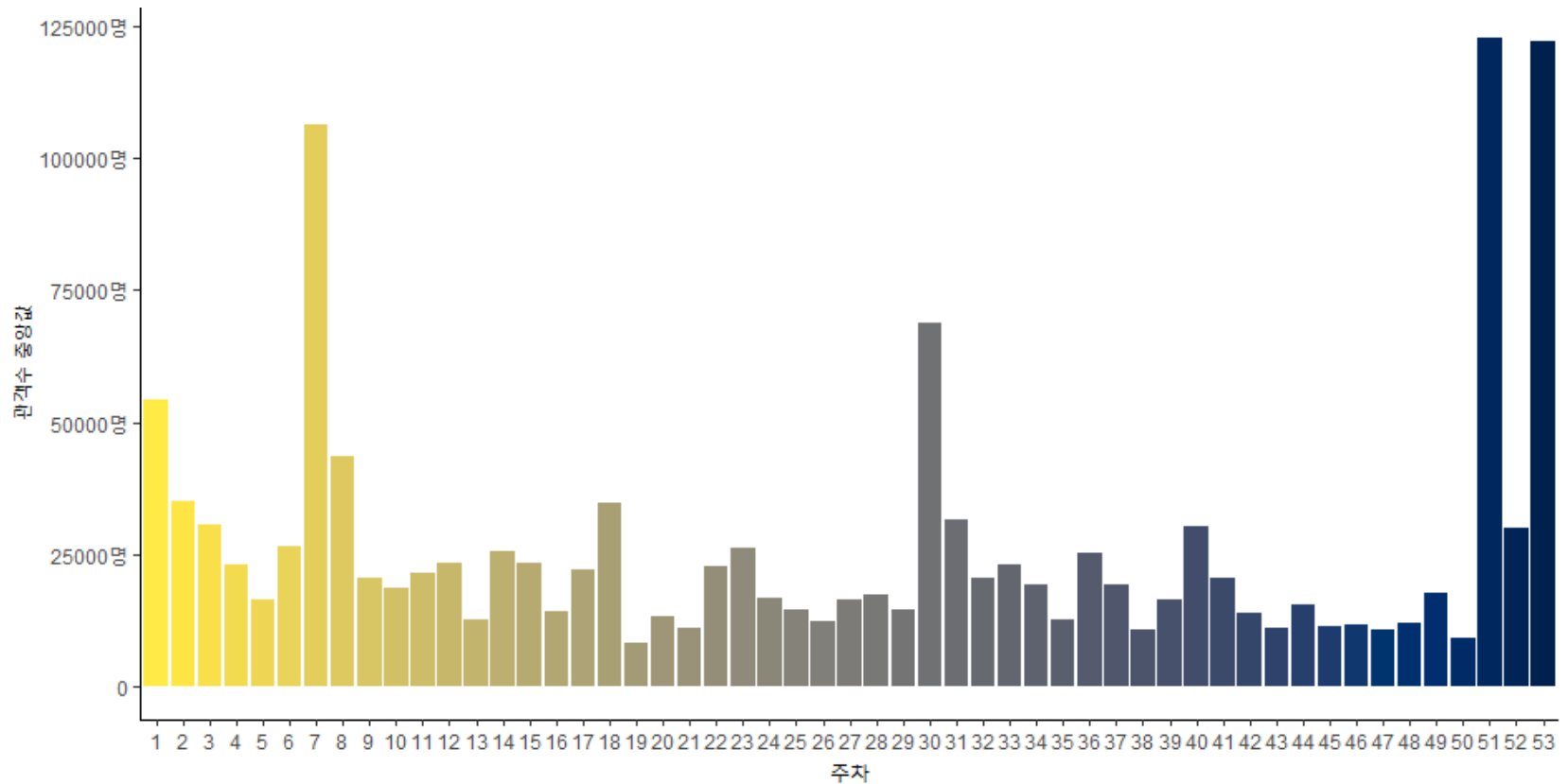


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉시기

개봉주차별 관객수 중앙값

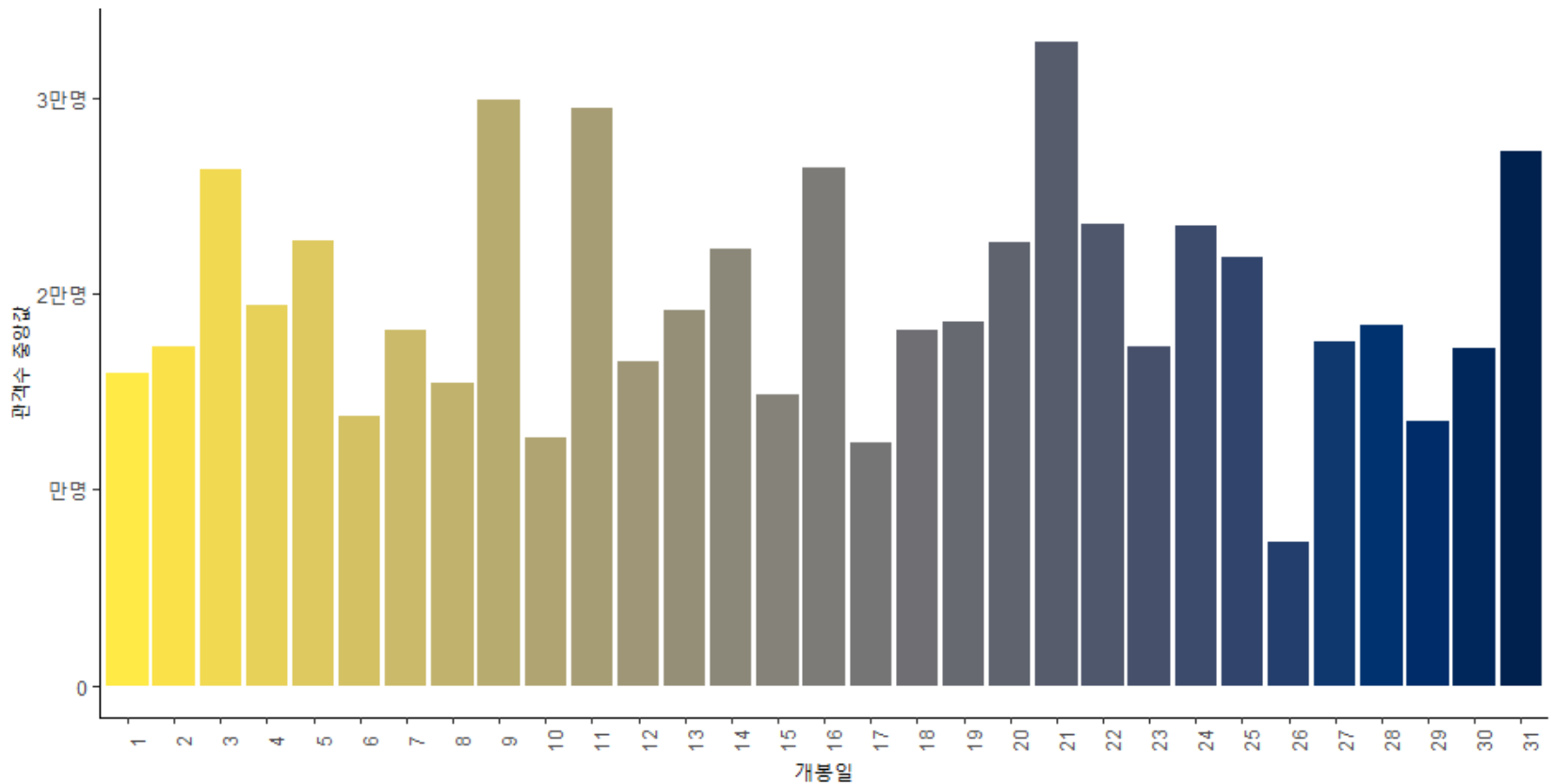


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉시기

개봉일별 관객수 중앙값

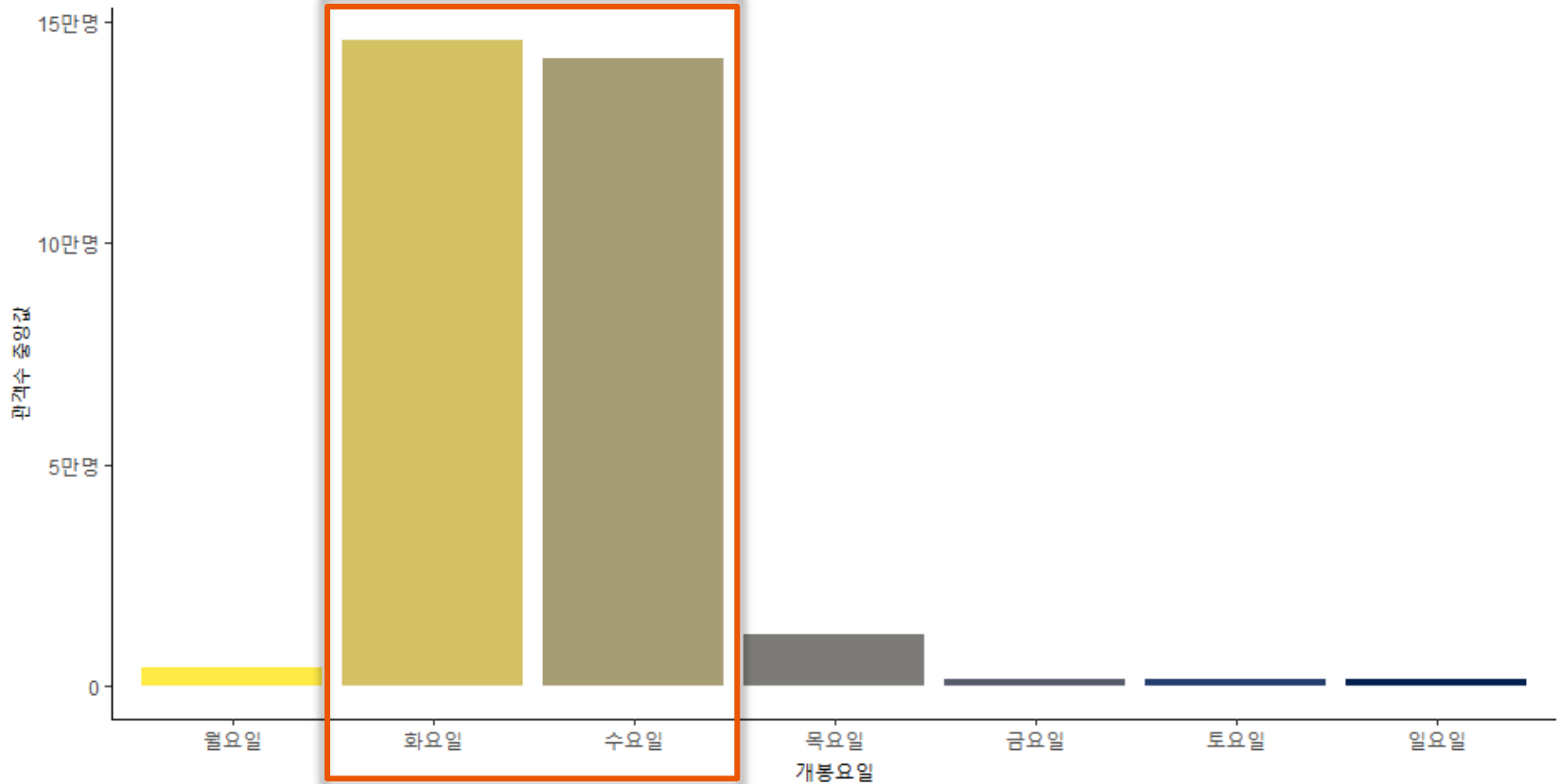


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉시기

개봉요일별 관객수 중앙값

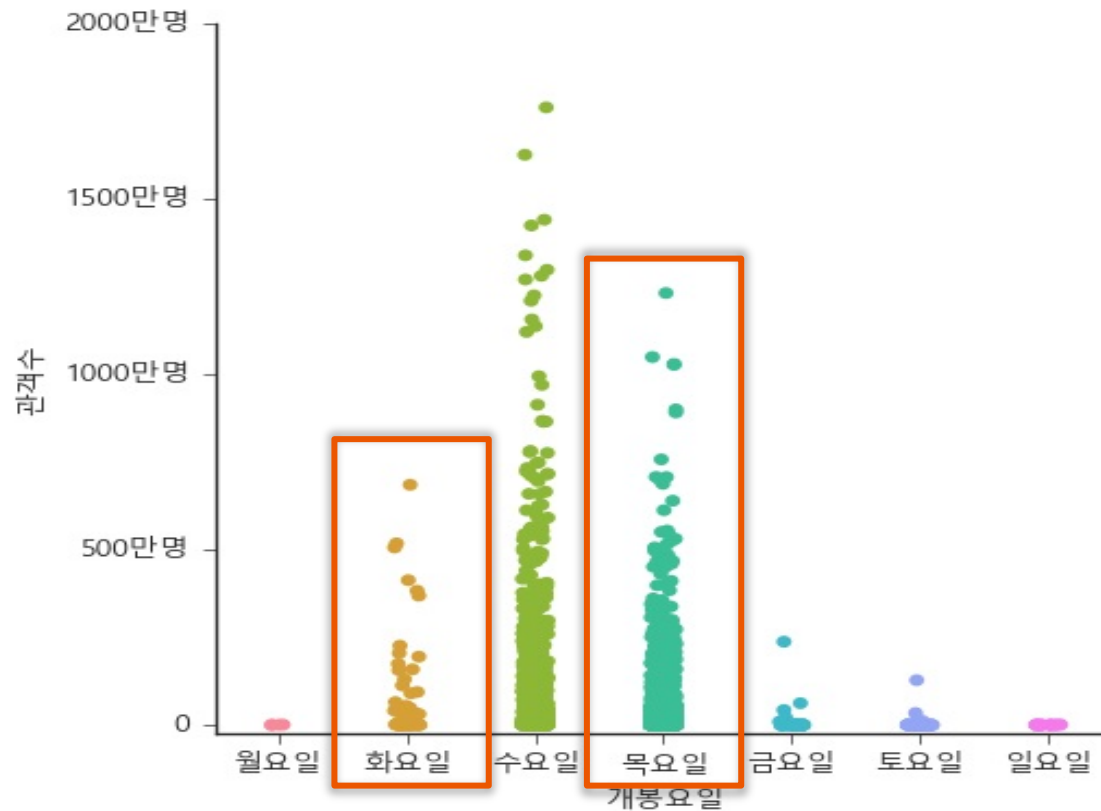


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉시기

개봉요일별 관객수 분포도

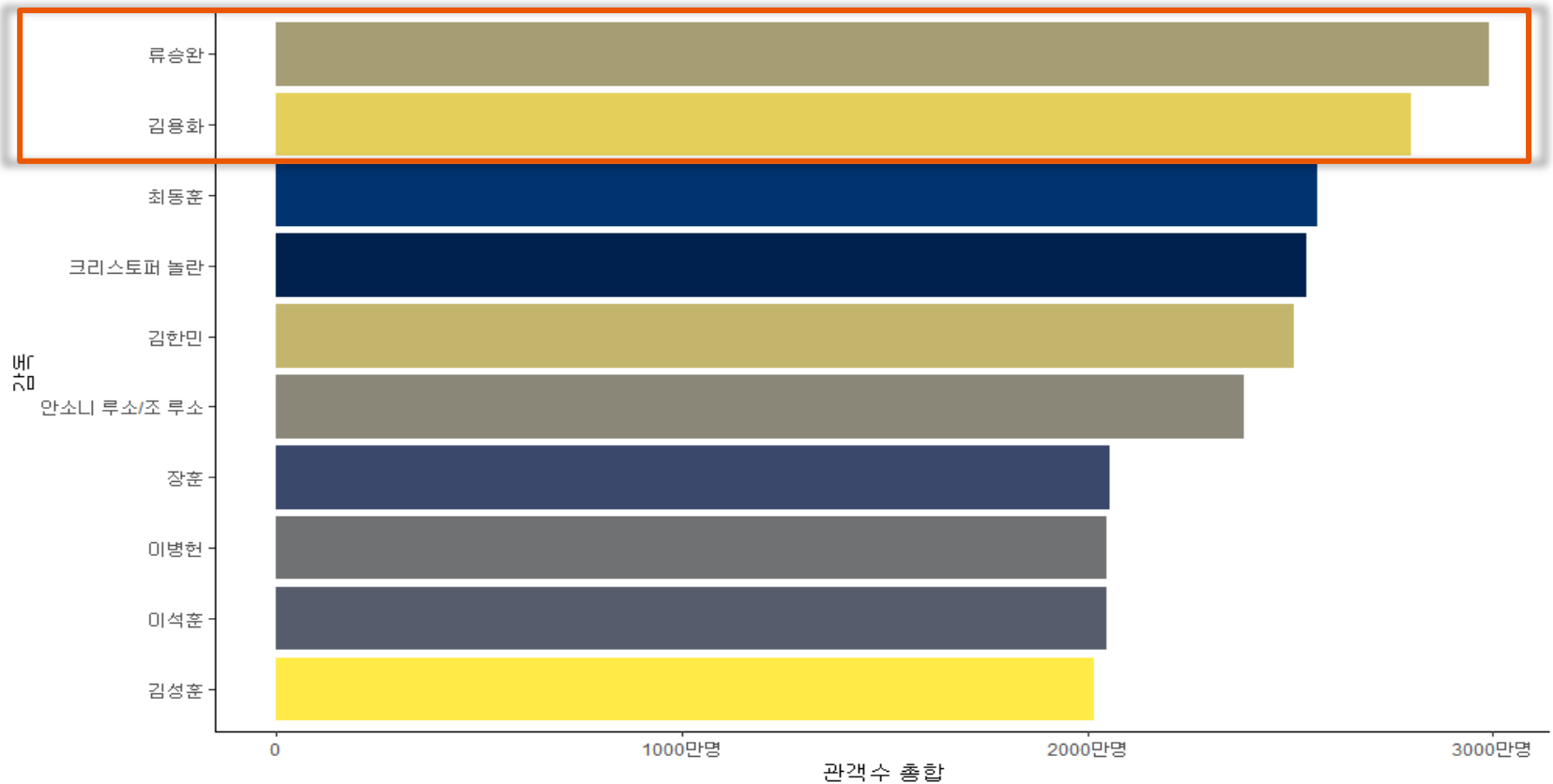


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

감독별 관객수 종합 상위 10위

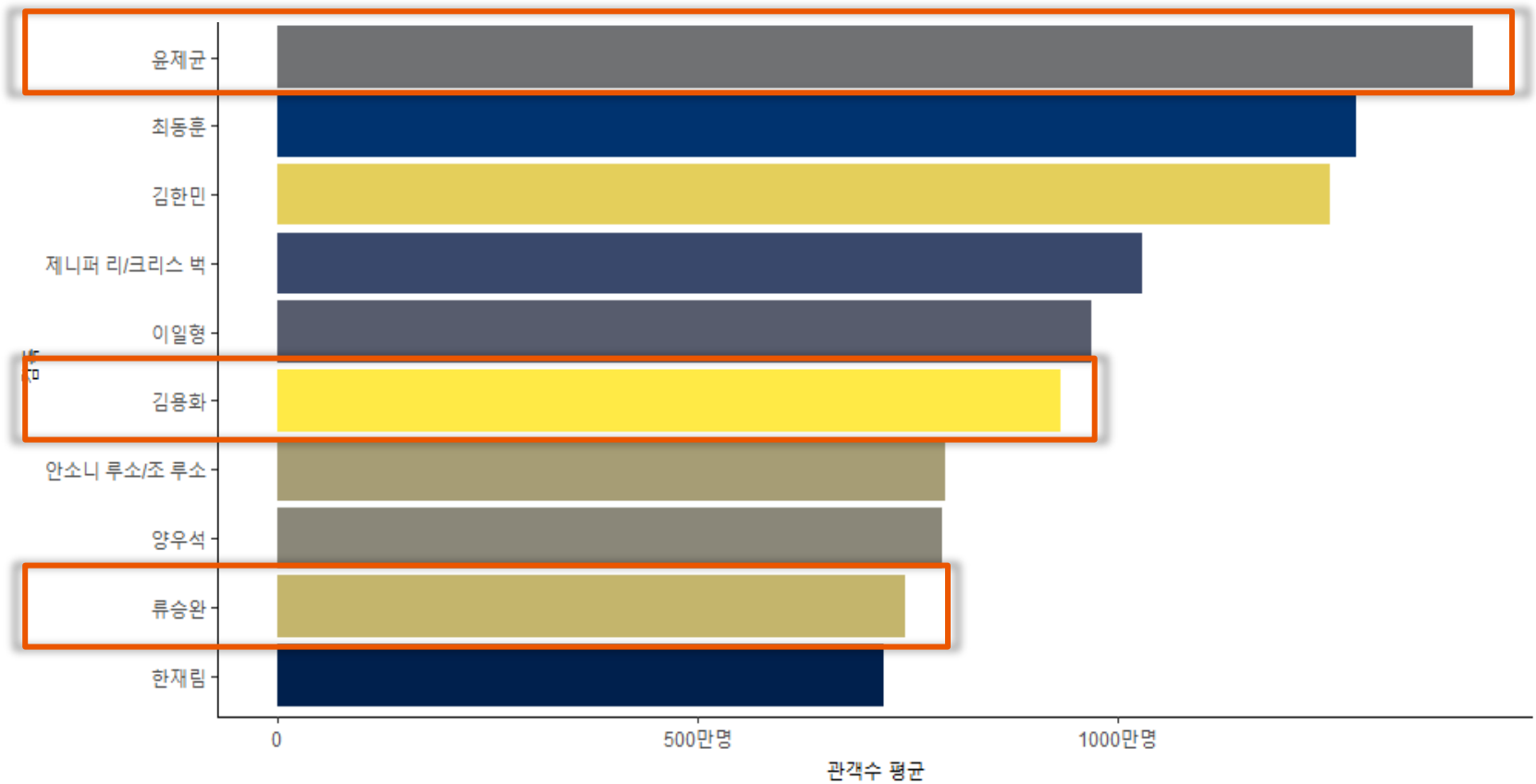


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

감독별 관객수 평균값 상위 10위

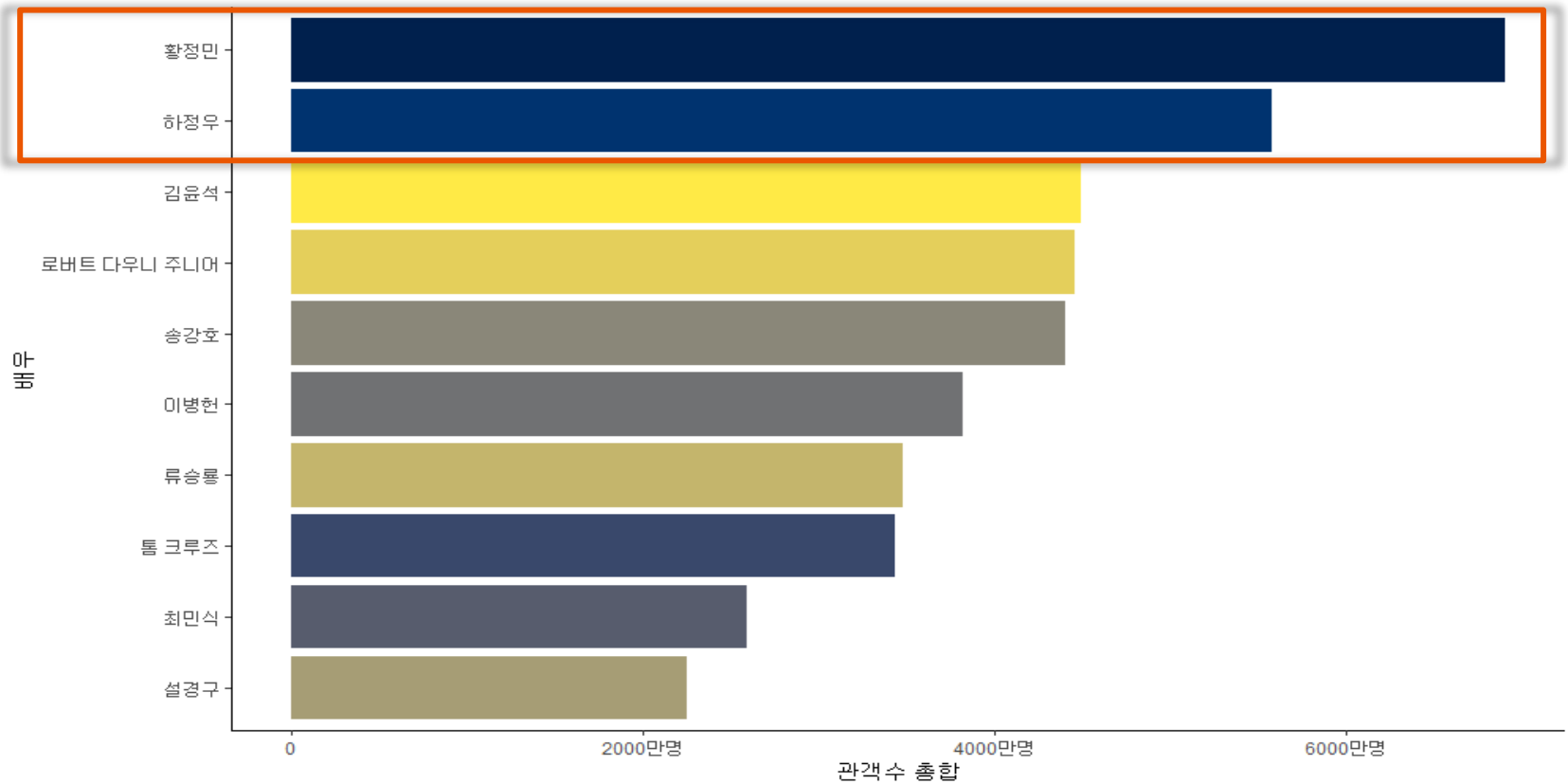


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

주인공 배우별 관객수 종합 상위 10위

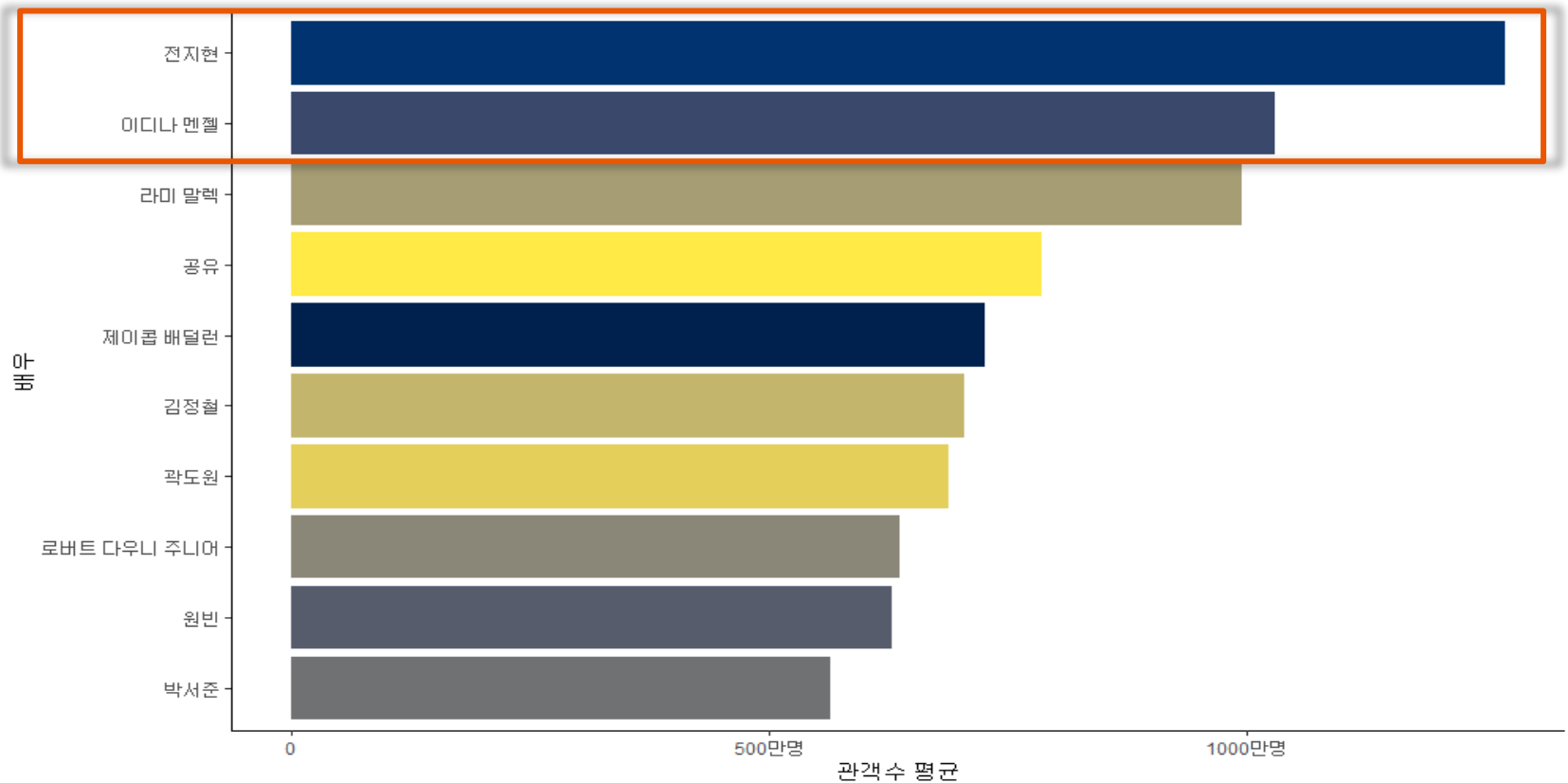


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

주인공 배우별 관객수 평균값 상위 10위

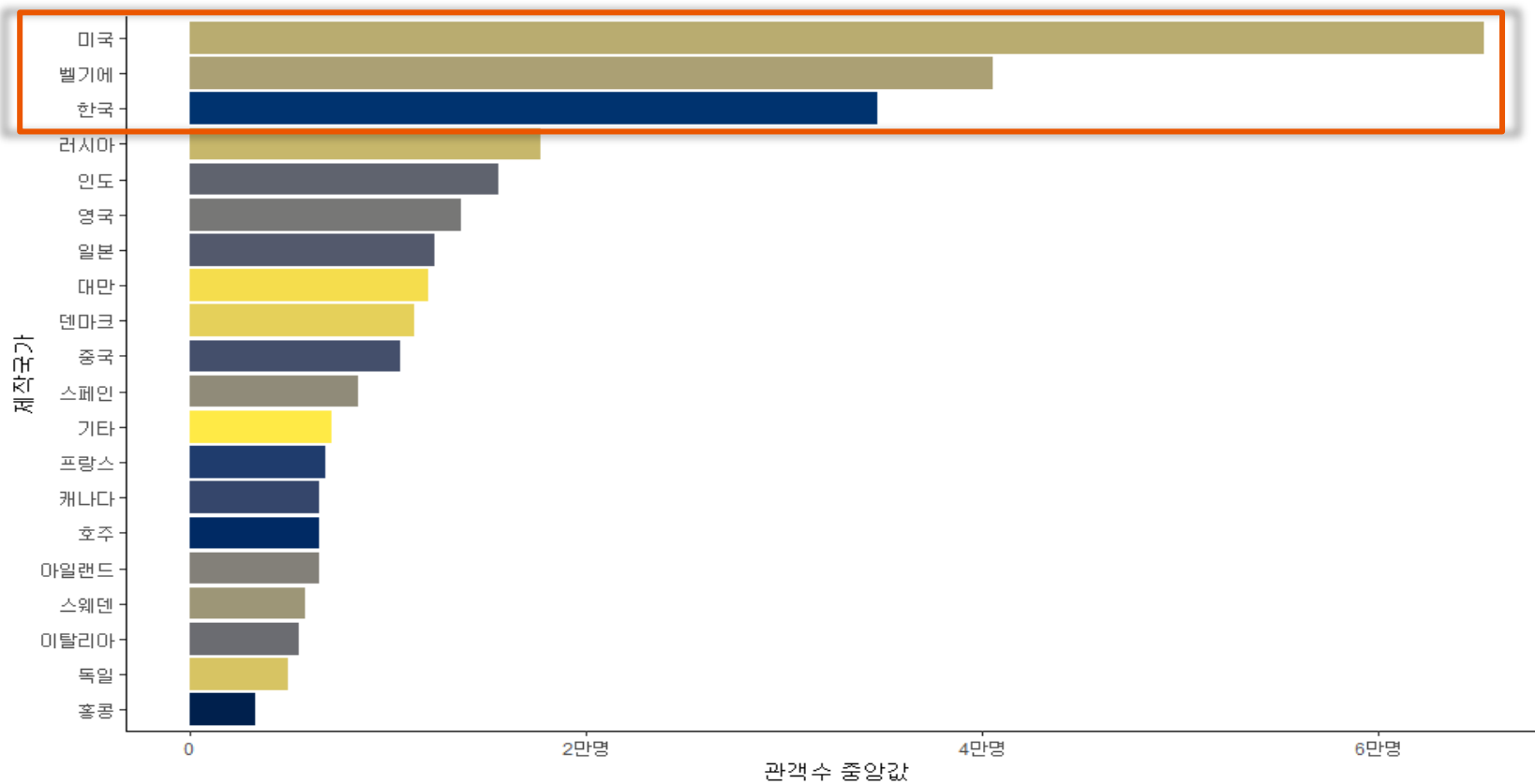


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

제작국가별 관객수 중앙값

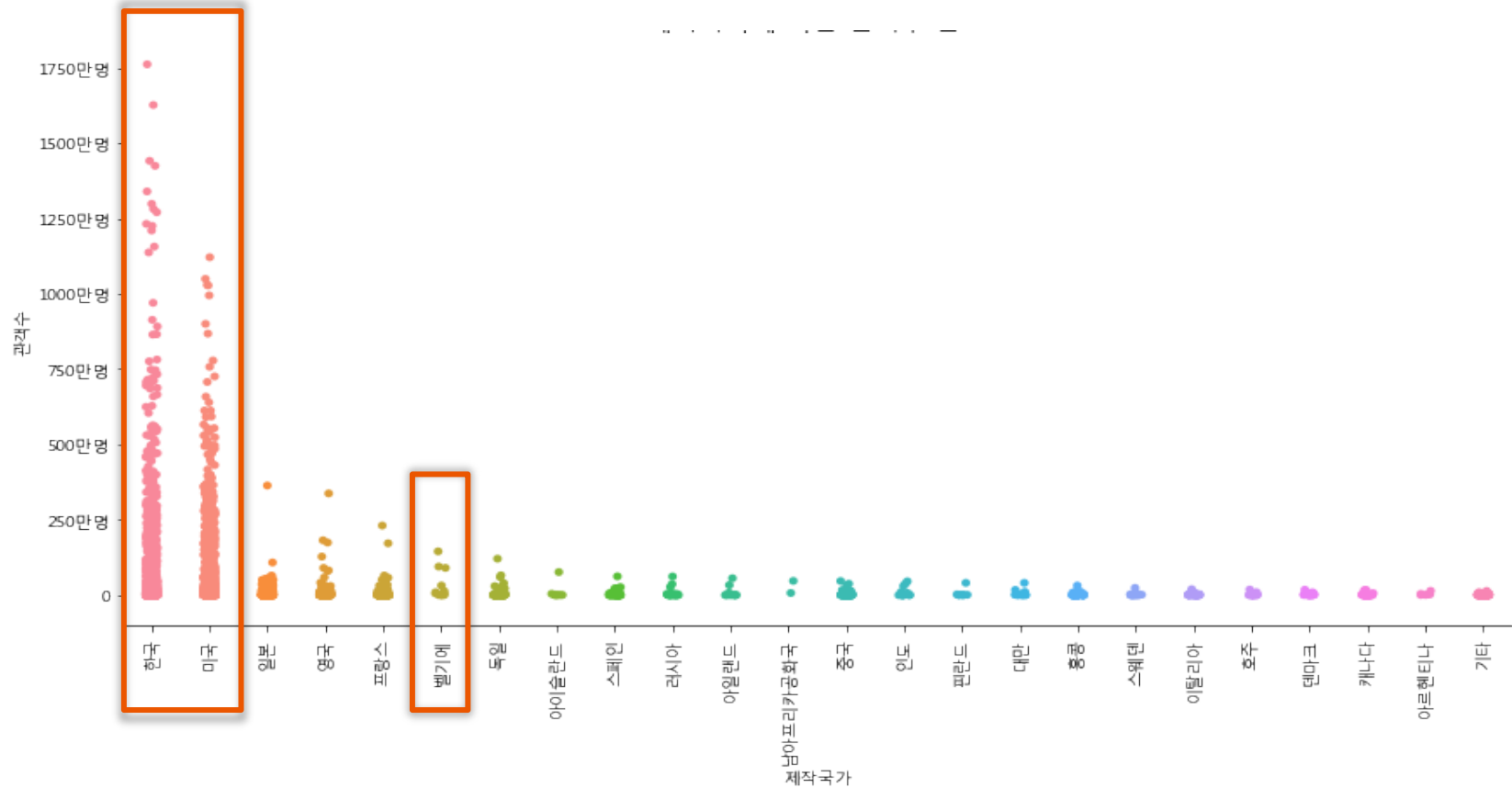


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

제작국가별 관객수 분포도

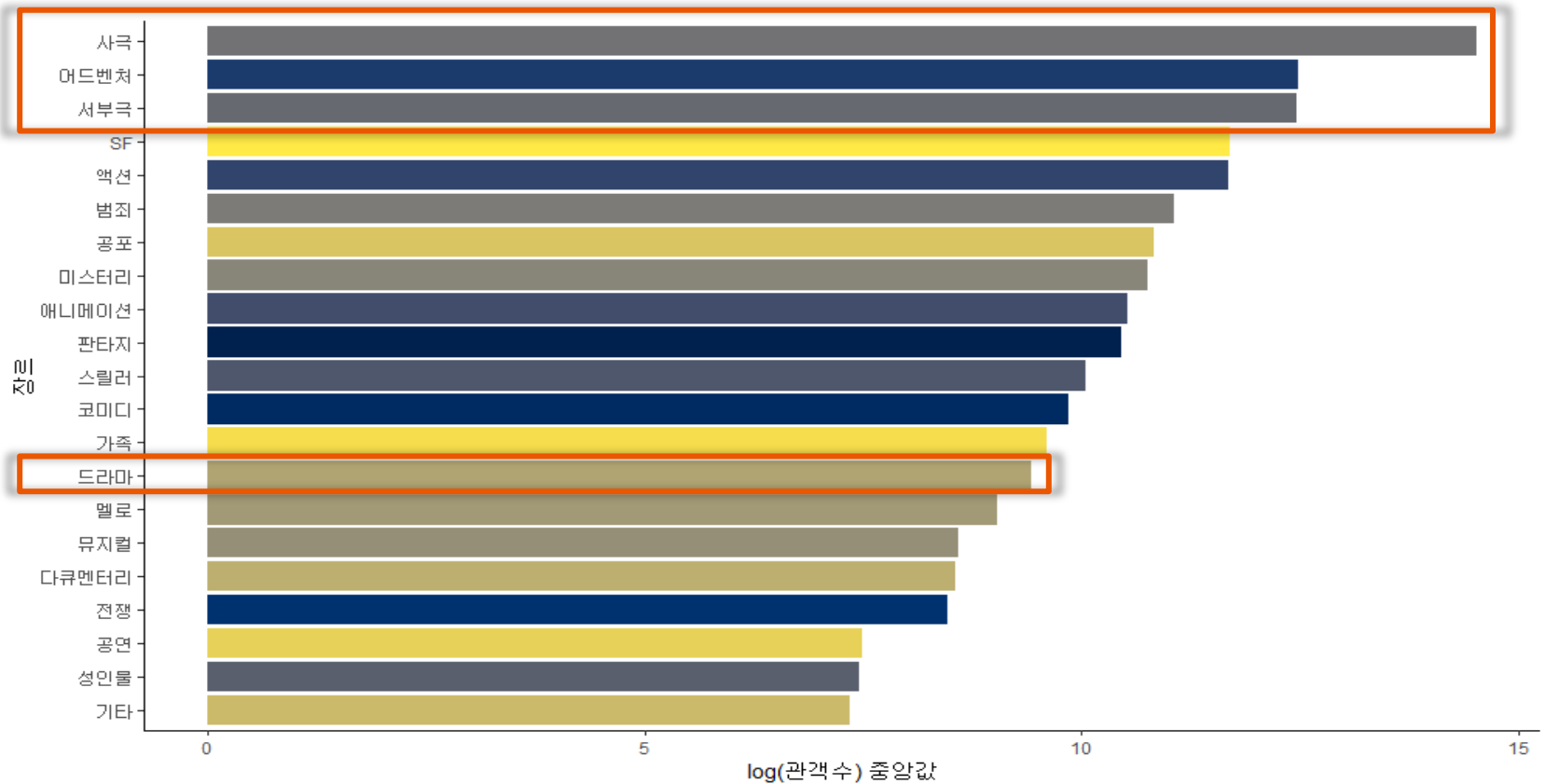


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

장르별 log(관객수) 중앙값

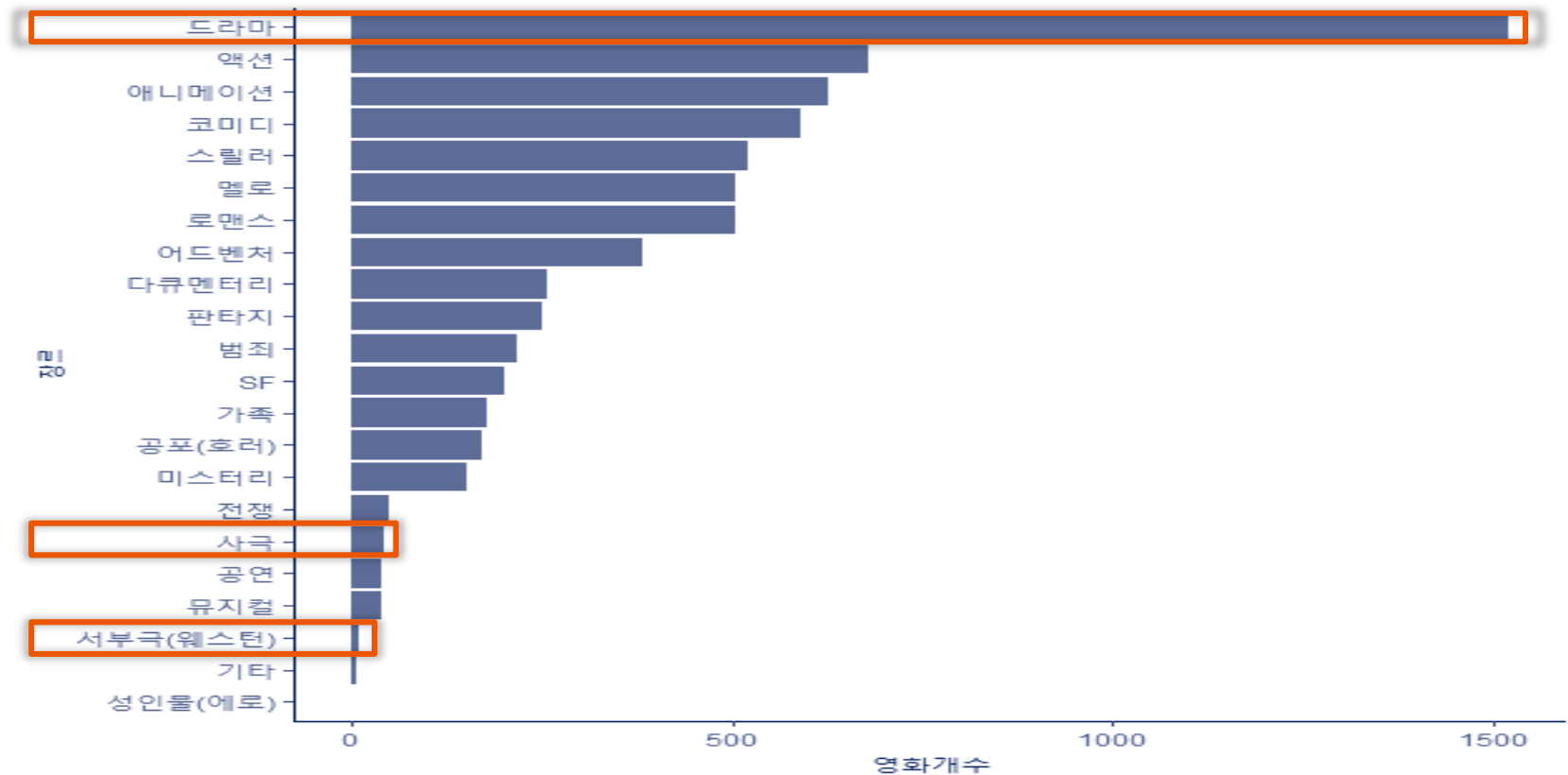


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

장르별 영화 개수

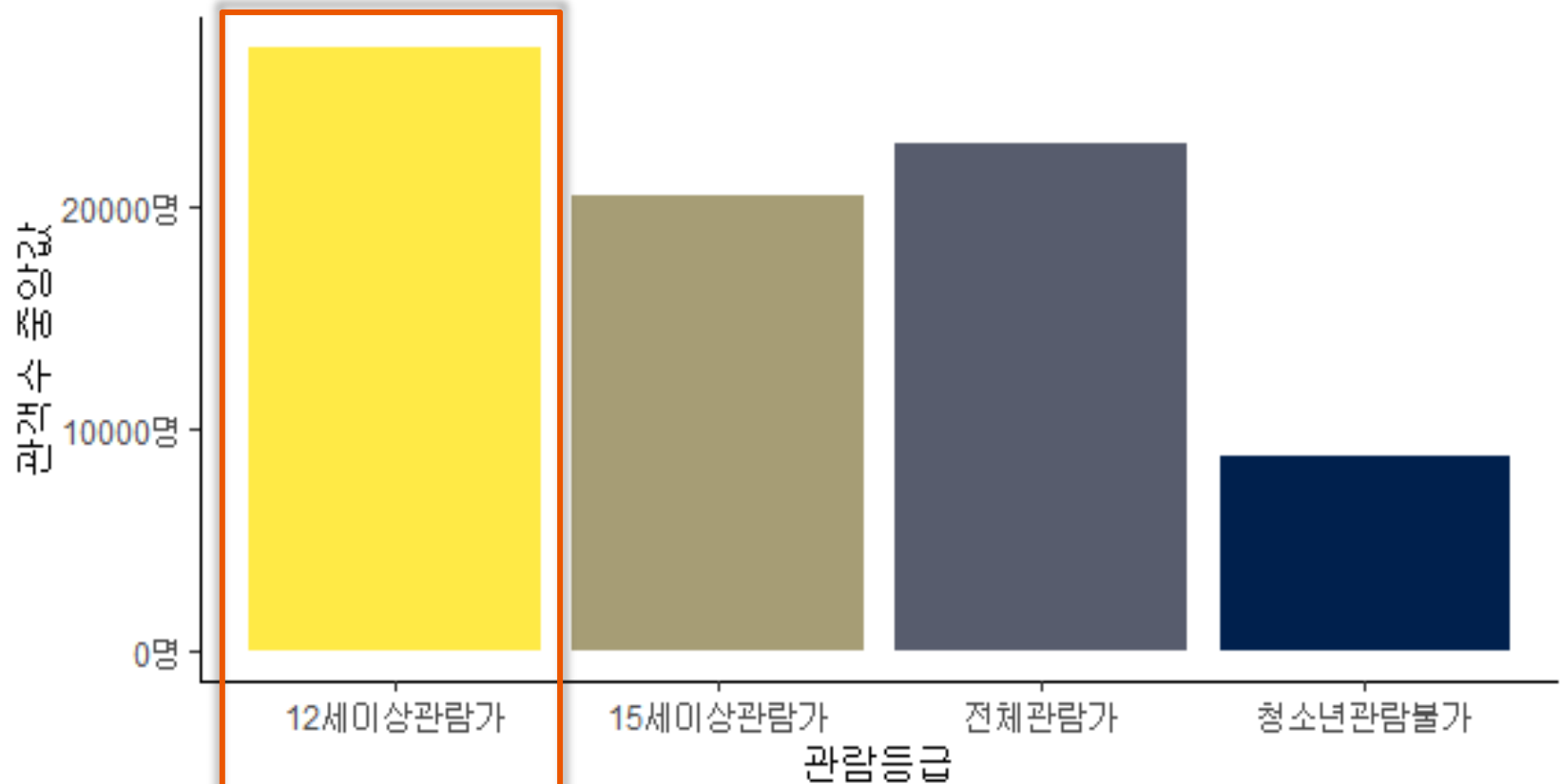


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

관람등급별 관객수 중앙값

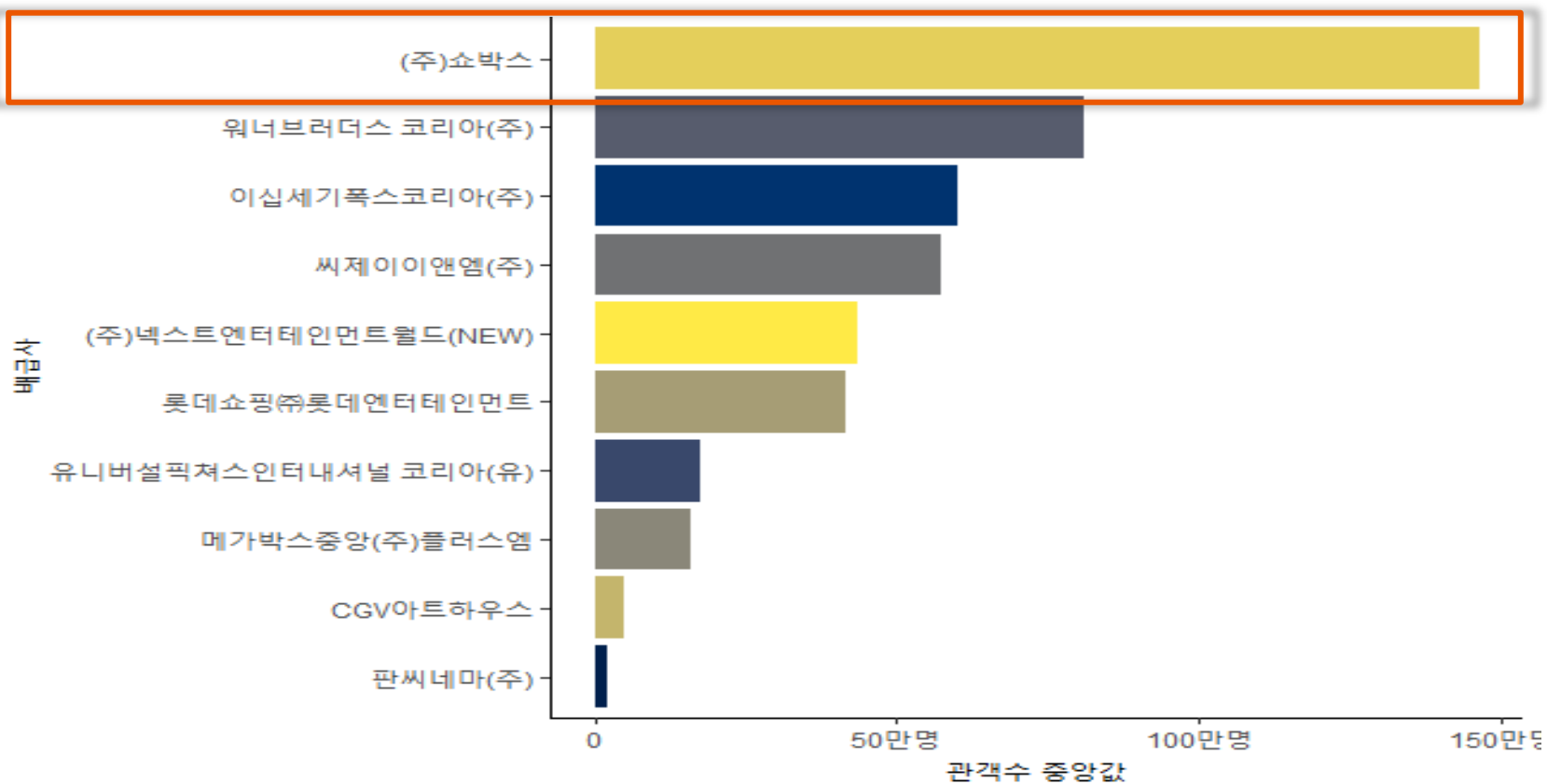


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

빈도수 상위 10위 배급사의 관객수 중앙값

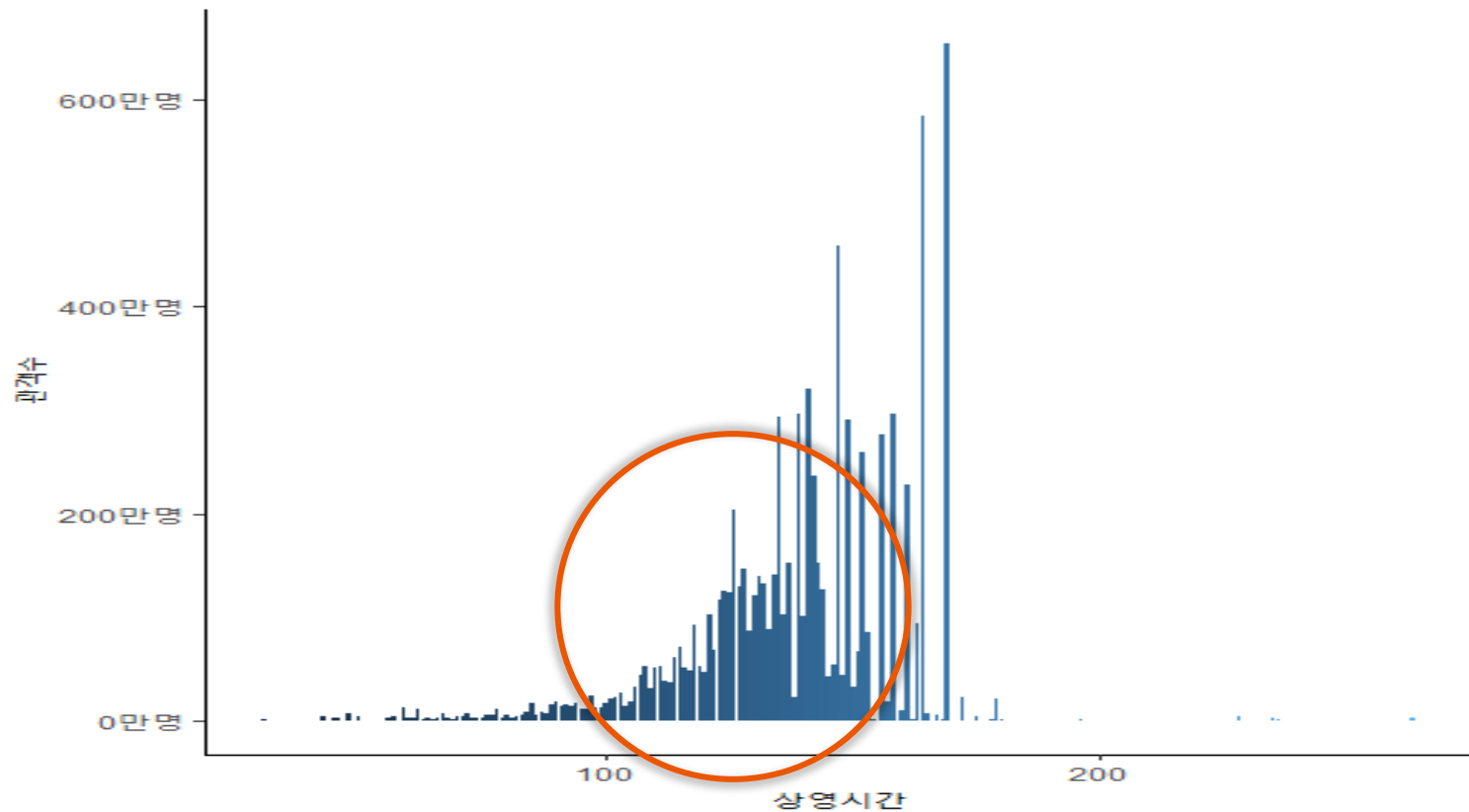


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

상영시간별 관객수

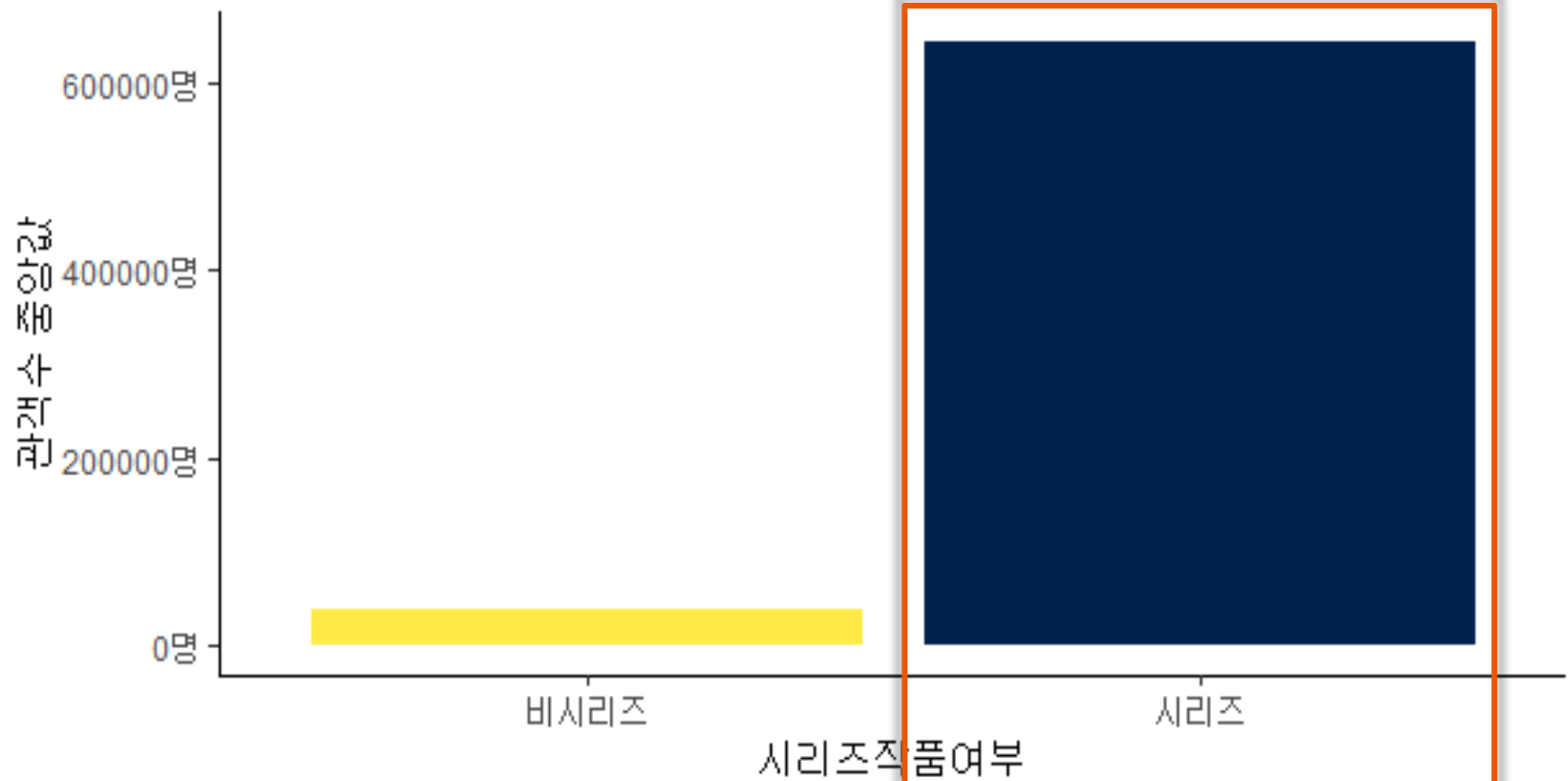


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

시리즈작품 유/무별 관객수 중앙값

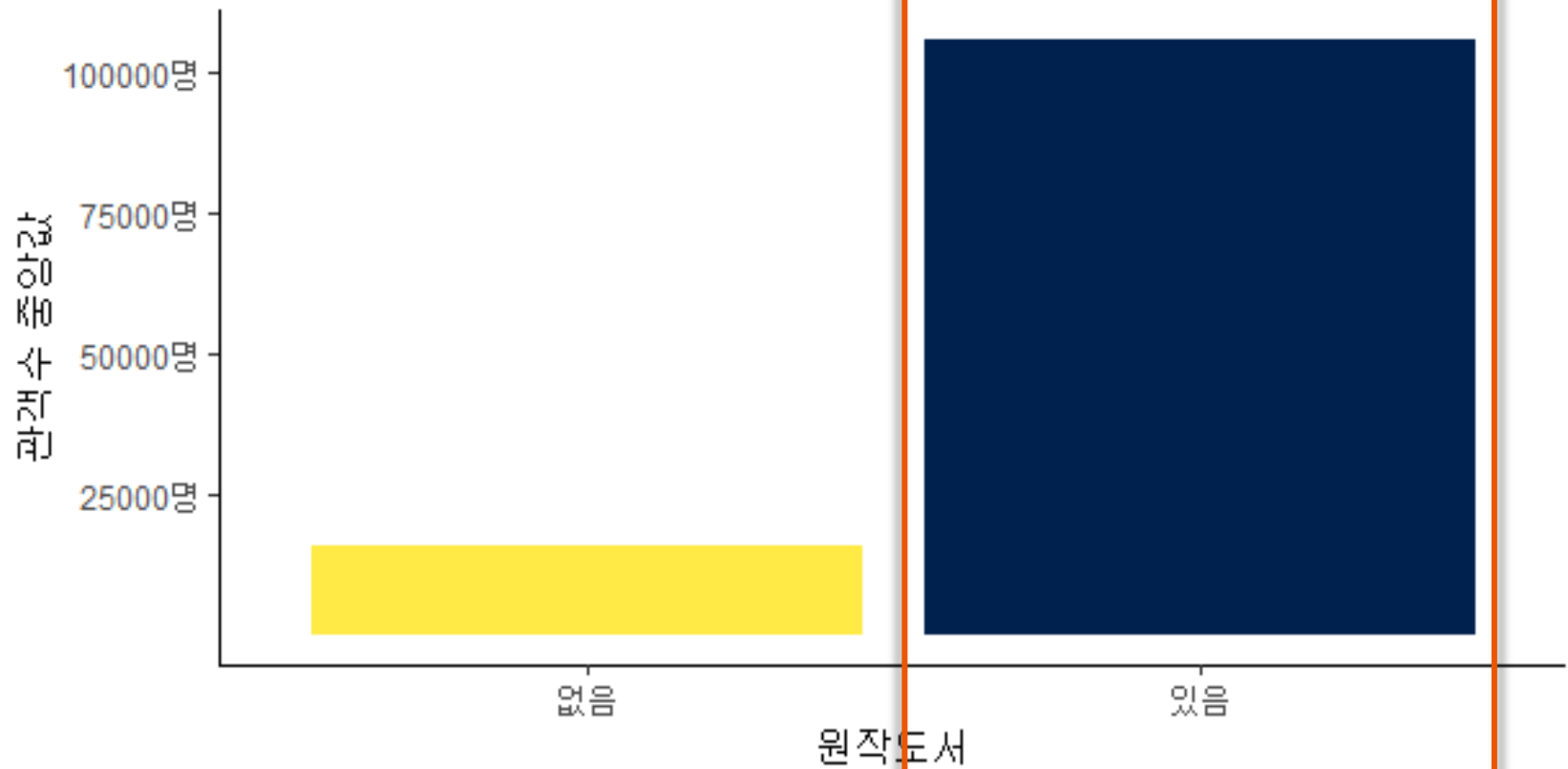


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

원작도서 유/무별 관객수 중앙값

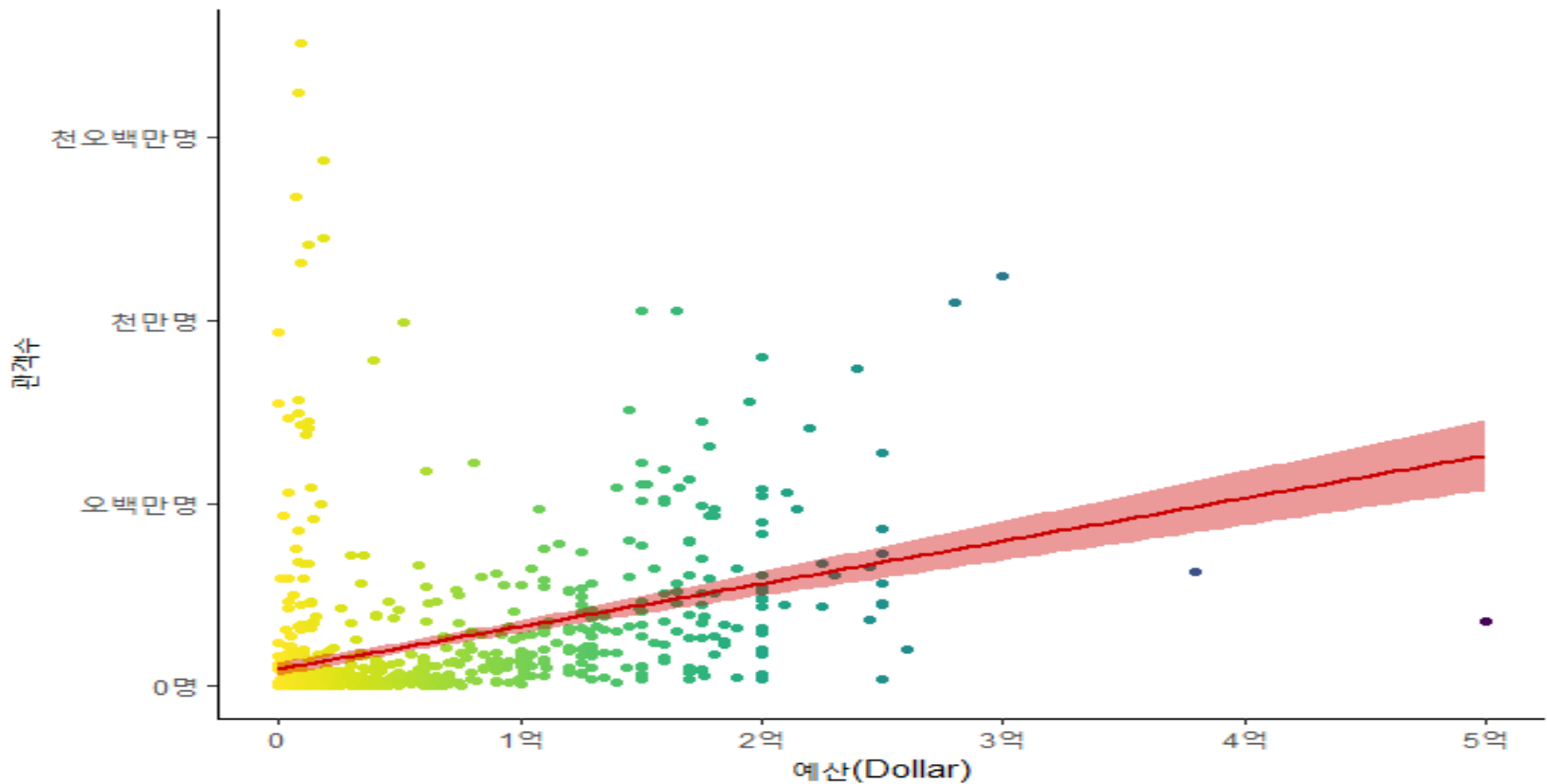


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

영화특성

예산에 따른 관객수

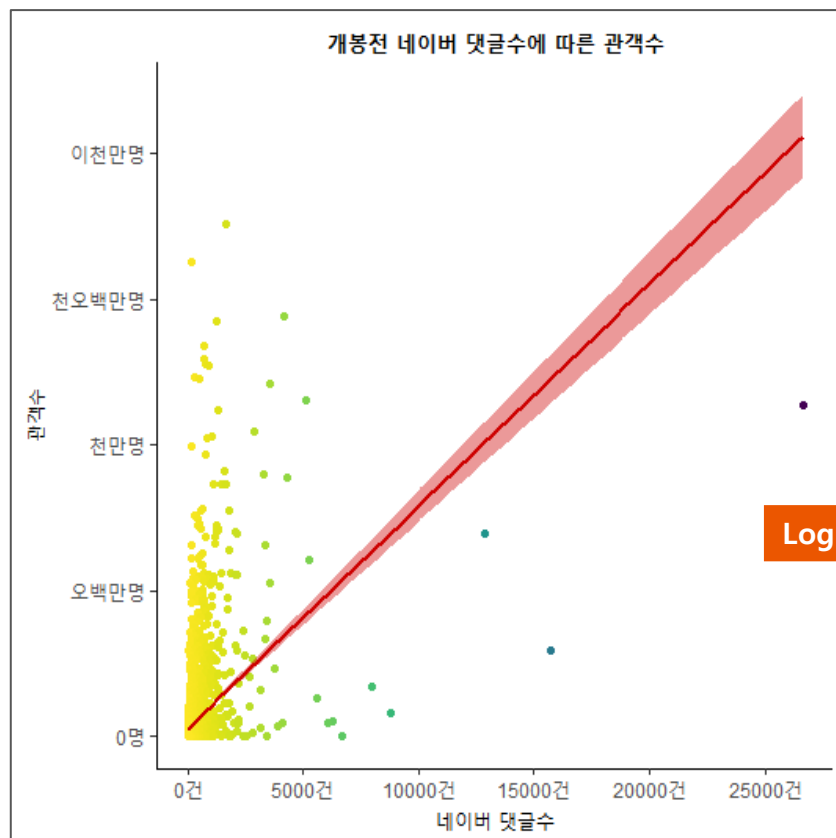


4. 연구 결과

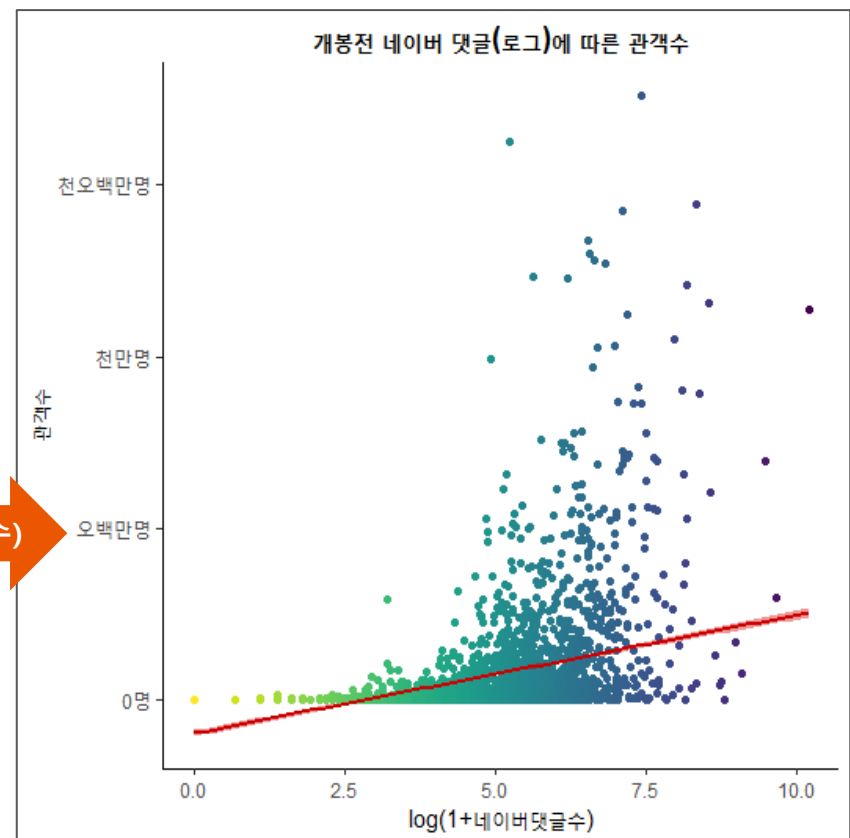
4.1 분석 결과 - 변수별 시각화

개봉 전 반응

개봉 전 댓글개수에 따른 관객수



Log(댓글 개수)

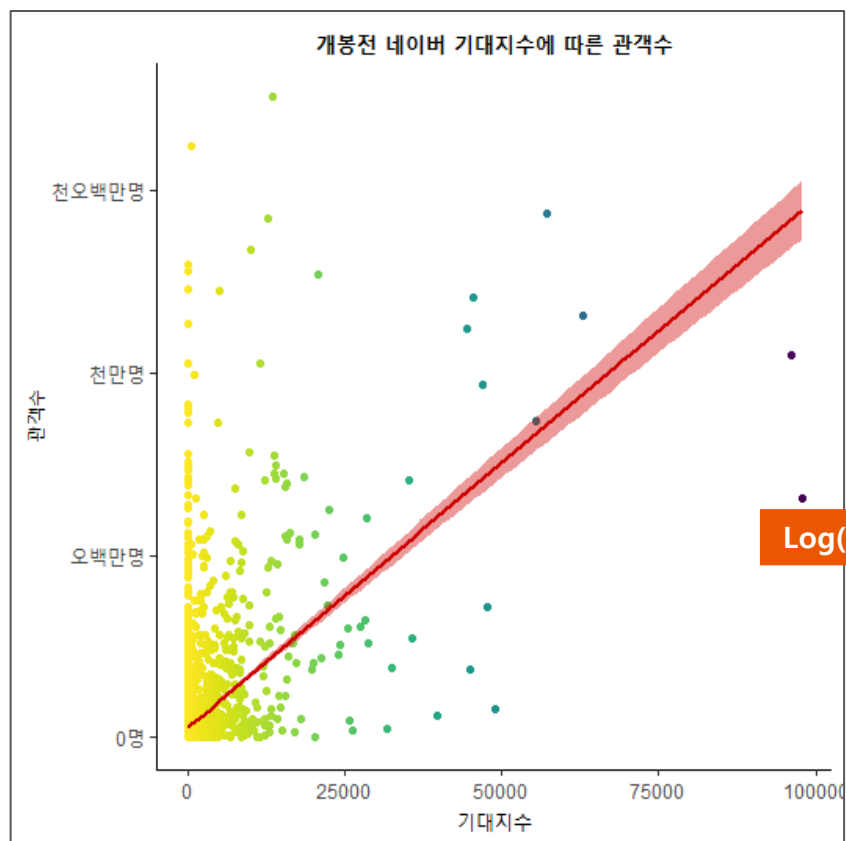


4. 연구 결과

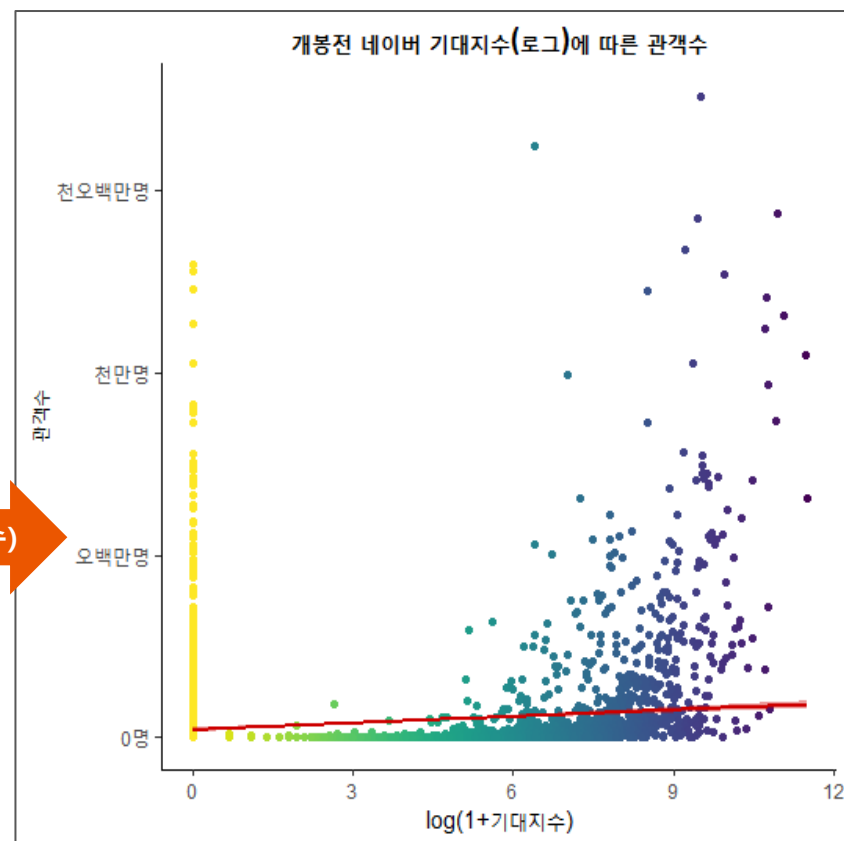
4.1 분석 결과 - 변수별 시각화

개봉 전 반응

개봉 전 기대지수에 따른 관객수



Log(기대지수)

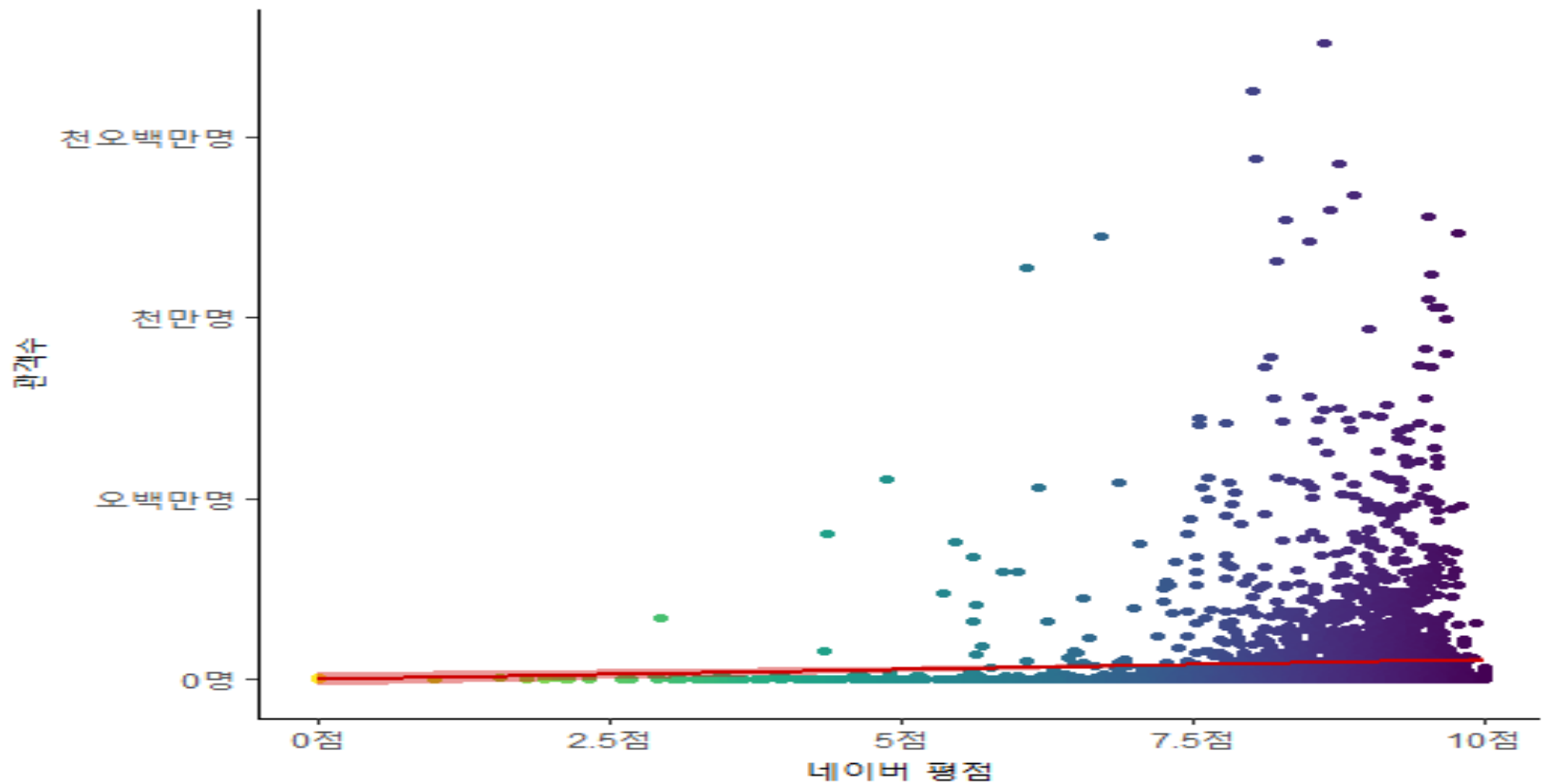


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉 전 반응

개봉 전 평점에 따른 관객수

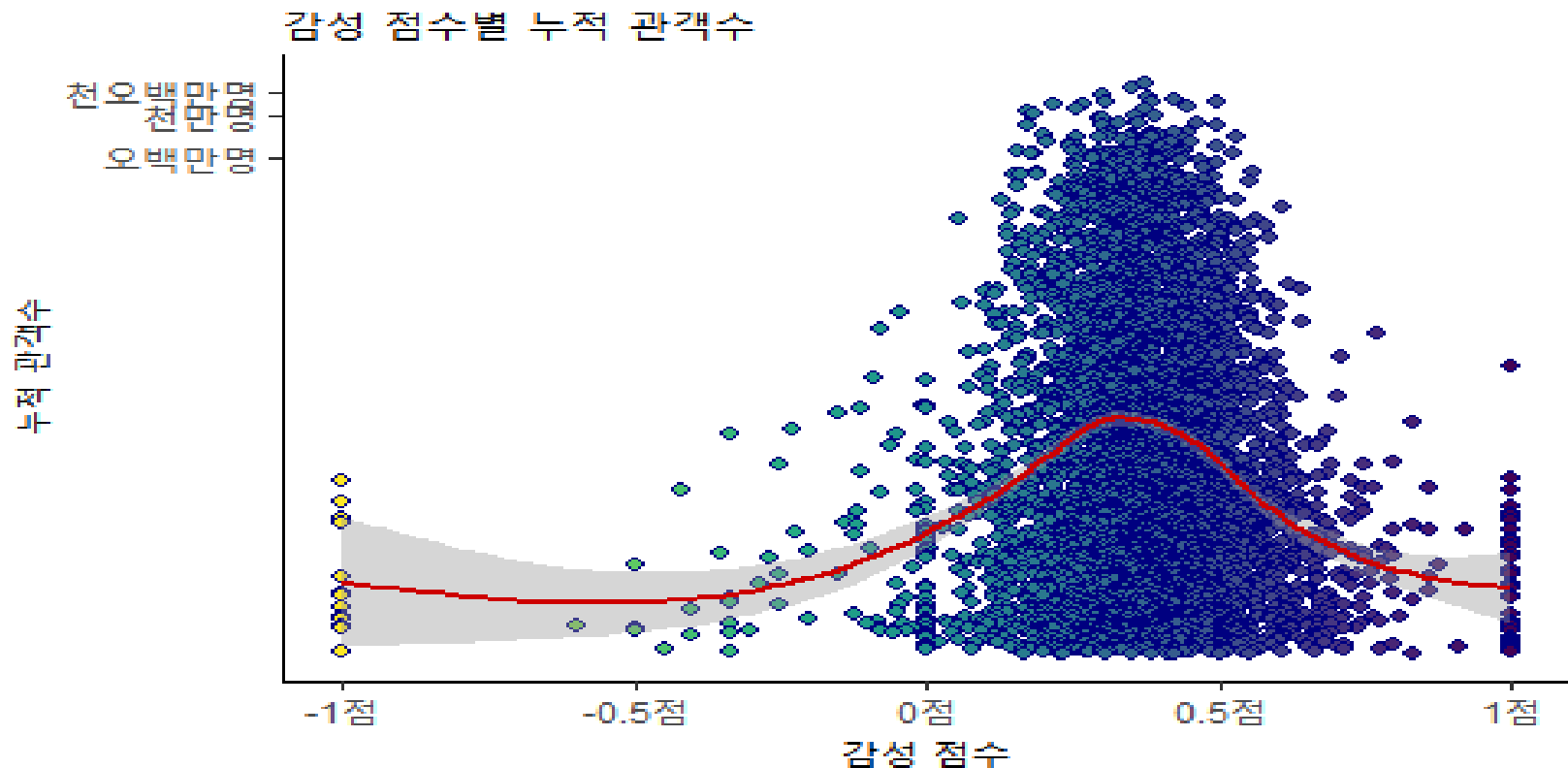


4. 연구 결과

4.1 분석 결과 - 변수별 시각화

개봉 전 반응

댓글 감성점수에 따른 $\log(\text{관객수})$



4. 연구 결과

4.1 분석 결과 - 분산 분석(Analysis of Variance, ANOVA)

* 귀무가설 : 독립변수에 따른 관객수에 차이가 없다.

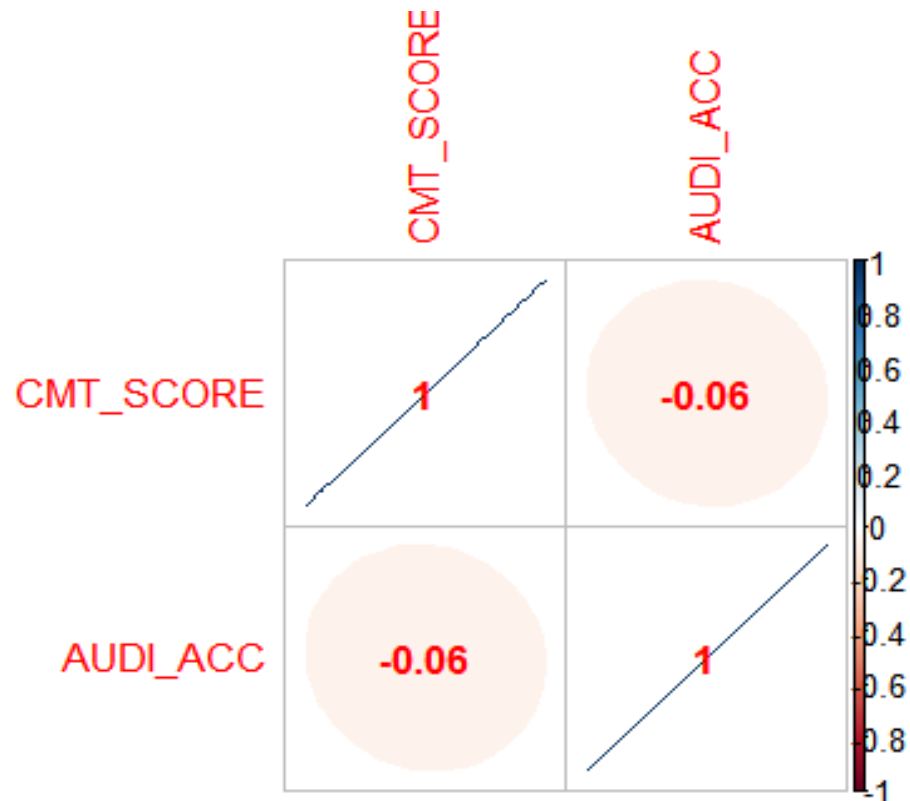
독립변수명	p-value	귀무가설 채택/기각
관람등급	< 2.2e-16	기각
감독	0.0292	
개봉일	< 2.2e-16	
개봉월	1.11E-05	
개봉요일	< 2.2e-16	
개봉분기	0.00302	
개봉주차	<2e-16	
제작국가	3.26E-05	
배급사	<2e-16	
장르	<2e-16	
주인공(배우)	0.264	
시리즈	<2e-16	
원작도서	1.38E-13	
개봉연도	0.888	채택
개봉일	0.974	

4. 연구 결과

4.1 분석 결과 - 상관 분석(Correlation Analysis)

개봉 전 반응

댓글 감성점수와 관객수의 상관관계



4. 연구 결과

4.1 분석 결과 - 상관 분석(Correlation Analysis)



4. 연구 결과

4.1 분석 결과 - 텍스트 마이닝(Text Mining)

영화특성

줄거리 빈도수 상위 100위 워드클라우드

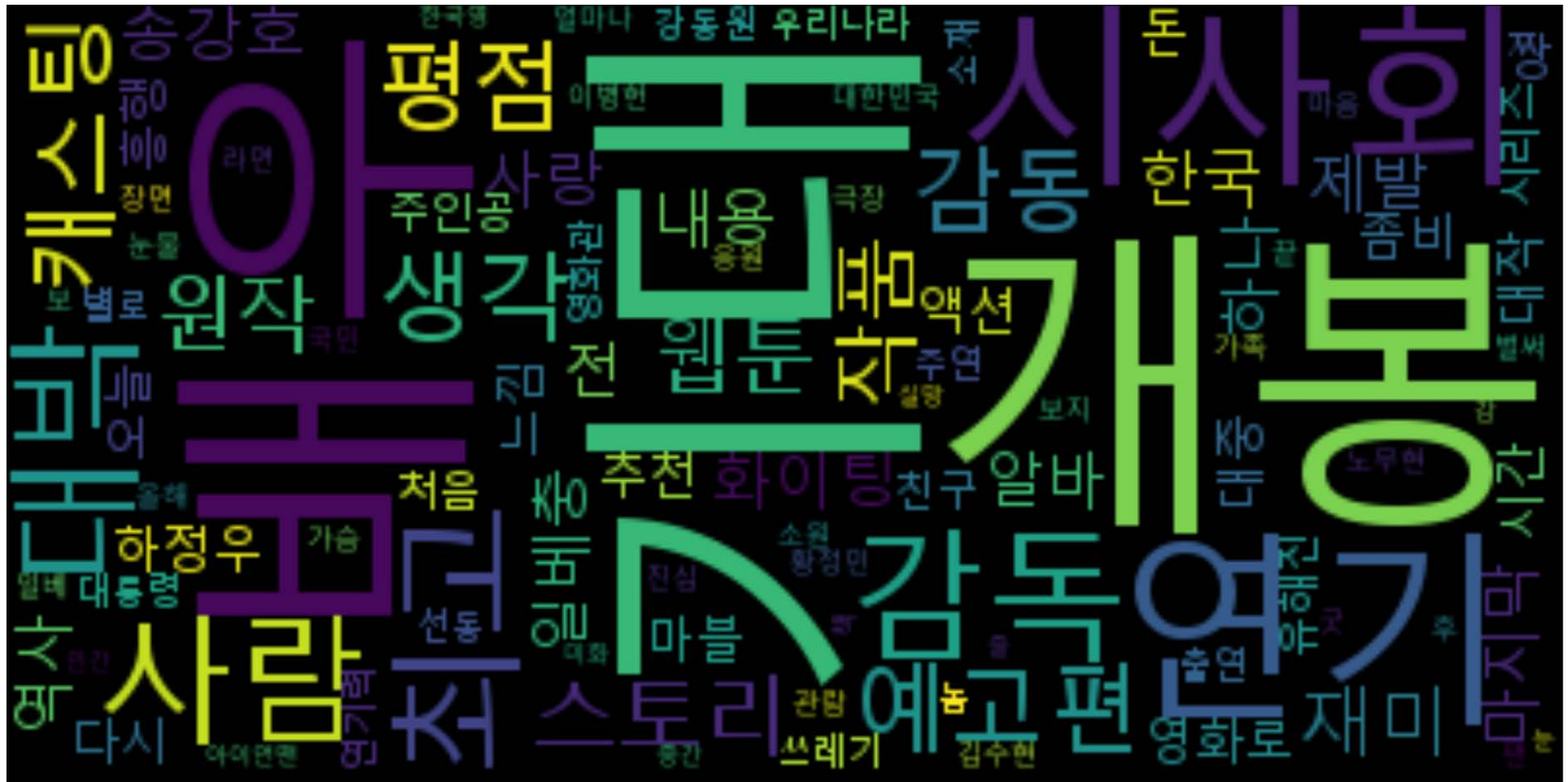


4. 연구 결과

4.1 분석 결과 - 텍스트 마이닝(Text Mining)

영화특성

맷글 빈도수 상위 **100위** 워드클라우드



(관객수 상위 244개 영화 댓글 대상)

4. 연구 결과

4.2 회귀 결과

Random Forest

독립변수 카테고리 세분화에 따른 정확도 향상

상위 빈도수 10개
배급사 외 기타 처리

Mean of squared residuals
0.3194209

% Var explained
69.96



상위 빈도수 30개
배급사 외 기타 처리

Mean of squared residuals
0.2797622

% Var explained
73.25

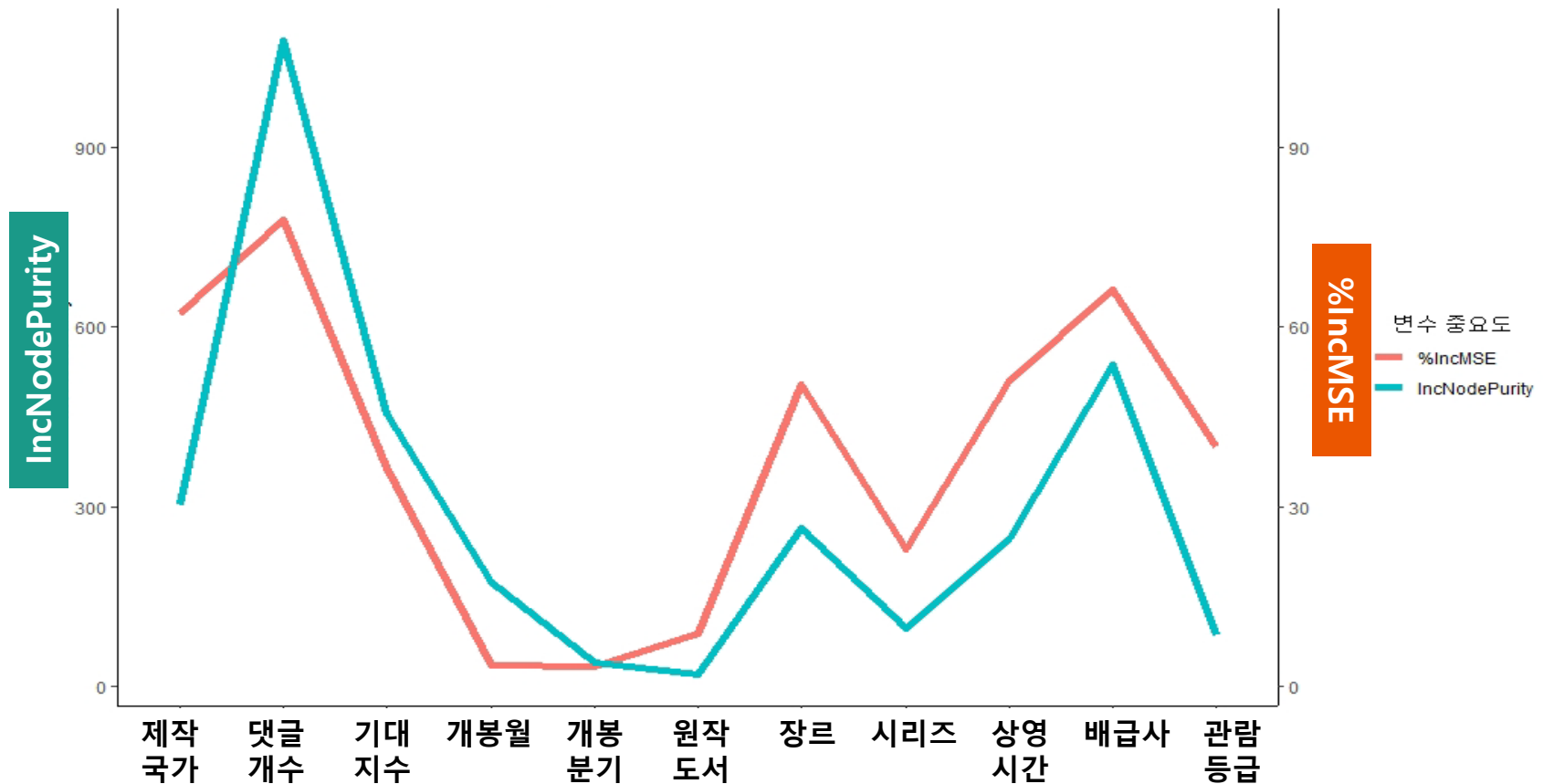


4. 연구 결과

4.2 회귀 결과

Random Forest

변수 중요도
(상위 빈도수 10개 배급사 외 기타 처리)

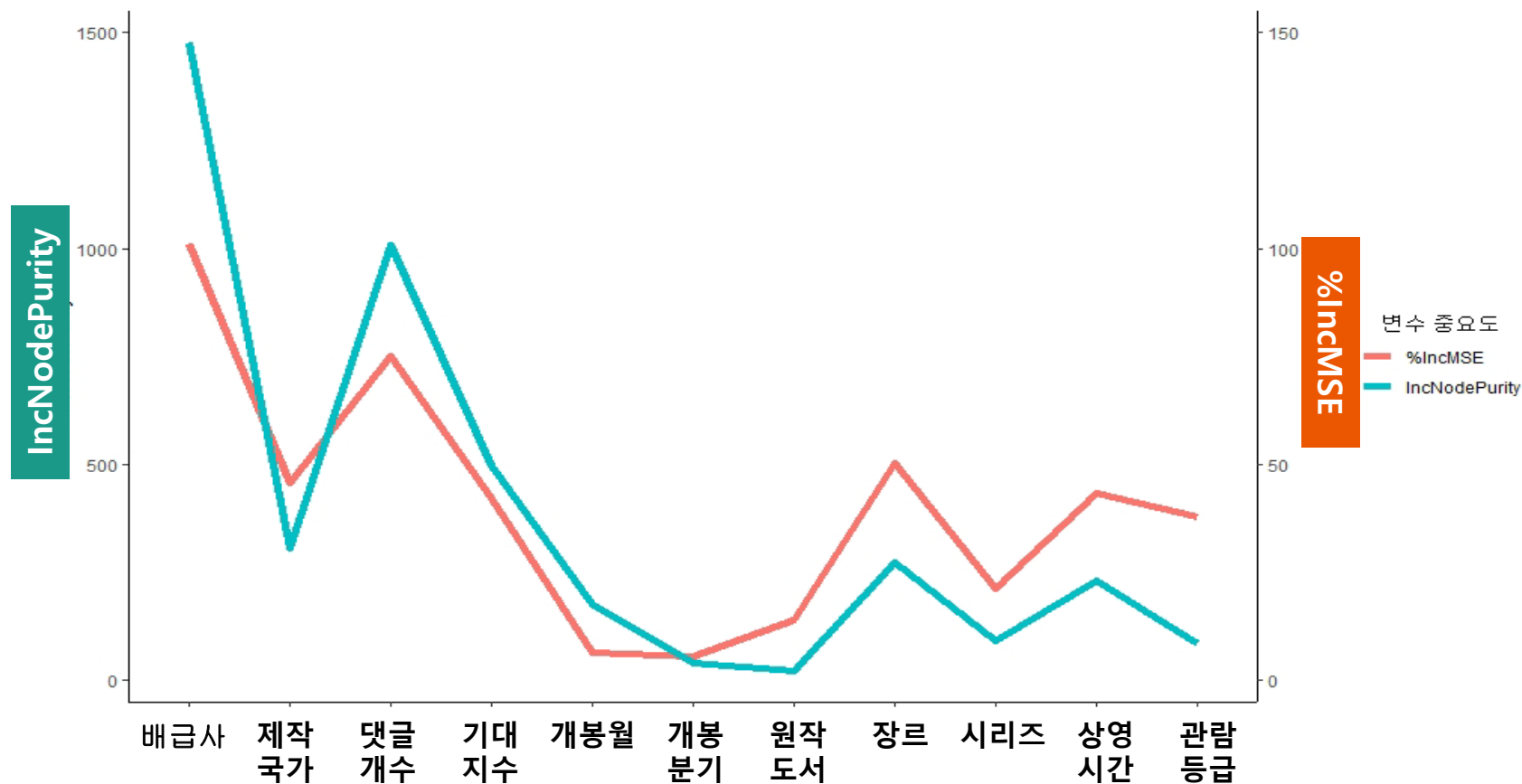


4. 연구 결과

4.2 회귀 결과

Random Forest

변수 중요도
(상위 빈도수 30개 배급사 외 기타 처리)

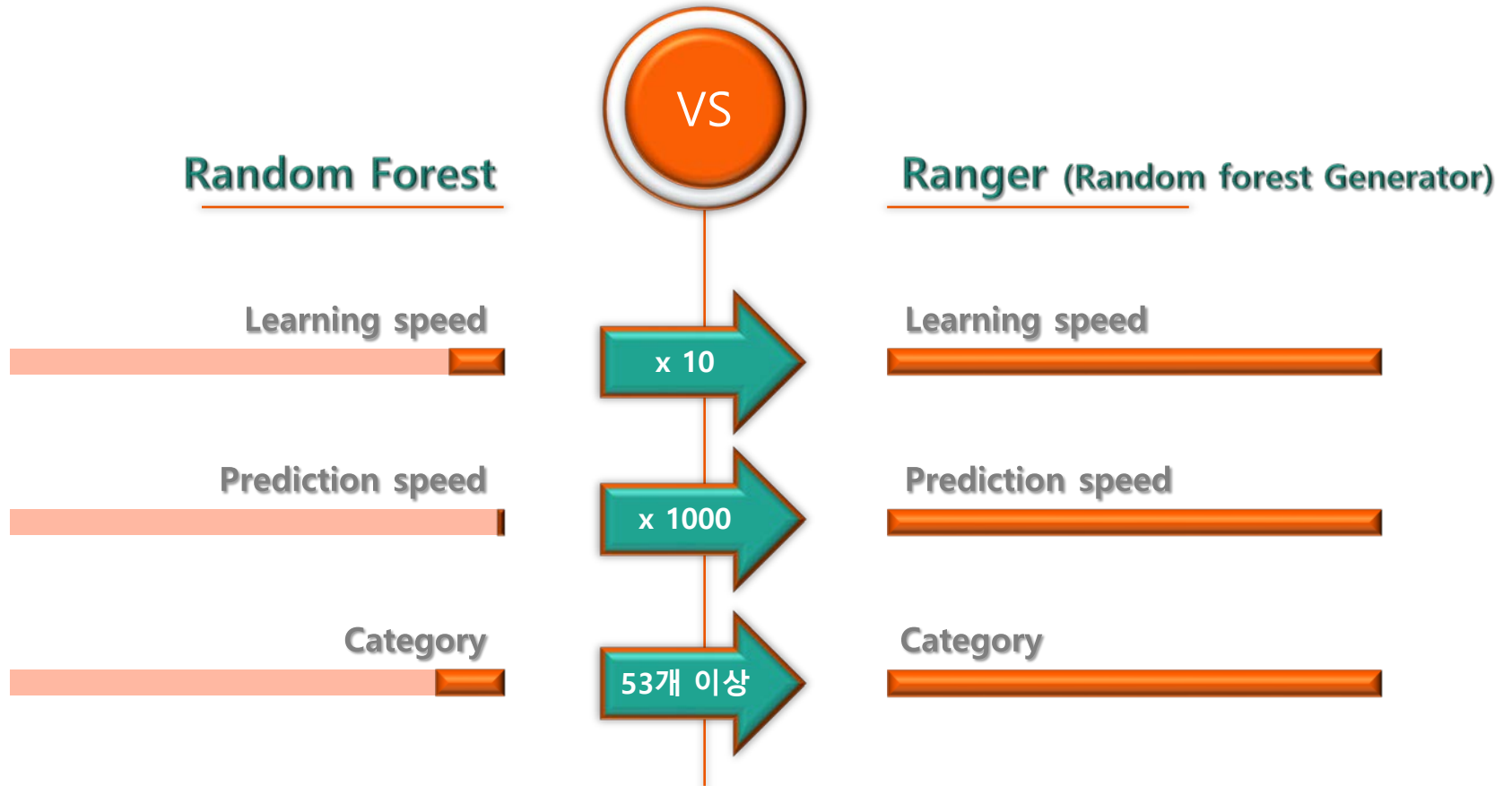


4. 연구 결과

4.2 회귀 결과

RF vs. Ranger

Random Forest와 Ranger의 성능 차이



4. 연구 결과

4.2 회귀 결과

Ranger

독립변수 추가에 따른 정확도 향상

mtry	splitrule	RMSE	Rsquared	MAE
2	variance	1.629516	0.6956181	1.359985
2	extratrees	1.740065	0.6529141	1.455992
67	variance	1.222334	0.7360995	0.949901
67	extratrees	1.212421	0.7400962	0.939942
132	variance	1.250245	0.7247678	0.969163
132	extratrees	1.218754	0.7376437	0.944625

독립변수 추가
(주인공, 감독)

mtry	splitrule	RMSE	Rsquared	MAE
2	variance	0.9693101	0.6794171	0.820884
2	extratrees	0.9763318	0.6643184	0.826862
39	variance	0.4932479	0.7805585	0.395363
39	extratrees	0.5166988	0.7582752	0.414305
765	variance	0.5035993	0.7634358	0.387089
765	extratrees	0.499741	0.766504	0.383093

4. 연구 결과

4.3 예측 결과

Random Forest

독립변수 카테고리 세분화에 따른 정확도 향상

영화제목	예측관객수 (상위 빈도수 10개 배급사 외 기타처리)	예측관객수 (상위 빈도수 30개 배급사 외 기타처리)	실누적관객수 (7월10일 기준)
스파이더맨: 파 프롬 홈	927,183	966,902	5,215,995
별의 정원	16,805	15,482	9,830
갯버스톤	12,875	10,525	5,905
칠드런 액트	8,134	11,276	11,823
RMSE의 평균값	1,076,611	 1,064,978	


영화선정 기준 : 7월 1일 ~ 7일 개봉 영화 중 선정

4. 연구 결과

4.3 예측 결과

RF vs. Ranger

Random Forest와 Ranger의 정확도 차이

영화제목	Random Forest 예측관객수	Ranger 예측관객수	실누적관객수 (7월10일 기준)
스파이더맨: 파 프롬 홈	966,902	2,212,724	5,215,995
별의 정원	15,482	17,937	9,830
칠드런 액트	11,276	5,999	11,823
겔버스톤	10,525	9,582	5,905
RMSE의 평균값	1,064,978	 755,219	

영화선정 기준 : 7월 1일 ~ 7일 개봉 영화 중 선정

4. 연구 결과

4.3 예측 결과

Ranger

독립변수 추가에 따른 정확도 향상

영화제목	예측관객수 (주인공, 감독 불포함)	예측관객수 (주인공, 감독 포함)	실누적관객수 (7월10일 기준)
스파이더맨: 파 프롬 홈	2,212,724	2,352,814	5,215,995
별의 정원	17,937	16,130	9,830
갤버스턴	9,582	12,257	5,905
칠드런 액트	5,999	6,240	11,823
RMSE의 평균값	755,219	 720,354	

영화선정 기준 : 7월 1일 ~ 7일 개봉 영화 중 선정

5. 결론 및 한계점

5.1 연구의 결론

01

분산분석 결과,
개봉요일 및 개봉연도는 관객수에 영향을 미치지 않았다.

02

상관분석 결과,
개봉 전 평점은 관객수에 영향을 미치지 않았다.

03

회귀분석 결과,
관객수에 대한 배급사의 변수중요도가 높았다.

04

회귀분석 결과,
Ranger모델의 정확도가 Random Forest모델보다 높았다.

5. 결론 및 한계점

5.2 연구의 한계점



예측 데이터의 기대 지수, 댓글 개수는 개봉일 전까지 누적 진행



출처별 데이터의 개수 및 표기법 상이



감성사전의 낮은 정확성

감사합니다.



<참고문헌>

- 백일, 김규곤, 최승배, 강창완. (2015). Stylometry를 이용한 영화 흥행 예측. Journal of the Korean Data Analysis Society (April 2015) Vol. 17, No. 2 (B), pp. 719-728
- 김병도, & 표태형. (2002). 개봉 전 영화의 수요예측모형. ", 경영논집, 제36권, 제1호(2002), pp.1-23
- 박지윤, 유인혁, & 강성우. (2017). 개봉 전후 트윗 개수의 증감률과 영화 매출간의 상관관계. 대한안전경영과학회지, 19(4), 169-182.
- 이오준, 박승보, 정다울, & 유은순. (2014). 소셜 빅데이터를 이용한 영화 흥행 요인 분석. 한국콘텐츠학회논문지, 14(10), 527-538.
- 장병희, 이양환, 김병선, & 남상현. (2009). 심리적 변인 활용을 통한 영화흥행 예측의 정교화: 1 주차 흥행실적을 중심으로. 한국언론학보, 53(4), 346-371.
- 김병선. (2009). 영화 유형에 따른 흥행 예측 요인 비교 연구: 2005~ 2007 년 국내 개봉 영화의 개봉 방식 및 상영 기간에 따른 유형 분류를 중심으로. 한국언론학보, 53(1), 257-287.
- 전희국, 현근수, 임경빈, 이우현, & 김형주. (2014). 영화 흥행 실적 예측을 위한 빅데이터 전처리. 정보과학회 컴퓨팅의 실제 논문지, 20(12), 615-622.
- 조승연, 김현구, 김범수, & 김희웅. (2014). 영화 흥행성과 예측을 위한 온라인 리뷰 마이닝 연구: 개봉 첫 주 온라인 리뷰를 활용하여. Information Systems Review, 16(3), 113-134.
- 문성민, 하효지, & 이경원. (2015). 영화의 흥행 성과와 리뷰 감정어휘와의 관계 분석. Design Convergence Study 53 Vol.14. no.4
- 이상훈, 조장식, 강창완, & 최승배. (2015). 텍스트 마이닝을 활용한 영화흥행 예측 연구. 한국데이터정보과학회지, 26(6), 1259-1269.
- Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018, April). A Data mining Technique for Analyzing and Predicting the success of Movie. In Journal of Physics: Conference Series (Vol. 1000, No. 1, p. 012100). IOP Publishing.
- Movie Success Prediction, Rakesh Parappa
- Liu, T., Ding, X., Chen, Y., Chen, H., & Guo, M. (2016). Predicting movie Box-office revenues by exploiting large-scale social media content. Multimedia Tools and Applications, 75(3), 1509-1528.
- Yoo, S., Kanter, R., & Cummings, D. (2011). Predicting Movie Revenue from IMDb Data. Stanford University.
- Akhter, M., Jackson, E., Rahman, M. A., & Chang, L. J. Predicting Movie Success Using Machine Learning Algorithms.
- Latif, M. H., & Afzal, H. (2016). Prediction of movies popularity using machine learning techniques. International Journal of Computer Science and Network Security (IJCSNS), 16(8), 127.