영화 관객 수에 영향을 미치는 요인 분석

- 회귀해석 중간보고서 (1반 18팀)-

2016110484 통계학과 오병찬

2016110495 통계학과 최정석

목 차

1. 들어가며

2019년 5월 25일 영화 '기생충' (봉준호 作)은 제 72회 칸 국제영화제에서 최고상인 황금 종려상을 수상했다. 이는 한국 영화 역사상 전무후무했던 수상이었으며, 한국 영화의 예술성이 세계적인 위치에 올랐다는 반증이기도 하였다.

영화 산업에 대한 발전과 더불어 영화는 우리 일상생활에서도 친숙하게 자리 잡고 있다. 현대 사람들에게 어떤 취미를 가지고 있는지에 대한 질문에 '영화 시청'은 빠질 수 없는 답변 중 하나이다. 실제 한국 갤럽에서 조사한 바에 의하면 영화 시청은 한국인이 좋아하는 취미 상위 10개 항목에 이름을 올리고 있었으며 2019년에도 TV 시청 항목과 함께 6위에 자리 매김을 하여 꾸준한 인기를 보이고 있다 [1]. 하지만 이러한 영화를 제작하는 데는 막대한 비용이 들며, 개봉 전 영화의 흥행을 판단하는 것은 상당히 어려운 일이다. 기존에 많은 연구들이 있었지만, 분석에서 중요한 부분을 차지하는 독립 변수의 수가 적고 제한되어 많은 부분을 포괄하는 결과가 도출되지 못 하였다고 판단한다. 최근 웹 검색의 중요성이 대두되고 있고 '크롤링'이라는 기술이 도입되었다. 이 기술을 접목시켜 기존에 분석들에서 수집할 수 없던 데이터들을 수집하여 다양한 변수를 이용하여 다방면에서의 회귀분석을 하고자 한다. 더 나아가 결과적으로 최근 개봉한 영화들을 바탕으로 관객수 예측을 하여 올바른 회귀 식이 도출되었는지 확인해보고자 한다.

2. 데이터 설명

1) 영화관 입장권 통합 전산망 (KOBIS)

KOBIS 홈페이지에서 공식 통계 데이터를 이용하였으며 이는 비상업적인 용도에 한하여 공용으로 사용할 수 있는 오픈 소스데이터이다. 영화에 대한 인식 변화, 영화에 대한 접근성 향상 등 분석 변수로 사용되지 않은 외부 요인에 대해서 최대한 통제하기 위해 비교적 오래된 데이터는 배제하기로 하였다. 또한 영화 흥행에 영향을 주는 변수를 찾음과 동시에 개봉할 영화에 대한 예측이 포함되어 비교적 최근 데이터인 2015년부터 2020년도 까지의 공식 통계 데이터를 분석에 사용하였다. <표 1>

[1] 한국인이 좋아하는 40가지 [문화편] (2019, Gallup)

<표 1> KOBIS data

변수 이름	속성	예시
영화명	문자	극한직업, 겨울왕국2, 1987
감독	문자	이병헌, 크리스 벅, 장준환
배급사	문자	씨제이앤엠㈜, 월트디즈니컴퍼니
개봉일	날짜	2019-01-23, 2017-12-17
영화 형태	문자	장편, 단편
국적	문자	한국, 미국, 네덜란드, 노르웨이
전국 스크린 수	숫자	1978, 2235, 935, 1444, 2142
전국 매출액	숫자	139,647,979,516 / 87,175,154,369
전국 관객 수	숫자	16,264,944 / 13,934,592 / 9,452,648
서울 매출액	숫자	31,858,660,536 / 16,753,103,620
서울 관객 수	숫자	3,638,287 / 15,445,294 / 1,412,367
장르	문자	코미디, 판타지, 액션, 어드벤처
등급	문자	15세이상관람가, 전체관람가
영화 구분	문자	일반영화, 독립영화

2) 크롤링

크롤링이란, 웹에서 유용한 정보를 특정 데이터베이스로 수집해오는 기술을 의미한다 [2]. KOBIS 등 공용 데이터 만을 분석에 이용하기엔 한계가 있었고, 더 많은 데이터를 수집하기 위해 크롤링을 실시하기로 하였다. '네이버 영화', '네이버 뉴스' 사이트를 대상 사이트로 결정하였고 타 사이트에 비교적 많은 데이터, 최근까지도 가장 활발하게 활동이 되어진다는 점이 선정 이유였다. Html 구조를 파악하여 파이썬 및 R 프로그램을 이용하여 코딩 작업을 하였다. 크롤링 한 데이터를 영리를 위한 목적이 아닌, 소장하고 활용하는 것은 불법이 아니라는 점을 확인하고 분석에 사용하였다. 크롤링 한 데이터는 다음과 같다. <표 2>

[2] 우리말샘『정보·통신』

<표 2> Crawling data

변수 이름	속성	최소값	최대값	예시	비고
주연배우	문자	NaN	NaN	엠마 왓슨, 이현승, 송강호	최대 3명
Score_ntz	숫자	0	10	8.5 / 7.83 / 5.8 / 9.0	영화 네티즌 평점
Score_crt	숫자	0	10	8.2 / 7.6 / 6.2 / 7.5	영화 전문가 평점
Freq_act	숫자	0	182736	124621 / 50012	배우 영향력
Freq_title	숫자	0	12355	6452 / 9421	개봉 전 영화 화제성

*** <표 2> 추가 설명

- Freq_act : 웹 포털에 주연 배우의 이름인, "배우 OOO" 로 검색을 하였을 때 나오는 뉴스 빈도수를 나타낸다. 배우의 인지도, 배우가 가진 티켓 파워를 수치화 하기 위해 만들어졌다.

- Freq title : 웹 포털에 영화 제목인, "영화 OOO" 로 검색을 하였을 때 나오는 뉴스 빈도수를 나타낸다. 영화가 언론에 얼마나 광고가 되고 이슈가 되었는지를 수치화 하기 위해 만들어졌다. 기간은 개봉일로부터 1년전에서 개봉일까지이다.

3) IMDb

IMDb는 Internet Movie Database의 약자로 영화, 배우, 텔레비전 드라마, 비디오 게임 등에 관한 정보를 제공하는 온라인 데이터베이스이다. 2014년 기준으로 6백만 이상의 인물의 정보를 소유하고 있는 영화 관련 데이터베이스 중 최대 규모이다. 이 사이트에서 제공하는 지표 중 "STARmeter"가 있는데 이는 배우, 감독 등 등록된 인사들에 대한 순위를 나타낸다. 이 순위는 포털사이트에서 검색된 빈도의 수, 참여한 작품에 대한 관객 수와 평가 그리고 수상 경력 등 다양한 부분에서의 정보를 종합적으로 계산하여 산출된다. 본 연구에서는 감독의 영향력을 수치화 하기 위해 이러한 STARmeter를 지표로 사용하였다.

<표 3> IMDb STARmeter data example

title	Director	STARmeter (Rank)
엑시트	이상근	144929
덩케르크	크리스토퍼 놀란	43
1987	장준환	57926

3. 전처리

1) 변수 생성

- 개봉일이 제일 빠른 날짜를 가진 영화를 기준으로 개봉일이 몇 주가 차이나는 지 계산하여 주(week) 변수를 추가하였다. 이 변수는 경쟁 점수 변수를 만드는데 사용된다.

- 국적 변수는 범주형 변수이다. (한국, 미국, 칠레, 프랑스…) 따라서 dummy 화를 진행하였다. 한국 영화의 경우 외국에서 만든 영화들과 차이가 있는지에 대한 분석을 위해 한국 영화를 기준변수(reference variable) 로 지정하였다. 이에 한국 영화 다음으로 많은 미국 영화를 dummy_US 변수를 사용하여, 나머지 국가들을 dummy_Others 변수를 사용하여 표현하였다.

- 영화 등급과 장르는 범주형 변수이다. 이 변수들을 모두 범주화하여 분산 분석(ANOVA)을 한다면 발생되는 더미변수가 너무 많아지게 된다. 따라서 회귀분석에 적합하게 이 두 변수의 내용을 공통적으로 포괄할 수 있는 새로운 변수 '자극성' 변수를 만들기로 하였다. 영화 등급이 '청소년관람불가'인 경우 자극적인 영화라고 지정하였다. 또한 등급에 상관없이 '범죄', '전쟁', ',스릴러', '공포' 장르인 영화 또한 자극적인 영화로 지정하였다. 마지막으로 '15 세이상관람가' 이면서 장르가 '미스터리', '사극', '액션', 'SF' 인 경우 자극적인 영화로 변수를 부여하였다. 15 세 이상 관람가인 사극 장르에는 대표적으로 봉오동 전투가 있는데, 이를 비롯한 많은 영화에서 전쟁 장면을 통한 잔인한 장면이 연출되었기에 자극성을 가진 영화라고 판단하였다. 가공 전 변수를 통한 분석과 가공 후 변수인 '자극성' 변수를 이용한 분석에서 설명 계수가 거의 차이를 보이지 않았다. 따라서 설명력에서 차이가 없을 때, 변수의 수를 최소화하는 '자극성' 변수를 사용하기로 하였다.

- 본 분석과 비슷한 주제에 대한 다양한 연구가 많이 있었다. 하지만 이러한 연구들에서 언급은 하였지만 실질적으로 변수로 추가하지 못한 요인이 '경쟁작의 영향' 이었다. 본 연구에서는 이러한 요인을 반영하고자 하였고 새로운 시도를 하였다. 앞서 만든 주 변수를 기준으로 앞 뒤 2 주씩을 실제 경쟁작의 영향을 받는 기간으로 설정하였고 그 기간안에 개봉일이 포함된 영화는 경쟁작으로 선정하였다[3]. 이때 관객수가 5 만명이 되지 않는 영화의 경우 다른 영화들과의 경쟁 구도가 만들어지지 않는다고 판단하여 제거하였다. 그렇게

[3] 복합 상영관 확산과 함께 전국 동시 개봉하는 광역 개봉 체계가 정착되면서 영화 소비 주기가 점차 짧아지는 실정임. 영화 평균 상영기간 : 12년 이후 30일. <영화산업 시장분석, 2017> 공정거래위원회, 미래산업전략 연구소

만들어진 그룹 내에서 실제 관객 수를 기준으로 매긴 순위를 그룹 크기로 나누었다. 1 에서 이 값을 빼고, 100 을 곱하여 범위는 0~100 이며, 100 에 가까울수록 경쟁력이 있음을 나타내는 경쟁 점수 변수를 만들었다.

- 계절이 흥행에 미치는 영향을 파악하기 위하여 봄, 여름, 가을, 겨울, 4 가지 요인으로 이루어진 계절 변수를 만들었다. 영화 개봉일 기준 3-5 월을 봄, 6-8 월을 여름, 9-11 월을 가을, 12-2 월을 겨울이라는 요인을 할당해주었다.

2) 변수 제거 및 대체

- 전국에 대한 데이터를 분석하기 때문에 서울 지역에 국한되어 있는 변수를 제거하였다.

- 배급사 변수는 범주형 변수로 회귀 분석에 직접적으로 사용할 수 없다. 따라서 '스크린 수' 확보가 '배급사'의 파워와 연관이 있다는 연구[4]에 기반하여 스크린 수를 이용하여 배급사의 영향력을 수치화 하였다. 실제 배급사 별 평균 스크린 수를 정렬해보았을 때, 대체적으로 유명한 배급사의 경우 스크린 수가 많은 경향을 확인하였다.

3) 데이터 전처리

- KOBIS 데이터 중 성인물은 일반적인 영화 상영 방식과 달라 제거하였다.

- 매출액의 분포를 고려하여 매출액이 50만원 이하인 경우, 관객 수에 영향을 주는 변수들을 찾는 본 연구의 의도와 어긋나는 데이터로 판단하여 제거하였다. 또한 일반적인 결과를 도출하기 위해 관객수가 5만 이상, 1000만 이하인 데이터를 분석에 이용하였다.

- 청소년관람불가 등급을 받은 영화들 중 걸러지지 않은 성인물이 있어, 청소년관람불가 등급이면서 매출액이 700만 이하인 데이터는 제거하였다.

- 네티즌 평점 또는 전문가 평점이 0점인 경우, 네이버 영화 사이트 자체에서 데이터가 없는 것이므로 제거하였다.

- 본 분석에서 배우의 영향력이 변수로 사용된다. 하지만 애니메이션의 경우 배우가 아닌 성우가 크롤링되어, 올바른 배우의 영향력을 알 수 없게 된다. 이에 애니메이션 항목은 제거하였다.

- 개봉일로부터 1년전부터 개봉일까지 네이버 뉴스에 검색된 뉴스 수 크롤링하였다. (ex) "영화

[4] 영화 흥행 결정 요인과 흥행 성과 예측 연구 (2011. 한국통계학회논문집 김연형 외 1명)

베테랑” 검색), 이 과정에서 영화 제목이 ‘내일’ 과 같은 경우, 영화에 대한 기대에 비례한 뉴스 수가 아닌 ‘내일’ 이라는 단어가 일상생활에서 익숙한 언어라는 점의 영향을 받아 영화 흥행과 관계없이 비정상적으로 많은 빈도수가 발견되었다. 따라서 관객수가 100만 이하이지만, 검색된 뉴스 수가 4000개 이상인 경우 본 분석에서 의도하지 않은 결과를 도출할 것이라 판단하여 제거하였다,

- STARmeter 지표를 이용한 감독 영향력 변수는 스펙트럼이 넓었다. 그 원인 중 하나가 분석 데이터 속에 데뷔작이 포함되어 있어 그 감독의 영향력이 매우 낮아 즉 순위가 비정상적으로 낮게 측정되는 경우가 있었다. 따라서 감독의 영향력을 제대로 판단할 수 없는 신인 감독인 경우 제외하였다.

- “Unbroken” 이라는 영화의 경우 배우 안젤리나 졸리가 감독을 맡았는데 그의 배우로서의 높은 STARmeter 순위가 감독으로서의 높은 영향력을 나타낸다고 볼 수 없다고 판단하였다. 따라서 배우가 감독으로 참여한 작품은 제외시켰다. 레미제라블의 주연인 러셀크로우가 감독으로 참여한 영화 “The Water Diviner”를 제외한 것이 그 예이다.

최종적으로 분석에 사용되는 데이터는 다음과 같다. <표 4>

<표 4> 최종 데이터셋

변수 이름	속성	범위	예시	비고
Title	문자	NaN	극한직업 / 공조	영화 제목
Score_ntz	숫자	0~10	8.5 / 9.4 / 7.2	영화 네티즌 평점
Score_crt	숫자	0~10	7.6 / 8.9 / 7.5	영화 전문가 평점
Audience	숫자	90~16264944	154720 / 25423	관객 수 (종속 변수)
Screen	숫자	40~2142	1108 / 711 / 462	상영 스크린 수 (배급사 영향)
Freq_act	숫자	0~182736	124756 / 84236	배우 영향력
Freq_title	숫자	1~12355	3458 / 9124	개봉 전 영화 화제성
Nation_US	더미	0,1	0 / 1	제작 국가가 미국일 시, 1
Nation_Others	더미	0,1	0 / 1	한국, 미국이 아닐 시 1
Provocation	더미	0,1	0 / 1	자극적인 영화일 시 1
Score_comp	숫자	2.7~100	95.8 / 50.5 / 12.6	경쟁력 점수
Score_direct	숫자	4~2184328	4208 / 515826	감독의 STARmeter 순위
Season	요인	NaN	Autumn,Winter	개봉일이 속한 계절

4. 실증 분석

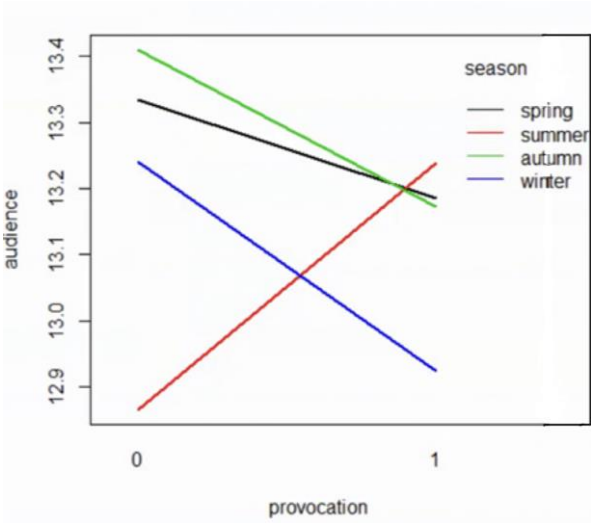
1) 산점도 (<부록 1>, <부록 2> 참고)

원 변수에 대한 산점도를 ggpairs 함수를 이용하여 시각적으로 확인하였다 <부록 1>. 그 결과 더 확실한 선형 관계를 위해 변수의 변환이 필요하다고 판단하였다.

Audience 변수의 단위는 천만명이 넘는 영화도 많아 다른 변수들과의 직접적인 비교가 불가하였다. 따라서 log 변환을 통해 스케일을 조정해주었다. score_direct 변수의 경우 데이터의 범위가 2 백만이 넘는 분산이 큰 변수이다. 따라서 log 변환을 통해 분산을 줄여주었다. 또한 screen, freq_act, freq_title 변수들의 단위가 다른 변수들에 비해 커서 조정이 필요하지만 audience 변수보다는 단위가 작아 log 보다 변환의 정도가 덜 한 sqrt 를 사용하였다. 이러한 변수 변환을 거친 후 변수들간의 산점도와 상관 계수는 <부록 2>에 첨부하였다.

2) 분산분석

범주형 변수들 사이에 유의미한 차이가 있는지 확인하기 위해 분산분석을 실시하였다. 먼저 계절에 대한 영향력을 생각해보았다. 추가적으로 계절 변수는 자극성 변수와 교호작용이 발생할 것이라고 예상했는데, 이는 여름에 무더위를 씻겨줄 공포 영화가 많이 개봉되고 이러한 영화들에 관객들이 많이 몰렸었던 경험에서 야기하였다. <그림 1>과 같이 교호작용 플랏에서 선들 간의 교점이 발생해 플랏 상으로는 교호작용이 있다고 판단할 근거를 보여주었다. 더 정확한 판단을 위해 두 변수의 교호작용 항을 추가하여 이원분산분석을 실시하였다 <표 5>.



<그림 1> Interaction plot of Provocation and Season

<표 5> Two Way Anova between Season and Provocation with Interaction

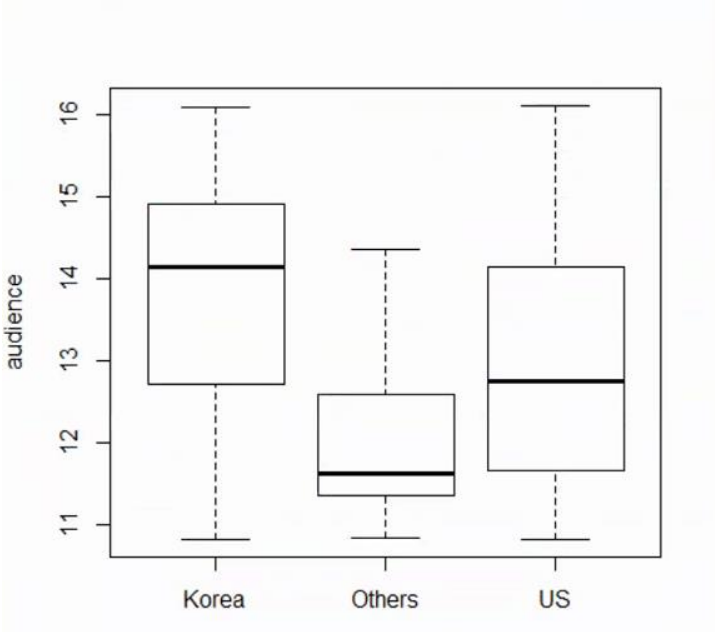
	Df	Sum sq	F value	Pr(>F)	Signify.
Season	3	6.0	0.893	0.445	
Provocation	1	1.0	0.432	0.511	
Season:Prov	3	8.6	1.286	0.279	
Residuals	484	1083.2			

<표 5>에서 볼 수 있듯이 테스트 결과 두 변수간의 교호작용은 존재하지 않았고, 이러한 교호작용을 제외한 후 이원분산분석을 실시하여도 계절과 자극성은 관객 수를 나타내는 변수를 나타내기에 유의하지 않은 변수임이 확인되었다. <표 6>

<표 6> Two Way Anova between Season and Provocation without Interaction

	Df	Sum sq	F value	Pr(>F)	Signify.
Season	3	6.0	0.891	0.445	
Provocation	1	1.0	0.431	0.512	
Residuals	484	1083.2			

다음은 영화를 제작한 국가에 대해 관객 수가 차이에 나는지 확인하고자 하였다. 한국과 미국 그리고 그 외 나라에 대해서 관객 수를 박스 플랏으로 나타내었다 <그림 2>. 한국과 미국의 영화의 경우 관객 수의 범위는 비슷하였다. 하지만 평균값과 중앙값의 경우 한국 영화가 조금 더 많은 관객 수를 나타냈는데, 이는 한국에 배급된 미국의 영화들 가운데 높은 관객 수를 기록하는 영화는 한국 영화에 비해 소수에 치우쳐져 있음을 말해준다. 이에 따라 한국 영화의 경우 미국 영화에 비해 관객 수의 평균값과 중앙값이 높게 나타났다. 또한 미국(US)과 다른 국가(Others)들의 비교에서, 미국이 다른 국가들에 비해 한국 영화 산업에서 주류를 이끌고 있음을 알 수 있다.



<그림 2> Boxplot between Nation

영화의 제작국가에 따라서 관객 수에 차이가 있다는 것을 확인하였고, 다중 비교에서 생길 수 있는 오류를 보정하기 위해 본페로니 교정을 이용하여 사후검정을 실시하였다.

<표 7> One Way Anova between Nation

	Df	Sum sq	F value	Pr(>F)	Signify.
Nation	2	171.8	45.3	<2e-16	* * *
Residuals	489	927.0			

<표 8> Post hoc analysis using Bonferroni

	Difference	pvalue	Signif.	LCL	UCL
Korea - Others	1.900	0.0	* * *	1.375	2.426
Korea - US	0.906	0.0	* * *	0.585	1.227
Others - US	-0.994	0.0	* * *	-1.500	-0.488

결과적으로 계절과 자극성에 대한 변수는 관객 수에 영향을 주지 않는 것을 확인하였고, 실제 영향을 주었던 국가 변수를 더미화하여 회귀분석에 이용하고자 하였다.

3) 회귀분석

독립 변수를 하나씩 추가/제거하여 종속 변수를 잘 예측하는 변수들을 선택하는 기법인 단계적 분석방법 (step wise regression)을 이용하였고, 이 변수들로 회귀분석을 하였다. 회귀분석을 통한 결과는 <부록 3> 과 같으며 다음과 같은 회귀식이 도출되었다.

$$log(audience) \sim 8.62 + 0.103*(Score\_ntz) + 0.065*(sqrt(screen)) + 0.007*(sqrt(freq\_title)) + 0.001*(sqrt(freq\_act)) + 0.256*(Nation\_US) + 0.275*(Nation\_Others) + 0.028*(Score\_comp)$$

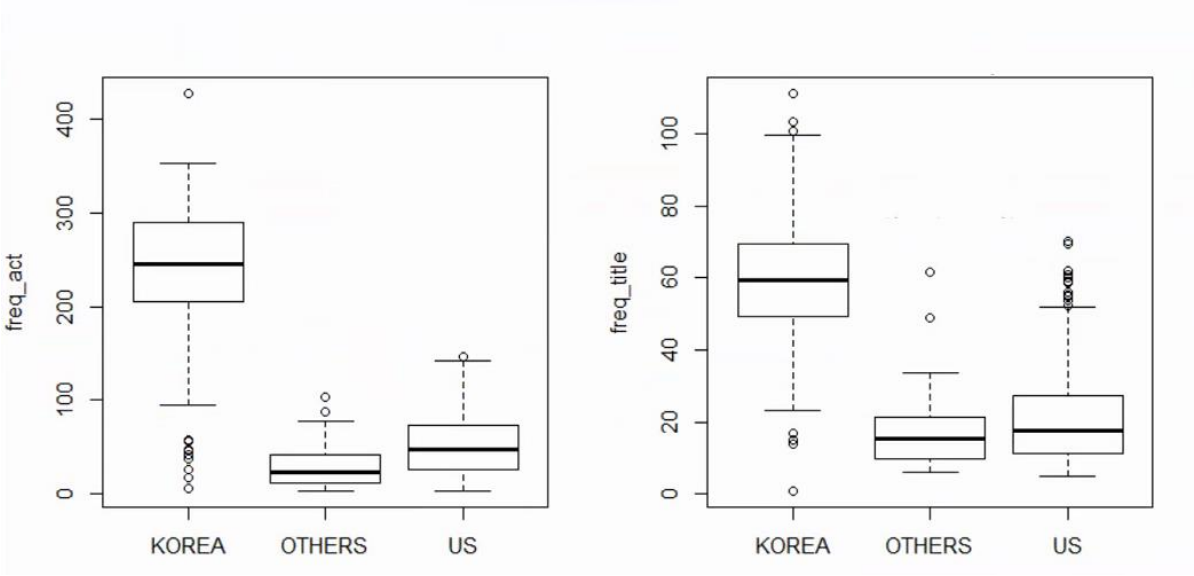
유의수준을 0.05 로 정하였을 때, Score_crt 변수, Provocation 변수, Score_direct 변수 그리고 Season 변수가 제외가 되었다. 설명계수는 0.916 이고 변수의 개수를 고려한 조정된 설명계수도 0.9147 로 낮지 않은 수치를 보여준다.

하지만 이는 앞서 분산분석을 통해 예상한 결과와 맞지 않는 회귀식이었다. <그림 2>를 통해 한국 영화의 평균적인 관객 수가 미국과 다른 국가의 영화에 비해 더 많다는 것을 확인했지만, Nation_US 와 Nation_Others 변수의 계수가 0.256 과 0.275 즉 양수로, 기준 변수인 한국 영화에 비해 외국의 영화인 경우 관객 수가 증가한다는 해석이 가능한 식이 도출되었기 때문이다. 따라서 독립 변수 간의 다중공선성이 발생하여 생긴 문제라 판단하여 확인하였다.

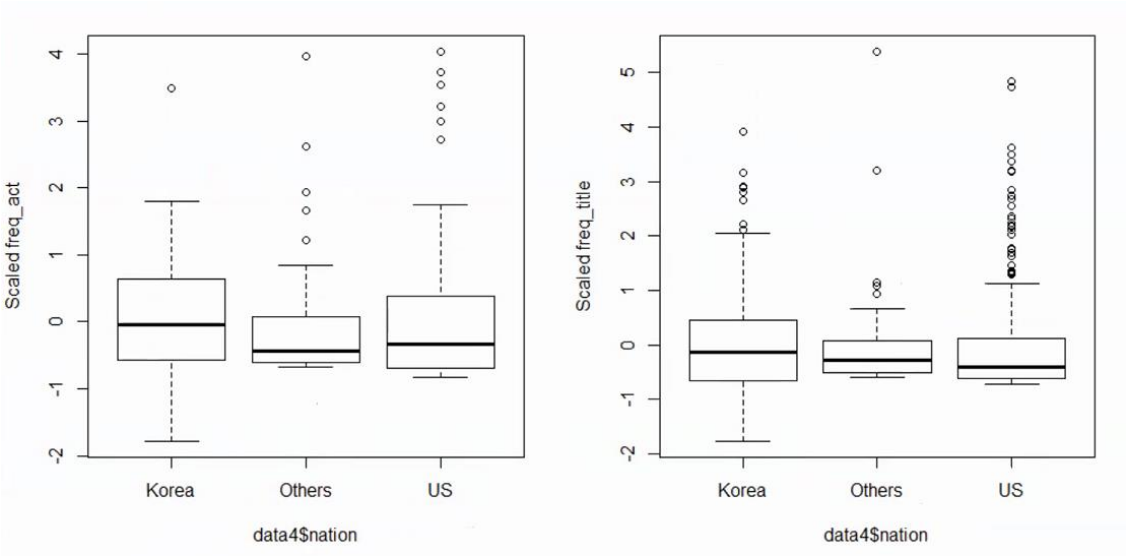
<표 9> Multicollinearity

변수 이름	Score_ntz	Sqrt(screen)	sqrt(freq_title)	sqrt(freq_act)
Vif	1.19	3.34	3.76	5.06
변수 이름	Nation_US	Nation_Others	Score_comp	
vif	5.31	2.86	3.02	

실제 freq_act 변수와 Nation_US 변수의 vif 값이 5 이상으로 다중공선성이 발생한 것을 알 수 있다. 추가적으로 배우의 영향력을 보여주는 freq_act 변수와 개봉 전 영화의 화제성을 보여주는 freq_title 변수를 국가 별로 비교를 해보자 한국의 배우, 영화에서 높은 값을 가진 것을 확인하였다 <그림 3>. 이는 빈도수를 크롤링 한 검색엔진으로 외국 사람보다 한국 사람의 이용률이 압도적으로 높은 네이버 포털 사이트를 사용하였기 때문에 발생한 문제였다. 따라서 국가별로 국가의 평균치를 빼고 국가의 표준편차로 나눠주는 표준화를 실시하였다. 그 결과 <그림 4> 처럼 국가 간의 차이를 어느정도 조정해줄 수 있었다.



<그림 3> freq_var between Nation



<그림 4> Scaled freq_var between Nation

이렇게 표준화 한 변수를 가지고 다시 회귀분석을 실시하였다. 이 과정을 통한 회귀식은 변수들이 더 유의하게 p_value 가 조정이 되었다. 이전에 예상하였던 대로 한국에 비해 다른 국가의 영화일 때 예측되는 관객 수가 적은 것을 알 수 있었고, 적은 정도는 미국일 때 보다 다른 국가일 때 더 커진다는 것 또한 반영이 되었다. <표 10>은 회귀분석 결과이다.

<표 10> Result of Regression Analysis with Scaled variables

변수 이름	Coefficients	Std.Error	t value	Pr(>ltl)	비고
(Intercept)	11.075	0.212	52.266	< 2e-16	* * *
Score_ntz	0.112	0.019	5.730	1.77e-08	* * *
Sqrt(screen)	0.069	0.004	15.931	< 2e-16	* * *
Scaled_title	0.010	0.023	4.135	4.19e-05	* * *
Scaled_act	0.006	0.022	2.753	0.006	* *
Nation_US	-0.663	0.047	-14.020	< 2e-16	* * *
Nation_Others	-1.156	0.088	-13.130	< 2e-16	* * *
Score_comp	0.762	0.032	23.654	< 2e-16	* * *
R-squared	0.9116	Adjusted R^2	0.9103	p-value	< 2.2e-16

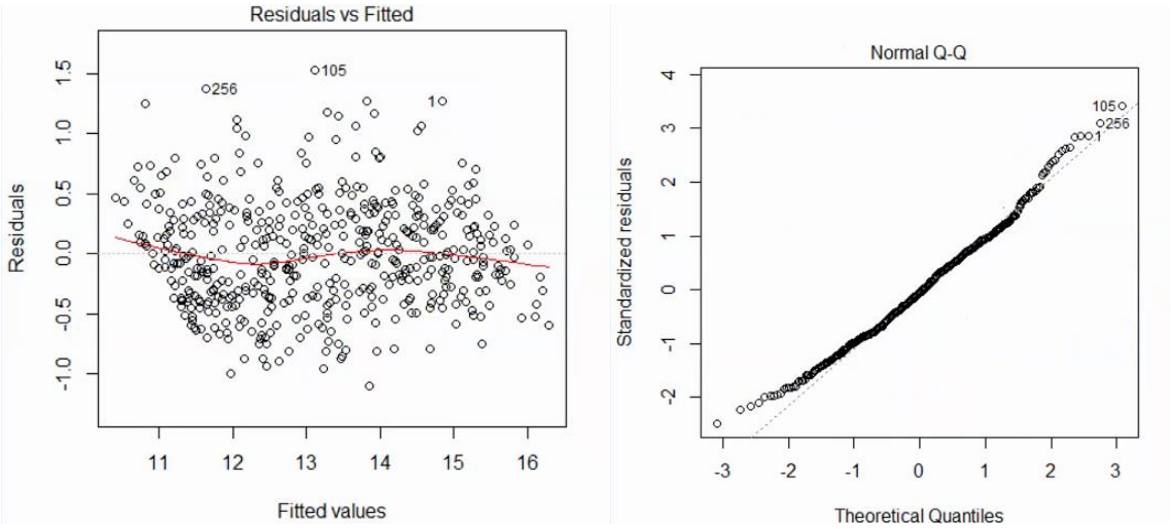
$$\boldsymbol{log}(\boldsymbol{audience}) \sim 11.075 + 0.112*(\boldsymbol{Score\_ntz}) + 0.069*(\boldsymbol{sqrt}(\boldsymbol{screen})) + 0.010*(\boldsymbol{scaled}(\boldsymbol{freq\_title})) + 0.006*(\boldsymbol{scaled}(\boldsymbol{freq\_act})) + (-0.663)*(\boldsymbol{Nation\_US}) + (-1.156)*(\boldsymbol{Nation\_Others}) + 0.762*(\boldsymbol{Score\_comp})$$

다음은 freq 변수에 대해 표준화를 한 후 실시한 회귀분석 결과를 바탕으로 다중공선성을 확인한 결과이다.

<표 11> Multicollinearity with Scaled variables

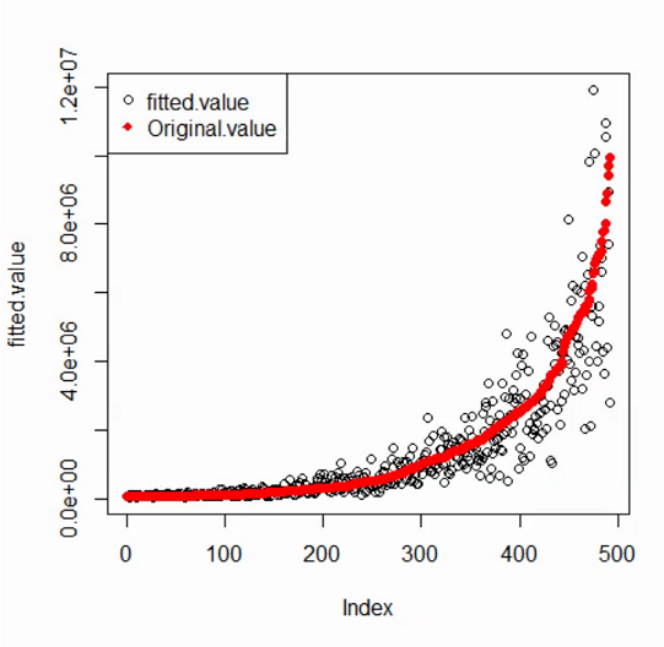
변수 이름	Score_ntz	Sqrt(screen)	Scaled_title	Scaled_act
Vif	1.18	3.24	1.32	1.18
변수 이름	Nation_US	Nation_Others	Score_comp	
vif	1.36	1.76	2.53	

3) 잔차의 정규성 및 추가 작업



<그림 5> Normal probability plot about Residuals

잔차 플랏을 확인해보면 평균 0 을 중심으로 퍼져있고 어느 부분에서 분산이 크거나 작지 않고 일정하게 퍼져있는 것을 확인할 수 있다. 또한 Q-Q 플랏에서도 정규성을 확인할 수 있었다.마지막으로 실제 데이터와 분석을 통해 만들어진 회귀식에 대한 추정값에 대한 플랏이다. 관객 수를 많이 보유한 영화의 경우 차이를 조금 보이긴 하지만 전체적으로 경향을 잘 표현하고 있다.



<그림 6> Fitted value and Original value

추가적으로 종속 변수인 관객 수를 예측하는데 각 독립 변수들의 영향력을 비교하는 차트도 <부록 4>에서 확인할 수 있다.

5. 해석

영화 관객 수의 영향을 미치는 요인을 정리해보면 '네티즌 평점', '스크린 수(배급사 영향)', '개봉 전 영화 화제성 ', '배우 영향력', '국가', '경쟁력' 변수라고 말할 수 있다.

네티즌 평점은 실제 종속변수인 관객 수의 관객들이 직접 작성하는 평점으로 대표성을 가지고 있다고 판단된다. 반면에 전문가 평점은 영화에 대한 전문적 평가에 대한 수치로 영화가 문학적, 기술적 등 다양한 전문가 관점에서의 완성도와 관객수는 필수적으로 비례하지는 않는다는 것을 알 수 있었다. 또 배급사의 영향력이 클수록, 출연하는 배우의 인지도가 높을수록, 개봉 전에 영화가 화제가 많이 될수록 예상되는 관객 수가 높다고 해석된다.

국가 변수를 살펴보면 기준이 되는 한국 영화에 비해 미국과 그 외 다른 국가의 영화일 때 예상되는 관객의 수가 적어지는 것을 볼 수 있다. 이는 앞에서 말했듯이 한국의 영화의 관객 분포가 평균이 중앙값보다 왼쪽으로 치우쳐진 분포이고 미국의 영화의 경우 이와 반대인 분포를 형성한 데서 기인한 현상이다. 결과적으로 한국 영화에 비해 인기있는 영화의 비중이 소수에 치우친 미국 영화일 때 평균적으로 관객 수가 낮게 예상된다는 결과가 도출된다. 그 외 국가의 영화의 경우 수집한 데이터에서 한국과 미국에 비해 평균적으로 낮은 관객 수를 보였으며 이는 아직 한국에서는 다른 국가들의 영화보다 미국의 영화가 주류를 담당하고 있음을 보여준다.

<부록 4>에서 독립변수의 영향력 차트에서도 확인할 수 있듯이 독립 변수 중 경쟁력 변수가 관객 수에 가장 큰 영향을 미친다는 것을 알 수 있다. 이는 인지도가 높은 배우가 출연하고 개봉 전 다양한 매체에서 홍보가 많이 된 영화라도 같은 시기에 개봉한 영화들과의 관계에서 우위를 차지하지 못한다면 상대적으로 관객수가 적게 예상된다고 해석할 수 있다.

추가적으로 제외된 편수를 살펴보자. 계절에 대해 관객 수의 추이는 크게 차이가 없었고, 감독에 대한 영향력도 예상과는 다르게 관객 수에 영향을 주지 않았다. 본 분석에 사용된 데이터는 이상치를 제거하였고 이 과정에서 어느 정도의 영향력이 있는 감독들의 집합에서는 감독의 영향력 외의 다양한 요인들의 관객 수에 영향을 보다 더 미친다는 것을 확인할 수 있었다. 영화의 자극성에 따른 관객 수의 차이도 확연한 차이가 없어 최종 회귀 모형에 포함되지 않았다.

6. 한계점과 기대효과

위 분석에서는 영화에 어떤 배우가 출연하는지에 따라 관객 수에 영향을 미칠 것이라고 판단하고 이 판단을 확인하는 분석을 진행하였다. 이 과정에서 앞에서도 언급했듯이 애니메이션 장르에는 주연배우 대신 목소리를 녹음한 성우가 크롤링이 되어 데이터가 만들어졌다. 따라서 본 분석은 애니메이션 장르를 제거한 모든 장르를 대상으로 한 분석이며 제외된 애니메이션 장르에 대한 분석은 추후 고민해야 하는 숙제이다.

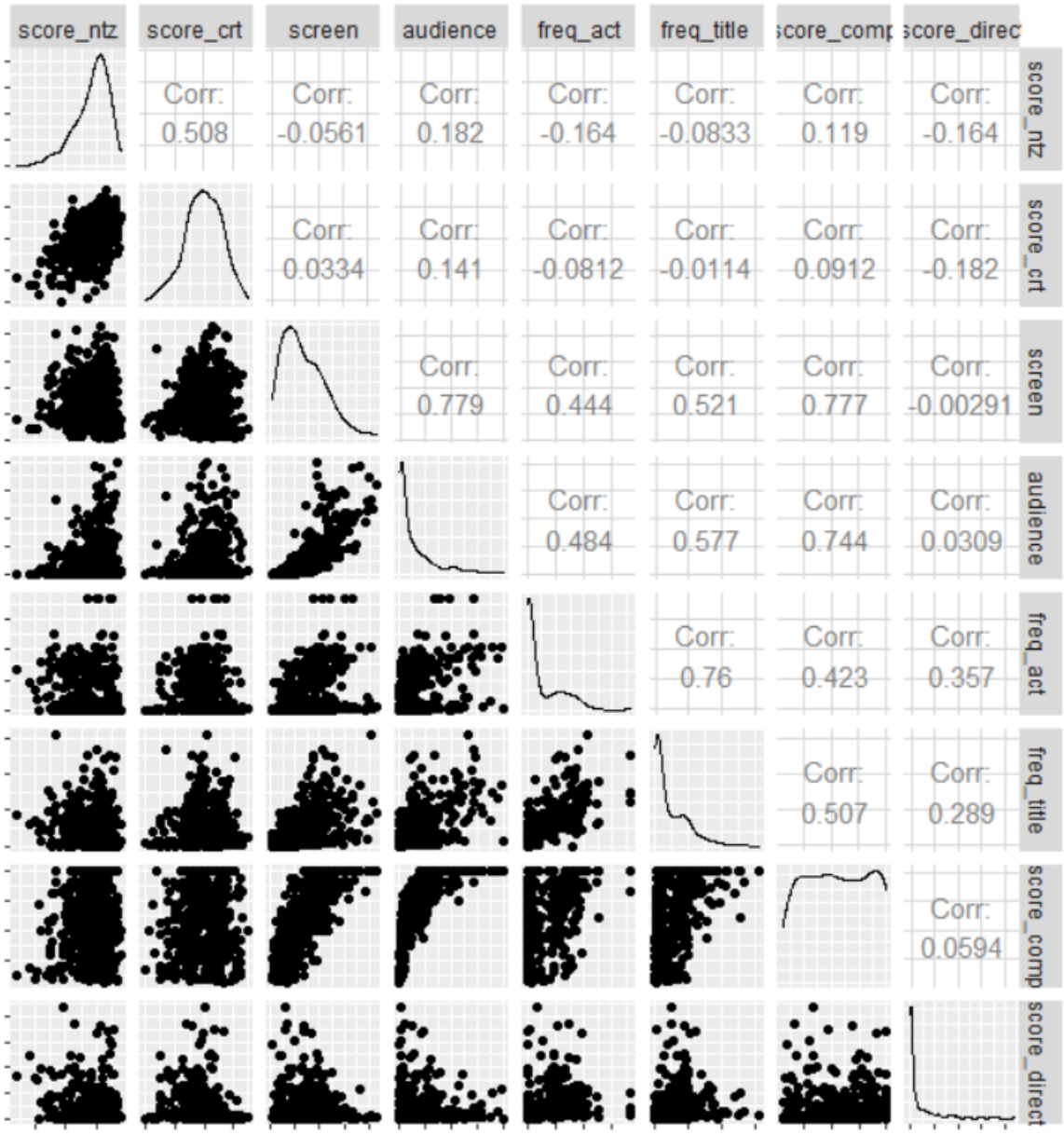
분석에 사용된 변수를 보면 '평점', '경쟁 점수' 등 개봉 하기 전에는 알 수 없는 변수들이 존재한다. 따라서 미개봉 영화에 대해서는 정확한 관객 수 예측을 하지 못 하고 이미 개봉이 되고 평가가 되어진 영화에 대해서 관객 수 확인이 가능하다. 예측 측면에서 적용하자면 정확한 값이 아닌 어떤 변수가 관객 수에 영향을 미칠 것이다 라는 추상적인 결론이 지어질 것이다.

예측 값과 많이 차이가 난 영화를 살펴보면 대표적으로 보헤미안 랩소디가 있는데 이 영화는 개봉전에도 많은 기대가 있었지만, 개봉 후 언론과 입소문을 타고 관객 수가 기하급수적으로 상승한 대표적인 영화이다. 따라서 본 분석의 독립 변수의 경우 영화 개봉 전에 얻을 수 있는 데이터를 가지고 관객 수를 예측하는 과정으로 이 경우에는 예외적으로 잘 적용이 되지 않는다.

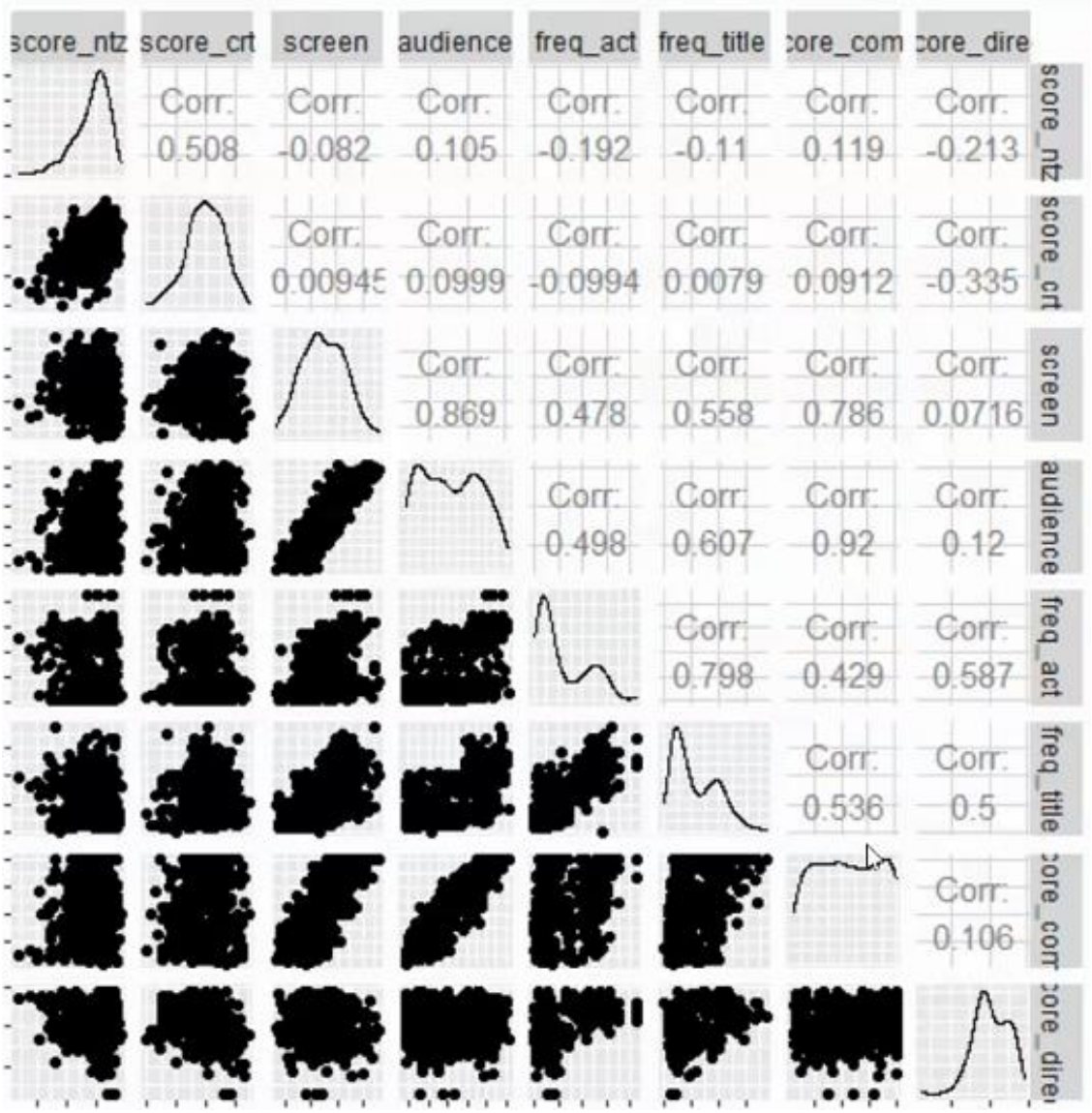
본 연구의 예측모형은 개봉 전 알 수 있는 영화의 특성만을 가지고 관객 수를 예측하는데 목적을 두고있다. 특히 기존에 독립 변수로 사용하기 어려웠던 다양한 변수들을 파생시켜 영화 관객 수를 예측하는데 있어 보다 우수함을 보이고 있다. 최근 잇따른 영화 산업의 부흥으로 많은 투자자들로 하여금 영화 산업을 고수익 투자 상품으로 보여지게 하고 있다. 하지만 규모 자체가 막대한 만큼 변동성에 대한 위험 부담 또한 필수불가결 적으로 동반된다. 단순히 영화의 작품성에 대해서 사람들이 판단하고 좋은 내용으로 만들어졌다면 많은 관객 수로 이끌것이라는 결론은 더 이상 적용시킬 수 없게 되었다. 이에 보다 분석적이고 데이터를 기반으로 한 근거있는 모형을 통해 예측 모형 개발을 만들었고 이는 보다 합리적이고 객관적인 의사결정의 한 잣대가 될 것이라 판단된다.

[부록]

(Appendix)



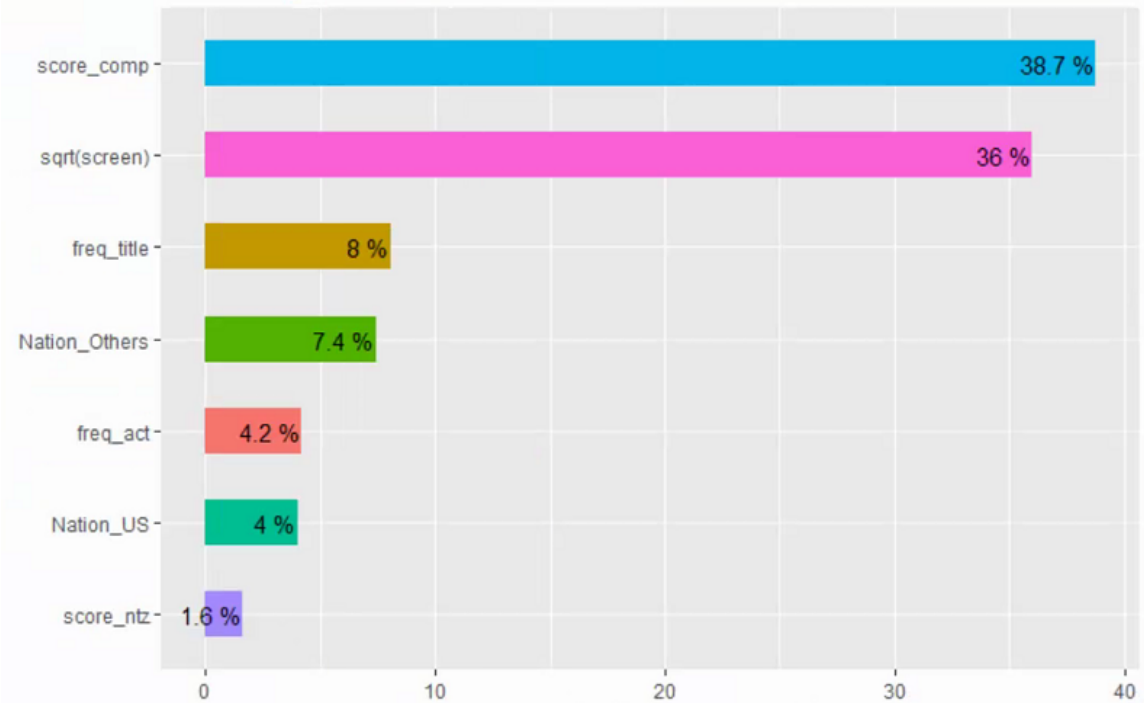
<부록 1> law data (ggpairs)



<부록 2> transformated data (ggpairs)

<부록 3> Result of Regression Analysis (1차)

변수 이름	Coefficients	Std.Error	t value	Pr(>ltl)	비고
(Intercept)	8.620	0.178	48.54	< 2e-16	* * *
Score_ntz	0.103	0.019	5.445	8.24e-08	* * *
Sqrt(screen)	0.065	0.004	15.163	< 2e-16	* * *
Sqrt(freq_title)	0.007	0.002	4.419	1.23e-05	* * *
Sqrt(freq_act)	0.001	0.001	2.793	0.005	* *
Nation_US	0.256	0.091	2.816	0.005	* *
Nation_Others	0.275	0.109	2.513	0.012	*
Score_comp	0.028	0.001	24.190	< 2e-16	* * *
R-squared	0.916	Adjusted R^2	0.9147	p-value	< 2.2e-16



<부록 4> Relative Importance of Predictor Variables