



Capstone Design 최종보고서					
1. 신청과제					
수행기간	2021학년도 1학기		교과목명	데이터분석모델링캡스톤디자인	
과제명	코로나 19 이후 다양한 사회 요인과 극장가 매출 간의 연관성 연구 및 매출 예측 모델 개발				
팀명	일석삼조		신청예산	총 1,200,000 원	
지도교수(학과)	정보통계학과		지도교수(성명)	박대우	
2. 참여학생(최소 2인 이상)					
구분	역할	성명	전공	학년	학번
팀장	자료수집, 시각화, 분석, 보고서 작성	임성수	정보통계학과	3	2019251169
팀원1	자료수집, 시각화, 분석, 보고서 작성	박영현	정보통계학과	3	2019251144
팀원2	자료수집, 시각화, 분석, 보고서 작성	이승현	정보통계학과	3	2019251124
팀원3	자료수집, 시각화, 분석, 보고서 작성	황성아	정보통계학과	3	2019251126
3. 과제타입(택1)					
<input checked="" type="checkbox"/> 일반형		<input type="checkbox"/> 융합형	<input type="checkbox"/> L2M 인재양성형	<input type="checkbox"/> 기업연계형	<input type="checkbox"/> 지역사회기여형
4. 결과물 종류(택1)					
유형	<input type="checkbox"/> 시제품 및 결과모형		<input type="checkbox"/> 하드웨어	<input type="checkbox"/> 기타()	
무형	<input type="checkbox"/> 인쇄물 및 영상		<input type="checkbox"/> 설계도면	<input checked="" type="checkbox"/> 보고서(조사, 분석결과 등)	
	<input type="checkbox"/> 소프트웨어(어플리케이션, 웹페이지 등)		<input type="checkbox"/> 기타 ()		
5. 참여업체(기업연계형만)					
기업(관)명		사업자등록번호		담당자명	
<p style="text-align: center;">위와 같이 Capstone Design 과제 최종보고서를 제출합니다.</p> <p style="text-align: center;">2021년 6월 14일</p> <div style="display: flex; justify-content: space-around; margin-top: 20px;"> <div style="text-align: center;"> <p>(대표학생)</p> <p>임성수</p> <p>(인도인서명)</p> </div> <div style="text-align: center;"> <p>(지도교수)</p> <p>박대우</p> <p>(인도인서명)</p> </div> </div>					
		연세대학교 원주LINC+사업단장 귀하			

Capstone Design 과제요약서

1. 예산집행결과 (홈페이지 확인 후 작성)

지원금 집행내역 (단위 : 원)				
지출항목	신청 예산	집행 액	잔액	비고
물품 · 제작비	960,000원		960,000원	
회의비	240,000원	158,900원	81,100원	
교통비				
멘토비				
합 계	1,200,000원	158,900원	1,041,100원	

2. 수행 과제

2.1 과제(작품)수행 개요

국내 코로나 19의 대유행이 본격화되며 사회 각 분야의 전반적인 경제적 피해가 잇따르는 가운데 정부는 생계의 어려움을 겪는 사업체를 지원하기 위한 각종 정책을 발표하였다. 이로 인해 일시적인 회복의 추세를 보이는 분야도 있지만 문화예술계와 같이 그렇지 못한 분야도 존재한다. 그 안에서도 중심 역할을 하는 영화계는 피해 실태에 대한 조사 및 정부의 지원 부재 그리고 사회적 거리두기 지침 강화에 따른 위축에 의해 빠르게 무너지고 있다.

지난 1년간 감소한 영화 산업의 총매출액은 69.2%로 이와 같은 피해 수치는 문화예술계 주산업의 현시점을 보여주며 장기화될 것으로 예측되는 코로나 19 팬데믹 상황에 맞추어 대책을 마련해야 함을 알리고 있다. 따라서 본 연구는 코로나19 발생 후 1년간 극장 간의 매출이 다양한 요인들과 어떠한 연관성을 보이는지 살피고 이를 바탕으로 도출된 전략에 부합하는 영화들의 매출을 예측하는 모형을 수립한다.

이번 연구를 통해 어떠한 요인들이 영화 매출에 영향을 미치는지 확인하게 되면 향후 여러 요인들을 그에 맞춰 조정하는데 활용할 수 있을 것이다. 또한 올바른 전략을 적용할 시 변화할 매출을 예측해봄으로써 새로운 영화 제작을 준비하는 사람들이 이후 상황을 고려해 제작 방향을 결정하는 데에 도움을 줄 것으로 예상된다.

3. 과제(작품)의 개발과정 및 역할

3.1 과제(작품)의 개발과정

✓ 자료 수집

본 과제의 최종 목표는 데이터를 통해 다양한 사회 요인과 매출 간의 연관성을 시각화 하고 데이터 분석 모형을 통해 코로나19 사태가 장기화되는 상황에서 문화예술계 분야의 회복 추세를 높이는 전략적 방안을 수립하는데 있다. 본 과제에서는 영화에 대한 정보와 코로나 확진자수의 정보를 독립적으로 수집하였다. 두 자료는 다음과 같이 구성된다.

- 자료1) 영화에 대한 정보는 영화관 입장권 통합전산망(KOBIS)에서 수집되었으며 주요 변수 들은 다음과 같다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	■ 개봉일만 월별																			
2	- 조회일: 2021-03-18																			
3	- 출처: 영화진흥위원회 통합전산망 (http://www.kobis.or.kr)																			
4	[검색조건] [조회기간 : 2020~2020] 국적: 전체 영화명: 전체 개봉일: 전체 영화구분: 전체 영화유형: 개봉영화:]																			
5																				
6	순번	영화명	감독	제작사	수입사	배급사	개봉일	영화유형	영화형태	국적	전국 스크린수	전국 매출액	전국 관객수	서울 매출액	서울 관객수	장르	등급	영화구분		
7	1	다만 약에서 구하소	홍원찬	(주)하이브미디어코프	씨제이이엔씨		2020-08-05	개봉영화	장편	한국	1,998	37,721,967,250	4,255,708	8,690,416,230	958,007	범죄	15세이상	일반영화		
8	2	반도	연상호	(주)영화사드림피터	(주)넥스트엔터		2020-07-15	개봉영화	장편	한국	2,575	34,059,520,180	5,810,212	7,996,305,840	895,123	액션	15세이상	일반영화		
9	3	#살아있다	조일형	영화사 집(주)퍼스펙티브	롯데컬처웍스		2020-06-24	개봉영화	장편	한국	1,882	15,965,329,900	1,903,703	3,885,303,840	476,332	드라마	15세이상	일반영화		
10	4	감칠비2: 정상회담	안우석	(주)스튜디오게니우스	롯데컬처웍스		2020-07-29	개봉영화	장편	한국	2,137	14,628,206,710	1,780,503	3,451,321,810	431,963	드라마	15세이상	일반영화		
11	5	오케이 마담	이철하	영화사울(주)	사나메가박스중앙		2020-08-12	개봉영화	장편	한국	1,288	10,790,021,200	1,205,199	2,112,388,380	231,120	코미디	15세이상	일반영화		
12	6	결백	박성현	(주)이디오펜	소니픽처스엔		2020-06-10	개봉영화	장편	한국	1,112	7,641,029,340	865,683	1,694,630,700	189,284	드라마	15세이상	일반영화		
13	7	네티	크리스토퍼 놀란	워너브러더스	워너브러더스		2020-08-26	개봉영화	장편	미국	2,228	6,496,393,320	714,412	2,050,431,900	216,041	액션	12세이상	일반영화		
14	8	침입자	손원평	(주)비메이엔터테인먼트	(주)에이스메		2020-06-04	개봉영화	장편	한국	1,365	4,806,314,000	531,212	1,029,368,660	110,923	미스터리	15세이상	일반영화		
15	9	1917	센 데데스	씨제이이엔씨(주)	스마일이		2020-02-19	개봉영화	장편	미국	932	4,161,124,420	464,824	1,642,180,560	176,208	드라마	15세이상	일반영화		
16	10	인버저블	리 워널	유니버설픽처스	유니버설픽처		2020-02-26	개봉영화	장편	미국	776	4,042,102,580	459,170	1,239,852,600	136,977	공포(호러)	15세이상	일반영화		
17	11	운위드 단 하루의 7	만 스캔론	윌트비즈	윌트비즈		2020-06-17	개봉영화	장편	미국	1,138	3,581,335,060	416,739	1,108,737,520	122,318	에이메이	전체관람	일반영화		
18	12	프라운 이스케이프	정진영	(주)비메이엔터테인먼트	(주)에이스메		2020-05-06	개봉영화	장편	영국	328	1,848,297,600	217,056	556,477,740	63,837	아드벤처	12세이상	독립/예술영화		
19	13	사라진 시간	정진영	(주)비메이엔터테인먼트	(주)에이스메		2020-06-18	개봉영화	장편	한국	874	1,651,188,580	185,653	388,762,580	43,314	미스터리	15세이상	일반영화		
20	14	발설: 세상을 바꾼 8	제이 로치	그린라레미디	(주)홀츠이스		2020-07-08	개봉영화	장편	미국	882	1,518,428,340	178,963	668,060,220	77,431	드라마	15세이상	독립/예술영화		
21	15	지푸라기라도 잡고	김용훈	(주)비메이엔터테인먼트	메가박스중앙		2020-02-19	개봉영화	장편	한국	996	1,232,453,270	139,777	358,470,900	39,768	에이메이	청소년관람	일반영화		
22	16	토를: 월드 투어	윌트 도른	데이비트 P	스유니버설픽처스	유니버설픽처		2020-04-29	개봉영화	장편	미국	344	1,168,000,320	152,159	303,995,980	37,939	에이메이	전체관람	일반영화	
23	17	다크 워터스	토드 해인즈	씨제이이엔씨(주)	(주)이수C&E		2020-03-11	개봉영화	장편	미국	575	1,096,657,740	127,251	383,981,180	43,314	드라마	12세이상	독립/예술영화		
24	18	에니멀 크래커	토니 밴크로프트	주식회사 디	(주)이수C&E		2020-08-05	개봉영화	장편	미국	473	1,023,705,340	129,006	151,065,470	17,842	에이메이	전체관람	독립/예술영화		
25	19	언더워터	윌리엄 유벵크	윌트도니	윌트도니		2020-05-27	개봉영화	장편	미국	594	937,112,800	109,207	259,889,160	29,571	스릴러	15세이상	일반영화		
26	20	정직한 후보	장유정	(주)스필름(주)홍필름	(주)넥스트엔터		2020-02-12	개봉영화	장편	한국	1,187	828,566,090	109,263	232,992,660	30,341	코미디	12세이상	일반영화		

<그림 1> KOBIS(영화관 입장권 통합전산망) 자료 예시

- 자료2) 코로나19 데이터(서울)에 대한 정보는 서울 열린 데이터 광장에서 제공하는 공공데이터로부터 수집되었다.

연번	확진일	환
1	33441	2021-04-07
2	33440	2021-04-07
3	33439	2021-04-07
4	33438	2021-04-07
5	33437	2021-04-07
6	33436	2021-04-07
7	33435	2021-04-07
8	33434	2021-04-07
9	33433	2021-04-07
10	33432	2021-04-07
11	33431	2021-04-07
12	33430	2021-04-07
13	33429	2021-04-07
14	33428	2021-04-07
15	33427	2021-04-07
16	33426	2021-04-07
17	33425	2021-04-07
18	33424	2021-04-07
19	33423	2021-04-07
20	33422	2021-04-07
21	33421	2021-04-07
22	33420	2021-04-07
23	33419	2021-04-07
24	33418	2021-04-07
25	33417	2021-04-07
26	33416	2021-04-07
27	33415	2021-04-07
28	33414	2021-04-07
29	33413	2021-04-07
30	33412	2021-04-07
31	33411	2021-04-07
32	33410	2021-04-07
33	33409	2021-04-07

<그림 2> 코로나19 정보

✓ 전처리

데이터분석에 앞서 수집된 영화 데이터의 전처리를 진행하였다. 수행된 전처리 과정은 다음과 같다.

- 1) 전국 매출액 50만원 이하의 수익을 거둔 영화들은 무의미한 영향을 끼칠 것으로 판단하여 해당 데이터는 제거한다.
- 2) 이전 데이터에서 중요하지 않은 변수들을 제거하고 중요한 변수들만 가려내기 위해 감독, 수입사, 영화형태를 삭제한다.
- 3) 전국 스크린 수, 전국 관객 수의 단위를 조정한다. 100개, 100명 단위로, 소수점 자릿수를 지정하고 100명 이하, 100개 이하의 단위는 0으로 지정한다.

- 4) 영화제작국가는 범주의 수준 수를 줄이기 위해 한국을 제외하고 대륙별로 구분한다.

<대륙별 국가들 목록>

- 한국
- 아시아: 대만, 우즈베키스탄, 일본, 중국, 태국, 터키, 필리핀, 홍콩
- 유럽: 네덜란드, 노르웨이, 덴마크, 독일, 러시아, 불가리아, 스위스, 스페인, 슬로바키아, 아일랜드, 영국, 우크라이나, 이탈리아, 크로아티아, 포르투갈, 폴란드, 프랑스, 핀란드, 헝가리
- 북아메리카: 미국, 캐나다
- 남아메리카: 브라질
- 오스트레일리아: 호주

- 5) 2019년, 관객점유율 기준 영화배급사 순위에 따라 배급사 자료 분류해 순서형 자료로 나타낸다.

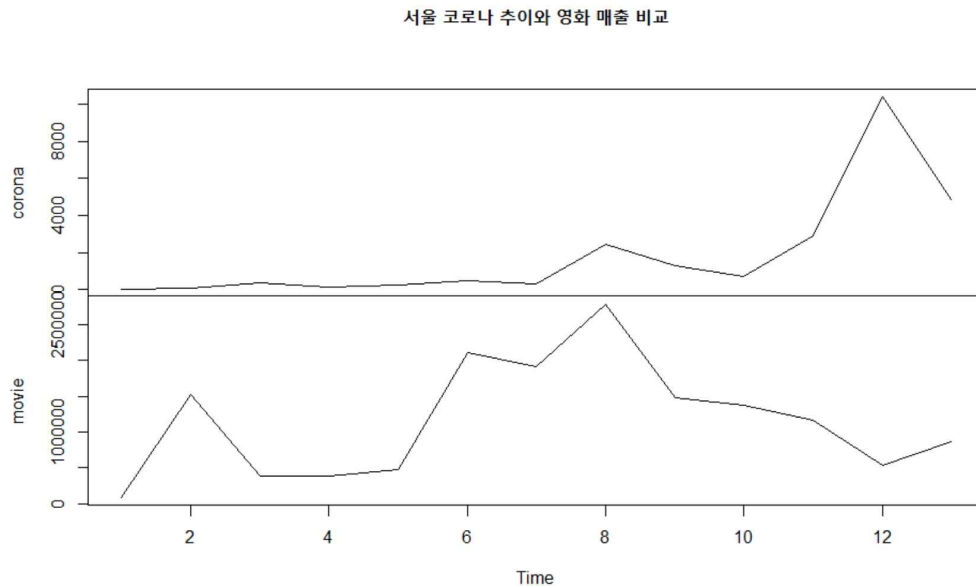
<배급사 분류 기준>

- 1위~5위: 대형 배급사로 2를 부여한다.
1위. 월트디즈니 컴퍼니코리아 - 30.2%
2위. CJ ENM 관련계열 - 28%
3위. 롯데컬처웍스 - 7.5%
4위. 쇼박스 - 5.1%
5위. 넥스트엔터테인먼트월드(NEW) - 4.8%
- 6위~10위: 중형 배급사로 1을 부여한다. (소니 메가박스 추가)
6위. 워너브라더스코리아 - 3.9%
7위. 이십세기폭스코리아 - 3.4%
8위. 유니버설픽처스인터내셔널 코리아 - 3.2%
9위. 메리크리스마스 - 1.8%
10위. 에이스메이커무비웍스 - 1.5%
- 나머지 순위는 0으로 부여한다.

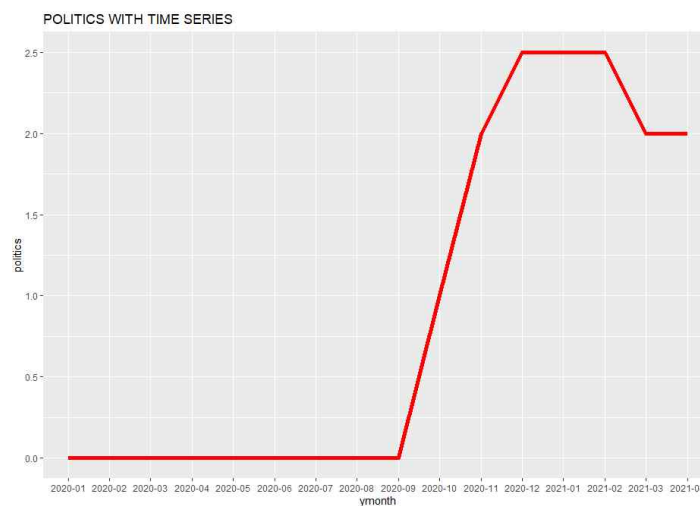
6) 매출액을 종속변수(response variable)로 설정한다.

✓ 탐색적 자료 분석 및 시각화

서울시 코로나19 상황에 따른 서울시 영화 매출의 변화를 알아보기 위하여 우선적으로 코로나19 추이와 서울시 영화 매출 2020년 01월 ~ 2021년 01월 자료를 '월' 기준으로 그래프로 나타내었다. 여기서 x축은 해당 숫자의 '월'을 의미한다.



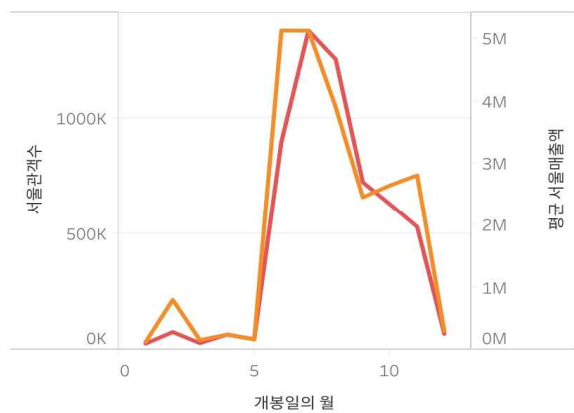
<그림 1> 서울시 월별 코로나19 추이와 영화 매출 비교



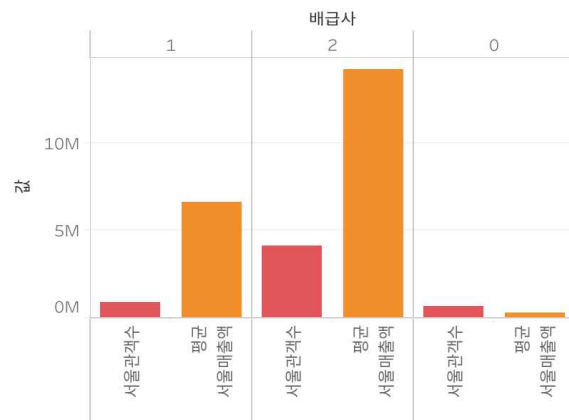
<그림 2> 서울시 정책 - 사회적 거리두기 단계 그래프

코로나19가 발생한지 얼마 되지 않은 2020년 1월부터 2020년 5월까지는 2020년 2월에 반짝 서울시 영화 매출이 오르기는 했지만 전체적으로 코로나 19 확진수가 상대적으로 적음에도 불구하고 코로나19에 대한 사람들의 불안감으로 인해 매출이 낮은 수치를 기록했다.

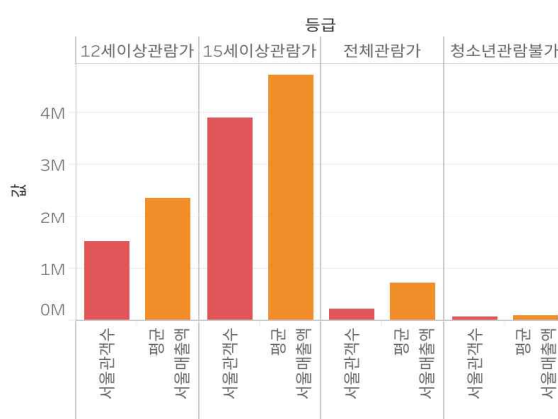
2020년 5월부터 2020년 8월까지의 코로나19에 익숙해져 상대적으로 매출이 급등한 것을 볼 수 있다. 하지만 2020년 8월에 코로나 확진자 수도 함께 증가하였고, 이로 인해 2020년 8월을 기점으로 매출도 함께 하락했음을 알 수 있다. 또한 2020년 11월부터는 코로나19 확진자가 급속도로 증가하며 다시 매출이 감소한 걸 볼 수 있다. 이로 보아 월별 코로나19 확진자 수가 약 4000명 미만인 부근에서는 영화 매출액과 코로나 확진자수가 약간의 비례관계를, 월별 코로나19 확진자 수가 약 4000명 이상인 부근에서는 영화 매출액과 코로나 확진자수가 반비례 관계를 가지는 것으로 보여진다. 이는 코로나19 확진자 수가 너무 많을 때에는 <그림2>와 같이 사회적 거리두기 단계가 높아짐에 따라 영화관에 대한 규제 등에 의해 영화관에 가는 사람들이 줄어들지만, 반대로 코로나 19 확진자 수가 약간 많을 시에는 영화관에 가는 등 사람들의 이동이 많아지면서 코로나 19 확진자 수가 늘어난 것이라고 유추해 볼 수 있다.



<그림 3> 개봉일(월)과의 관계



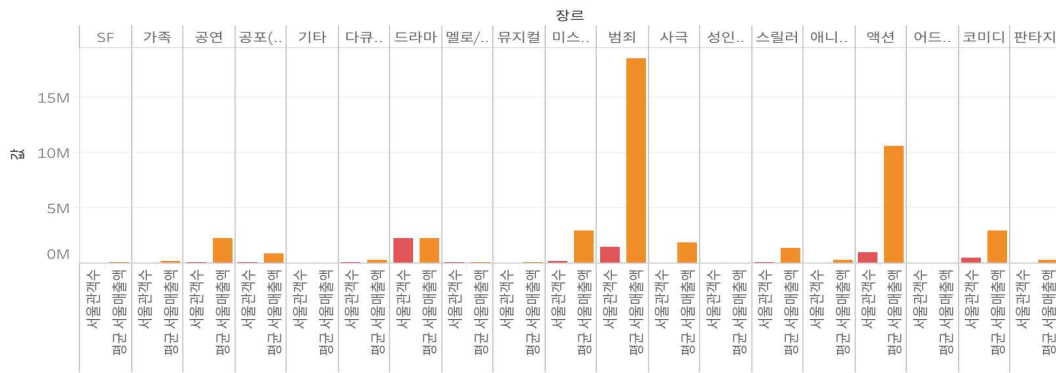
<그림 4> 배급사와의 관계



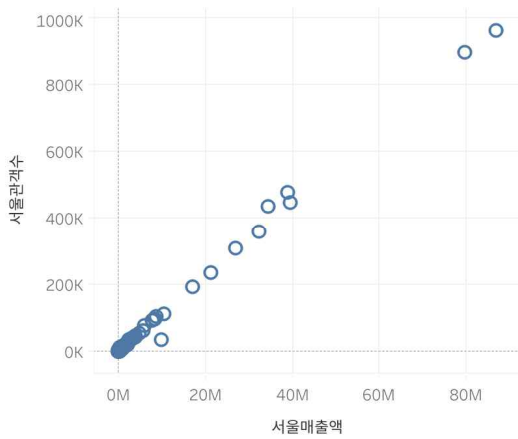
<그림 5> 등급과의 관계



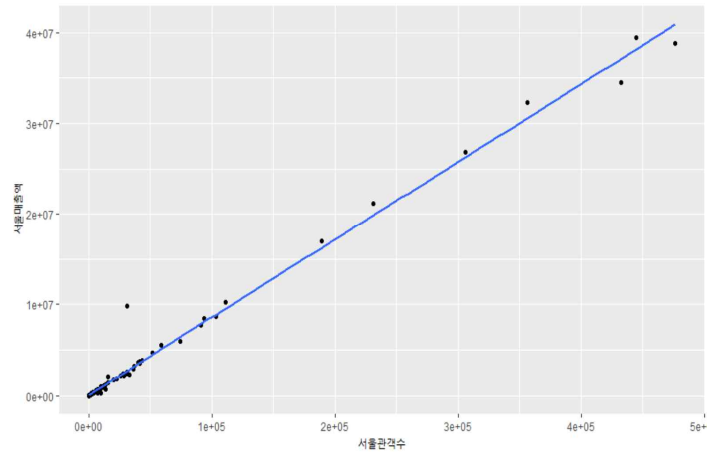
<그림 6> 영화구분과의 관계



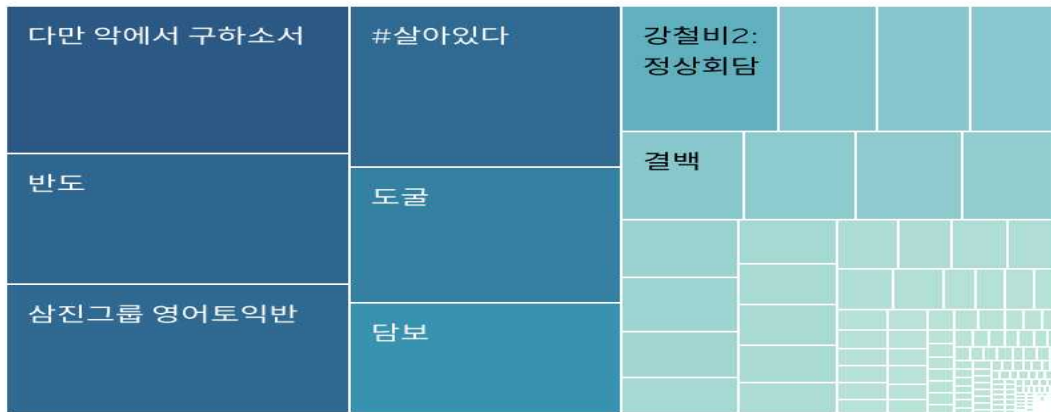
<그림 7> 장르와의 관계



<그림 8> 서울관객수와의 관계



<그림 9> 서울매출액과 서울관객수의 산점도 및 추정 회귀선



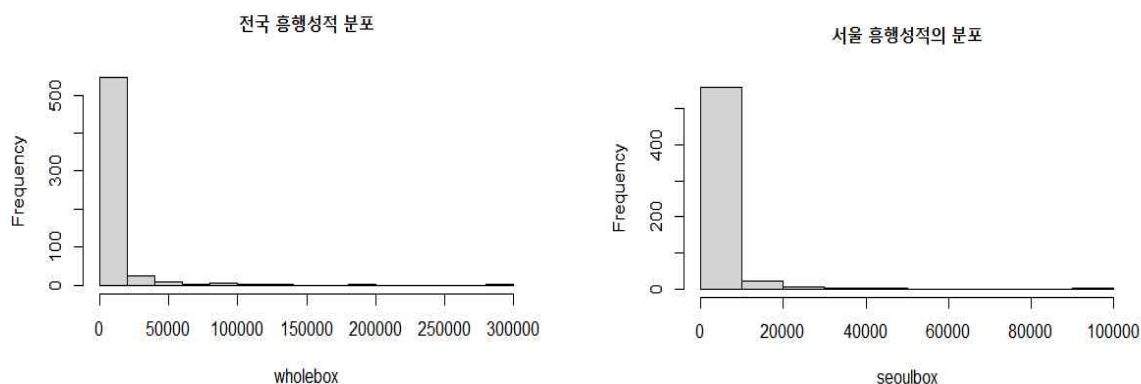
<그림 10> 서울매출액과 영화명

코로나19 상황에서 여러 요인과 매출을 비교하기 위해 평균 서울 매출액과 나머지 변수들의 관계를 통계 소프트웨어 R로 구현한 후, tableau를 통해 시각화 대시보드를 나타내었다. <그림 3>의 개봉일(월)과의 관계를 살펴보면 평균 서울 매출액은 5월에 급증하여 6~8월쯤 최고를 찍고 12~5월의 매출액이 가장 적은 것을 알 수 있다. <그림 4>의 배급사와의 관계를 살펴보면 평균 서울 매출액은 대형배급사>중형배급사>소형배급사 순으로 많은 것을 알 수 있으며, 소형 배급사에 비해 대형배급사의 평균 서울 매출액이 압도적임을 알 수 있다. <그림 5>의 등급과의 관계를 살펴보면 15세이상관람가 > 12세이상관람가 > 전체관람가 > 청소년관

람불가 순으로 서울 평균 매출액이 높은 것을 알 수 있다. 또한, 15세이상, 12세이상관람가가 거의 대부분의 매출을 차지하는 것도 알 수 있다. <그림 6>의 영화구분과의 관계를 살펴보면 독립/예술영화에 비해 일반영화의 서울 평균 매출이 압도적으로 높았다. <그림 7>의 장르와의 관계를 살펴보면 범죄와, 액션 순으로 영화의 매출이 가장 높다. 그리고 나머지는 공연, 공포, 드라마, 미스터리, 사극, 스릴러, 코미디에 비교적 고르게 분포되어 있다. <그림 8>은 서울 관객수와 서울 매출액은 결정계수가 1에 가까우므로 선형관계를 보여준다. 여기서 선형 회귀식은 $\text{서울매출액} = -1916.9 + 88.6 * (\text{서울관객수})$ 으로 구해진다. <그림 10>은 가지고 있는 데이터 중 서울 매출액이 높은 영화의 몇 가지의 순위를 보여준다. <그림 9>은 매출액과 관객수가 가장 큰 2개의 데이터는 영향점으로 파악이 되어 나머지 변수들의 관계를 제대로 보기 위해 이를 제외하고 나타낸 그래프이다.

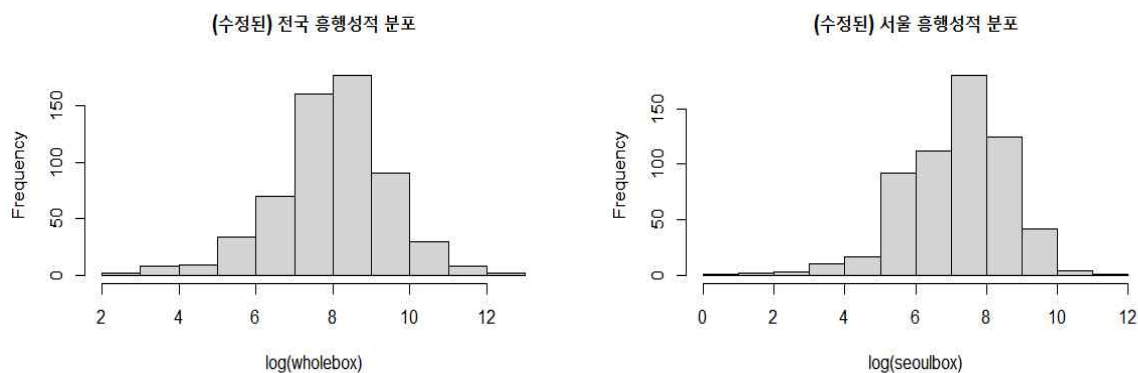
실제적인 흥행 정도를 파악하기 위하여 흥행성적과 기타 변수들 간의 연관성 분석을 위해 영화 관련 자료에서 (매출액/전국스크린수)로 '흥행성적' 변수를 생성하였다. 또한 매출액이 전국과 서울로 나누어진다는 점을 고려하여 이를 각각 '전국흥행성적', '서울흥행성적'으로 구분해 분석을 진행하였다. 전국스크린수가 10개 미만인 영화에서 상대적으로 스크린 수가 많은 영화보다 흥행성적에 있어 과장된 결과가 나타남을 확인할 수 있었다. 따라서 본 분석은 전국스크린 수가 10개 미만인 데이터를 제외하고 진행되었다.

전국흥행성적과 서울흥행성적의 분포를 히스토그램으로 표현한 결과는 아래의 <그림 11>, <그림 12>와 같다. 두 그래프 모두 극단적인 우향 왜곡을 띄고 있으므로 로그 변환을 통해 다음과 같이 수정된 그래프 <그림 13>, <그림 14>를 얻었다.



<그림 11> 전국 흥행성적 히스토그램

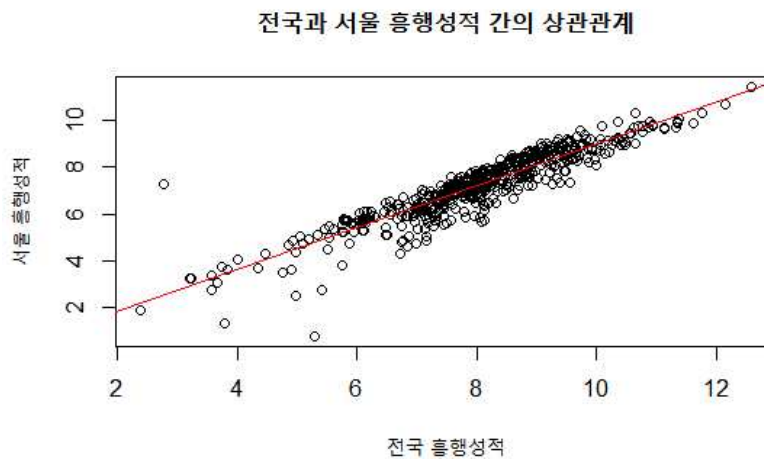
<그림 12> 서울 흥행성적 히스토그램



<그림 13> 수정된 전국 흥행성적 히스토그램

<그림 14> 수정된 서울 흥행성적 히스토그램

위의 결과를 토대로 전국흥행성적과 서울흥행성적 간의 유의한 관련성이 있는지 확인하고자 산점도를 그린 후 선형 관련성이 보여 회귀분석을 진행해보았다. 추정 회귀식은 서울 흥행성적 $= \beta_0 + \beta_1 \times \text{전국흥행성적}$ 이며, $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$ 검정을 진행한 결과 유의확률이 0에 가까워 유의수준 5% 하에서 귀무가설을 기각하였다. 또한 결정계수가 0.8495로 측정되어 둘 사이에 유의한 선형 관련성이 있음을 확인하였다. 따라서 향후 서울 데이터만을 가지고 전국 흥행성적을 추론하는데 이 결과를 활용할 수 있을 것이다. 산점도에 추정된 회귀직선(서울흥행성적 $= 0.06496 + 0.89360 \times \text{전국흥행성적}$)을 표시한 결과는 <그림 15>와 같다.



<그림 15> 전국과 서울 흥행성적의 산점도 및 추정회귀직선

배급사와 흥행성적 간의 연관성을 살피기 위해 회귀분석을 진행했고 검정 결과 전국과 서울 흥행성적 모두 배급사와 유의한 선형 관계에 있다고 판단했다. 추정된 선형회귀식은 각각 <수식 1>, <수식 2>와 같다. 그러나 결정계수가 각각 0.1182, 0.0678로 작으므로 향후 다른 연구에 활용 시에 배급사의 구분을 세분화하거나 더 많은 양의 자료를 사용하면 설명력을 높일 수 있을 것으로 기대된다. 직관적인 이해를 돕기 위해 tableau의 분석 기능을 활용한 결과는 <그림 16>과 같다.

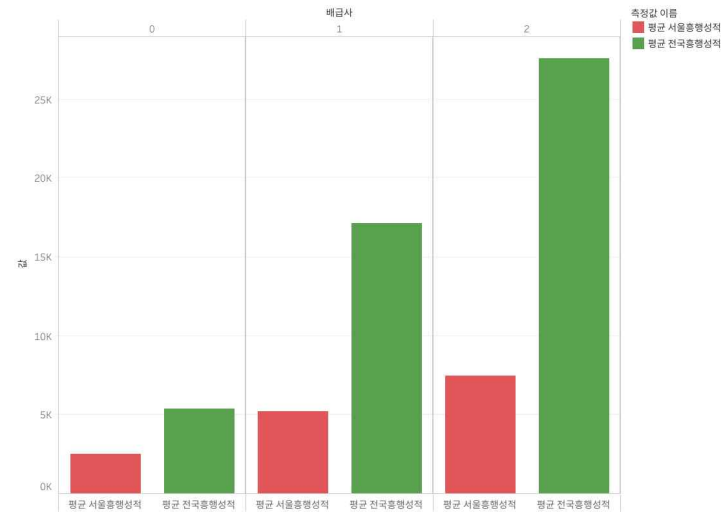
$$y(\text{전국흥행성적}) = 5396.7 + 11206.9 \times x(\text{배급사}) \quad (\text{단}, x = 0, 1, 2)$$

<수식 1> 배급사와 전국흥행성적 간의 추정 선형회귀식

$$y(\text{서울흥행성적}) = 2508.2 + 2525.7 \times x(\text{배급사}) \quad (\text{단}, x = 0, 1, 2)$$

<수식 2> 배급사와 서울흥행성적 간의 추정 선형회귀식

배급사별 전국, 서울 흥행 성적

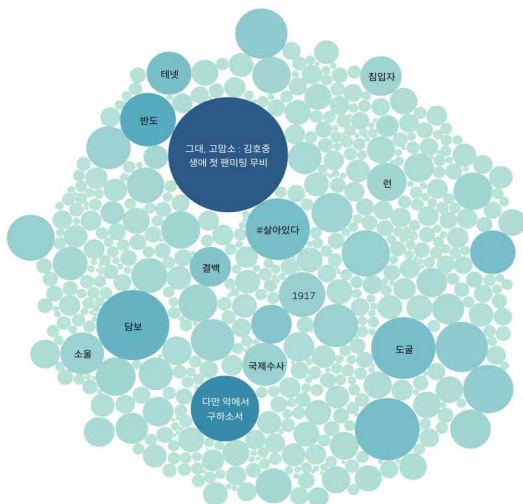


<그림 16> 배급사별 전국, 서울 흥행 성적

배급사별 평균 전국, 서울 흥행성적을 막대그래프로 나타낸 결과, 배급사의 규모가 커질수록 흥행 성적이 높았음을 알 수 있다. 특히 평균 전국 흥행 성적은 평균 서울 흥행 성적과 비교했을 때 배급사가 올라갈수록 가파르게 증가하는 경향을 보였다. 따라서 배급사의 영향은 서울보다 전국 흥행 성적에 더 많은 영향을 보였음을 알 수 있다.

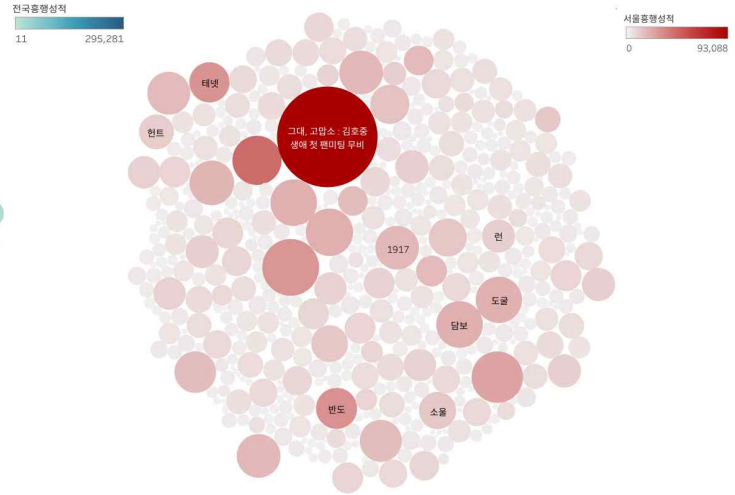
<그림 17>과 <그림 18>는 장르와 흥행 성적 간의 연관성을 알아보기 전에 한 해 동안 개봉한 영화 중 흥행 성적이 높았던 영화가 무엇인지 파악하기 위해 tableau를 활용하여 영화명과 흥행 성적 간의 관계를 간단히 시각화해본 결과이다. 전국과 서울 모두 「그대, 고맙소 : 김호중 생애 첫 팬미팅 무비」가 가장 높은 흥행 성적을 거두었고 그 뒤로 「반도」, 「테넷」과 같은 대규모 액션 장르와 「도굴」과 같은 범죄, 「런」과 같은 미스터리, 「1917」과 같은 드라마 장르가 뒤를 이었음을 확인할 수 있다.

영화명별 전국 흥행 성적



<그림 17> 전국 흥행성적으로 살펴본 흥행 영화

영화명별 서울 흥행 성적

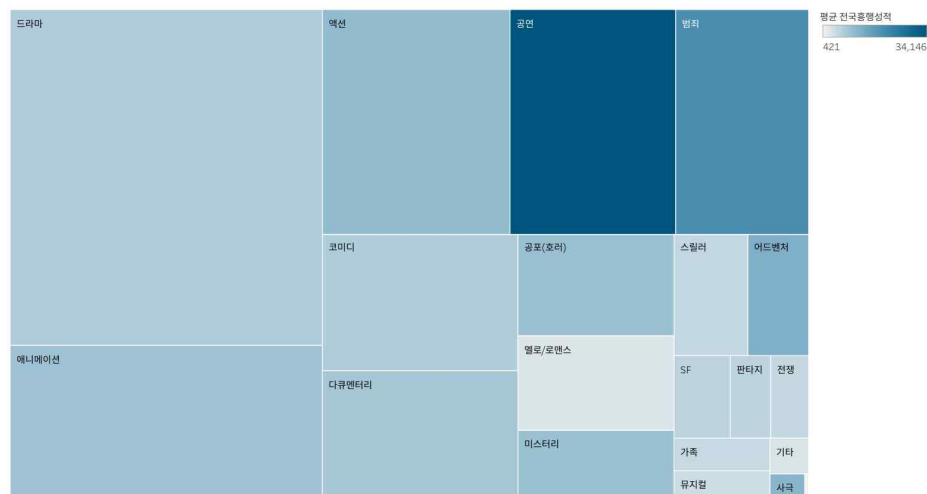


<그림 18> 서울 흥행성적으로 살펴본 흥행 영화

<그림 19>과 <그림 20>는 구체적으로 이 영화들의 장르와 흥행 성적의 연관성을 파악하기 위해 tableau로 활용해 나타낸 결과이다. 평균적으로 드라마 장르의 영화가 가장 많이 제작되었지만, 흥행 성적에서는 공연이 가장 높고 그 다음으로 범죄, 액션, 다큐멘터리, 애니메이션

선, 드라마 등이 흥행을 거두었다. 이것이 공연 장르 영화 전반의 흥행을 의미하는지 확인한 결과 상위 2개의 영화(「그대, 고맙소 : 김호중 생애 첫 팬미팅 무비」, 「미스터트롯: 더 무비」가 전체 흥행 성적의 87.35%를 차지하는 것으로 나타났다. 따라서 이는 공연 장르의 흥행성보다 지난 한 해 '트로트'라는 음악적 유행이 불러온 흥행으로 보는 것이 타당하다고 판단했다.

장르별 전국 흥행 성적



<그림 19> 장르별 전국 흥행 성적

장르별 서울 흥행 성적



<그림 20> 장르별 서울 흥행 성적

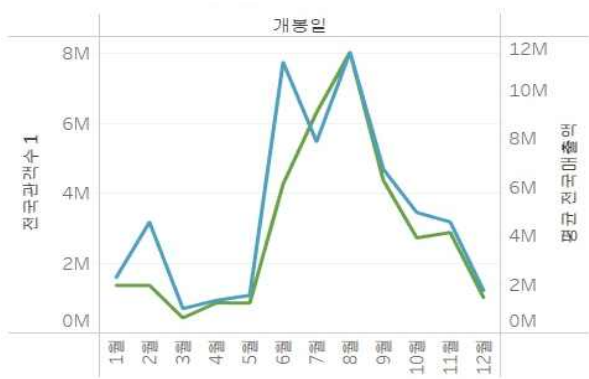
개봉일과 흥행 성적 간의 연관성을 살펴보기 위해 각 데이터의 개봉 월과 흥행 성적을 하나의 선형 그래프에 표현했다. 그 결과 두 그래프가 비슷한 형태로 그려졌다. 전국과 서울 모두 6월과 10월에 가장 높은 흥행 성적을 보였다. 따라서 서울 영화에 관한 데이터가 전국 영화 데이터로 확장할 수 있을 것으로 여겨진다. 전국 영화 매출액과 나머지 영화 관련 변수들의 관계는 다음과 같다 <그림 21>.

개봉월별 전국, 서울 흥행 성적

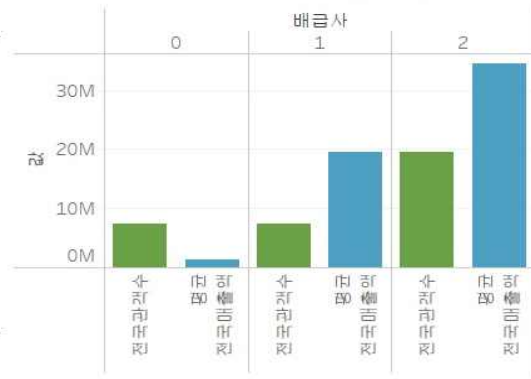


<그림 21> 개봉월별 전국, 서울 흥행 성적

<그림 22 - 31>은 평균 전국 매출액과 나머지 변수들의 관계를 시각화한 결과를 보여준다. <그림 22>는 개봉월과 평균 전국 매출액, 그리고 전국 관객수로 이뤄진 그래프이다. 전체적으로 전국 매출액과 관객수는 비슷한 양상을 보이며 5월에 관객수 및 매출액이 급증하여 8월에 고점을 찍고 계속 감소한다. <그림 23>에서 배급사와의 관계를 살펴보면 배급사가 대형(2), 중형(1), 소형(0) 순으로 전국 평균 매출액이 많은 것을 볼 수 있다. <그림 24> 등급과의 관계에서는 15세 이상 관람가, 12세 이상 관람가, 전체관람가, 청소년 관람불가 순으로 전국 관객수와 평균 전국 매출액 모두 많은 것을 알 수 있다. <그림 25> 영화 구분과의 관계는 일반영화가 관객수와 매출액에서 독립/예술 영화를 압도하는 것을 볼 수 있다. <그림 26> 장르에 따른 전국매출액 비교를 살펴보면 드라마, 액션, 범죄 순으로 매출액이 높은 것을 볼 수 있으며, 그 세 장르가 대부분의 매출액을 차지한다고 할 수 있다. <그림 27> 전국 관객수와 매출액 사이의 관계를 살펴보면 전국 매출액과 전국관객수는 선형 관련성을 띄는 것을 볼 수 있다. <그림 28>은 tableau를 사용하여 전국관객수를 원의 크기로, 매출액을 원의 색으로 시각화 한 것이다. '다만' 악에서 구하소서'와 '반도'가 크기와 색 모두 다른 영화들보다 크고 진한 것을 확인할 수 있다. <그림29>는 전국 매출액과 스크린수 간의 상관관계를 나타낸 그래프로 선형 관계로 나타내기 위해 양변에 로그 변환을 취한 그래프는 <그림30>과 같다. <그림31>은 전국 매출액과 전국 관객의 관계를 나타냈다. 양의 상관관계를 가지며 결정계수는 0.99로 선형 관계이다.



<그림 22> 개봉일(월)과의 관계



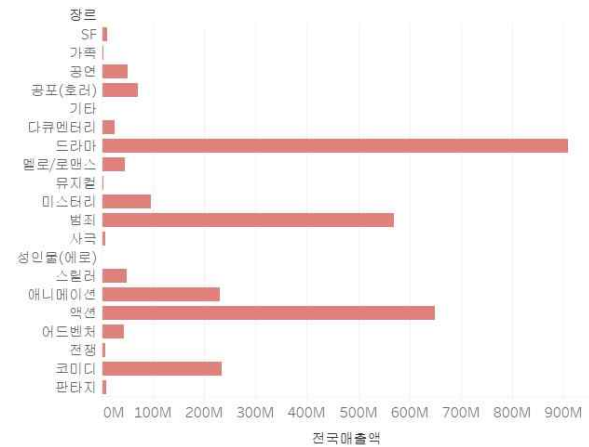
<그림 23> 배급사와의 관계(전국)



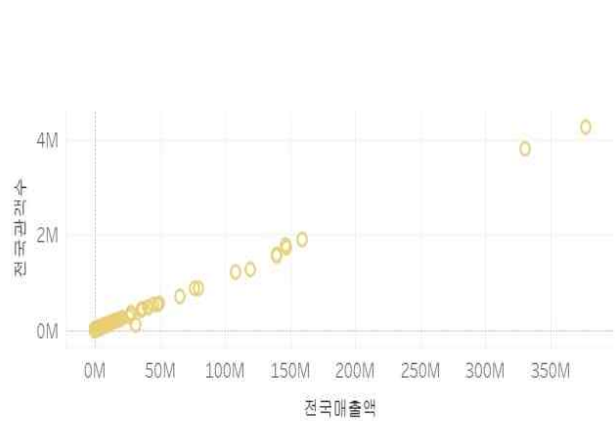
<그림 24> 등급과의 관계



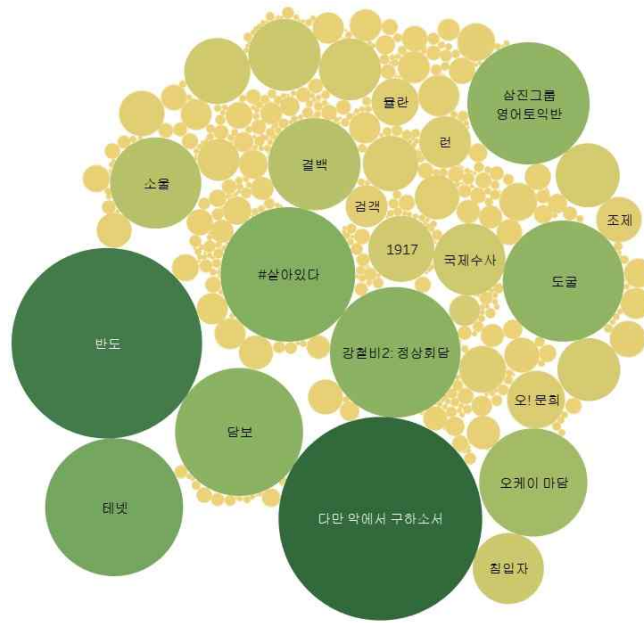
<그림 25> 영화 구분과의 관계



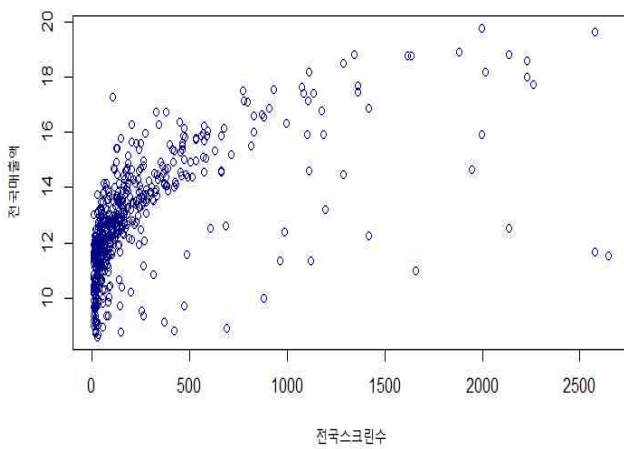
<그림 26> 장르에 따른 전국매출액 비교



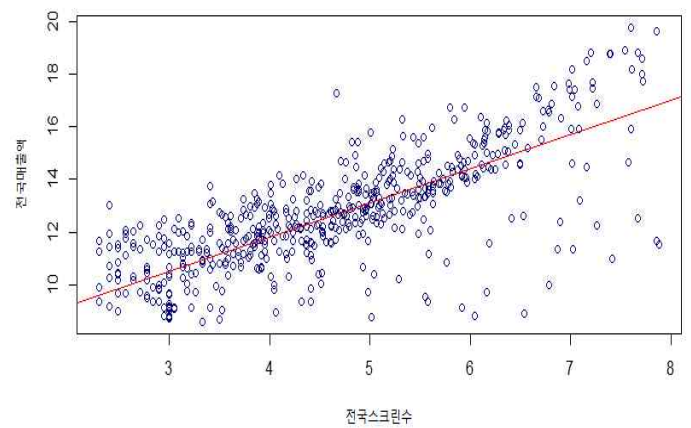
<그림 27> 전국 관객수와 매출액 사이의 관계



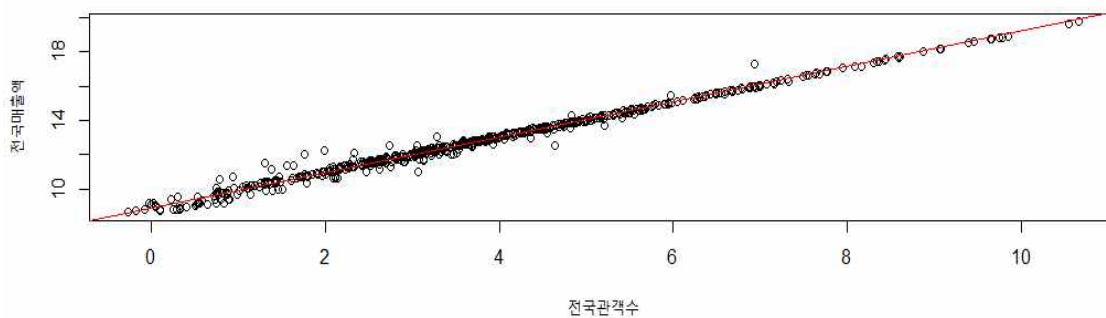
<그림 28> 전국 관객수와 매출액



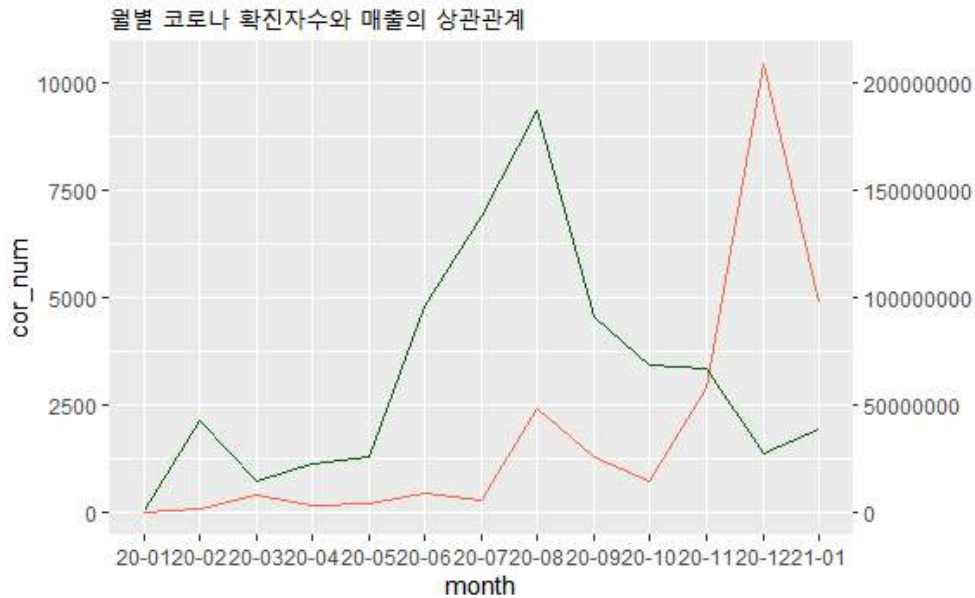
<그림 29> 전국매출액과 스크린수 간의 상관관계



<그림 30> (수정된)전국매출액과 스크린수 간의 상관관계



<그림 31> 전국 매출액과 전국 관객수



<그림 32> 서울 월별 코로나 확진자수와 전국 매출의 상관관계

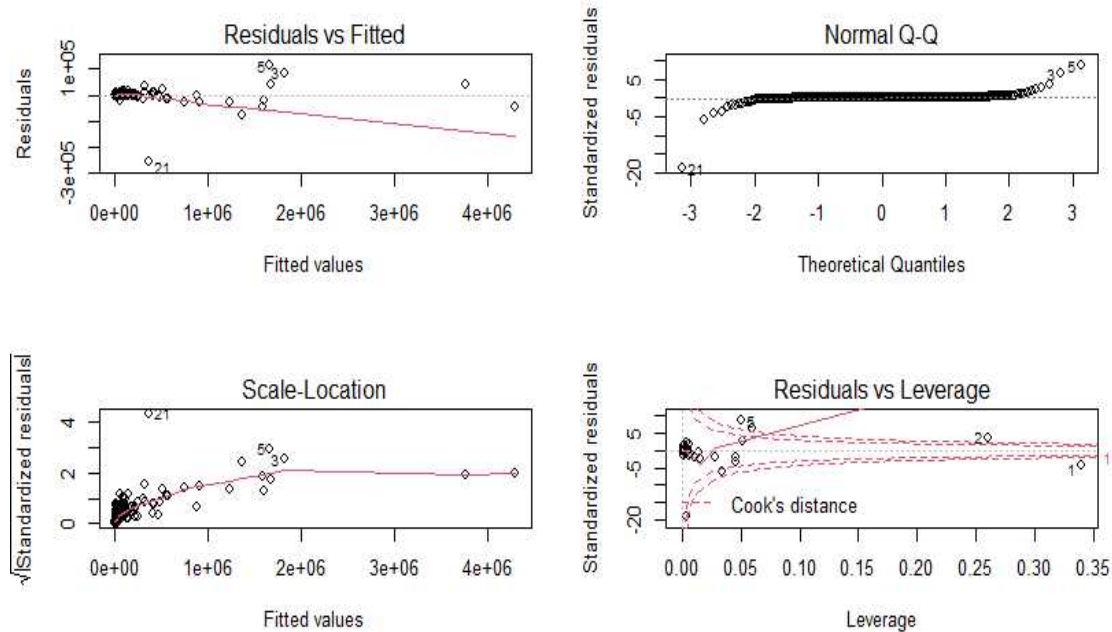
<그림32>에서 보여주는 서울시 월별 코로나19 확진자수와 전국 영화 매출은 <그림1>의 서울시 영화 매출과 거의 동일한 분포를 가지는 것으로 분석되었다. 따라서 추후의 분석과 전국 매출을 예측하는 모델을 구축하기 위해 서울 코로나 19 데이터와 전국 영화 매출에 관한 데이터의 비교로 확장하겠다.

✓ 단순회귀분석

회귀분석(regression analysis)은 수집된 자료의 분석을 통해서 추출된 모집단이 가진 특징이나 모집단의 현상을 추정·예견하는 통계적 추론기법이다. 회귀분석의 주된 목적은 독립변수와 종속변수의 연관성을 확인하는 데에 있다. 회귀분석은 고려되는 변수의 수에 따라 단순회귀분석, 다중회귀분석 등으로 구분할 수 있다. 본 연구에서 사용할 단순회귀분석은 독립변수 1개, 종속변수 1개로 두 변수의 관계를 선형관계로 가정하는 경우이다. 회귀분석을 이용하여 구축된 모형을 회귀모형이라고 한다.

$$\text{<수식3> } Y = \beta_0 + \beta_1 X + \epsilon$$

여기서 X 는 독립변수, Y 는 종속변수, β_0 과 β_1 는 회귀계수를 의미한다. 오차항 ϵ 은 평균이 0인 정규분포를 가정한다. 기울기에 해당하는 β_1 는 독립변수(X)가 종속변수(Y)에 미치는 영향을 나타낸다. 회귀분석에서 결정계수(R^2)의 값은 종속변수의 총 변동에 대한 독립변수들의 설명력의 크기를 나타내는 척도로서 0부터 1까지의 범위를 갖는다. R^2 의 값이 1에 가까울수록 독립변수에 의한 종속변수의 설명력이 크고, 회귀식의 적합도가 높다는 것을 의미한다.



<그림 33> 전국매출액과 전국관객수 회귀분석 등분산성 가정 확인

```
> # 3. 독립성 확인 (상한 이상 값 나올 경우, 독립성 확인)
> states %>% lm(전국관객수...12 ~ 전국매출액...9,..) %>% residuals %>% durbinwatsonTest
[1] 1.835261
> # 4. 정규성 확인 (p-value > 0.05 이면 정규성 가정 가능)
> states %>% lm(전국관객수...12 ~ 전국매출액...9,..) %>% residuals %>% shapiro.test

shapiro-wilk normality test

data: .
W = 0.18146, p-value < 2.2e-16
```

<그림 34> 전국매출액과 전국관객수 독립성, 정규성 가정 확인

회귀분석을 진행함에 앞서, 반드시 확인해야 하는 기본 가정 선형성, 등분산성, 독립성, 정규성 네 가지가 있다. 여러 변수 중 전국매출액과 전국관객수를 예시로 들어 기본 가정 성립 과정을 보여줄 것이다. 선형성은 독립변수와 종속변수의 산점도를 그려 파악하는 것으로, <그림27>에서 확인하였다. 등분산성은 독립변수와 잔차의 산점도를 그려 확인할 수 있으며, 이는 <그림33>을 보아 만족함을 알 수 있다. <그림34>에서 독립성은 잔차에 대한 Durbin watson 검정을 통해 상한 이상의 값이 나와 독립성이 확인되었다. 하지만 Shapiro-wilk normality test를 통한 정규성 검정에서는 정규성을 만족하지 못하는 것으로 나타났다. 여기서 분석의 편의를 위해 정규성을 가정하고 진행하도록 한다. 이는 다른 변수에서도 동일하게 적용된다.


```
> summary(lm(전국매출액...9~전국관객수...11, data=movie))

Call:
lm(formula = 전국매출액...9 ~ 전국관객수...11, data = movie)

Residuals:
    Min       1Q   Median       3Q      Max
-9609048  -45904   -7868    305 22370120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1742.72   49346.07  -0.035   0.972
전국관객수...11  8755.55     16.14 542.429 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1178000 on 589 degrees of freedom
Multiple R-squared:  0.998,    Adjusted R-squared:  0.998
F-statistic: 2.942e+05 on 1 and 589 DF, p-value: < 2.2e-16

> |
```

<그림 35> 전국매출액에 대한 전국관객수 회귀분석 결과

```
> summary(lm(전국매출액...9~전국스크린수, data=movie))

Call:
lm(formula = 전국매출액...9 ~ 전국스크린수, data = movie)

Residuals:
    Min       1Q   Median       3Q      Max
-98409448  -2323382  1654597   3320512 303916971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4190880   1005160  -4.169 3.51e-05 ***
전국스크린수   38786     2127  18.236 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21060000 on 589 degrees of freedom
Multiple R-squared:  0.3609,    Adjusted R-squared:  0.3598
F-statistic: 332.6 on 1 and 589 DF, p-value: < 2.2e-16

> summary(lm(전국매출액...9~장르, data=movie))

Call:
lm(formula = 전국매출액...9 ~ 장르, data = movie)

Residuals:
    Min       1Q   Median       3Q      Max
-31624083  -4956793  -3348971  -585053 345575180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1010377    8257360   0.122  0.90266
장르가속      -655110    12386040  -0.053  0.95784
장르공연      2880903    10983315   0.262  0.79319
장르공포(호러) 2560242    10113159   0.253  0.80024
장르기타      -959710    13484212  -0.071  0.94329
장르다큐멘터리 -334630     9280471  -0.036  0.97125
장르드라마     4068243     8484878   0.479  0.63179
장르멜로/로맨스 -352956     8827489  -0.040  0.96812
장르뮤지컬     -597871    11997666  -0.050  0.96027
장르미스터리    5916936    10811422   0.547  0.58440
장르범죄      30634116    10298737   2.975  0.00306 **
장르사극       5410830     27386564   0.198  0.84345
장르성인물(에로) -1001960    17189065  -0.058  0.95354
장르스릴러     1316354    10032575   0.131  0.89566
장르애니메이션 2599044     8879062   0.293  0.76985
장르액션     12251775     9060855   1.352  0.17686
장르어드벤처  5108364    12868166   0.397  0.69153
장르전쟁      -82351     12386040  -0.007  0.99470
장르코미디     4077679     9110796   0.448  0.65464
장르판타지     224479     12868166   0.017  0.98609
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26110000 on 571 degrees of freedom
Multiple R-squared:  0.04774,    Adjusted R-squared:  0.01605
F-statistic: 1.507 on 19 and 571 DF, p-value: 0.07704
```

<그림 36> 전국매출액에 대한 전국스크린수, 장르 회귀분석 결과

```
> summary(lm(전국매출액...9~등급, data=movie))

Call:
lm(formula = 전국매출액...9 ~ 등급, data = movie)

Residuals:
    Min       1Q   Median       3Q      Max
-7756337 -7152030 -4790672 -573260 369457331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6627419    2185979   3.032  0.00254 **
등급15세이상관람가  1134923    2811806   0.404  0.68663
등급전체관람가    -4232352    3227803  -1.311  0.19030
등급청소년관람불가 -6044944    3366287  -1.796  0.07305 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26230000 on 587 degrees of freedom
Multiple R-squared:  0.01206, Adjusted R-squared:  0.007011
F-statistic: 2.389 on 3 and 587 DF, p-value: 0.06791
```

<그림 37> 전국매출액에 대한 등급 회귀분석 결과

```
> summary(lm(전국매출액...9~배급사, data=movie))

Call:
lm(formula = 전국매출액...9 ~ 배급사, data = movie)

Residuals:
    Min       1Q   Median       3Q      Max
-34875082 -1267306 -1125186 -578905 342285546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1287853    1076810   1.196   0.232
배급사      16823137    1711295   9.831 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24420000 on 589 degrees of freedom
Multiple R-squared:  0.141, Adjusted R-squared:  0.1395
F-statistic: 96.64 on 1 and 589 DF, p-value: < 2.2e-16

> summary(lm(전국매출액...9~개봉일, data=movie))

Call:
lm(formula = 전국매출액...9 ~ 개봉일, data = movie)

Residuals:
    Min       1Q   Median       3Q      Max
 -6205710 -5226618 -4555065 -3772133 372090843

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.103e+08  1.835e+08  -0.601   0.548
개봉일       7.227e-02  1.150e-01   0.629   0.530

Residual standard error: 26340000 on 589 degrees of freedom
Multiple R-squared:  0.0006703, Adjusted R-squared: -0.001026
F-statistic: 0.3951 on 1 and 589 DF, p-value: 0.5299

> summary(lm(전국매출액...9~국적, data=movie))

Call:
lm(formula = 전국매출액...9 ~ 국적, data = movie)

Residuals:
    Min       1Q   Median       3Q      Max
 -9349696 -9042830 -3432974 -569278 367864466

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    536278    5859422   0.092   0.927
국적북아메리카  4380607    6266802   0.699   0.485
국적브라질     -136347    26851247  -0.005   0.996
국적아시아     827416    6535036   0.127   0.899
국적유럽       371543    6336557   0.059   0.953
국적한국       8818929    6115412   1.442   0.150
국적호주       369042    11507677   0.032   0.974

Residual standard error: 26200000 on 584 degrees of freedom
Multiple R-squared:  0.01918, Adjusted R-squared:  0.009101
```

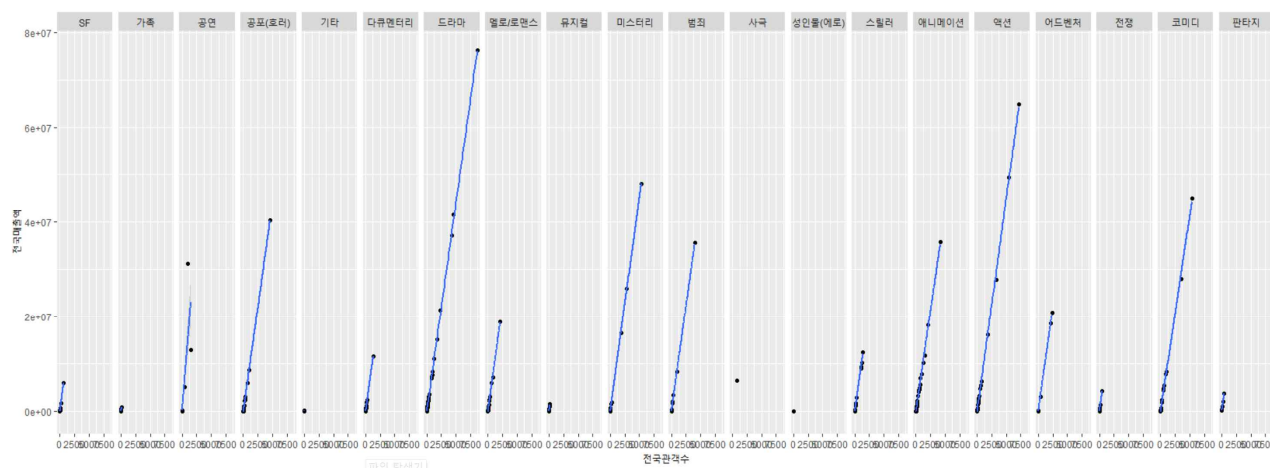
<그림 38> 전국매출액에 대한 배급사, 개봉일, 국적 회귀분석 결과

<그림35>부터 <그림38>까지는 전국매출액에 대한 각 영화 요인간의 유의성을 알아보기 위해 단순회귀분석이 진행된 결과창이다. 종속변수는 모두 전국 매출액으로 설정하였고, 독립변수는 각각 순서대로 전국 관객수, 전국스크린수, 장르, 등급, 배급사, 개봉일, 국적으로 설정

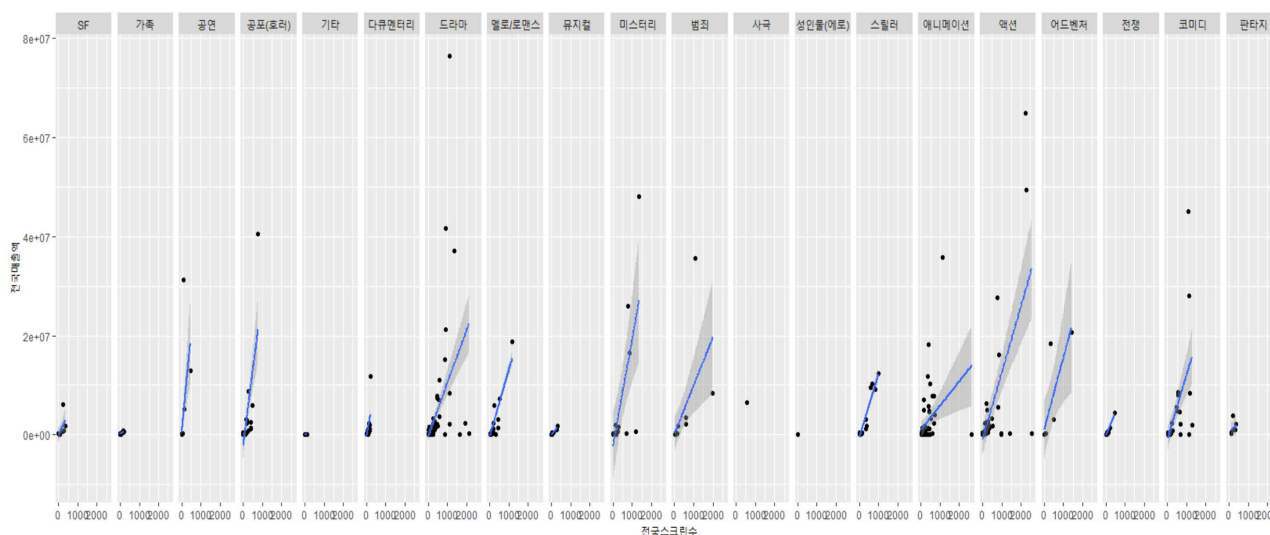
하였다. 여기서 단순회귀분석을 적용하기 위한 가정 4가지 선형성, 등분산성, 독립성, 정규성은 만족하는 것으로 가정한다.

단순회귀분석을 진행하고 각각의 p값을 확인해 본 결과 전국관객수, 전국스크린수, 장르: 범죄, 등급, 배급사, 국적이 ($p < 0.1$)로 유의수준 10% 하에서 유의함을 알 수 있다. 따라서 전국매출액에 영향을 주는 영화 요인은 전국관객수, 전국스크린수, 장르, 등급, 배급사, 국적으로 결정한다.

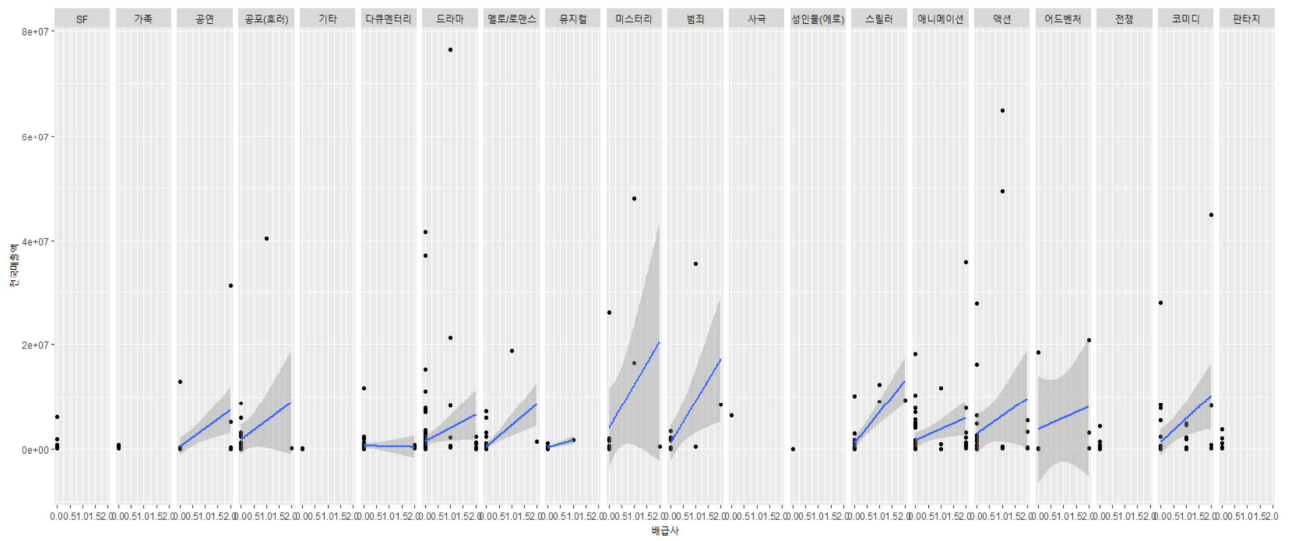
위에서 얻은 유의한 변수들을 바탕으로 관계를 쉽게 눈으로 파악하기 위해 그래프로 나타내고자 한다. 범주형 자료는 따로 분리하였고, 나머지 변수들은 전국매출액에 영향을 주지 않으므로 제외한다.



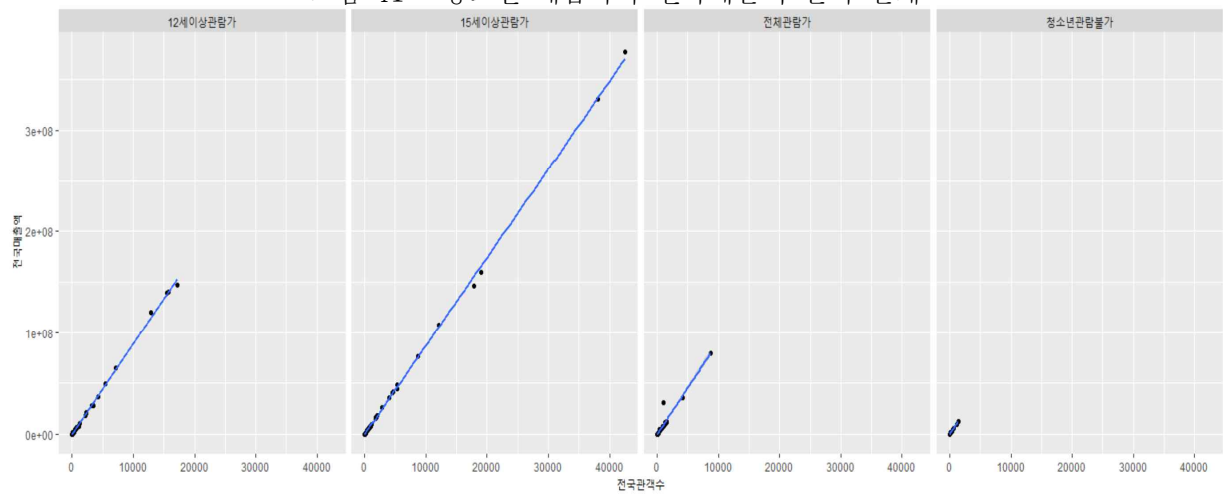
<그림 39> 장르별 전국관객수와 전국매출액 간의 관계



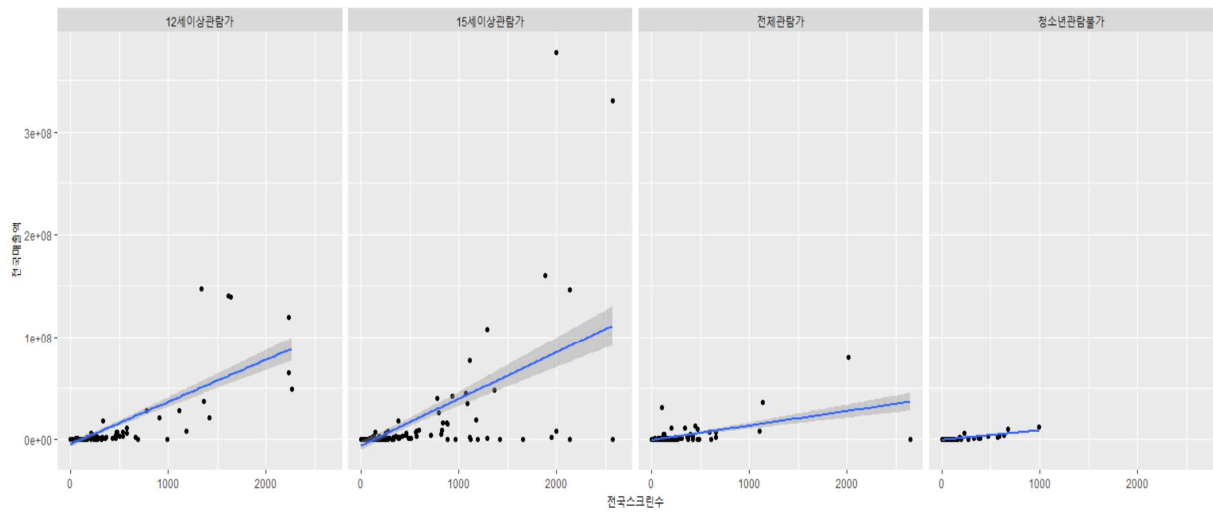
<그림 40> 장르별 전국스크린수와 전국매출액 간의 관계



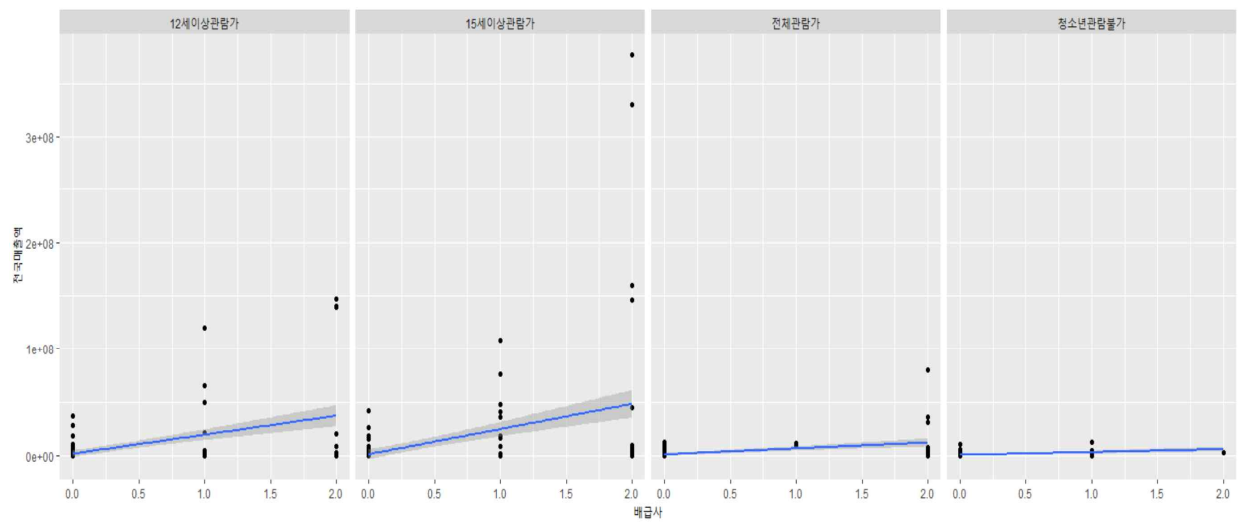
<그림 41> 장르별 배급사와 전국매출액 간의 관계



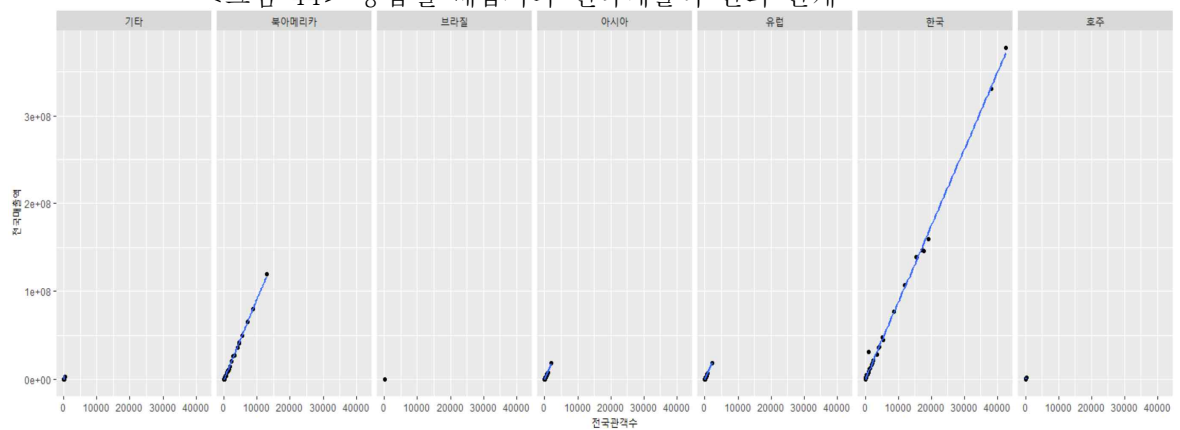
<그림 42> 등급별 전국관객수와 전국매출액 간의 관계



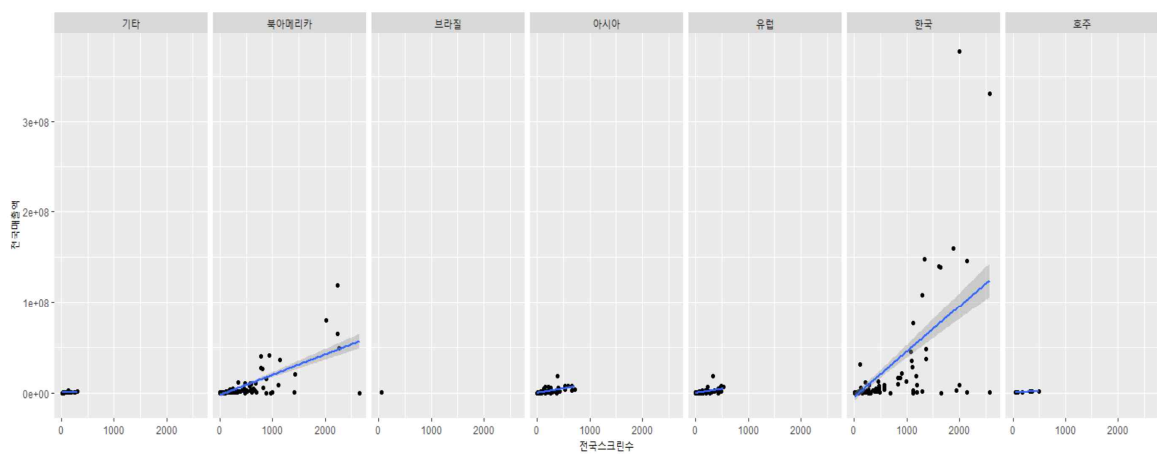
<그림 43> 등급별 전국스크린수와 전국매출액 간의 관계



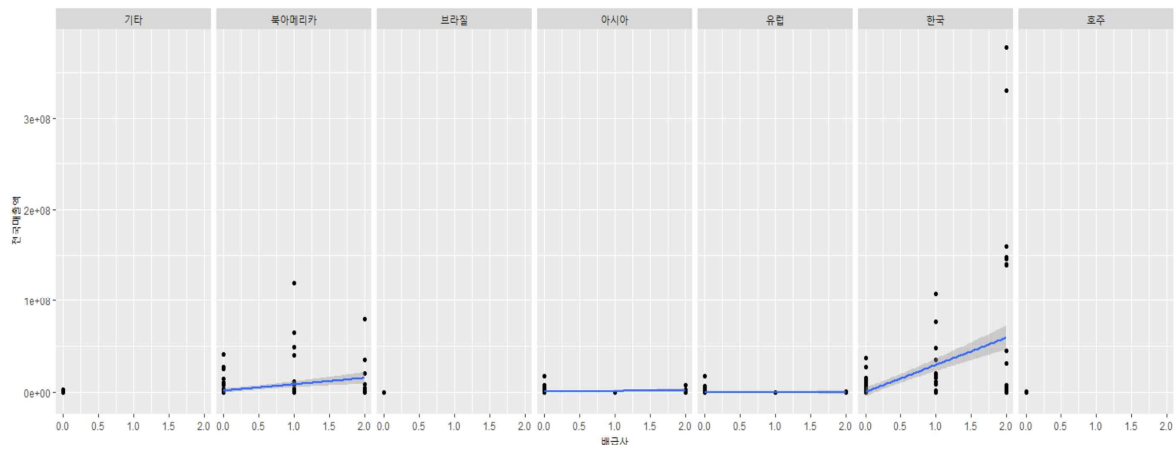
<그림 44> 등급별 배급사와 전국매출액 간의 관계



<그림 45> 국적별 전국관객수와 전국매출액 간의 관계



<그림 46> 국적별 전국스크린수와 전국매출액 간의 관계



<그림 47> 국적별 배급사와 전국매출액 간의 관계

다른 전국매출액들보다 값이 지나치게 큰 이상치 10개를 제거 한 후, 각 장르, 등급, 국적별로 그래프로 나타내었다. 장르별, 등급별, 국가별에서 비슷한 결과가 나타났는데, 전국관객수는 전국매출액과 강한 양의 선형관계를 띠었으며, 단순회귀분석 결과의 결정계수 값 (0.998)을 뒷받침한다. 이는 매우 높은 확률로 전국 관객수가 증가할수록 전국매출액이 증가함을 알 수 있다. 전국스크린수와 배급사도 대부분 전국매출액과 양의 선형관계를 띠었다. 이는 전국 관객수보다는 약하지만 전국스크린수가 증가할수록, 배급사가 소형에서 대형으로 갈수록 대체적으로 전국매출액이 증가함을 알 수 있다.

✓ 전국매출 예측 식 구축

위에서 선택한 유의한 변수 전국관객수, 전국스크린수, 장르, 등급, 배급사, 국적을 이용하여 전국 매출을 예측하는 식을 구축한다. 이를 위해 위에서 진행한 전처리에 추가로 전처리를 진행하고, 각 장르별로 다중회귀모형을 통해 식을 구하였다.

전국매출 예측을 위한 추가로 진행한 전처리 과정은 다음과 같다.

- 1) 장르는 코미디, 애니메이션, 로맨스, 드라마, 범죄와 같은 주요 장르들과 나머지(others)로 구분한다.
- 2) 국적은 한국, 북아메리카, 나머지로 구분한다.
- 3) 등급은 19세 이상과 나머지로 구분한다.

전국매출 예측 식을 구하기 위한 다중회귀분석의 식은 아래와 같다.

$$\text{<수식4> } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

다중회귀분석은 독립변수가 2개 이상이고, 종속변수는 1개로 이들 사이에 직선관계로 가정하는 경우이다. 다중회귀분석은 단순회귀분석의 연장선이라 할 수 있다. $\beta, \beta_1, \beta_2, \dots, \beta_p$ 를 회귀 계수라 부르며, 단순회귀분석에서와 마찬가지로 회귀분석에서 결정계수(R^2)의 값은 종속변수의 총 변동에 대한 독립변수들의 설명력의 크기를 나타내는 척도로서 0부터 1까지의 범위를

갖는다. R^2 의 값이 1에 가까울수록 독립변수에 의한 종속변수의 설명력이 크고, 회귀식의 적합도가 높다는 것을 의미한다. 하지만, 다중회귀분석에서는 독립변수의 수가 증가하면 결정계수의 값이 항상 커진다. 이러한 단점을 보완하기 위해, 다중회귀분석에서는 수정된 결정계수를 이용한다.

```
> summary(dramalm)

Call:
lm(formula = sales ~ distribution + view + num_screen + country,
    data = drama)

Residuals:
    Min       1Q   Median       3Q      Max
-764520  -17840    6549   25005  473071

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -24545.31   13859.95  -1.771  0.07843 .
distribution   -81444.98  23626.32  -3.447  0.00072 ***
view             8175.81    69.61 117.448 < 2e-16 ***
num_screen      115.29     89.88   1.283  0.20140
countryNorthAmerica 43955.19  21907.21   2.006  0.04645 *
countryothers   29140.29  17324.96   1.682  0.09448 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101000 on 164 degrees of freedom
Multiple R-squared:  0.9946,    Adjusted R-squared:  0.9945
F-statistic: 6069 on 5 and 164 DF,  p-value: < 2.2e-16
```

<그림 48> 장르가 드라마일 때 전국매출 예측

```
> summary(romancelm)

Call:
lm(formula = sales ~ distribution + view + num_screen + country +
    grade, data = romance)

Residuals:
    Min       1Q   Median       3Q      Max
-226198    -896     165     5465  152963

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   92869.72   34090.04   2.724  0.00833 **
distribution   84698.51   33167.73   2.554  0.01309 *
view           9028.27     67.27 134.202 < 2e-16 ***
num_screen    -377.94    136.01  -2.779  0.00718 **
countryNorthAmerica -101086.05  39523.50  -2.558  0.01296 *
countryothers  -136167.86  32239.61  -4.224 7.88e-05 ***
grade         -93303.77   33919.53  -2.751  0.00775 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54000 on 63 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9995
F-statistic: 2.485e+04 on 6 and 63 DF,  p-value: < 2.2e-16
```

<그림 49> 장르가 로맨스일 때 전국매출 예측

```
> summary(actionlm)
```

Call:

```
lm(formula = sales ~ view + num_screen + country, data = action)
```

Residuals:

Min	1Q	Median	3Q	Max
-2012095	-209481	-15422	104399	5998558

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-928585.54	419596.92	-2.213	0.03179 *
view	8688.57	34.72	250.211	< 2e-16 ***
num_screen	961.33	285.85	3.363	0.00154 **
countryNorthAmerica	847152.07	445129.73	1.903	0.06315 .
countryothers	759583.96	468705.63	1.621	0.11179

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1055000 on 47 degrees of freedom

Multiple R-squared: 0.9996, Adjusted R-squared: 0.9995

F-statistic: 2.768e+04 on 4 and 47 DF, p-value: < 2.2e-16

<그림 50> 장르가 액션일 때 전국매출 예측

```
> summary(animationlm)
```

Call:

```
lm(formula = sales ~ distribution + view + num_screen + country,
    data = animation)
```

Residuals:

Min	1Q	Median	3Q	Max
-780211	-44474	-21464	51375	898069

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6669.61	134388.92	0.050	0.961
distribution	-68704.02	57672.98	-1.191	0.239
view	8412.79	86.73	97.004	<2e-16 ***
num_screen	-264.08	240.54	-1.098	0.277
countryNorthAmerica	-191174.98	144598.19	-1.322	0.192
countryothers	41051.30	124673.61	0.329	0.743

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 225200 on 52 degrees of freedom

Multiple R-squared: 0.9959, Adjusted R-squared: 0.9955

F-statistic: 2507 on 5 and 52 DF, p-value: < 2.2e-16

<그림 51> 장르가 애니메이션일 때 전국매출 예측


```
> summary(comedylm)
```

Call:

```
lm(formula = sales ~ distribution + view + num_screen + country,
    data = comedy)
```

Residuals:

Min	1Q	Median	3Q	Max
-2206765	-53300	-6767	95232	1567939

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-70713.81	279615.52	-0.253	0.8017
distribution	105207.69	205734.01	0.511	0.6121
view	8923.29	61.15	145.930	<2e-16 ***
num_screen	-967.13	455.19	-2.125	0.0404 *
countryNorthAmerica	219057.66	288392.91	0.760	0.4523
countryothers	118451.62	296327.93	0.400	0.6917

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 631700 on 37 degrees of freedom
Multiple R-squared: 0.9989, Adjusted R-squared: 0.9988
F-statistic: 6758 on 5 and 37 DF, p-value: < 2.2e-16

<그림 52> 장르가 코미디일 때 전국매출 예측

```
> summary(criminallm)
```

Call:

```
lm(formula = sales ~ distribution + view + num_screen + country +
    grade, data = criminal)
```

Residuals:

Min	1Q	Median	3Q	Max
-770720	-101830	13111	88016	1921203

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	30060.8	336269.3	0.089	0.930
distribution	691708.4	523123.5	1.322	0.206
view	8867.3	19.6	452.384	<2e-16 ***
num_screen	-459.7	591.3	-0.777	0.449
countryNorthAmerica	-195757.0	422624.1	-0.463	0.650
countryothers	-74219.0	427038.2	-0.174	0.864
grade	188182.9	367395.5	0.512	0.616

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 599200 on 15 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 0.9999
F-statistic: 6.886e+04 on 6 and 15 DF, p-value: < 2.2e-16

<그림 53> 장르가 범죄일 때 전국매출 예측

```
> summary(otherslm)

Call:
lm(formula = sales ~ distribution + view + num_screen + country,
    data = others)

Residuals:
    Min       1Q   Median       3Q      Max
-2306754 -280927    9225   135022 19625040

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    331180.9   225053.8    1.472   0.1429
distribution    952876.1   228534.9    4.169 4.75e-05 ***
view           9486.3     266.2   35.637 < 2e-16 ***
num_screen     -2248.1     781.4   -2.877   0.0045 **
countryNorthAmerica -373002.6  305962.1  -1.219   0.2244
countryothers  -295536.6   277126.9  -1.066   0.2877
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1566000 on 179 degrees of freedom
Multiple R-squared:  0.9352,    Adjusted R-squared:  0.9334
F-statistic: 517 on 5 and 179 DF,  p-value: < 2.2e-16
```

<그림 54> 장르가 나머지일 때 전국매출 예측

<그림48>부터 <그림54>까지는 각 장르별에서 회귀분석 결과를 보여준다. 장르가 드라마, 로맨스, 액션일 때는 다양한 변수들의 전국매출액과 유의한 것으로 파악되었다. 해당 장르별 추정된 회귀식은 다음과 같다.

[장르가 로맨스일 때 전국매출 예측 식]

$$\text{전국매출액} = 92869.72 + (84698)(\text{배급사}) + (9028.27)(\text{전국관객수}) + (-377.94)(\text{전국스크린수}) \\ + (-101086.05)(\text{미국}) + (-136167.86)(\text{나머지국가}) + (-93303.77)(\text{등급})$$

[장르가 드라마일 때 전국매출 예측 식]

$$\text{전국매출액} = -24545.31 + (-81444.98)(\text{배급사}) + (8175.81)(\text{전국관객수}) \\ + (115.29)(\text{전국스크린수}) + (43955.19)(\text{미국}) + (29140.29)(\text{나머지국가})$$

[장르가 액션일 때 전국매출 예측 식]

$$\text{전국매출액} = -928585.54 + (8688.57)(\text{전국관객수}) + (961.33)(\text{전국스크린수}) \\ + (847152.07)(\text{미국}) + (759583.96)(\text{나머지국가})$$

로맨스, 드라마, 액션을 제외한 다른 장르에서는 전국매출액에 유의하게 영향을 줄 것으로 기대되는 요인이 나타나지 않았다. 통계적인 추론을 이끌어내기 위해서는 충분한 수의 데이터가 필요할 것으로 보인다.

본 과제에서 분석된 자료의 한계점을 보완한다면 추후의 영화계 매출에 관한 연구에 더욱 발전된 결과를 기대할 수 있을 것으로 사료된다.

3.2 과제(작품)의 팀원간 역할

구분	성명	전공	역할	참여도(%)
대표학생	임성수	정보통계학과	자료수집, 시각화, 분석, 보고서 작성	100
팀원	박영현	정보통계학과	자료수집, 시각화, 분석, 보고서 작성	100
팀원	이승현	정보통계학과	자료수집, 시각화, 분석, 보고서 작성	100
팀원	황성아	정보통계학과	자료수집, 시각화, 분석, 보고서 작성	100

4. 결론

4.1 과제(작품)의 개발 결과

코로나19 시기의 영화산업을 분석해 본 결과, 전국매출액에 영향을 주는 영화 요인은 전국관객수, 전국스크린수, 장르, 등급, 배급사, 국적으로 발견되었다. 전국관객수와 전국 스크린수가 많을수록, 배급사는 소형에서 대형으로 갈수록 매출이 증가하였다. 장르는 드라마, 범죄, 액션 영화의 평균 전국 매출이, 등급은 15세 이상 관람가 영화의 평균 전국매출이 가장 높았으며, 국가는 한국과 북아메리카 영화의 평균 전국매출이 가장 높은 것으로 나타났다. 데이터분석을 통해 얻어진 활용가능한 예측 식은 다음과 같다.

• 장르가 드라마일 때 전국매출 예측 식:

$$\text{전국매출액} = -24545.31 + (-81444.98)(\text{배급사}) + (8175.81)(\text{전국관객수}) \\ + (115.29)(\text{전국스크린수}) + (43955.19)(\text{미국}) + (29140.29)(\text{나머지국가})$$

• 장르가 로맨스일 때 전국매출 예측 식:

$$\text{전국매출액} = 92869.72 + (84698)(\text{배급사}) + (9028.27)(\text{전국관객수}) + (-377.94)(\text{전국스크린수}) \\ + (-101086.05)(\text{미국}) + (-136167.86)(\text{나머지국가}) + (-93303.77)(\text{등급})$$

• 장르가 액션일 때 전국매출 예측 식:

$$\text{전국매출액} = -928585.54 + (8688.57)(\text{전국관객수}) + (961.33)(\text{전국스크린수}) \\ + (847152.07)(\text{미국}) + (759583.96)(\text{나머지국가})$$

4.2 과제(작품)의 활용방안 및 기대효과

본 연구는 문화 예술계의 회복을 위한 전략 수립에 많은 기여를 할 것으로 기대된다. 또한 본 연구의 분석 방법 및 전략 수립 방안은 영화뿐만 아니라 다른 문화 예술계로의 확장성을 가진다. 따라서 실무자가 본 연구의 결과를 바탕으로 효율적인 의사결정을 내릴 수 있을 것으로 기대된다. 또한, 통계학을 접하지 못한 연구자들이 새롭게 수집되는 자료를 적용할 수 있는 도구를 제공하여 문화 예술계에서의 전략 수립에 기초가 된다는 점에서 기대되는 효과가 크다. 하지만 본 연구에서는 데이터 수집의 어려움으로 데이터 개수가 적어 전국매출을 예측하는 데에 한계점이 존재한다. 따라서 추후의 연구에서는 이러한 한계점을 개선하여 더 발전된 결과를 기대한다.

4.3 결과물 및 특허 출원/등록 결과

문화예술계는 사회구성원의 인식에 큰 영향을 미치지만 관련 연구가 매우 미흡하다. 따라서 본 연구는 문화예술계의 빠른 회복을 위한 전략 수립에 큰 기여를 할 것으로 예상되므로 관련 분야의 논문으로 투고할 수 있을 것으로 예상된다. 또한 연구에 사용되는 중요 자료와 프로그램을 패키지 형태로 제공함으로써 타 연구자들이 본 연구를 확장할 수 있도록 한다.