Big Data Analysis and Visualization
Project

이름: 백영민, 박희정, 김유나
학번: 201511093, 201511085, 201511043

Abstract

본 프로젝트에서는 기존 영화 데이터를 이용하여 영화의 흥행에 대해 전체적으로 분석해보고자 한다. 영화진흥원에서 기간별 데이터를 통해 2004 년부터 2017 년 까지의 영화 별 데이터를 얻었다. 우리는 영화 흥행에 영향을 주는 여러 요소 중 장르, 날짜(계절), 감독, 배우를 골라 이에 대해 중점적으로 분석했다. 또한 영화진흥원에서 제공하는 일별 데이터를 2013 년도 6 월~2 월, 2016 년도 6 월~2 월 데이터를 통해 영화의 흥행이 결정되는 시기와 경쟁 작에 따른 영화의 관객 수 변화 또한 알아보았다. 실제 장르, 날짜, 감독, 배우는 영화의 흥행에 크게 영향을 미치는 것으로 확인되었고, 각 요소끼리의 상관관계도 존재함을 밝혔다. 또한 경쟁 작 또한 영화의 흥행과 연관이 있다는 것을 볼 수 있었다. 이는 앞으로의 상영할 영화 또는 앞으로 제작할 영화들에게 시사점을 선사할 것이다.

1. Introduction

한국에 1998 년에 국내 최초의 멀티플렉스(CGV)가 생겼고, 그 이후 한국에서의 영화산업은 다른 산업들보다 더 급격하게 성장했다. 2013 년도에는 전국 관람객이 2 억명을 돌파했고, 2007 년도부터 1 인당 영화 관람 횟수 변화를 살펴보면 2007 년 당시 미국과 호주보다 대략 1 회 이상 떨어지던 수치가 2013 년에는 한국만 4 회를 넘어서며 호주, 미국, 영국, 프랑스를 꺾은 문화의 나라로 자리매김 하는 모습을 보였다.

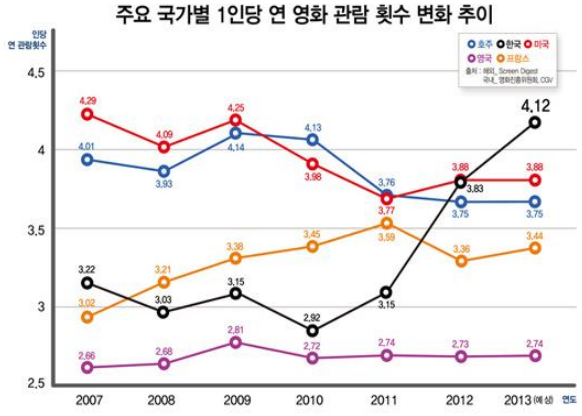

그림 1. 주요 국가별 1 인당 연 영화 관람 횟수 변화 추이

그러나 커져버린 영화 산업만큼 영화의 수도 크게 증가하였고, 따라서 영화의 흥행 또한 힘들어졌다. 영화진흥위원회에 따르면 2015 년 극장에서 개봉한 한국영화의 평균 수익률은 총 제작비 기준 3.99%이고, 총 비용의 3.4%에 불과한 것으로 집계됐다.

2015년도의 손익분기점을 넘긴 영화는 전체 영화의 0.27 밖에 되지 않았다. 우리는 이를 빅데이터를 이용해 영화에 영향을 주는 요소들을 찾아, 흥행한 영화와 그렇지 않은 영화들과 우리가 찾은 요소들의 상관관계를 비교하였다. 우리는 이를 통해 흥행한 영화들의 특징을 얻고자 했다.

2. Methodology and Result

.

1. *전체데이터 분석*

2004 년부터 2017 년까지 총 누적관객수가 1000 명이 넘은 영화를 대상으로, 어떤 요소가 영화의 흥행에 가장 많은 영향을 끼치는지 알아보고자 한다.

1.1 데이터

영화진흥위원회의 기간별 박스오피스에서 얻어온 데이터 사용.

1.2 데이터 전처리

기존의 데이터는 약 15000 개의 영화에 대한 정보를 담고 있음. 그러나 상영횟수가 0 회이거나 스크린 수가 0 인 이름만 올라와 있는 영화가 존재하여 총 누적관객수가 10000 명이상인 영화를 대상으로 분석을 진행하기로 하였다.

영화에 출연한 배우들은 최대 10 명으로 줄였고, 장르도 최대 1 개에만 속하도록 하였다.

개봉날짜 기준으로 개봉 년도, 개봉 월, 계절 정보를 추가하였다. 계절은 12~2 월→겨울, 3~5 월→봄, 6~8 월→여름, 9~11 월→가을로 정했다.

결과적으로, 전 처리를 마친 데이터의 columns 는 영화 명, 개봉일, 매출액, 매출액점유율, 누

적매출액, 누적관객수, 스크린 수, 배우, 감독, 장르, 국적, 등급으로 구성되어 있다. 배우는 분석의 편의성을 위해 배우 1, 배우 2… 배우 10 로 한 명씩 따로 분리하였다.

1.3 분석방법

영화 흥행에 가장 많은 영향을 줄 수 있는 요소로 장르, 개봉시기(계절), 감독/배우를 뽑았다. 이 요소들을 중점적으로 분석한 후에 기타 요소들의 분석을 진행하였다. 먼저 각 요소의 영향을 분석하고 각 요소들이 합쳐졌을 때, 어떤 영향을 미치는지 분석하고자 한다.

1.4 장르요소 분석

장르별로 영화당 누적관객수가 어떻게 달라지는 지 분석을 진행한다. 영화들을 총 누적관객수가 50 만명 미만, 50 만~100 만, 100 만~300 만, 300 만 이상인 영화들로 나누어 각 범위에 장르들이 어떻게 분포하고 있는지 알아본다.

1.5 개봉시기(계절)요소 분석

개봉시기(개봉 년도, 개봉 월, 개봉계절)에 따라 총 누적관객수가 어떻게 달라지는지 분석을 진행한다. 이 결과와 다른 요소들을 합하여 분석을 진행한다.

1.6 감독요소 분석

감독에 따라 총 누적관객수, 감독이 영화 한편당 평균적인 누적관객수를 분석한다. 그러나 같은 감독이 찍은 영화의 수는 대부분 1~10 사이에 분포하므로, 표본의 수가 적어 이후의 분석은 의미를 찾기 힘들 것으로 예상된다.

1.7 배우요소 분석

각 배우 별로 출연한 영화의 총 누적관객수, 영화 한편당 누적관객수를 분석한다. 영화들을 총 누적관객수가 50 만명 미만, 50 만~100 만, 100 만~300 만, 300 만 이상인 영화들로 나누어 각 범위의 영화들에 배우들이 얼마나 출현했는지 분석을 진행한다.

1.8 여러 요소들을 종합하여 분석

1.4~1.7 각 요소들의 분석 결과를 토대로 각 요소들을 종합하여 추가적인 분석을 진행한다.

2. *전체데이터 분석 결과*

2.1 장르요소 분석

각 장르별 총 누적관객수의합으로 분석을 진행했으나 총 누적관객수는 영화 수에 비례할 것이므로, 영화당 평균관객수로 분석을 진행하였다. 또한 관객 수 범위 별로 각 장르의 영화가 어떻게 분포하고 있는지 분석하였다. 이 결과는 표 1 과 같다.

장르	영화당관객수비율	count_50	count_100	count_300	count_나머지
사극	2.653087e+06	13	6	10	11
SF	1.358943e+06	52	18	21	12
전쟁	1.326652e+06	10	1	5	2
액션	1.289596e+06	252	63	95	60
범죄	1.149642e+06	66	16	24	13
어드벤처	1.102500e+06	22	7	13	2
판타지	8.093170e+05	25	9	10	3
미스터리	7.631379e+05	40	5	13	3
스릴러	6.048127e+05	102	16	30	3
코미디	5.984298e+05	271	40	47	17
멜로/로맨스	5.375281e+05	173	30	33	7
드라마	5.215124e+05	691	71	75	37
애니메이션	4.382896e+05	313	42	33	9
공포(호러)	3.645966e+05	139	18	13	2
다큐멘터리	1.590659e+05	88	1	2	1

표 1. 장르별 영화당 평균 관객수

영화당 관객수 비율은 (해당 장르의 총 관객수/해당 장르의 총 영화 수)로 계산 count_50 는 해당장르에서 총 관객 수가 50 만 이하인 영화의 수, 계산 count_100 은 해당장르에서 총 관객수가 50 만 초과 100 만 이하인 영화의 수, count_300 은 해당장르에서 총 관객 수가 100 만 초과 300 만 이하인 영화의 수, count_나머지는 해당장르에서 총 관객 수가 300 만 초과인 영화의 수이다.

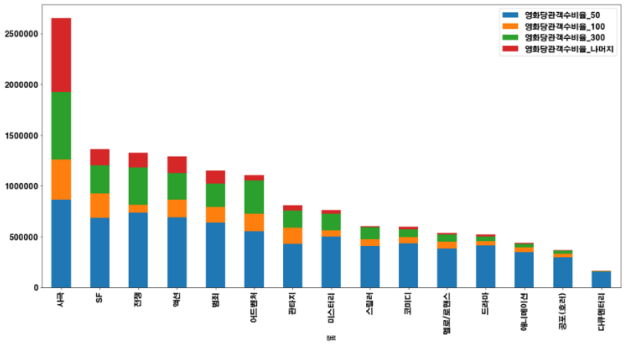

그림 2. 장르별 총 누적관객수의 분포

표 1 의 결과를 그래프로 나타내면 그림 2 와 같다. 사극장르가 영화당 평균 누적관객수가 가장 높았고, 다큐멘터리장르가 가장 낮았다. 흥행작의 개수로만 보자면, 액션장르에 흥행작이 가장 많음을 알 수 있다. 흥행의 기준을 어떻게 잡느냐에 따라 다르겠지만, 전체 중에 많은 관객 수가 차지하고 있는 비율이 높은 장르를 선택하는 것이 흥행 가능성을 조금 더 높여줄 수 있을 것이라고 생각한다. 이는 각 장르별 영화 제작비와 연관시켜 생각해 볼 필요가 있을 것 같은데, SF/액션과 같은 전형적인 제작비가 많이 들어가는 장르인 경우, 엄청난 흥행을 하지 않으면 손해를 볼 수 있다. 그러나 그림 2 에서 보이듯이 SF/액션 장르의 영화 중

300 만 이상의 누적관객수를 기록한 영화는 전체에 비해 적은 비중을 차지한다. 반면에 비교적 적은 제작비가 들어가는 사극 장르의 경우 그림 1 의 결과에서 영화당 평균 관객수도 높을 뿐만 아니라, 300 만 이상의 누적관객수를 기록한 영화의 비율도 가장 높다. 따라서 장르 자체의 요인만 고려해보자면, SF/액션 장르에 비해 사극장르를 개봉했을 때 손익분기점을 넘길 확률이 높다고 볼 수 있다.

2.2 개봉시기요소 분석

개봉 년도에 따른 누적관객수는 그림 3 와 같다. 개봉 년도가 최근에 가까워 질수록 증가하는 추세를 보인다. 이는 스크린 수 자체가 증가하고, 영화에 대한 관심이 높아졌기 때문으로 생각된다.

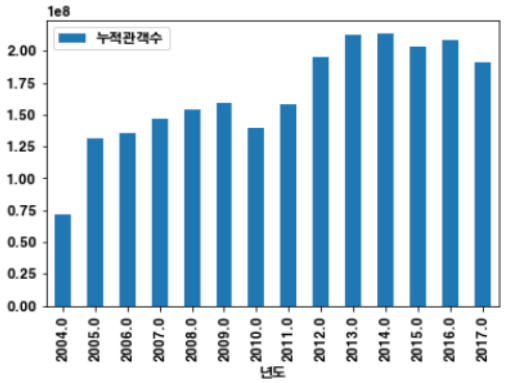

그림 2. 개봉 년도와 누적관객수

개봉 달에 따른 평균 누적관객수는 그림 4 와 같다. 7 월, 12 월이 특히 관객수가 많은 것을 볼 수 있는데, 이는 방학시즌과 연말에 관객수가 많아지는 것을 뜻한다. 3 월달은 일년 중 가장 관객수가 적다.

달에 따른 개봉 영화 수와 그 중 흥행한 영화의 수는 그림 5 와 같다. 이 그래프와 그림 4 를 연결시켜보면 흥미로운 결론을 얻을 수 있는데, 7 월과 12 월은 개봉 영화 수는 비교적 적지만 누적관객수가 가장 많다. 흥행한 영화가 많으면 개봉 영화 수에 비해 더 많은 관객수를 기록할 것이므로, 7 월과 12 월은 흥행작이 많을 것이다. 이는 그림 5 와 일치하는데, 7 월과 12 월은 다른 달에 비해 흥행작의 수가 많음을 알 수 있다. 반면에 2,3,4,11 월은 개봉한 영화 수는 많으나 총 누적관객수는 적다. 이러한 달에는 흥행작이 적은 것으로 볼 수 있는데, 그림 5 에 따르면 이러한 가설이 일치하는 것을 확인할 수 있다. 따라서 개봉 시기 자체의 요인만 고려해본다면, 7 월과 12 월이 관객 수도 많고, 개봉 작의 수가 적어 경쟁작도 적으므로 최적의 시기로 고려된다. 하지만 같은 시기에 개봉한 영화 중에 흥행작이 많으므로, 경쟁에서 도태될 수 있다. 이는 뒤의 경쟁 작에 의한 요인 분석에서 조금 더 다루도록 한다.

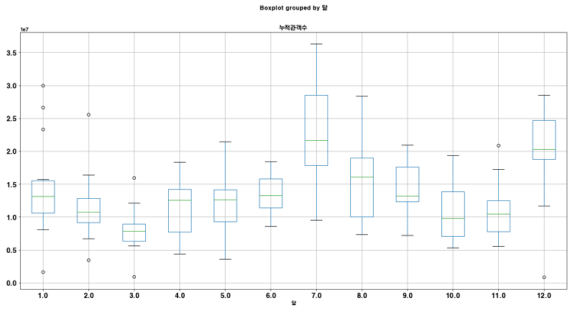

그림 4. 개봉 달과 누적 관객수

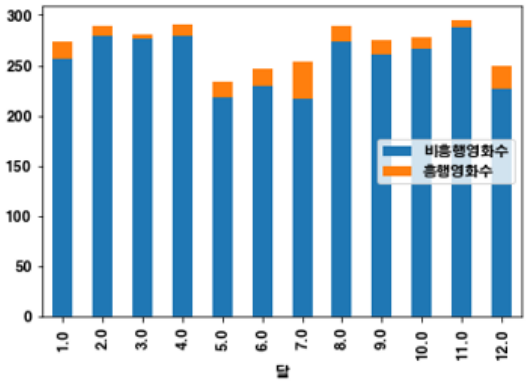

그림 5. 개봉 달과 흥행영화 수

2.3 감독요소 분석

특정 감독이 찍은 영화의 평균 누적관객수는 표 2 와 같다.

감독	누적관객수
최동훈	7.676431e+06
김한민	5.559939e+06
마이클 베이	4.836449e+06
크리스토퍼 놀란	4.403688e+06
이석훈	4.391531e+06
류승완	4.227352e+06
김성훈	3.814861e+06
데이빗 예이츠	3.241509e+06
이준익	3.060716e+06
매튜 본	2.921589e+06
김지운	2.880058e+06
강우석	2.676571e+06

표 2. 감독과 누적 관객수 평균
(영화 작품 수가 5 편 이상인 감독 선별)

실제로 감독은 영화의 많은 부분에 관여하므로 작품의 흥행에 많은 영향을 끼칠 것이다. 하지만 같은 감독이 찍을 수 있는 영화의 수에는 한계가 있으므로, 표본의 수가 매우 적다. 또한 감독은 어떤 장르의 영화를 찍느냐에 따라 결과가 다를 것이다. 따라서 이 결과만을 이용하여 상위권의 감독이 영화를 찍는다고 흥행을 보장할 수는 없다. 따라서 뒤에서 장르 요소와 함께 고려하여 감독요소의 결론을 이끌어내고자 한다.

2.4 배우요소 분석

특정 배우가 출연한 영화의 총 누적관객수, 영

화당 평균 관객수는 표 3 과 같다. 이는 출연영화 수가 10 편이상인 배우들을 대상으로 분석한 결과이다. 그림 5 는 특정 배우가 출연한 영화들의 총 관객수 합과 그 중 흥행작에 속해있는 비율이 얼마나 되는지 나타낸다. 그림에서 붉은색은 300 만이상의 영화, 초록색은 100 만~300 만의 영화, 주황색은 50 만~100 만의 영화, 파란색은 50 만 이하의 영화이다. 관객수가 많은 순서로 배우들을 나열했기 때문에 대부분 배우들이 붉은 부분이 가장 많다. 특히 송강호 같은 경우는 붉은 부분의 비율이 특히 높은 것을 볼 수 있는데 이는 표 3 의 영화당 관객수 평균이 가장 높은 것과 일치하는 결과라고 볼 수 있다.
배우도 감독과 마찬가지로 어떤 장르의 영화를 소화해내느냐에 따라 결과가 달라질 것이다. 단순히 영화당 관객수가 높은 배우가 출현한다고 그 영화의 흥행을 보장할 수는 없다. 따라서 뒤에서 다른 요소와 함께 분석하도록 한다

	배우	누적관객수	출연영화수	영화당관객수
1	송강호	87586680.0	16.0	5.474168e+06
93	이정재	54625744.0	11.0	4.965977e+06
772	오달수	162149020.0	45.0	3.603312e+06
228	정진영	53335974.0	15.0	3.555732e+06
4	로버트 다우니 주니어	56269980.0	16.0	3.516874e+06
8	최민식	41710626.0	12.0	3.475886e+06
3516	라미란	62211552.0	18.0	3.456197e+06
3716	김원해	37543710.0	11.0	3.413065e+06
95	곽도원	43425976.0	13.0	3.340460e+06
27	강동원	52755381.0	16.0	3.297211e+06
290	제레미 레너	42296967.0	13.0	3.253613e+06
83	유해진	111542970.0	35.0	3.186942e+06
0	황정민	104193263.0	33.0	3.157372e+06
1623	정만식	46721220.0	15.0	3.114748e+06

표 3. 배우와 누적관객수

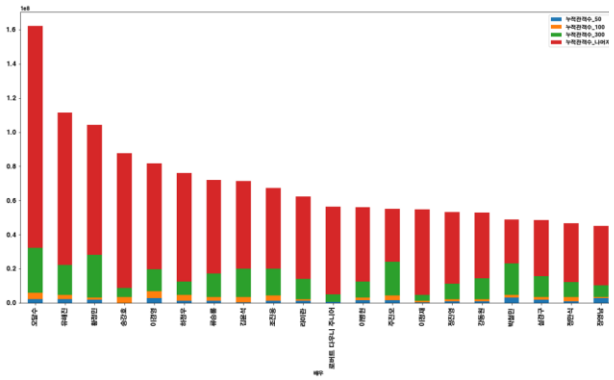

그림 6. 배우와 흥행작 비율

2.5 여러 요소들을 종합하여 분석
위의 분석들로부터 얻은 결론들로 여러 장르들을 함께 고려하여 다른 분석들을 진행한다.

2.6 최근(2015~2017 년도)에 어떤 장르가 주를 이루는지 보고자 한다. 그림 7 은 장르별로 전체 관객수 중 2015~2017 년도의 관객수의 비율을 그래프로 나타낸 것이다. 그림 8 은 장르별로 전체 개봉 영화 수 중 2015~2017 년도의 영화 수가 이루는 비율을 그래프로 나타낸 것이다.
2004~2014/2015~2017 년의 비율이므로 기준을 3/14 로 잡을 수 있다. 또한 영화 관람객의 수는 꾸준히 늘고 있는 추세이므로 실제 기준은 이보다 높아야 할 것이다. 이러한 기준보다 높으면 최근에 주를 이루는 장르라고 볼 수 있고, 이보다 낮으면 최근 관객수가 조금 줄어드는 추세인 장르라고 볼 수 있다. 액션, 애니메이션, 범죄장르가 대표적으로 최근에 주를 이루는 장르이고 SF, 판타지, 서부극 등은 최근에 관객수가 줄어드는 경향을 보여주고 있다.
그림 7 과 그림 8 을 함께 고려해서 보면 액션장르의 경우, 영화 수에 비해 관객수가 많은 것을 확인할 수 있다. 반면, 애니메이션 장르의 경우, 영화 수에 비해 관객수가 적다. 영화 수에 비해 관객수가 많은 것이 흥행할 확률이 높은 장르라고 볼 수 있다.
최근에 주를 이루면서, 흥행할 확률도 높은 장르인 경우 상대적으로 더 많은 관객수를 기대해 볼 수 있을 것이다.
통계자료에 의하면 액션/어드벤처 장르의 평균 투자 수익률이 101.93%로 가장 높았고, 코미디(21.68%), 범죄/스릴러(14.28%), 공포/미스터리(7.6%) 장르 순으로 흑자수익률을 기록했지만, 애니메이션(-74.11%), 다큐멘터리(-38.34%), 멜로/로맨스(-35.73%), 사극/시대극(-18.49%), 드라마(-12.52%)는 적자를 기록했다. 이 통계자료는 위 분석과 일치하는 결과를 보여준다

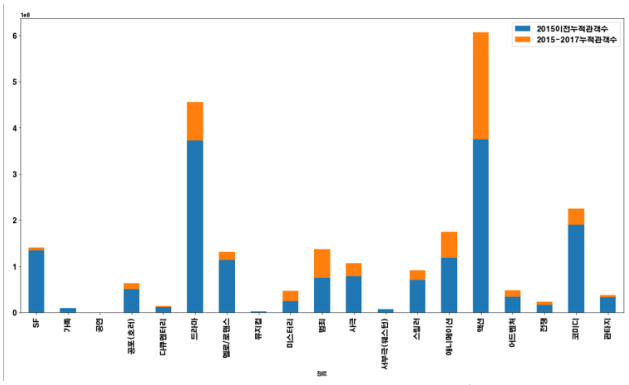

그림 7. 장르별 최근 관객수 비율

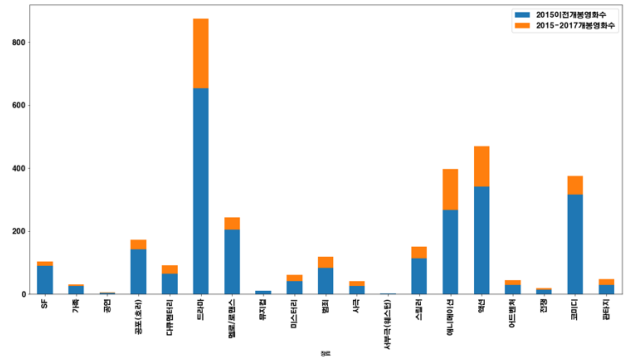

그림 8. 장르별 최근 개봉 영화 수 비율

2.7 계절과 장르
각 장르와 계절의 관계를 파악해보고자 한다. 계절별로 각 장르들이 차지하는 비율을 나타낸 그래프가 그림 9 이다. 이 비율은 특정 장르,

계절의 관객수를 해당 계절의 총 관객수로 나누어 계산하였다. 따라서 실제로 특정 계절의 관객수에서 해당 장르의 관객수가 차지하는 비율이라고 볼 수 있다.
각 계절별 비율이므로 총 합은 1 이나 그래프에 고려되지 않은 장르들 때문에 1 이되지 않는다. 이 장르들은 상대적으로 차지하는 비율이 낮아서 고려대상에서 제외하였다.
이 자료를 이용해 특정 장르의 영화를 개봉할 때, 흥행 확률이 높은 계절에 대해 알 수 있다. 이는 단순 관객 수의 합이므로, 개봉영화 수에 비례하기에 우리는 영화당 평균 관객수로 비율을 구했고 그 비율은 그림 10 과 같다.
이를 통해 각 계절별로 어떠한 장르를 개봉하는 것이 흥행 가능성을 높이는 방법인지 알아볼 수 있다.

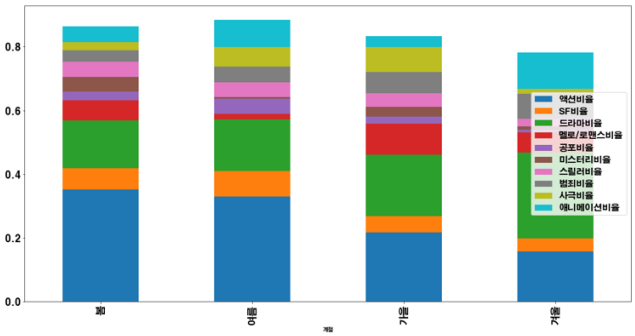

그림 9. 계절별 장르의 비율

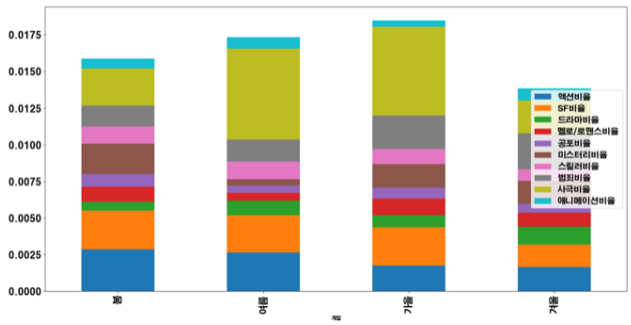

그림 10. 계절별 장르의 관객수평균 비율

2.8 장르와 배우

장르와 배우는 매우 밀접한 관계를 갖는다. 오달수가 멜로 장르를 찍거나 송강호가 코미디 장르를 찍는다면 상대적으로 배우가 출연한 다른 장르의 영화에 비해 흥행이 어려울 것이다. 표 4 는 장르와 배우의 영화당 관객수를 나타낸 것이다. 이를 바탕으로 특정 장르의 영화와 어울리는 배우들을 찾을 수 있을 것이다.
그림 11 은 배우 분석에서 상위에 있던 배우 4 명을 대상으로 그 배우들의 장르별 평균 누적관객수를 조사한 것이다. 그래프를 보면 배우마다 관객수가 높은 장르가 있다는 것을 다시 한 번 확인할 수 있다. 특히 하정우 같은 경우 장르에 따라 큰 편차를 보이는데, 코미디나 드라마장르와 하정우는 어울리지 않는다는 것을 알 수 있다. 이처럼 각 장르와 어울리는/어울리지 않는 배우들을 찾을 수 있으므로, 영화 제작자의 입장에서는 이러한 자료들을 토대로 배우를 선택해야 할 것이다.

배우	장르	누적관객수	출연영화수	영화당관객수
로버트 다우니 주니어	액션	44813047.0	7.0	6.401864e+06
정진영	드라마	32649968.0	6.0	5.441661e+06
오달수	액션	51769397.0	10.0	5.176940e+06
제레미 레너	액션	41218154.0	8.0	5.152269e+06
이경영	액션	24762755.0	6.0	4.127126e+06
박성웅	범죄	20531297.0	6.0	3.421883e+06
조진웅	액션	20062096.0	6.0	3.343683e+06
김영애	드라마	19900937.0	6.0	3.316823e+06
황정민	드라마	35843825.0	11.0	3.258530e+06
유해진	드라마	52093147.0	16.0	3.255822e+06
오달수	드라마	41296176.0	13.0	3.176629e+06
차태현	코미디	18570119.0	6.0	3.095020e+06
톰 크루즈	액션	33181638.0	11.0	3.016513e+06
장영남	드라마	21049282.0	7.0	3.007040e+06
크리스 헴스워스	액션	29685176.0	10.0	2.968518e+06

표 4. 배우와 장르에 따른 영화당 관객수
(출현 영화 수 5 편 이상인 배우 대상)

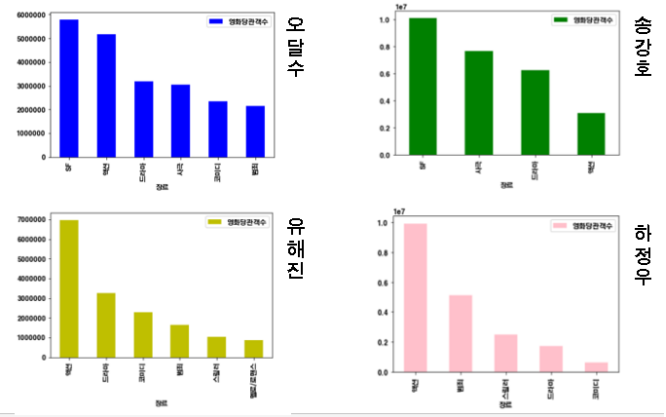

그림 11. 상위 배우 4 명의 장르별 평균 누적관객수

2.9 장르와 감독

배우와 마찬가지로 감독도 특정 장르와 시너지 효과가 뛰어날 것으로 생각한다. 표 5 는 감독과 장르에 따른 평균 누적관객수를 보여준다. 이를 통해 특정 장르와 잘 어울리는 감독을 찾아볼 수 있다.

감독	장르	누적관객수	영화수	영화당관객수
최동훈	액션	31754399.0	3.0	1.058480e+07
봉준호	SF	20267188.0	2.0	1.013359e+07
조스 웨든	액션	17569366.0	2.0	8.784683e+06
윤제균	드라마	16797012.0	2.0	8.398506e+06
김한민	사극	25097219.0	3.0	8.365740e+06
장훈	드라마	20548113.0	3.0	6.849371e+06
이환경	코미디	13183797.0	2.0	6.591898e+06
마이클 베이	SF	18014012.0	3.0	6.004671e+06
김성훈	액션	11267936.0	2.0	5.633968e+06
류승완	액션	29632846.0	6.0	4.938808e+06
황동혁	드라마	14222677.0	3.0	4.740892e+06
크리스토퍼 놀란	액션	9185212.0	2.0	4.592606e+06

표 5. 감독과 장르에 따른 영화당 관객수
(감독: 영화 수 2 편 이상)

2.10 전체데이터 분석 결론

전체데이터에서 각 요소가 어떻게 영화의 흥행에 영향을 주는지 알아보았다. 마지막으로 어떤 요소가 영향력이 큰지 대략적으로 파악해보고자 한다.

그림 12 는 각 요소 별 흥행비율의 분포를 boxplot 으로 나타낸 것이다. 이때 흥행비율은 각 다음과 같은 코드로 계산하였다.

```
movie_drop_월=pd.DataFrame(movie_drop['영화명'].groupby(movie_drop['달']).count()
movie_drop_월.columns=['달전체']
movie_흥행_월=pd.DataFrame(movie_흥행['영화명'].groupby(movie_흥행['달']).count()
movie_흥행_월.columns=['흥행달']
흥행_월=pd.merge(movie_drop_월,movie_흥행_월,right_index=True,left_index=True,how
흥행_월=흥행_월.fillna(0)
흥행_월['흥행비율']=흥행_월['흥행달']/흥행_월['달전체']
흥행_월=흥행_월.where(흥행_월['흥행비율']>0).dropna()
흥행_월['class']='월'
흥행_월.reset_index(inplace=True)
흥행_월=pd.DataFrame(흥행_월,columns=['class','흥행비율'])
흥행_월
```

이때 평균은 총 흥행비율의 평균이 아니라 흥행비율이 0 보다 큰 표본의 평균에 해당한다. 각 요소 별로 흥행비율이 0 인 표본의 수가 다르므로 평균은 의미가 없다. 이 그래프는 각 장르별 대략적인 분포를 알 수 있다. 넓게 분포할수록 그 요소에 따라 흥행비율의 차이가 크다고 볼 수 있다. 따라서 배우와 감독이 개봉 시기와 장르보다 영화의 흥행에 조금 더 영향을 많이 준다. 특정 영화의 개봉에 있어 각 요소를 이 비중에 따라서 고려한다면 흥행확률을 더 높일 수 있을 것이다.

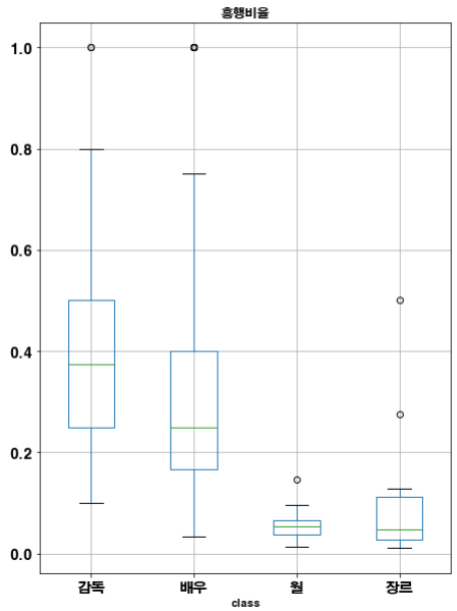

그림 12. 각 요소 별 흥행비율의 분포
(흥행 비율이 0 인 표본 제외)

3. 일별데이터 분석

3.1 영화의 흥행은 일주일안에 결정된다.

영화의 흥행이 일주일 안에 결정이 난다는 영화계의 일반적인 통설을 검증하고 그 기준이 일주일인 이유와 일주일 단위로 데이터를 분석하였을 때 흥행한 영화와 그렇지 못한 영화의 차이를 알아보고자 한다.

3.2 일별 데이터로부터 경쟁작이 영화 점유율에 미치는 영향 분석

경쟁작이 영화 점유율에 미치는 영향을 분석하여 특정 영화의 적당한 개봉 시기를 예측하고자 한다.

3.3 데이터

일별로 해당 날짜에 상영된 영화에 대한 정보를 담고 있는 데이터로 2013 년 여름,겨울, 2016 년 여름,겨울,가을의 데이터로 분석을 진행하였다.

3.4 데이터 전처리

전체 데이터에서 일별 매출액 점유율이 5%를 한번이라도 넘은 적이 있는 영화들의 데이터로 줄인다.

개봉 후 일수를 계산하여 새로운 column 을 만들었다. 전체 데이터에서 개봉 후 일수가 0~21 인 데이터로 줄인다.

3.5 개봉 후 일수 별 관객수 분석

개봉 후 5 주동안 일수 별 관객수 총합의 변화를 분석한다. 일수 별로 관객수의 변동이 가지는 규칙과 일주일 단위로 생기는 변화에 대해 살펴본다.

3.6 개봉 후 첫째 주, 둘째 주의 일주일관객수/누적관객수 비교 분석

개봉 후 첫째 주, 둘째 주의 일주일관객수/누적관객수 비율과 각각 값을 비교한 분석을 진행한다. 흥행을 한 영화라면 관객수 비율이 개봉 후 둘째 주에도 어느 정도 유지될 것이라는 가설을 세웠다. 따라서 (둘째 주의 일주일관객수/누적관객수)/(첫째 주의 일주일관객수/누적관객수)를 누적관객수 300 만 명 이상인 영화, 50 만 명 이상 300 만 명 미만인 영화, 50 만 명 미만인 영화로 나누어 비교한다면 그 값이 점점 줄어들 것이라 예상하였다.

3.7 일별 매출액점유율 변화 분석

주 별로 일주일관객수/누적관객수의 비율을 비교하는 것 외에도 특정 기간 동안 일별로 데이터를 비교하여 추이를 보는 것이 필요하다. 이에 따라 개봉 후 일주일/이주일 동안의 일별 매출액점유율 변화를 분석한다. 분석은 누적관객수 300 만 명 이상인 영화, 50 만 명 이상 300 만 명 미만인 영화, 50 만 명 미만인 영화로 나누어 비교하는 방향으로 진행한다. 흥행을 하지 못한 영화일수록 관객들이 점점 영화를 찾지 않는 경향이 강해질 것이기 때문에 추세선의 기울기가 관객수가 적어질수록 점점 줄어들 것이라 예상하였다.

3.8 경쟁작의 정의

우리의 분석에서 경쟁작은 기존영화들에 대해 새로 개봉하는 모든 영화로 정의한다.

3.9 새로운 영화가 개봉했을 때, 기존에 상영중인 영화의 분석

특정 영화가 개봉한 날짜에 대해서 그 날짜에 상영중인 기존 영화들의 점유율 변화를 분석한다. 개봉한 날짜 기준으로 이전주의 매출액 점

유율 평균과 다음주의 매출액 점유율 평균을 비교하여 점유율 변화를 구한다. 이 때 개봉 후 일수에 따른 매출액 평균 점유율 변화에 따라 점유율을 보정해준다.(일주일 단위로 상영횟수가 줄어들어 점유율이 떨어지므로) 개봉 후 일주일 뒤에는 1.697 만큼, 개봉 후 2 주 뒤에는 3.34 만큼 보정을 해준다.

위의 과정으로 얻은 데이터로 개봉영화의 장르, 기존 영화의 장르, 순위, 총관객수 등의 요인에 따라 기존영화의 매출액점유율 변화를 분석한다.

3.10 새로운 영화가 개봉했을 때, 개봉영화의 일주일 매출액 점유율 평균 분석

개봉영화의 일주일 매출액 점유율 평균을 추가로 계산한다. 이는 실제로 영화가 개봉 후 일주일 동안의 점유율 평균을 나타낸다. 이를 통해 개봉영화가 어떤 요인에 영향을 받는지 알아보고자 한다. 앞의 분석과 동일하게 장르가 상당한 영향을 미칠 것으로 생각되어 장르에 관한 분석을 진행하였다.

4. *일별데이터 분석 결과*

4.1 누적관객수/총관객수 분석

개봉 후 일수에 따른 영화의 누적관객수/총관객수를 분석한 결과는 다음과 같다. 개봉 후 5 일 부근에서 가장 급격한 상승곡선을 보이며 그 이후에는 그래프의 모양이 완만하게 상승한다. 개봉 후 약 일주일 안에 평균적으로 그 영화의 70% 정도의 관람객이 동원되는 것을 알 수 있다. 총 관객수의 대부분이 영화를 관람하는 시기가 일주일이라는 점을 미루어 봤을 때 영화의 흥행이 일주일 안에 결정이 난다는 통설이 설득력 있음을 알 수 있다.

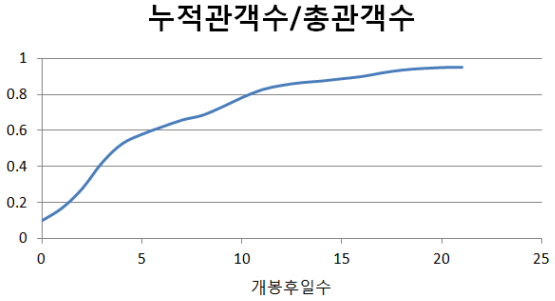

그림 13. 개봉 일 수에 따른 누적관객수/총관객수

4.2 개봉 후 일수 별 관객수 분석

개봉 후 5 주동안 일수 별 관객수 총합의 변화를 분석한 결과는 다음과 같다. 먼저 일주일 단위로 관객수의 총합이 줄어든다. 특히 첫 주에서 둘째 주로 넘어갈 때 가장 많이 떨어진다는 것을 알 수 있다.

이 분석에서 주목해야 할 점은 바로 일주일 단위로 그래프의 모양이 비슷하다는 것이다. 그래프는 개봉 후 3~4 일, 10~11 일, 17~18 일에 값이 급격히 증가하는 양상을 보인다. 그 이유는 기존 목요일 개봉 관례를 깨고 경쟁 영화를 선점하기 위한 배급사들과 제작사들이 하루 일찍 개봉해 경쟁작과의 경쟁에서 우위를 차지하기 위한 전략이자 영화관람 지원 정부 정책이 집중되는 '문화가 있는 날'이 수요일인 점을 겨냥하여 그 효과를 극대화하기 위한 목적으로 수요일에 영화를 개봉하는 추세 때문이다. 실제로 영화진흥위원회가 조사한 '2016 년 한국영화 동반성장 이행협약 모니터링 보고서'에 따르면 지난해 흥행순위 상위 30 편 중 25 편이 목요일 개봉 관례를 깨고 수요일에 개봉했다. 이렇듯 대부분의 영화가 수요일에 개봉한다고 가정하면 개봉 후 3~4 일, 10~11 일, 17~18 일은 주말이 되는데, 주말에 영화관람이 압도적으로 증가하는 것은 자명하다.

따라서 요일 별로 총 관람객 수의 변동이 너무 크기 때문에 개봉 후 일수 별로 분석을 진행할 때 관람객수에 관한 데이터를 사용하는 것은 적절치 못하다고 판단하여 다음에 소개할 개봉 후 일주일/이주일 동안의 변화 분석에서는 매출액점유율 데이터를 사용하였다.

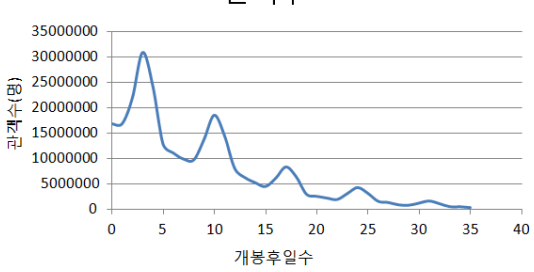

그림 14. 개봉일수에 따른 관객수 변화

4.3 개봉 후 첫째 주, 둘째 주의 일주일관객수/누적관객수 비교 분석

흥행 정도가 다른 영화 별로 개봉 후 첫째 주, 둘째 주의 일주일관객수/누적관객수 비율을 구하고 두 값을 비교한 값, 즉 (둘째 주의 일주일관객수/누적관객수)/(첫째 주의 일주일관객수/누적관객수)을 구하여 다음 표로 정리하였다. 흥행 정도에 따른 영화 분류는 누적관객수 300 만 명 이상인 영화, 50 만 명 이상 300 만 명 미만인 영화, 50 만 명 미만인 영화로 기준을 잡아 나누었다. 분류 결과 300 만 명 이상인 영화는 16 개, 50 만 명 이상 300 만 명 미만인 영화는 67 개, 50 만 명 미만인 영화는 185 개가 나왔다.

첫째 주 대비 둘째 주의 일주일관객수/누적관객수 비율이 관람객이 줄어들수록 작아짐을 확인할 수 있다. 즉, 흥행을 한 영화일 수록 관객수 비율이 개봉 후 둘째 주에도 어느 정도 유지된다는 것을 확인하였다. 그 추이를 다음 그래프로 정리하였다.

	300 만 명 이상	50 만 명 이상 300 만 명 미만	50 만 명 미만
1 주차	0.257	0.356	0.462

2 주차	0.618	0.749	0.809
2 주차 /1 주차	2.40	2.10	1.75

표 6. 영화 총 관객수 주차 별 관객수 비율

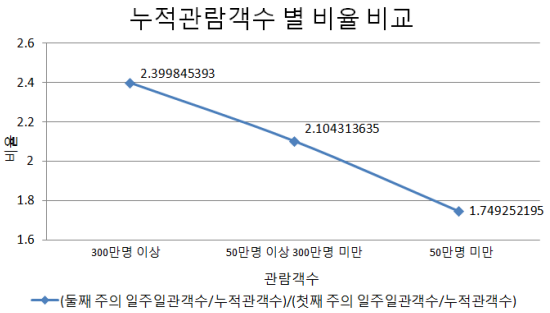

그림 15. 영화 총 관객수에 따른 누적관람객수 별 비율 비교

4.4 일별 매출액점유율 변화 분석

개봉 후 일주일/이주일 동안의 일별 매출액점유율 변화를 분석하였다. 마찬가지로 영화의 흥행 여부에 따른 기간 내의 추이를 살펴보기 위해 데이터를 누적관객수 300 만 명 이상인 영화, 50 만 명 이상 300 만 명 미만인 영화, 50 만 명 미만인 영화로 나누어 비교하였다. 앞서 언급한 바와 같이 개봉 후 일수 별로 분석을 진행할 때 관람객수 데이터를 사용하는 것은 적절치 못하므로 매출액점유율 데이터를 사용하였다. 그 결과는 아래 그래프로 정리하였다.

예상했던 바와 같이 매출액점유율 그래프의 추세선의 기울기가 관객수가 적어질수록 점점 줄어드는 모습을 보였다. 일주일 동안 분석한 누적관객수 300 만 명 이상의 그래프 기울기만 양수이고 300 만 명 미만의 그래프의 기울기는 음수라는 것도 확인할 수 있다. 이주일 동안 매출액점유율을 분석한 그래프에서 세 그래프 모두 6-7 일을 기점으로 그래프의 모양이 급격히 하락하는 것을 보아 영화의 흥행은 일주일 안에 결정 난다는 것을 다시 한번 확인할 수 있었다.

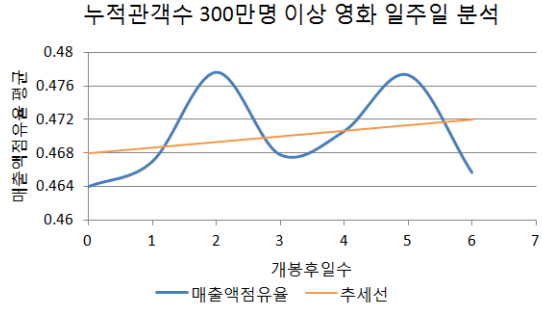

표 7. 누적관객수 300 만 명 이상 영화 개봉 후 기간에 따른 관객수 변화(1 주차)

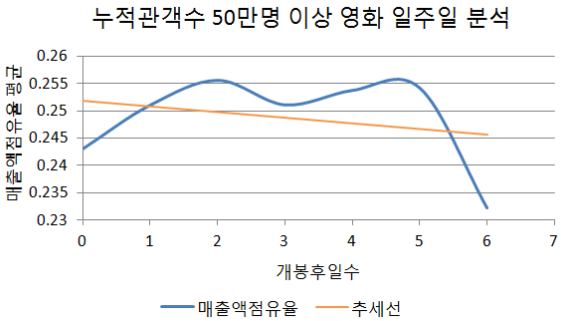

표 8. 누적관객수 50 만명 이상 영화 개봉 후 기간에 따른 관객수 변화(1 주차)

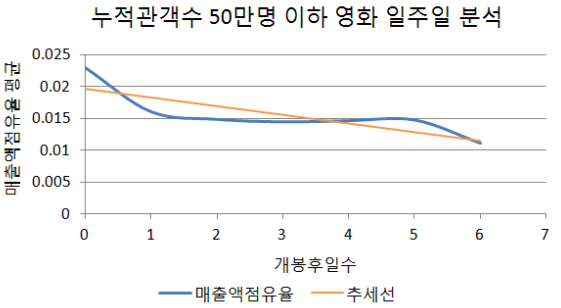

표 9. 누적관객수 50 만명 이하 영화 개봉 후 기간에 따른 관객수 변화(1 주차)

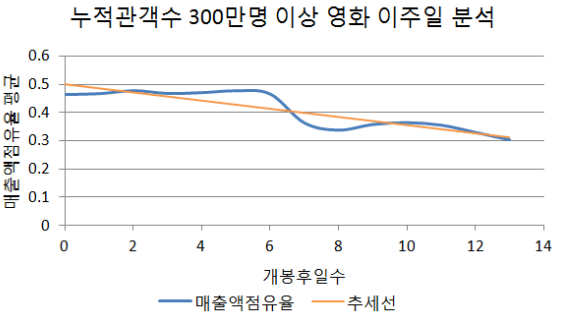

표 10. 누적관객수 300 만명 이상 영화 개봉 후 기간에 따른 관객수 변화(2 주차)

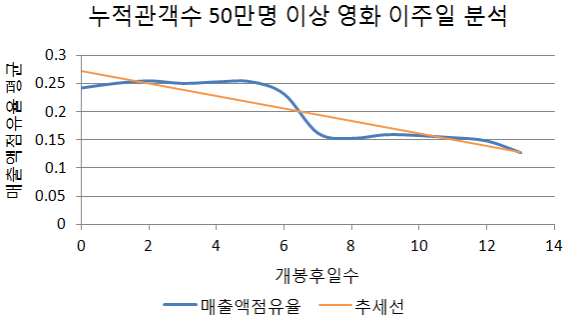

표 11. 누적관객수 50 만명 이상 영화 개봉 후 기간에 따른 관객수 변화(2 주차)

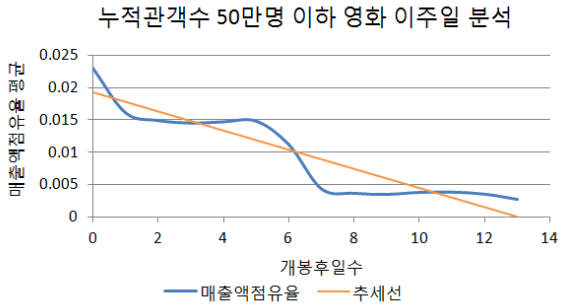

표 12. 누적관객수 50 만명 이하 영화 개봉 후 기간에 따른 관객수 변화(2 주차)

4.5 새로운 영화가 개봉했을 때, 기존에 상영중인 영화의 분석

4.5.1 누적관객/총 관객에 따른 분석

누적관객/총 관객은 특정 날짜까지의 누적관객수가 총 관객수 중에 비율을 얼마나 차지하는지 나타내기 위해 만든 column 이다. (0 에 가까울수록 개봉한지 얼마 되지 않은 영화이고 1 에 가까울수록 개봉한지 오래된 영화이다.)

기존영화의 누적관객/총 관객의 범위에 따라 기존영화의 점유율평균을 분석하였다. 결과는 표 13 과 같다. 표에서 0.1 이면 누적관객/총 관객의 수치가 0~0.1 사이에 있는 데이터들의 점유율평균을 계산한 것이다. 표 13 에 따르면, 누적관객/총 관객<=0.6 이면 이전주에 비해 점유율이 증가한다. 또한 장르별로 누적관객/총 관객이 0.6~0.7 가 되는 개봉 후 일수 평균을 구해보았을 때, 대부분 7~8 일정도였으므로 개봉 후 7~8 일까지는 점유율이 증가한다. 이는 총관객수 중 60%가 영화를 보기 전까지 혹은 개봉 후 약 7~8 일 전까지는 다른 영화의 개봉에 기존 영화의 점유율이 영향을 거의 받지 않는다는 것을 뜻한다.

	가능한 누적관객/총관객	점유율평균
0	0.1	42.449585
1	0.2	6.048831
2	0.3	30.433310
3	0.4	0.318584
4	0.5	33.237211
5	0.6	4.921015
6	0.7	-6.540750
7	0.8	-27.194736
8	0.9	-41.386043
9	1.0	-69.059133

표 13. 누적관객/총관객에 따른 점유율변화

4.5.2 장르에 따른 기존영화의 점유율변화 분석

장르는 영화의 흥행에 상당한 영향을 미친다. 따라서 기존영화/개봉영화의 장르에 따라 기존영화의 점유율이 어떻게 변화하는지 분석한다. 결과는 표 14 와 같다. 왼쪽은 개봉 영화의 장르에 따른 기존 영화의 점유율 변화, 오른쪽은 기존 영화의 장르에 따른 기존 영화의 점유율 변화를 나타낸다.

표 14 에서 SF 가 개봉했을 때 기존영화의 점유율 변화가 가장 적고, 사극이 개봉했을 때 가장 크다. 또한 특정 영화가 개봉했을 때, 멜로/로맨스영화의 점유율변화가 가장 적고, SF 의 점유율 변화가 가장크다. 이는 위에서 분석하였던 장르별 누적관객/총관객이 0.6~0.7 이 되는 개봉 후 일수의 결과에서 멜로/로맨스는 13 일 SF 는 5.65 일인 것과 연결할 수 있다. 우리의 데이터는 개봉 후 일수가 0~21 까지만 존재한다. 맬로 로맨스는 약 13 일간 점유율변화가 천천히 떨어지므로 점유율 변화 평균이 높게 나타나고, SF 영화는 5~6 일 이후에는 점유율 변화가 급격히 떨어지므로, 점유율 변화 평균이 낮게 나타난다.

개봉장르	점유율 변화
SF	-14.962246
판타지	-28.512214
액션	-31.005245
스릴러	-31.585397
공포(호러)	-32.030370
가족	-37.134277
애니메이션	-39.579543
멜로/로맨스	-41.192822
드라마	-42.198881
어드벤처	-44.914655
범죄	-47.140494
전쟁	-47.991899
코미디	-49.577250
미스터리	-54.260230
사극	-63.461373

장르	점유율 변화
멜로/로맨스	7.874992
전쟁	-10.149014
드라마	-22.840928
애니메이션	-23.018248
범죄	-29.808682
어드벤처	-30.646808
사극	-31.117584
코미디	-39.925863
공포(호러)	-42.687965
스릴러	-48.693136
미스터리	-51.076083
판타지	-52.569078
액션	-56.197812
SF	-61.047030

표 14. 장르에 따른 기존영화의 점유율변화

개봉장르와 기존영화의 장르를 합쳐서 이에 따른 기존 영화의 점유율변화를 비교해보면 표 15 와 같다.

이 표를 보고 특정 장르의 영화를 개봉할 때 개봉시기를 잘 정해야 한다는 점을 알 수 있다. 예를 들어 액션영화가 개봉했을 때, 기존영화의 점유율이 적게 떨어진다. 또한 액션영화는 다른 장르가 개봉했을 때, 점유율이 가장 많이 떨어진다. 따라서 액션 영화는 다른 영화의 개봉에 영향을 많이받지만 개봉시 기존 영화에 영향을 미치는 부분은 적으므로, 개봉시기를 잘 선택해야한다는 결론을 얻을 수 있다.

개봉영화와 기존영화의 장르가 같을 경우도 생각해볼 수 있다. 직관적으로 기존에 있는 장르와 같은 장르의 영화가 개봉하면 기존에 있는 영화의 점유율이 많이 떨어질 것이라고 생각할 수 있다. 이 분석의 결과는 표 16 에 나타난다. 앞의 예측과는 달리 장르에 따라 분포가 다양하게 나타났다. 멜로/로맨스나 어드벤처장르 같은 경우 같은 장르가 개봉해도 점유율 변화가 거의 없으나 SF 나 공포장르는 같은 장르가 개봉하면 점유율 변화가 매우 크다. 따라서 SF 나 공포, 미스터리장르는 주변에 같은 장르의 같은 영화가 개봉하는지 확인 후 개봉날짜를 결

정할 필요가 있다.

개봉장르	장르	점유율 변화
애니메이션	드라마	-13.575622
액션	애니메이션	-16.511996
	어드벤처	-17.050419
애니메이션	애니메이션	-19.083384
액션	드라마	-20.367237
	범죄	-25.551471
드라마	애니메이션	-26.993366
	드라마	-31.831875
액션	공포(호러)	-33.256551
	액션	-43.747305
	코미디	-43.829359
애니메이션	액션	-51.861571
액션	스릴러	-54.716614
어드벤처	액션	-56.695310
드라마	액션	-62.042787
코미디	액션	-73.412582
범죄	액션	-74.243600

표 15. 개봉장르와 기존장르에 따른 점유율 변화

개봉장르	점유율 변화
멜로/로맨스	-3.861490
어드벤처	-4.028883
애니메이션	-19.083384
드라마	-31.831875
스릴러	-35.164452
액션	-43.747305
코미디	-44.653494
미스터리	-57.142857
공포(호러)	-85.420895
SF	-91.883797

표 16. 개봉/기존장르가 같을 때 점유율 변화

4.5.3 총 관객수, 기존영화의 순위에 따른 점유율 변화

현재 기존의 영화가 한창 흥행 중 일 때, 다른 영화가 개봉해도 점유율변화가 적을 것이라는 예측을 할 수 있다.

표 17 은 기존영화의 총관객수에 따른 매출액 점유율 변화를 나타낸다. 표 18 은 기존영화의 순위에 따라 점유율 변화를 나타낸다.

앞의 예측과 동일한 결과를 얻을 수 있었다. 총 관객수가 많을수록, 개봉한 영화가 개봉하기 전 기존 순위가 높을수록 그 영화가 현재 흥행 중이라고 볼 수 있다. 흥행한 영화일수록 점유율변화가 적으므로, 흥행한 영화는 다른 영화의 개봉에 크게 영향을 받지 않는다고 볼 수 있다

	매출액점유율평균	총관객수
0	-1.658467	3000000
1	-14.024689	2000000
2	-23.967681	1000000
3	-33.452668	500000

표 17. 기존영화의 총관객수와 점유율 변화

	순위	점유율 변화
0	1.0	37.454684
1	2.0	1.219767
2	3.0	-17.564985
3	4.0	-28.405094
4	5.0	-31.777130
5	6.0	-41.812294
6	7.0	-51.060731

표 18. 기존영화의 순위와 점유율 변화

4.6 새로운 영화가 개봉했을 때, 개봉영화의 일주일 매출액 점유율 평균 분석

특정 영화가 개봉했을 때, 그 영화의 일주일 점유율 평균은 약 20%이다. 만약에 기존 영화에 같은 장르가 존재할 때 점유율 평균을 구하면, 16%이다. 따라서 기존 영화에 앞으로 개봉할 영화와 같은 장르가 있다면 없는 경우보다 점유율을 적게 올릴 확률이 크다.

개봉장르에 따라 일주일 점유율 평균을 구한 것이 표 19 이다. 이 표에 따르면, 초반에 많은 점유율을 올려야 하는 SF,전쟁 장르의 영화가 점유율이 높음을 알 수 있다. 앞의 결론에서 영화의 흥행여부는 개봉 후 7~8 일이 결정할 수 있으므로, 일주일 점유율 평균은 크면서 누적관객/총 관객이 0.6 이되는 지점은 느리게 만들 수 있는 장르가 수치상 흥행을 할 수 있는 최적의 장르가 될 것이다.

총관객수가 100 만 이상인 영화들을 대상으로 기존 영화장르와 개봉영화 장르에 따른 일주일 매출액 평균을 나타낸 것이 표 20 이다. 총관객수가 100 만 이상인 영화들을 선택한 이유는 우리의 목표가 흥행이기 때문에 어느 정도 흥행한 영화들을 대상으로 했다.

이를 통해 기존장르에 어떤 장르가 있을 때, 어떤 장르의 영화가 개봉하는 것이 점유율 평균을 높일 수 있는지 알아볼 수 있고, 이를 통해 영화 개봉 시기를 정할 수 있다. 예를 들면 액션장르를 개봉하고 싶을 때에는 기존 영화의 장르가 드라마, 범죄, 액션인 시기를 선택하면 된다. 현재는 표본의 수가 적어 이 표가 정확도를 신뢰할 수 없지만, 조금 더 많은 데이터를 추가한다면 신뢰할만한 결론을 얻을 수 있을 것이라 생각한다.

개봉장르	일주일매출액평균
전쟁	0.367857
가족	0.343583
SF	0.334391
범죄	0.321332
스릴러	0.231496
사극	0.224943
멜로/로맨스	0.223934
액션	0.218480
어드벤처	0.216378
드라마	0.206410
판타지	0.169377
코미디	0.152834
공포(호러)	0.128129
애니메이션	0.120589
미스터리	0.116359

표 19. 개봉장르와 일주일 점유율 평균

개봉장르	장르	일주일매출액평균_y
드라마	액션	0.426595
액션	드라마	0.349243
	범죄	0.349167
	액션	0.310784
범죄	액션	0.307595
액션	애니메이션	0.306667
애니메이션	액션	0.206714

표 20. 개봉장르/기존장르와 일주일 점유율 평균

4.7 경쟁작 분석 결과

우리의 초기 목표는 경쟁작 분석을 통해 어떤 영화의 개봉시기를 정할 때, 어떤 시기가 가장 적당한지 예측하는 것이었다. 우리는 개봉영화/기존영화의 장르에 주 초점을 맞춰서 적당한 개봉시기에 대해 생각해보았다. 분석을 진행하면서 생각보다 영화의 흥행에는 더 많은 요소들이 복합적으로 영향을 미친다는 것을 느꼈다. 우리의 데이터 수가 적어서 발생한 문제일 수도 있지만, 데이터 사이의 일관성을 찾기가 상당히 힘들었고, 얻은 결과를 신뢰하기 위해서는 더욱 많은 데이터와 조금 더 정밀한 분석방법이 필요할 것 같다. 머신러닝과 같은 기술을 이용하여 조금 더 정량화된 분석 또한 해볼 만하다고 생각한다.

3. Discussion and Conclusion

우리는 기간별 데이터를 통해 장르, 개봉시기, 감독, 배우가 영화 흥행에 영향을 미친다는 것을 확인했다. 그리고 각각의 요소끼리도 영향을 미친다는 것을 확인할 수 있었다. 따라서 이를 통해 영화 제작사나 감독들은 흥행을 위해 고려해야 할 요소들과 마케팅적인 전략 등을 수립할 수 있을 것이라 생각한다. 그러나 우리는 제작비에 관한 자료가 없어 흥행 영화를 관객수 300 만으로 잡았는데 제작비를 통해 흥행여부를 결정하고, 제작비와 영화의 상관관계를 조사한다면 더 정확한 흥행 영화에 대한 고찰이 이루어졌을 것이다. 또한 초반 스크린 수에 의해 관객수가 많이 변동되었을 것으로 예상되는데 이를 고려하지 않은 상태로 데이터 분석을 진행하였다. 이를 고려하여 데이터를 분석하고, 추가적으로 초반 스크린 수를 통해 홍보비용에 따른 영화 관객수의 변화 등을 알아본다면 더 정확한 정보를 제공할 수 있을 것이다.

또한 이렇게 분석한 요소들을 가지고 각각의 가중치를 부여할 수 있는 머신 러닝 기법을 통해 깊이 분석하게 된다면 한층 더 정확하고 중요한 데이터를 제공할 수 있을 것이라 생각된다.

우리의 데이터 중 배우 분석을 보면 배우의 경우 여자 배우가 순위권에 거의 존재하지 않는 것을 확인할 수 있다. 배우의 남, 녀에 따른 추가적인 분석이 진행된다면 재미있는 분석 결과가 나올 것으로 예상된다.

지금까지는 한국 내에서의 영화 흥행에 대해서 알아보았다. 그러나 현재 한국은 영화 포화 시장으로 더 많은 이윤창출을 위해 영화 시장 범위를 넓힐 필요가 있다. 이 보고서에서 서술한 방식과 유사하게 해외의 데이터를 통해 해당 나라의 영화 분석과 해외에서의 흥행한 한국 영화 분석 등이 이루어진다면 지속가능한 영화 산업이 이루어질 것이다.