국내 역대 박스오피스 영화 데이터 분석 프로젝트

박하영

01 프로젝트 개요
프로젝트에 대한 간단한 소개

02 데이터 수집(크롤링)
목표 데이터와 데이터 수집 과정

03 데이터 분석
데이터 분석 결과와 해석

01

프로젝트 개요

01 프로젝트 개요

서버구축 , 데이터 베이스

AWS, 리눅스 환경 이용해서 서버 생성
MongoDB 데이터 베이스 구축

데이터 크롤링

http://www.kobis.or.kr/
영화관입장권통합전산망 페이지에서
영화정보 데이터 크롤링

데이터 분석

Python을 활용하여 데이터 분석

전체 과정에 대한 코드는 아래 URL에서 확인할 수 있습니다.

https://github.com/hayoungmon/Crawling-EDA-Movie-data-

02

데이터 수집(크롤링)

01 크롤링 목표

1.역대 국내 개봉영화
2.연도별 역대 박스오피스 50위권 (2004 - 2019년까지)

위 조건을 만족하는 영화의 각종 데이터 수집

02-1 크롤링 전략

박스오피스

박스오피스
- 일별
- 주간/주말
- 월별
- 연도별
- 역대
- 기간별

해외 박스오피스
예매율
좌석점유율
상영점유율
스크린점유율
체인영화관별 상영현황

연도별 박스오피스

박스오피스 › 박스오피스 › 연도별

- [박스오피스-연도별 박스오피스]코너는 2004년 이후 연도별 영화상영관 연동율에 따라 수집된 발권데이터를 집계한 것으로, 과거 연도별 박스오피스는 [공식통계-연도별 박스오피스] 코너를 참고하시기 바랍니다. (재개봉 등 누적상영에 따른 수치의 변동이 있을 수 있음)
- 이월작의 경우 개봉년도와 상영년도로 구분되어 관객수 및 매출액이 나누어 조회되므로, 최종 누적스코어 등은 해당영화의 상세정보 클릭 후 팝업창의 "통계정보"를 참고하시기 바랍니다.
- 매일 24시 이후 전환/제공되는 [전일자 통계정보]는 상영마감 및 보정처리 등의 사유로 익일 오전까지 계속 업데이트 되며, 일마감 후 데이터보정 등의 사유로 통계정보는 변동 될 수 있음을 참고하시기 바랍니다.
- 통계이용안내
  ① [박스오피스], [테마통계]코너는 연도별 영화상영관 연동율에 따라 실시간 수집된 발권데이터를 전일기준까지 반영한 통계정보입니다.
  ② [공식통계]코너는 영진위에서 매년 발표하는 "한국영화연감"의 영화별 흥행기록을 참고한 것입니다.
  - 한국영화연감(1971~2010) 통계를 기준으로 정리한 것이며, 2011년부터는 통합전산망을 기준으로 일정한 주기(매월, 매년)로 마감 처리하여 산출되는 통계정보입니다.
  - 통계마감 주기(일별, 년별)에 따라 공식통계 수치는 추후 변동될 수 있습니다.

엑셀

조회기간 2019 · 영화구분 --전체-- · 국적 --전체-- · 지역 --전체-- 조회

총 50건

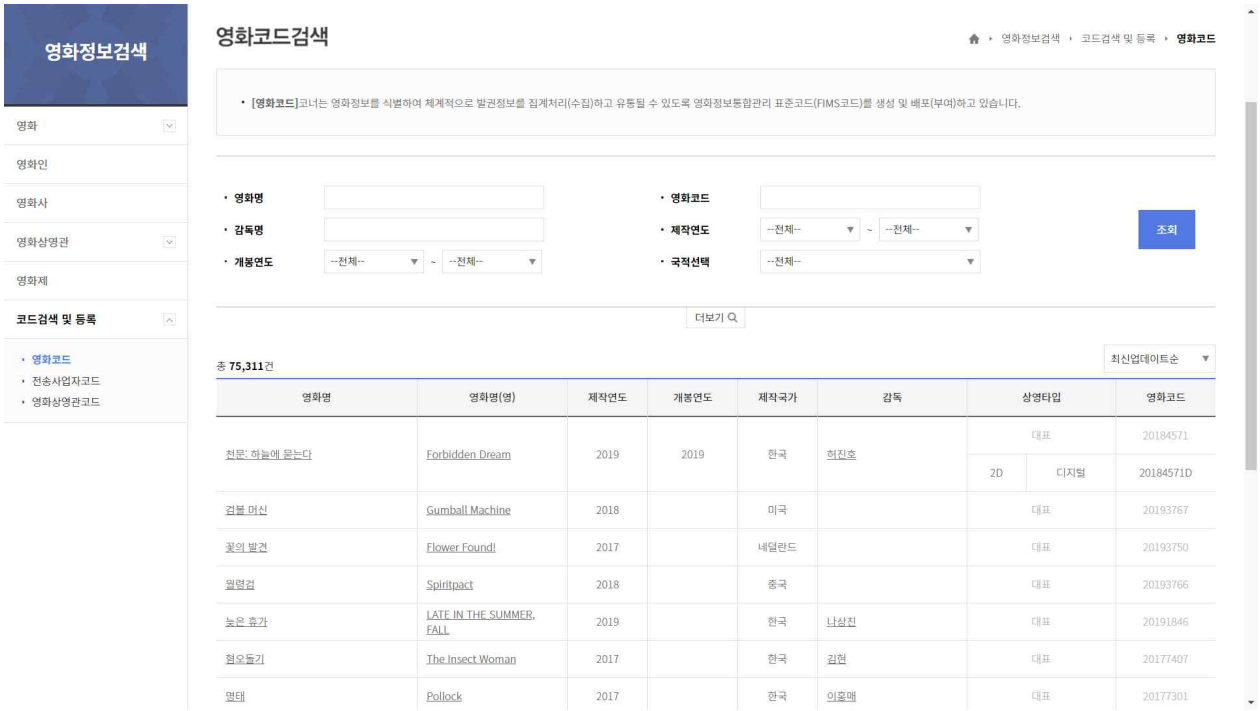
순위	영화명	개봉일	매출액	매출액 점유율	관객수	스크린수	상영횟수
1	극한직업	2019-01-23	139,651,845,516	7.4%	16,265,618	2,003	292,584
2	어벤져스: 엔드게임	2019-04-24	122,182,694,160	6.5%	13,934,592	2,835	242,001
3	겨울왕국 2	2019-11-21	109,107,997,490	5.8%	13,074,053	2,648	275,800
4	알라딘	2019-05-23	106,955,138,359	5.7%	12,552,283	1,409	266,469

"http://www.kobis.or.kr/kobis/business/stat/boxs/findYearlyBoxOfficeList.do"

영화이름 , 개봉시기 , 매출 , 관객 정보 수집

audience	date	sales	title	year
1,025,817	2019-12-11	8,555,856,330	쥬만지: 넥스트 레벨	2019

02-2 크롤링 전략



"http://www.kobis.or.kr/kobis/business/mast/mvie/searchUserMovCdList.do"

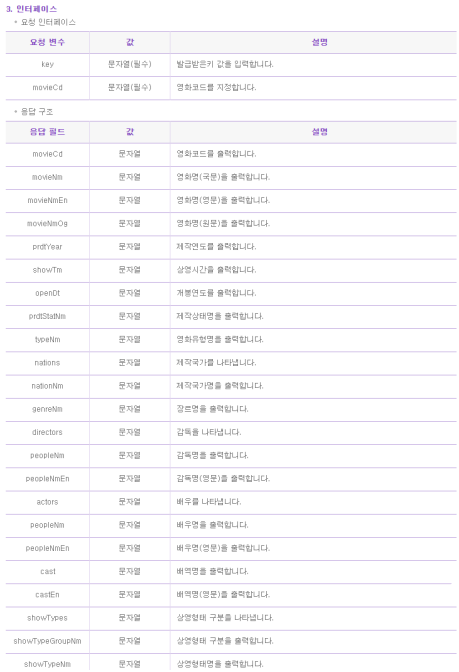
앞에서 수집한 영화이름 , 개봉시기를 파라미터로 넘겨서 영화 코드 수집

code	title
20198681	쥬만지: 넥스트 레벨

02-3 크롤링 전략

http://www.kobis.or.kr/kobisopenapi/homepg/apiservice/searchServiceInfo.do

Kobis에서 제공하는 api 서비스를 이용 → 앞에서 얻은 영화코드를 이용해 영화 정보 수집



3. 인터페이스

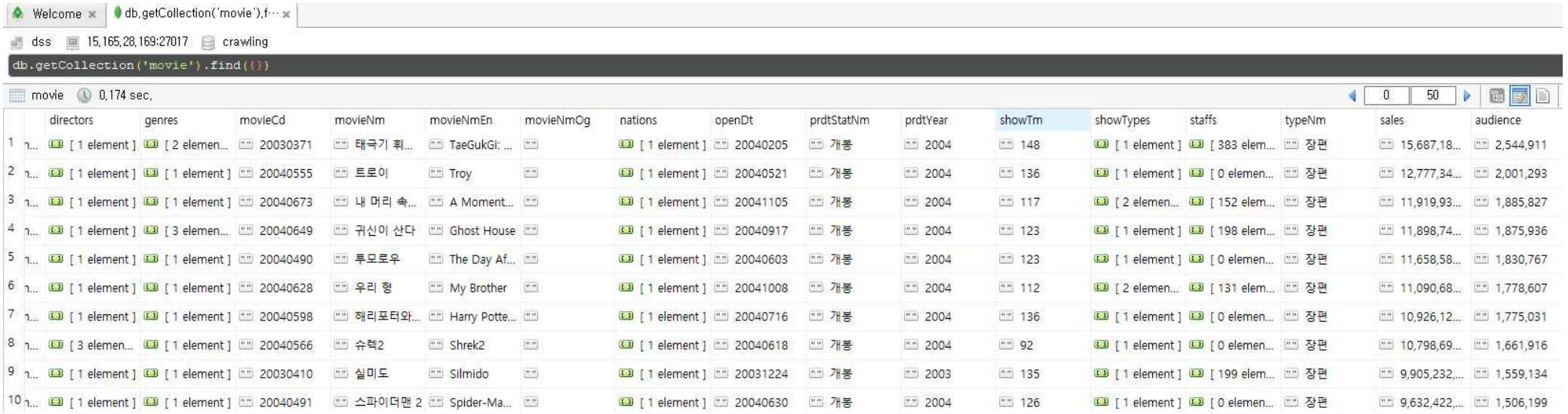
- 요청 인터페이스

요청 변수	값	설명
key	문자열(필수)	발급받은키 값을 입력합니다.
movieCd	문자열(필수)	영화코드를 지정합니다.

- 응답 구조

응답 필드	값	설명
movieCd	문자열	영화코드를 출력합니다.
movieNm	문자열	영화명(국문)을 출력합니다.
movieNmEn	문자열	영화명(영문)을 출력합니다.
movieNmOg	문자열	영화명(원문)을 출력합니다.
prdtYear	문자열	제작연도를 출력합니다.
showTm	문자열	상영시간을 출력합니다.
openDt	문자열	개봉연도를 출력합니다.
prdtStatNm	문자열	제작상태명을 출력합니다.
typeNm	문자열	영화유형명을 출력합니다.
nations	문자열	제작국가를 나타냅니다.
nationNm	문자열	제작국가명을 출력합니다.
genreNm	문자열	장르명을 출력합니다.
directors	문자열	감독을 나타냅니다.
peopleNm	문자열	감독명을 출력합니다.
peopleNmEn	문자열	감독명(영문)을 출력합니다.
actors	문자열	배우를 나타냅니다.
peopleNm	문자열	배우명을 출력합니다.
peopleNmEn	문자열	배우명(영문)을 출력합니다.
cast	문자열	배역명을 출력합니다.
castEn	문자열	배역명(영문)을 출력합니다.
showTypes	문자열	상영형태 구분을 나타냅니다.
showTypeGroupNm	문자열	상영형태 구분을 출력합니다.
showTypeNm	문자열	상영형태명을 출력합니다.

02-4 데이터베이스에 저장


Welcome × db.getCollection('movie'),f··· ×

dss 15,165,28,169:27017 crawling

```
db.getCollection('movie').find({})
```

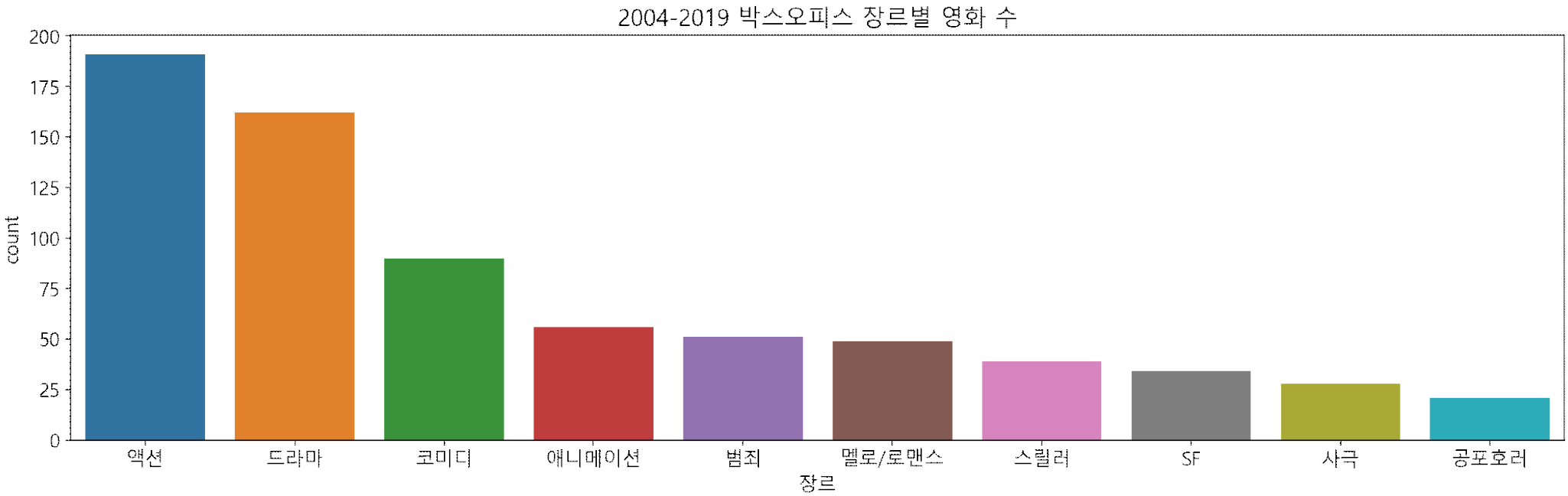
movie 0.174 sec. 0 50

	directors	genres	movieCd	movieNm	movieNmEn	movieNmOg	nations	openDt	prdtStatNm	prdtYear	showTm	showTypes	staffs	typeNm	sales	audience
1	[1 element]	[2 elemen...	20030371	태극기 휘...	TaeGukGi: ...		[1 element]	20040205	개봉	2004	148	[1 element]	[383 elem...	장편	15,687,18...	2,544,911
2	[1 element]	[1 element]	20040555	트로이	Troy		[1 element]	20040521	개봉	2004	136	[1 element]	[0 elemen...	장편	12,777,34...	2,001,293
3	[1 element]	[1 element]	20040673	내 머리 속...	A Moment...		[1 element]	20041105	개봉	2004	117	[2 elemen...	[152 elem...	장편	11,919,93...	1,885,827
4	[1 element]	[3 elemen...	20040649	귀신이 산다	Ghost House		[1 element]	20040917	개봉	2004	123	[1 element]	[198 elem...	장편	11,898,74...	1,875,936
5	[1 element]	[1 element]	20040490	투모로우	The Day Af...		[1 element]	20040603	개봉	2004	123	[1 element]	[0 elemen...	장편	11,658,58...	1,830,767
6	[1 element]	[1 element]	20040628	우리 형	My Brother		[1 element]	20041008	개봉	2004	112	[2 elemen...	[131 elem...	장편	11,090,68...	1,778,607
7	[1 element]	[1 element]	20040598	해리포터와...	Harry Potte...		[1 element]	20040716	개봉	2004	136	[1 element]	[0 elemen...	장편	10,926,12...	1,775,031
8	[3 elemen...	[1 element]	20040566	슈렉2	Shrek2		[1 element]	20040618	개봉	2004	92	[1 element]	[0 elemen...	장편	10,798,69...	1,661,916
9	[1 element]	[1 element]	20030410	실미도	Silmido		[1 element]	20031224	개봉	2003	135	[1 element]	[199 elem...	장편	9,905,232,...	1,559,134
10	[1 element]	[1 element]	20040491	스파이더맨 2	Spider-Ma...		[1 element]	20040630	개봉	2004	126	[1 element]	[0 elemen...	장편	9,632,422,...	1,506,199

03

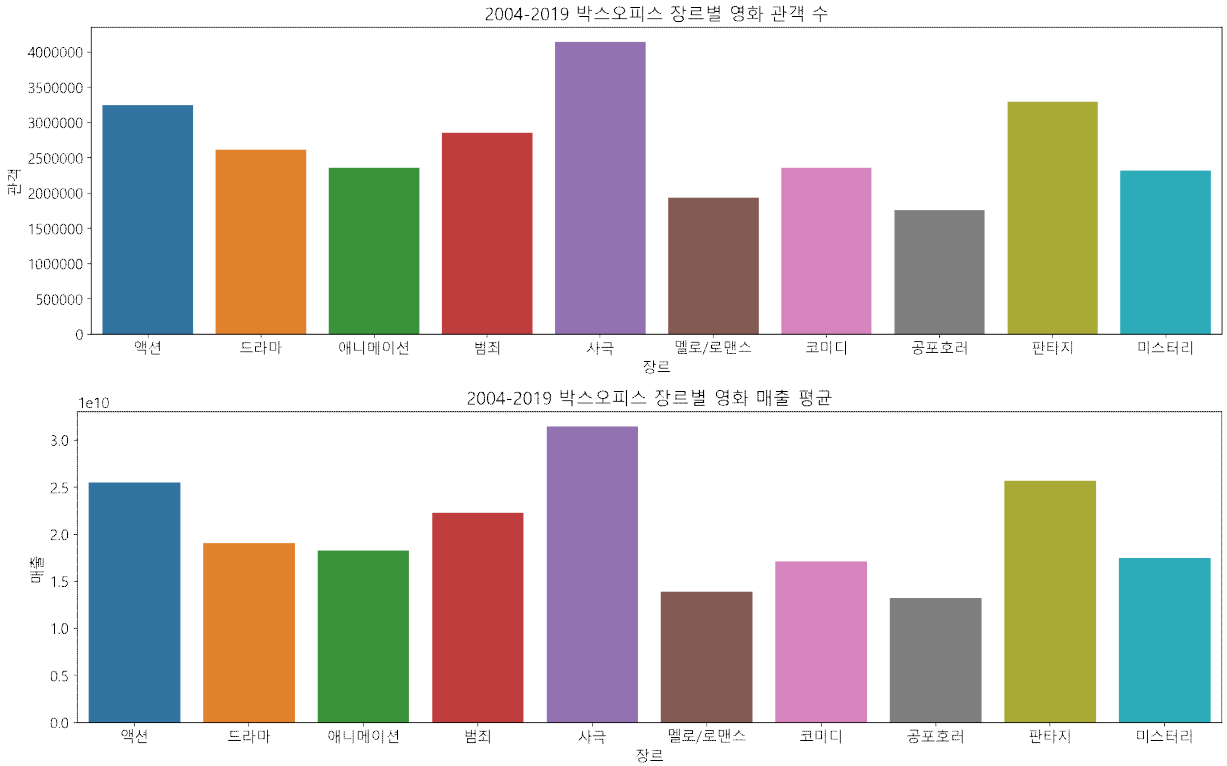
데이터 분석

03-1 장르별 분석



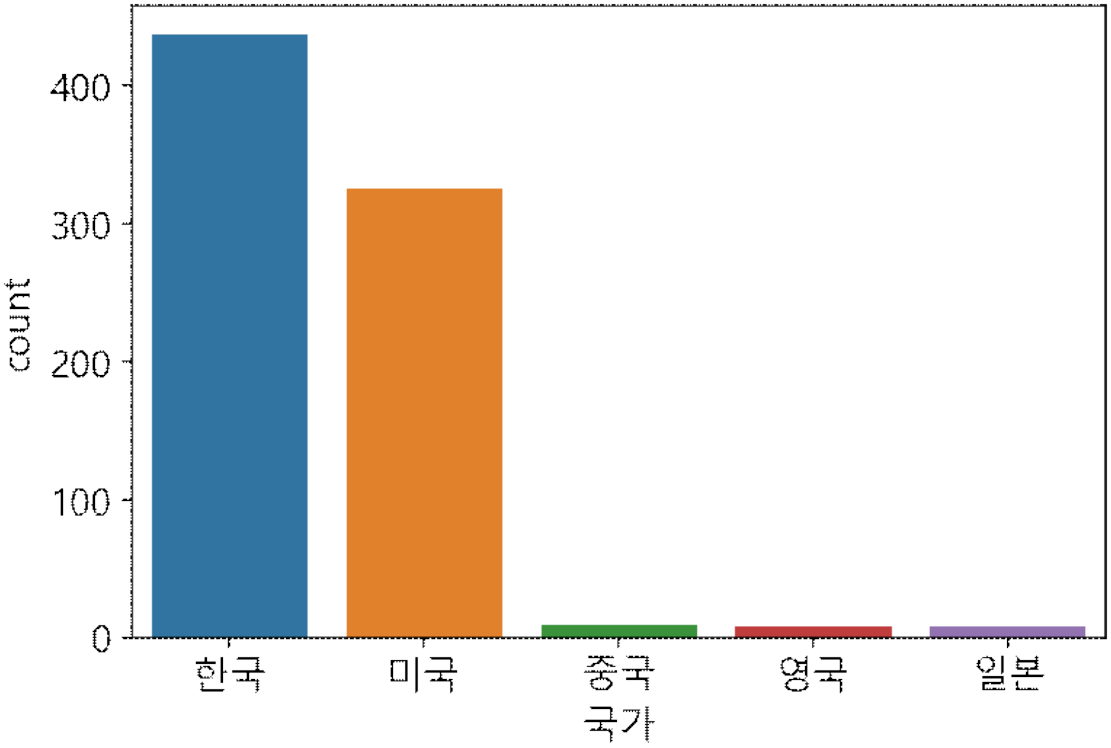
장르별로 영화를 구분해 보면 액션,드라마,코미디 등의 순으로 개수가 많다.

03-2 장르별 분석



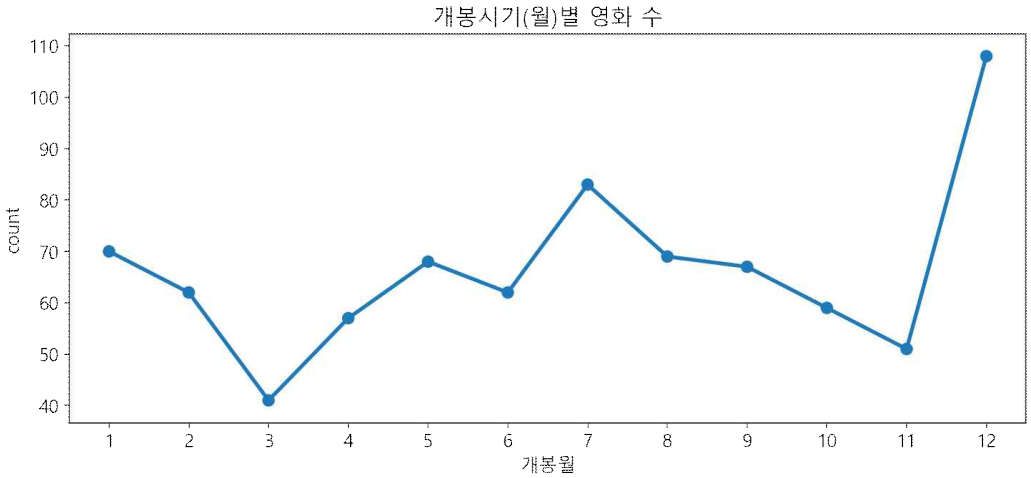
영화 개수와 평균 영화 관객,매출의 양상은 다르게 나타난다.
개수 자체는 적었지만 사극이나 판타지, 미스터리 장르의
평균 관객이나 매출이 높은 점이 눈에 띈다.

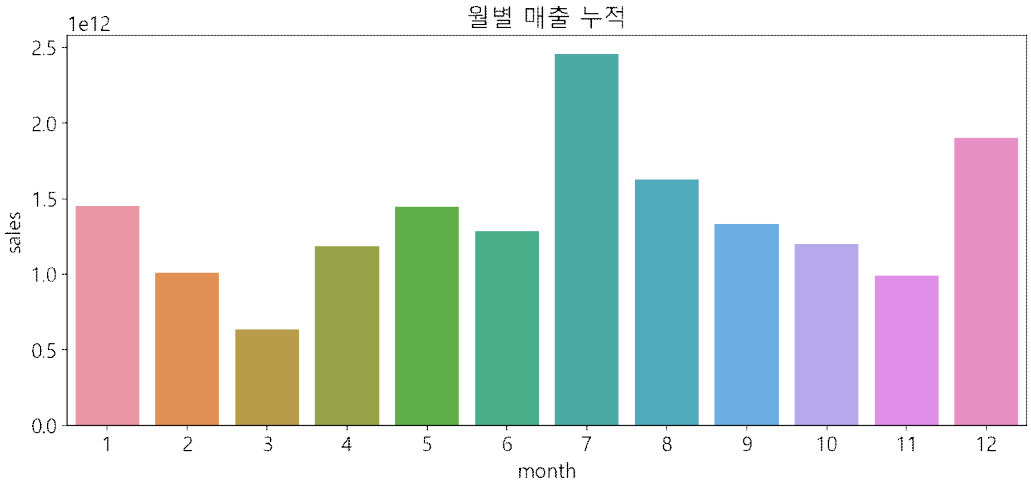
03-3 국가별 분석



국내 박스오피스이기 때문에 국내 영화가 많은 점은 예상할 수 있었다.
해외 영화의 경우 미국영화가 대다수이고 오히려 같은 동양권인 중국이나 일본영화는 국내 박스오피스에 진입한 영화가 적다. 전반적으로 해외영화를 선호하는 국내 대중의 경향을 확인할 수 있다.

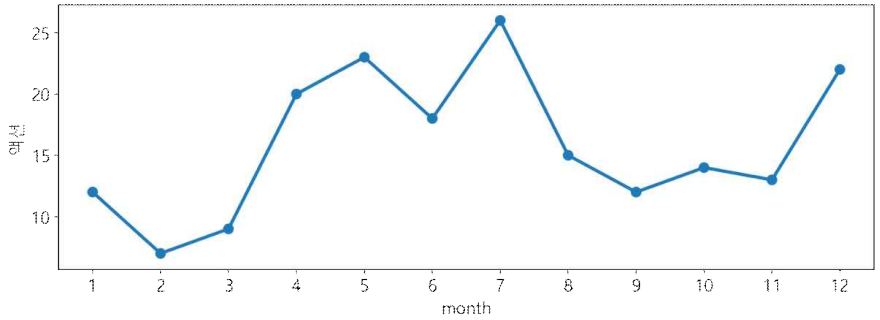
03-4 개봉시기별 분석

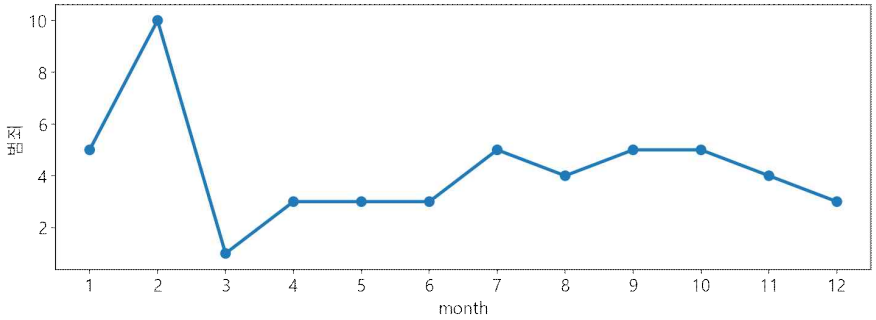


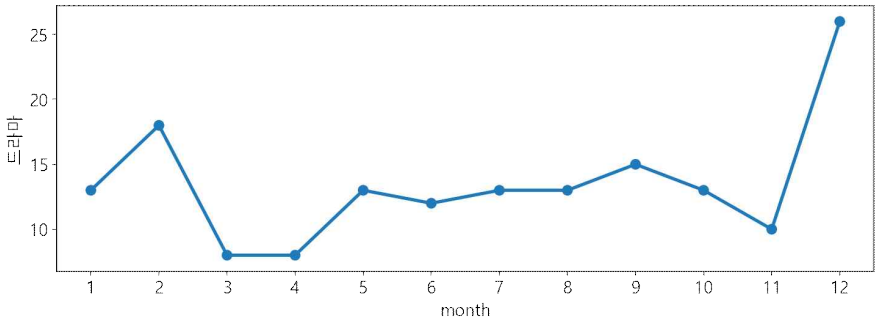


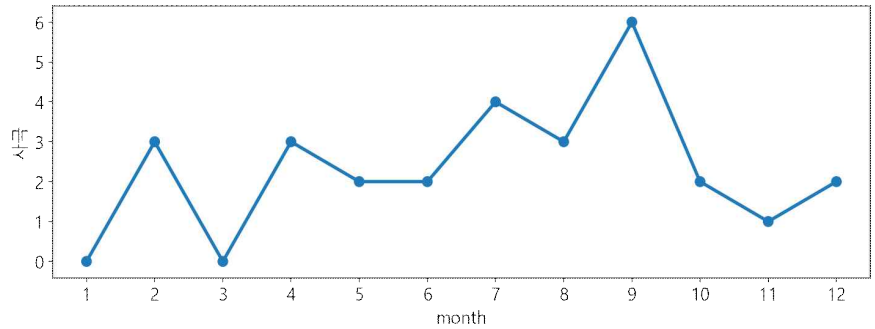
개봉시기에 따라 영화 수에 어느정도 차이가 있는 것을 확인할 수 있다.
그중 7월과 12월에 유독 많은 점이 눈에 띈다. 유추를 해보면 방학시기인 7월과, 연말 시기에 영화 수요 층 자체가 늘어나기도 하고 개봉하는 영화도 많기 때문에 박스오피스권 영화가 나올 가능성도 높아진다고 볼 수 있다.

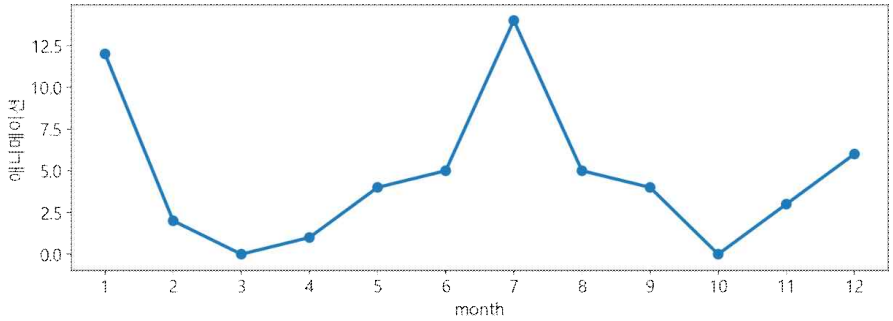
03-5 개봉시기별 분석

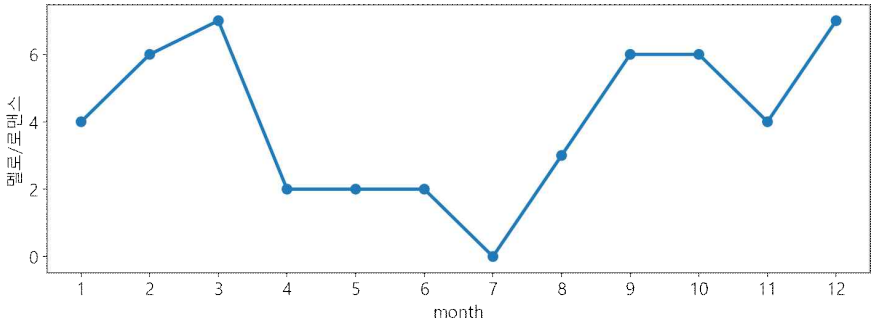




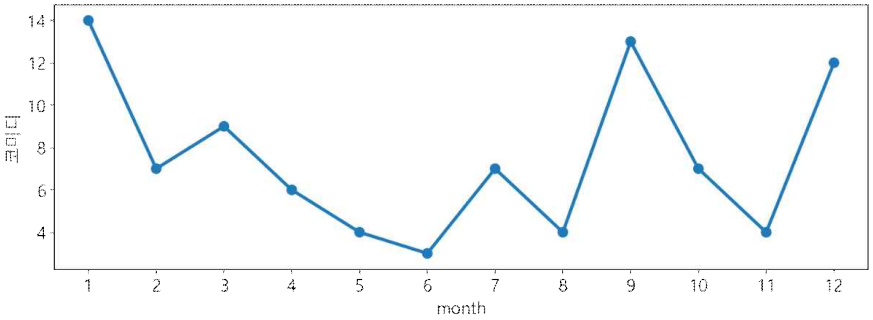


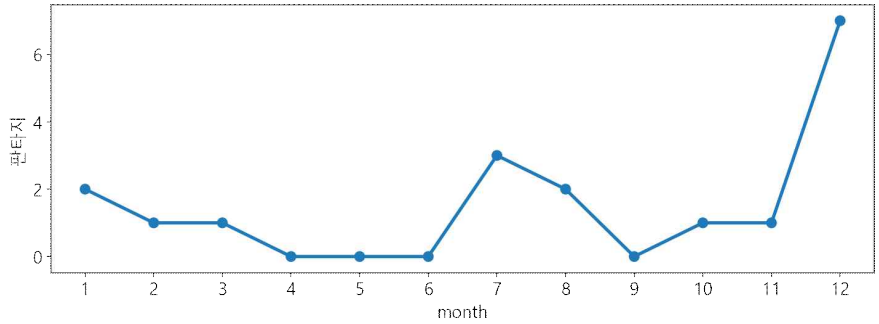


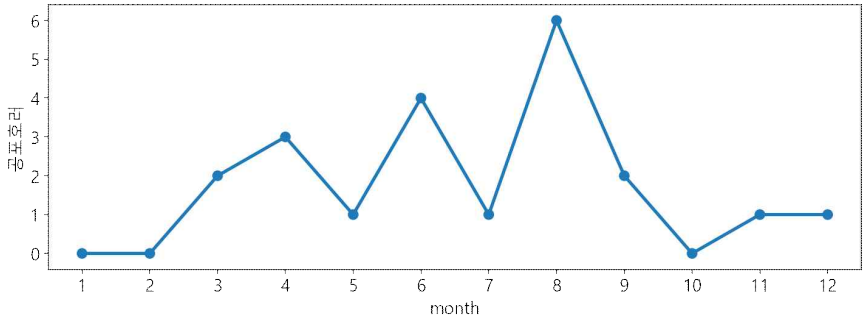


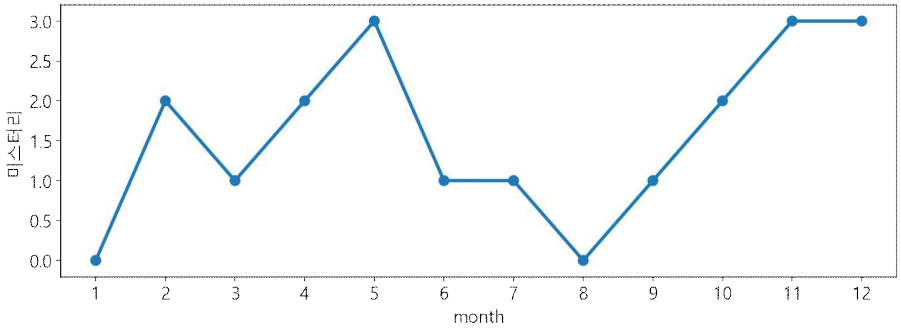


03-6 개봉시기별 분석



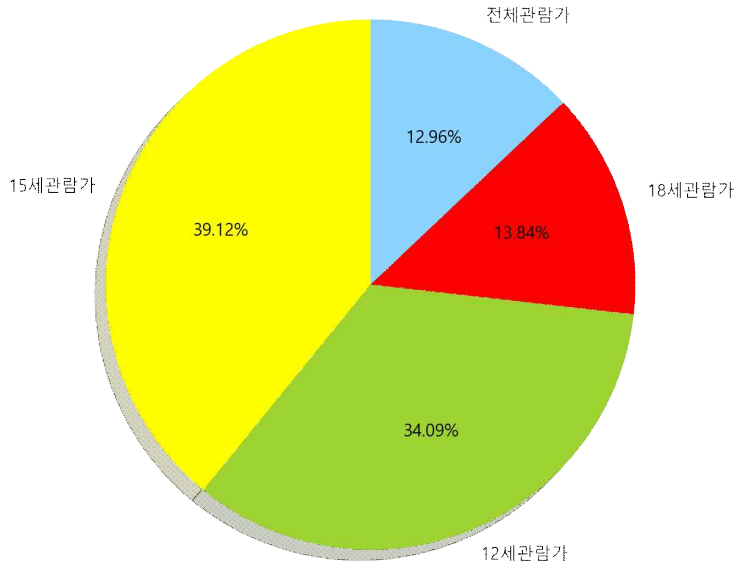






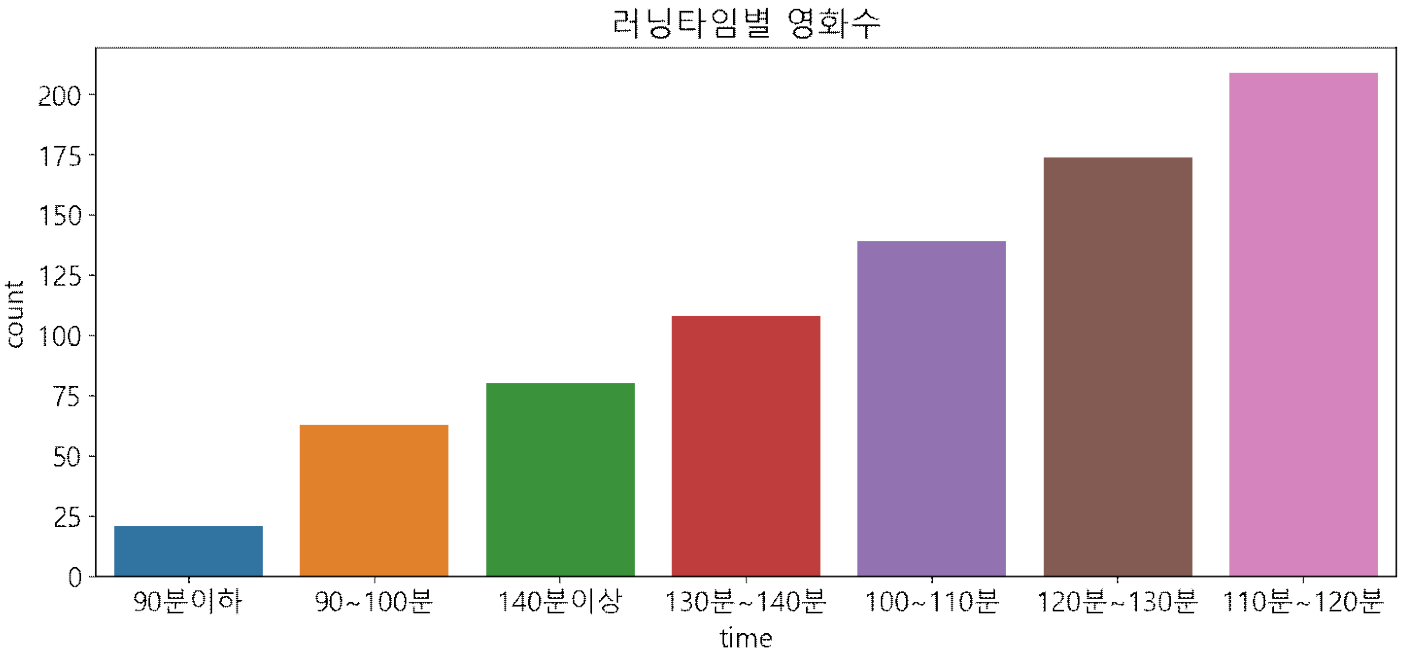
장르를 구분해서 개봉시기별 분석을 해보면 장르적 특성에 따라서 차이가 있는 것을 확인할 수 있다.
예를 들어 가족이나 저연령층이 많이 보는 애니메이션의 경우 방학시기인 7월이 많고, 분위기가 중요한
멜로,드라마,로맨스장르의 경우 겨울시기가 많은 경향을 보이는 것을 확인할 수 있다.

03-1 관람이용가 분석

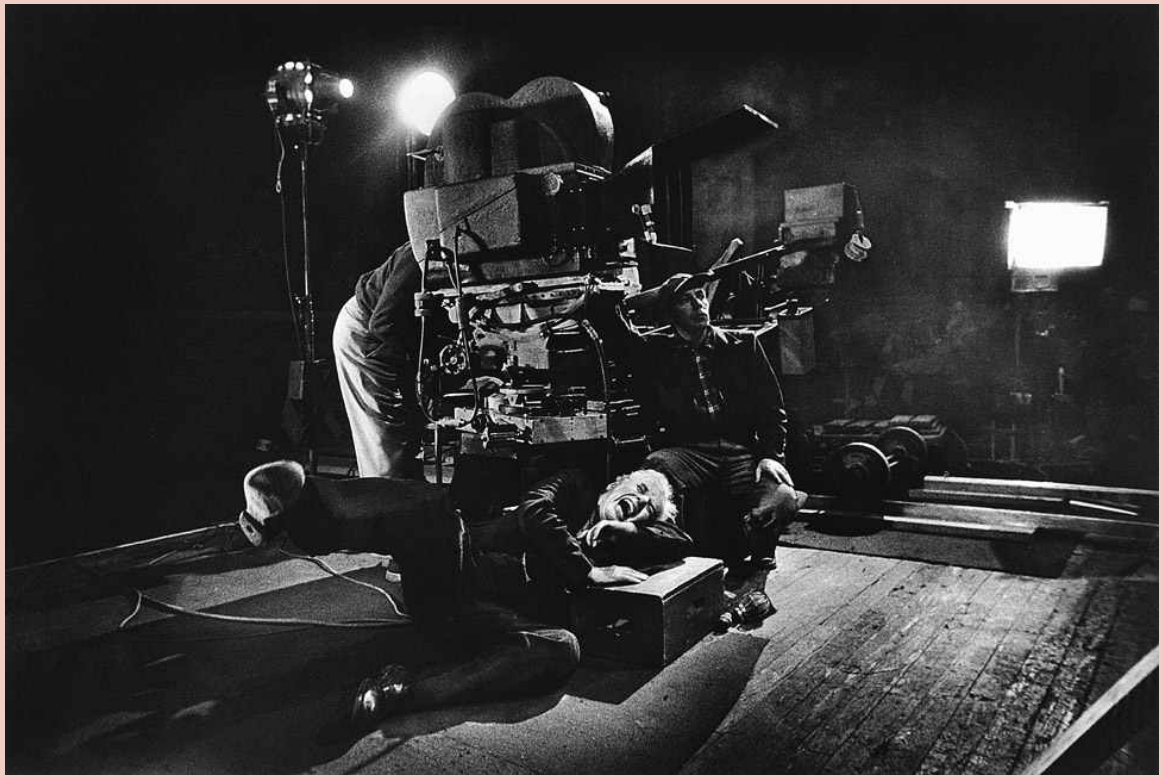


뚜렷한 특징을 나타내는 것으로 보이지는 않지만, 12,15세 관람가의 비중이 70% 정도로 가장 높고
전체관람가,18세관람가가 30%정도의 비중을 차지하고 있다. 일반적으로 개봉하는 영화의 관람이용가와 비슷한 비율을 보인다.

03-1 러닝타임별 분석



러닝 타임별로 나눠보면 120분 이하까지는 러닝타임이 길수록 박스 오피스권 영화가 많지만
그 이상의 경우 길어질 수록 줄어드는 양상을 보인다. 러닝타임도 영화의 흥행에 영향을 주는 요소로 보인다.



Thank You