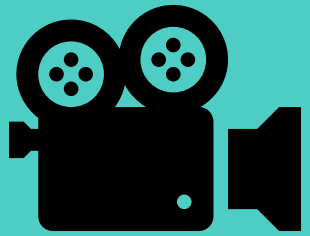
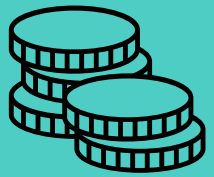
영화 매출에 영향을 주는 요인 분석





2016112564 정용희
2016112578 서은교

Index

- 주제 선정 이유
- 데이터 추출
- 전처리
- 데이터 시각화
- 분석의 결론 & 한계점

주제를 선정하게 된 이유

관객수에 영향을 주는 요인 분석

축적된 boxoffice 데이터를 기반하여 관객수에 영향을 주는 요인을 분석해서 새로 영화를 출시할 경우에 어떤 방법으로 영화를 출시하면 가장 많은 관객을 유치할 수 있을 지에 대한 insight를 주기 위하여 선정하였다.

데이터 추출

2018년 01월 01일 ~ 2018년 12월 30일 까지의
boxoffice 집계 데이터

데이터 column 설명
- 영화의 해당일의 순위
- 영화명
- 개봉일자
- 매출액
- 매출액점유율
- 매출액증감(전일대비)
- 누적매출액
- 관객수
- 관객수증감(전일대비)
- 관객수증감율(전일대비)
- 누적관객수
- 스크린수
- 상영횟수
- 대표국적
- 국적
- 제작사
- 배급사
- 등급
- 장르
- 감독
- 배우

```
In [2]: data_path = '20180101-20181230_박스오피스'

        base_dir = os.path.join(data_path)

In [3]: data = os.listdir(base_dir)

        data_csv = ['./20180101-20181230_박스오피스/'+i for i in data if 'csv' in i]
        data_csv
        # 데이터 폴더 안의 모든 csv 데이터를 불러온다.

Out[3]: ['./20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-01-07.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-01-14.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-01-21.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-01-28.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-02-04.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-02-11.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-02-18.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-02-25.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-03-04.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-03-11.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-03-18.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-03-25.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-04-01.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-04-08.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-04-15.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-04-22.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-04-29.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-05-06.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-05-13.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-05-20.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-05-27.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-06-03.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-06-10.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-06-17.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-06-24.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-07-01.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-07-08.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-07-15.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-07-22.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-07-29.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-08-05.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-08-12.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-08-19.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-08-26.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-09-02.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-09-09.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-09-16.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-09-23.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-09-30.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-10-07.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-10-14.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-10-21.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-10-28.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-11-04.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-11-11.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-11-18.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-11-25.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-12-02.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-12-09.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-12-16.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-12-23.csv',
         './20180101-20181230_박스오피스/KOBIS_일별_박스오피스_2018-12-30.csv']
```

전처리

- 컬럼명 재설정
- 빈 행 삭제
- Objcect로 돼 있는 수치형 변수들 형 변환
- 애매한 시청 등급 재분류하여 ‘fin_등급’ column 생성
- 날짜에서 개월 정보 추출하여 '월’ column 생성
- 다중 장르에서 대표 장르 추출

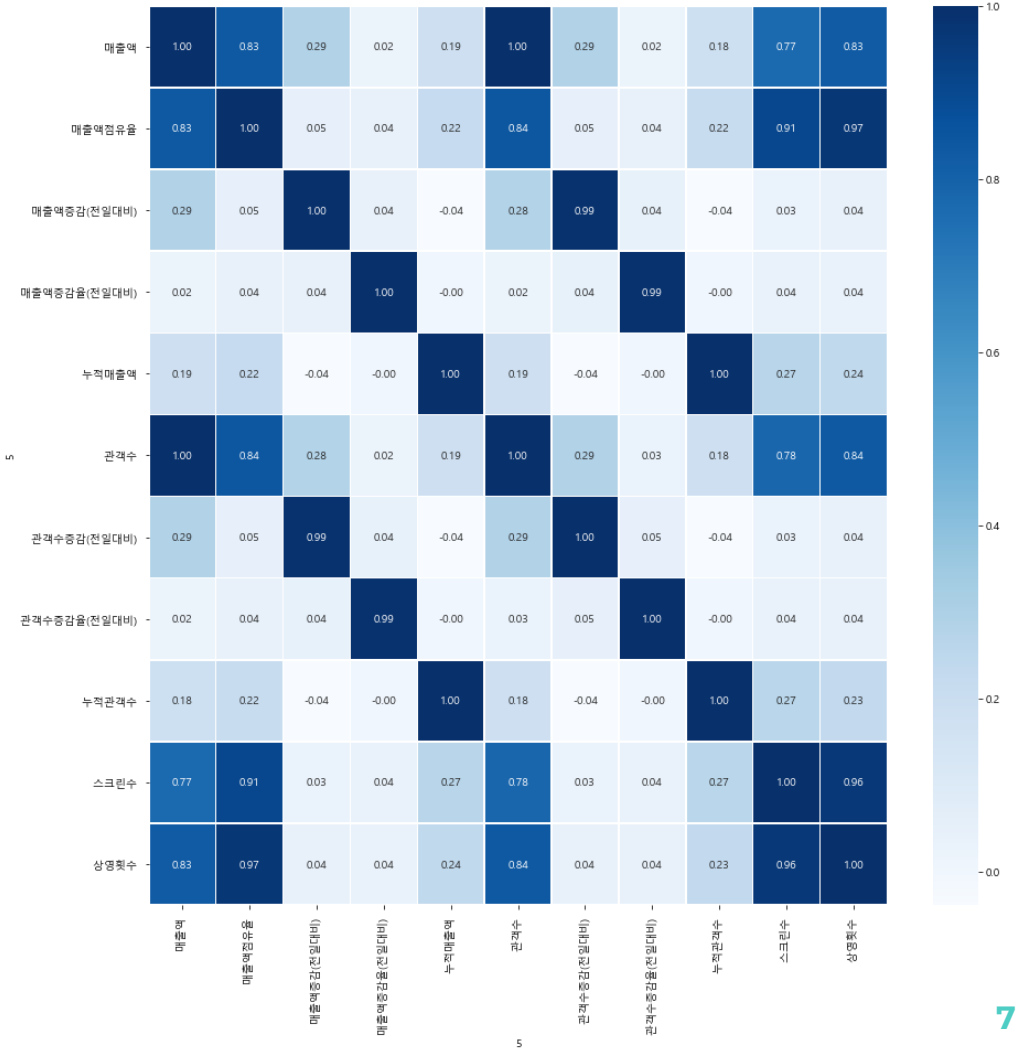
데이터 시각화

1. 상관계수
2. 배급사별 집계 시각화
3. cps 높은 배급사 10개들의 출시 장르 비중
4. 모든 배급사들의 출시 장르 비중
5. 월별 관객수, 매출액, 상영횟수 시각화 1,2
6. 주요 5개 장르들의 월별 추이
7. 비주류 장르들의 월별 추이
8. 등급별 영화의 수
9. 영화 등급별 관객수

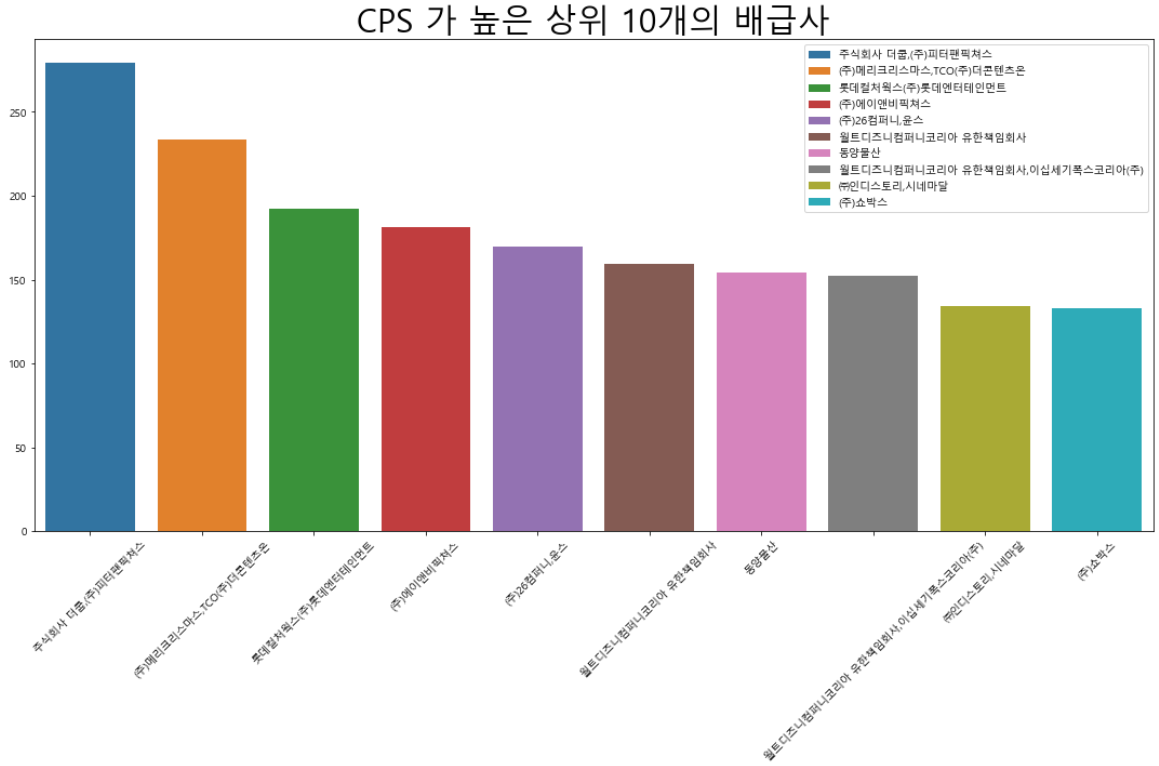
상관계수

각 변수들의 상관계수를 분석한 결과 스크린수와 관객수, 매출액 등은 강한 양의 상관관계를 가짐을 알 수 있었다.

스크린 당 관객을 CPS(customer per person)라 칭하였고 스크린 점유 대비 많은 관객을 유치한 배급사가 어딘지 시각화해보기로 했다.

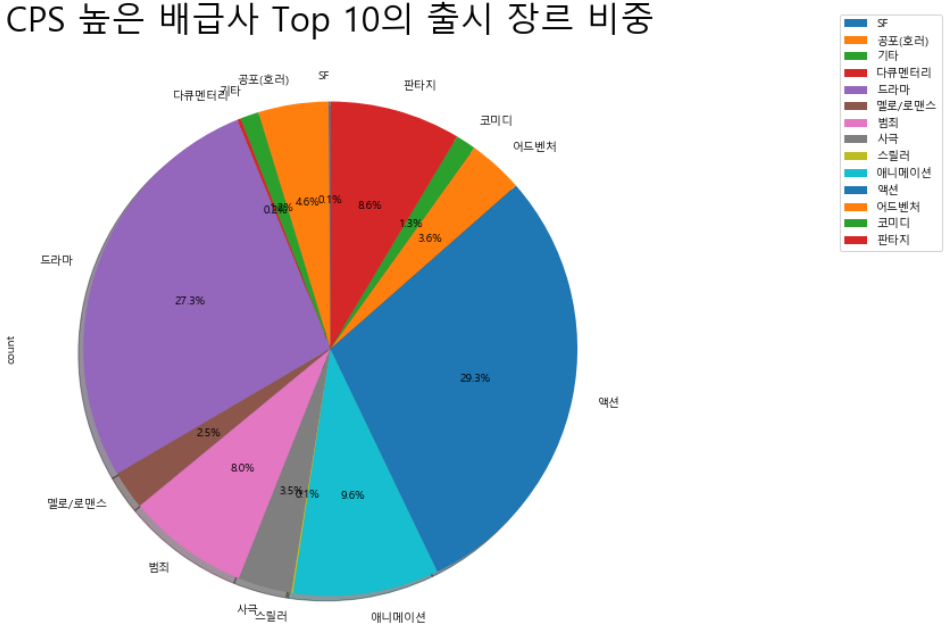


배급사별 집계 시각화



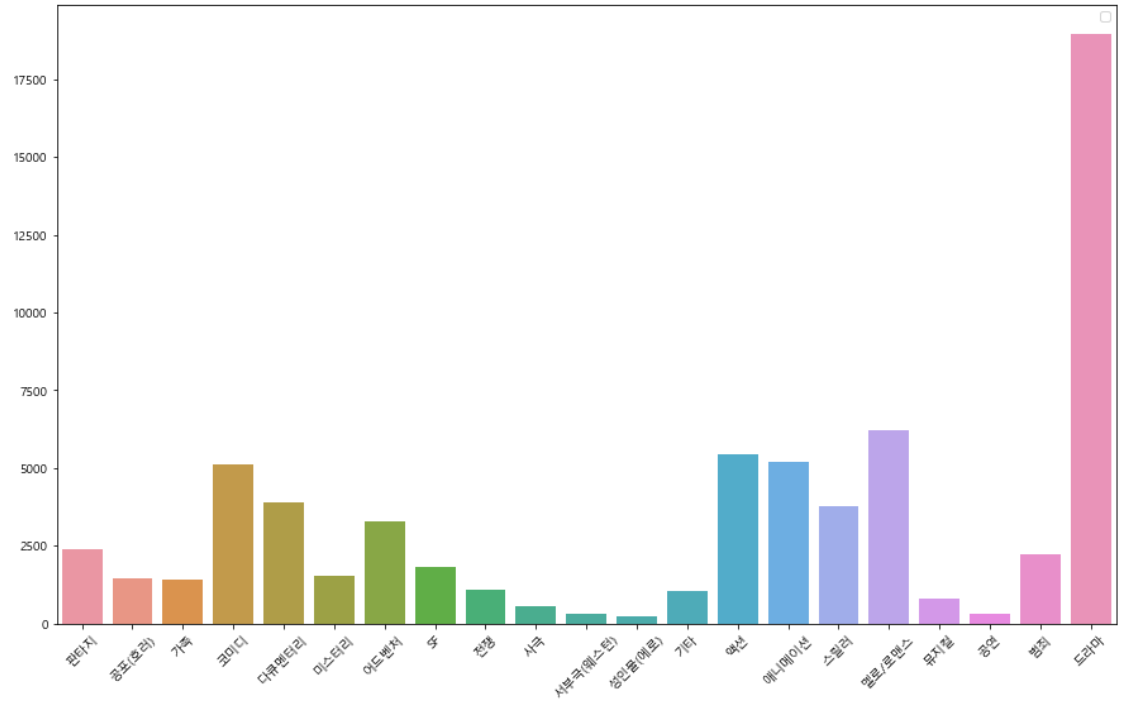
cps가 높은 상위 10개의 배급사는 다음과 같다.

CPS 높은 배급사 Top 10의 출시 장르 비중



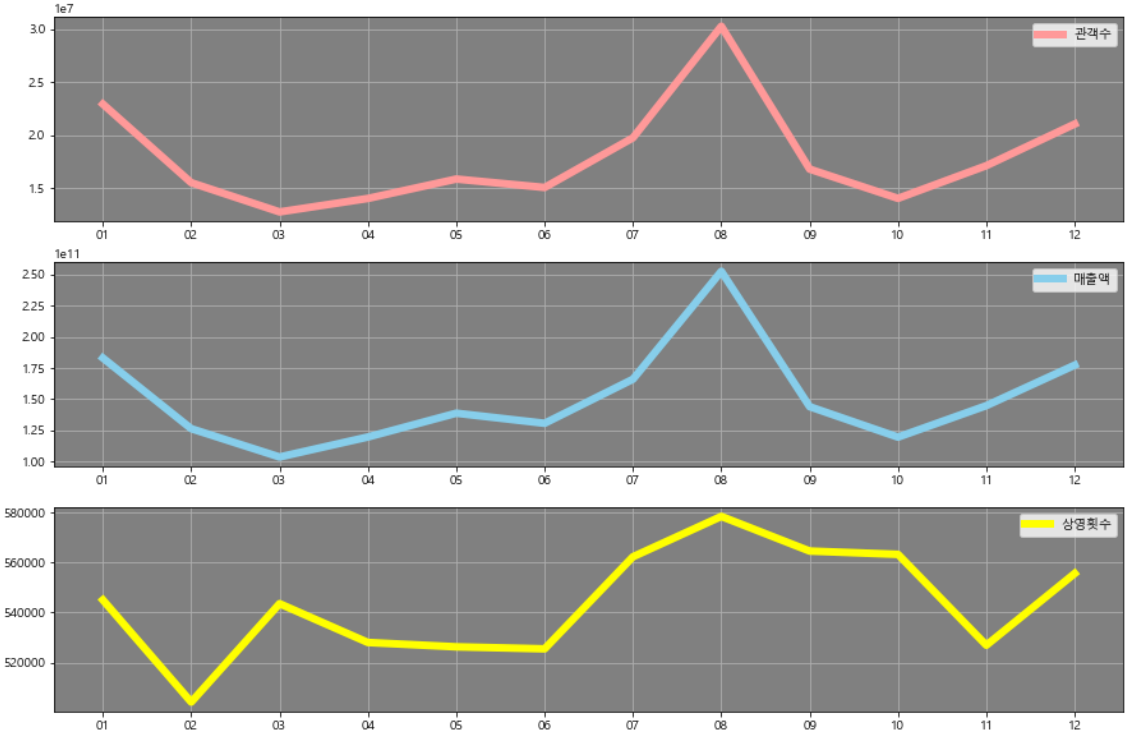
cps가 높은 상위 10개 배급사들의 장르들을 분석한 결과 액션과 드라마가 과반수 이상으로 가장 많았고 애니메이션과 판타지등이 뒤를 이었다.

모든 배급사 출시 장르 비중



데이터에 포함된 모든 배급사들을 분석한 결과 액션의 비중이 가장 높았던 cps top10 배급사들의 분석 결과와 달리 드라마의 비중이 압도적으로 높았다. 또한 멜로/로맨스, 코미디 등 cps top 10 배급사들에서 다소 비주류였던 장르들의 비중이 컸다.

월별 관객수, 매출액, 상영횟수 시각화 1



월별 관객수, 매출액, 상영횟수를 시각화하였다.
세가지 그래프 모두 비슷한 양상을 띄고 있었다.
먼저 세가지 그래프들은 모두 1월부터 6월까지 대체로 하락세를 보였고 6월부터 상승하여 8월에 정점을 찍고 가을까지 하락하다가 다시 상승했다.

1년중 8월에는 관객수, 매출액, 상영횟수 모두 가장 많았다.
관객수와 매출액은 1월부터 점차 감소하여 3월에 최하를 기록했고 6월과 10월에도 저조했다.
이와 달리 상영횟수는 1월부터 2월까지 감소하다가 3월까지 다시 증가하며 3월에 적지 않은 상영횟수를 기록했고 관객수와 매출액이 매우 저조한 편이었던 10월에 높은 상영횟수를 기록했다.

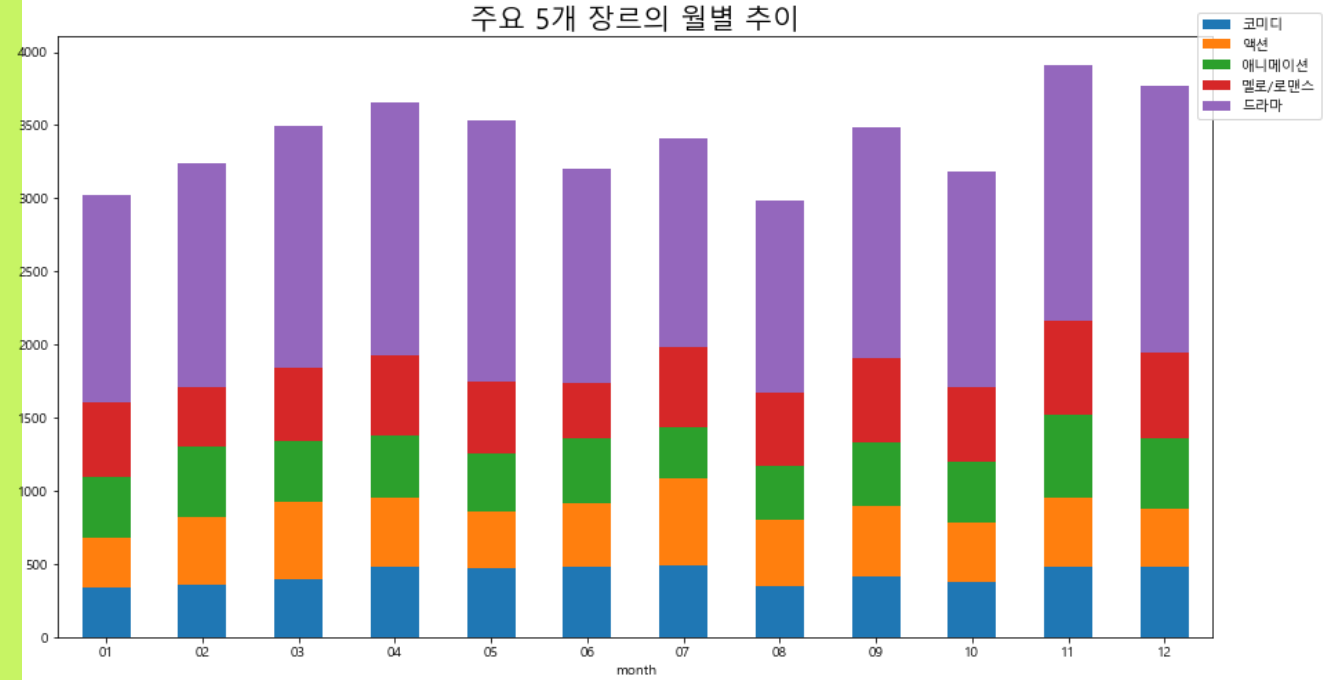
위의 결과로 보아 8월에 가장 많은 수익을 창출했고, 상영횟수가 많은 달에는 대부분 많은 관객을 유치했지만, 3월과 10월에는 많은 상영횟수에도 불구하고 관객을 많이 유치하지 못했다.

월별 관객수, 매출액, 상영횟수 시각화 2

X축이 관객수 y축이 상영횟수, 점의 색으로 월을 구분한 산점도로 관객수,상영횟수,월간의 관계를 분석했다.
대체로 8월 전후 영화들의 관객수와 상영횟수가 많다는 것을 알 수 있다.

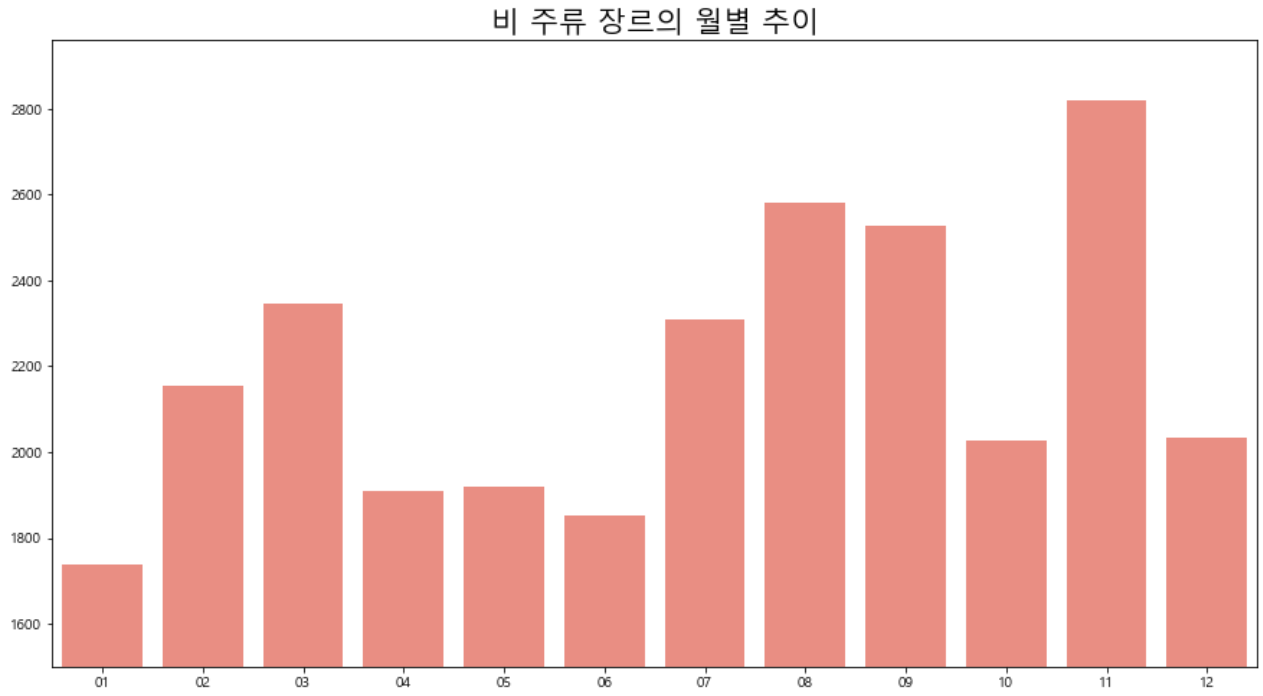
분포가 위로 볼록한 모양인데
상영횟수 5000회이하, 관객수150000명 이하 지점까지 영화들이 밀집해있고 분포가 가파르나 이 지점 이후부터는 분포가 분산되어 있고 완만하게 증가하는 모양이다.

주요 5개 장르들의 월별 추이



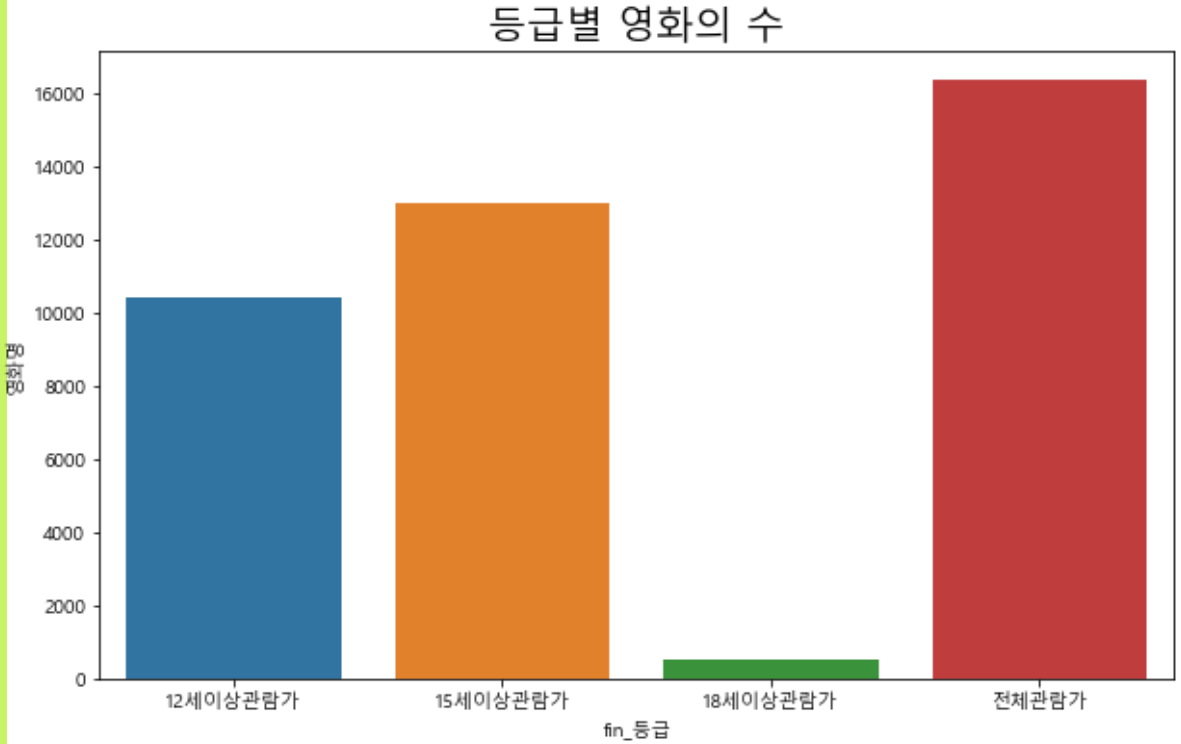
주류를 이루었던 코미디,액션,애니,멜로,드라마 5개의 장르의 월별 추이를 막대그래프로 시각화했다. 장르마다 월별 점유율은 1년 내내 거의 비슷한 모습을 보였다. 드라마의 점유율이 1년동안 가장 컸고 나머지 장르들의 비율은 1년내내 모두 비슷했다.

비주류 장르의 월별 추이



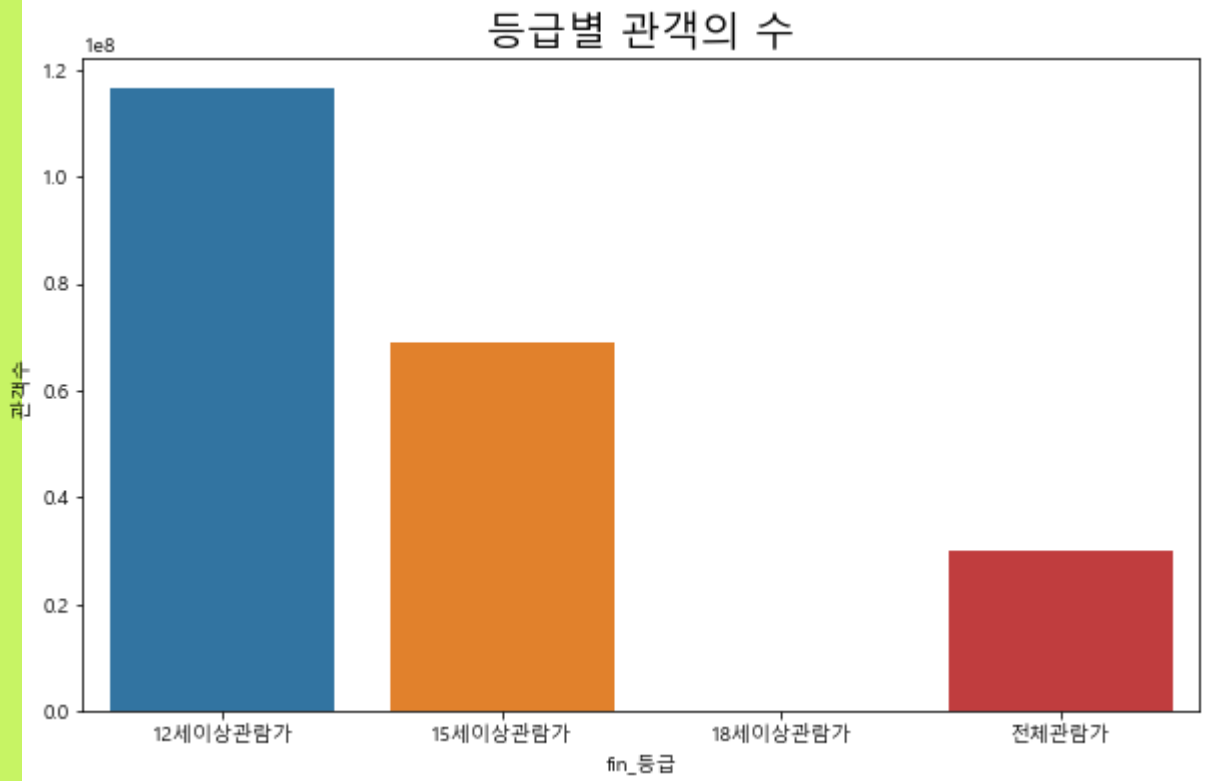
주요 5개 장르들을 제외한 나머지 비주류 장르들의 월별 추이를 막대그래프로 시각화했다.
1년동안 변화가 큰데 1월 최하를 기록했고 11월에 최고를 기록했다.
1년동안 들쭉날쭉한 모습을 보였다.

등급별 영화의 수



기존 데이터에서 애매하게 분류되어 있던 시청등급을 12세이상관람가, 15세이상관람가, 18세이상관람가, 전체관람가 총 4개로 재분류했다.
등급별 영화의 수를 막대그래프로 시각화했다.
등급별 영화의수는 전체관람가가 가장 많았고 15세이상관람가, 12세이상관람가, 18세이상관람가가 뒤를 이었다.
18세이상관람가 영화의 수는 나머지에 비해매우 적었다.

영화 등급별 관객의 수



등급별 관객의 수를 살펴보면 12세이상관람가의 관객수가 압도적으로 많았고 15세이상관람가와 전체관람가가 뒤를 이었다. 이를 통해 등급별 영화의 수와 등급별 관객의 수는 다소 다른 양상을 띄고 있는 것을 알 수 있었다. 18세이상관람가는 비교가 무의미할 정도로 관객의수가 적었다.

분석의 결론
한계점

분석의 결론

박스오피스 데이터 최종 분석 결과

스크린수와 관객수, 매출액 등은 강한 양의 상관관계를 가짐을 알 수 있었다.

장르별 분석:
cps가 높은 상위 10개 배급사들의 장르들을 분석한 결과 [액션]장르의 비율이 약 30%로
가장 많이 차지했는데 **이를 통해 [액션]장르의 영화가 관객을 가장 많이 유치했다고 유추해볼 수 있었다.**
장르별 스크린 점유율은 1년간 [드라마]가 가장 높았다.
비주류 장르의 월별 추이는 11월에 가장 높고 1월에 가장 낮았다.

등급별 영화의 수는 전체관람가, 15세이상관람가, 12세이상관람가, 18세이상관람가 순으로 많았고
관객의 수는 **12세이상관람가**, 15세이상관람가, 전체관람가, 18세이상관람가 순으로 많았다.

8월 전후 시기에 관객수와 매출액이 가장 높았다.

분석의 한계점

2018년도의 1년치 boxoffice 데이터만 사용했기 때문에, 일반화하기에 데이터가 다소 부족하고 해마다의 특수한 상황을 고려할 수 없기 때문에, 분석으로 도출한 결과가 일반적으로 통용될 것이라고 확신할 수 없다.

[액션장르]의 영화가 관객을 가장 많이 유치했다는 결론의 기준인 CPS는 현재 통용되고 있는 비교 지표가 아닌 임의로 생성한 지표임으로 분석 결과의 정확성에 대한 오류가 있을 수 있다.

데이터에서 중복되는 의미를 가지는 featur가 많고 (ex. 매출액, 매출액 증감율, 매출액 점유율) 데이터 자체가 시사하는 정보가 많지 않아, 분석의 방향이 단조로운 경향이 있다.

Thanks!

Any questions?