**□ 모듈프로젝트 보고서**

-클라우드 ai 융합 전문가 양성 과정-

|  |
| --- |
| 모듈프로젝트 1  < 역대 흥행 한국 영화 분석> |

2023-01-18

**1조**

이름: 이용희, 지승엽

<목차>

1. 프로젝트 주제 및 목표

- 주제

- 목표

2. 데이터 크롤링

- 타겟 사이트 선정

- 크롤링

3. 데이터 정제 및 시각화

- 준비과정

- 요소별 정제 및 시각화

4. 결론 및 기대효과

1. 프로젝트 주제 및 목표

주제

- 한국 영화의 각종 데이터(상영 시기, 배우, 장르 등)를 추출하여 흥행한 영화에 대해 분석

목표

- 영화 제작사 측에서 여러가지 데이터를 보고 흥행 요소에는 어떤 것들이 있는지 파악 가능

- 데이터 수집(크롤링) 및 정제, 시각화를 함으로써 수업시간에 배웠던 내용 복습 및 실력 향상

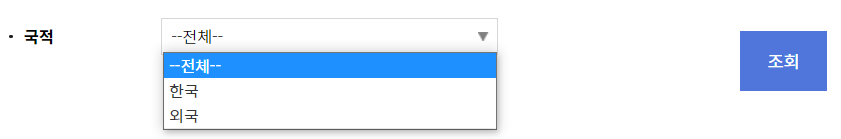
2. 데이터 크롤링

2-1. 타겟 사이트 선정

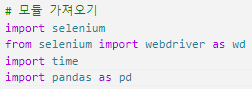
- 역대 박스오피스 관객 수 기준 상위 200위를 보여주는 사이트인 영화권입장권 통합전산망(<https://www.kobis.or.kr/kobis/business/main/main.do>)으로 선정

- 한국 영화만 크롤링 기법으로 데이터 수집

2-2. 크롤링



- 모듈 불러오기, 타겟 사이트 접속, 국적 값 선택, 조회 버튼 클릭







- 영화 목록

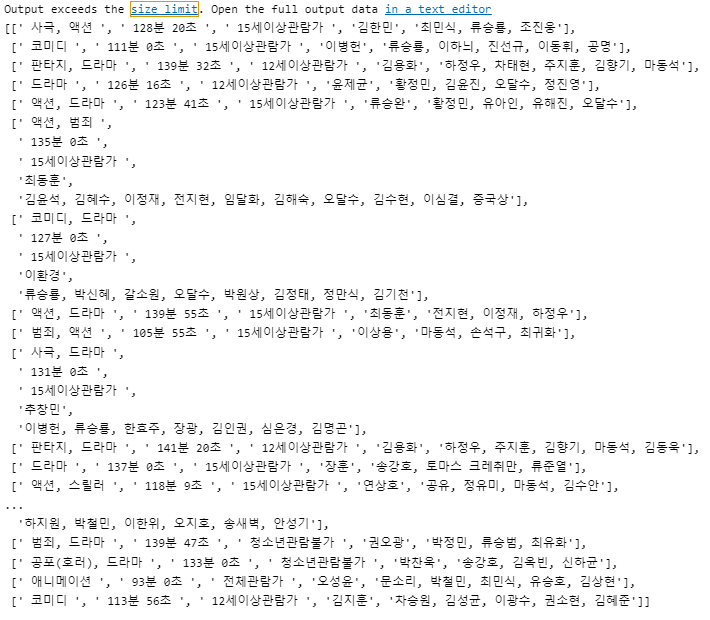


- 각 영화를 클릭하면 세부 정보가 나온다.

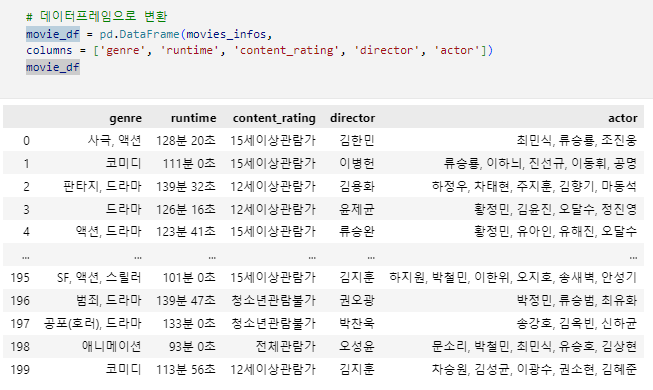


- 각 링크를 클릭하여 정보 추출(장르, 상영시간, 관람 등급, 감독, 배우)





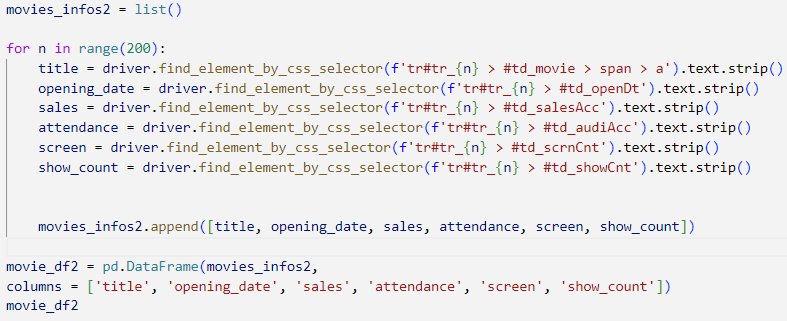
- 추출한 정보를 데이터프레임으로 변환

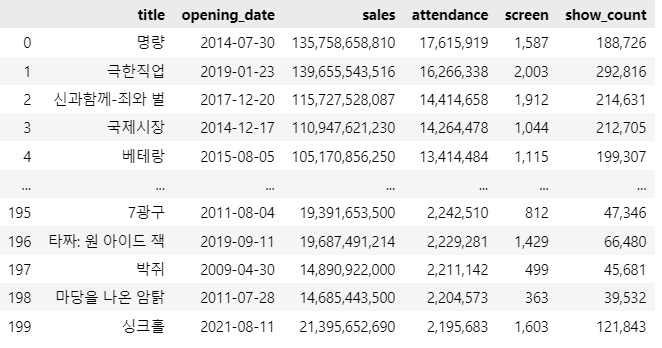




- 첫 화면에서의 정보를 엑셀로 다운받았으나 로드가 되지 않았다. (직접 다시 크롤링 후 추후에 데이터 병합)

- 첫 화면에서의 크롤링 후 데이터프레임으로 변환



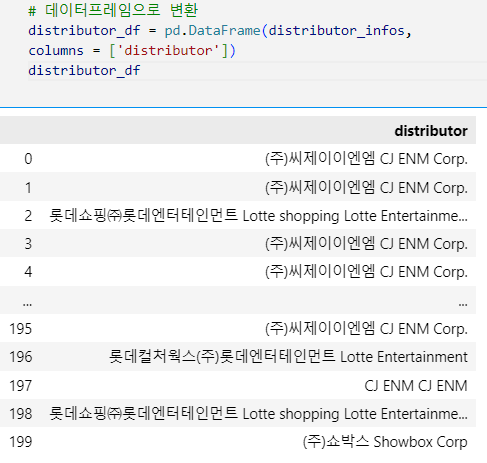


- 두 개의 데이터프레임 저장



- 배급사 추가 추출( 추후에 병합 ), 데이터프레임으로 변환, 저장







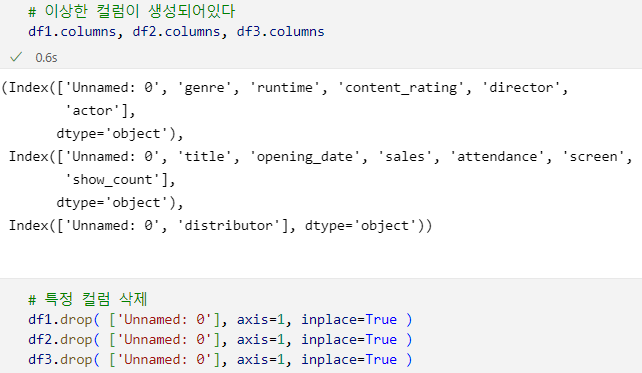
3. 데이터 정제 및 시각화

3-1. 준비 과정

- 모듈 불러오기, 폰트 다운로드, 저장해둔 csv 파일 업로드



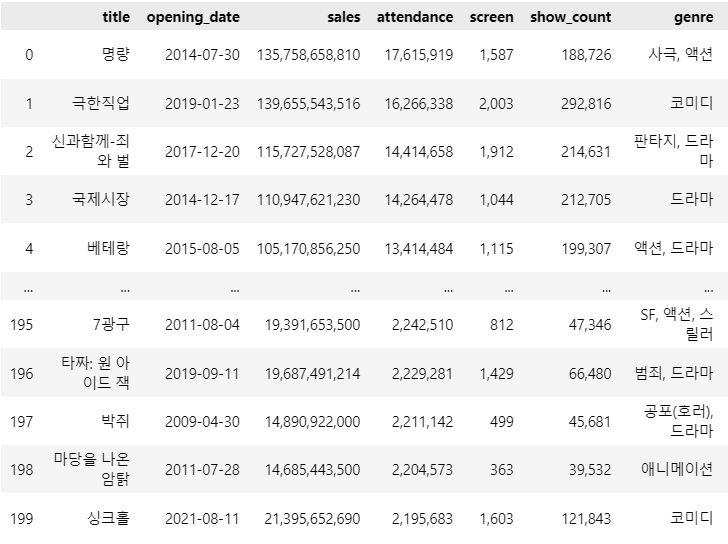
- 이상한 컬럼이 생성되어 삭제

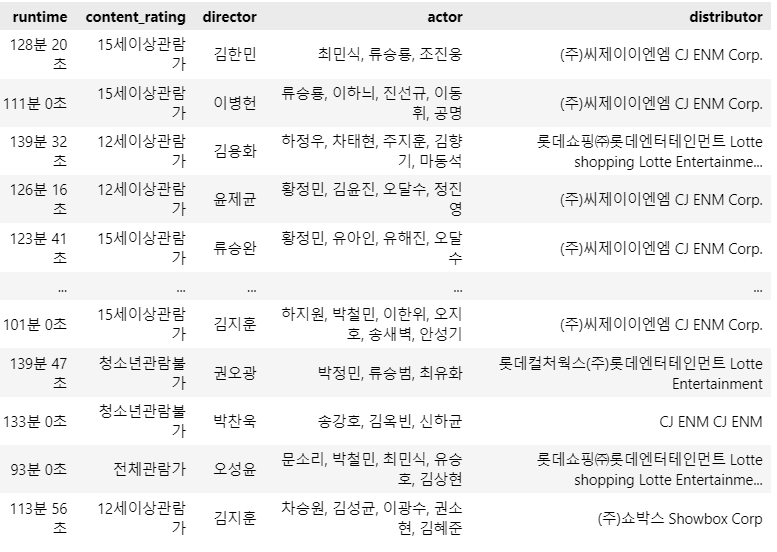


- 3개의 데이터프레임 병합





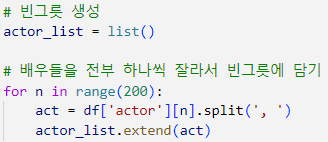




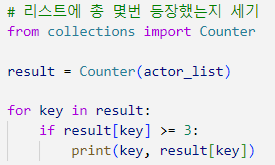
-> 최종적으로 우리가 사용할 데이터 프레임이다. 관객수 기준 상위 200위 영화들의 데이터이다. ( 컬럼이 너무 많아 데이터프레임이 너무 길어서 두 장의 사진을 세로로 붙였다.)

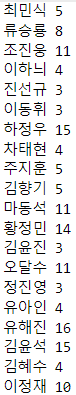
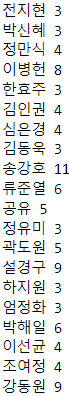
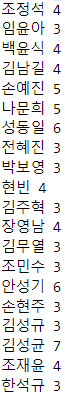
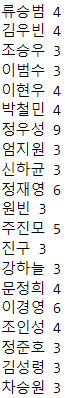
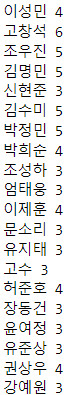
3-2. 첫번째 요소 : 배우

- 배우들을 각각 하나씩 잘라서 리스트에 담기



- 리스트에서 총 3번 이상 출연한 배우들 출력(100명)

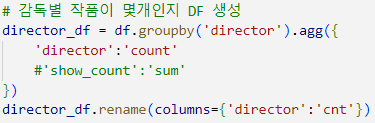


* 200위 안의 영화에 출연한 배우들은 총 900명이며 1번~2번 출연한 배우들이 상당수를 이루고 있기 때문에 3번 이상 출연한 배우들 100명을 출력하였다.

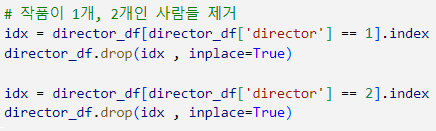
3-3. 두번째 요소 : 감독

- 감독 별로 상위 200위 중 몇 개의 영화를 제작했는지 확인



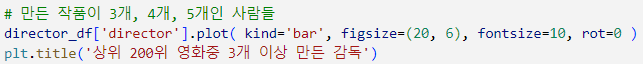


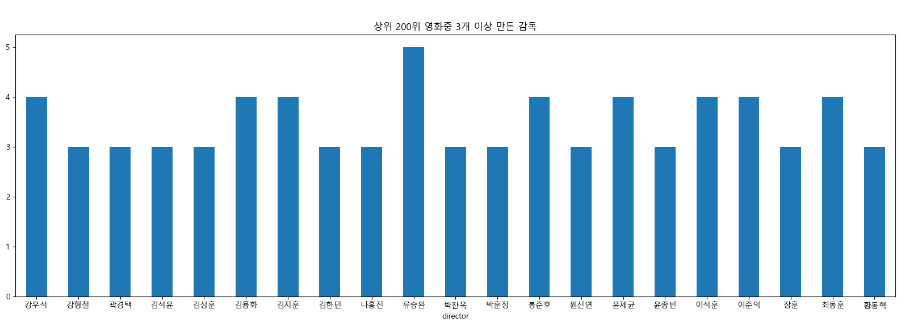
- 작품이 1개, 2개인 감독들 제거



- 상위 200위 영화 중 3개 이상 만든 감독





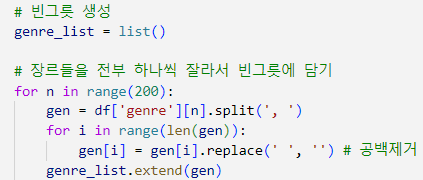
* 200위 안의 영화를 만든 감독은 총 123명이며 1개~2개를 만든 감독들이 상당수를 이루고 있기 때문에 3개 이상 만든 감독들을 바 차트로 표현하였다.

3-4. 세번째 요소 : 장르

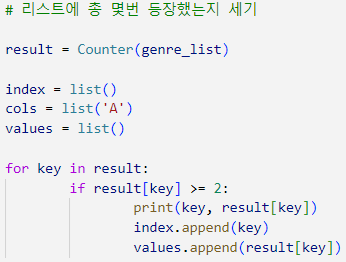
- 장르 종류 확인



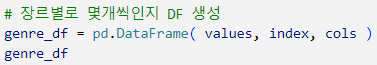
- 문자열 정제



- 리스트에 담은 후 총 몇 번 등장했는지 체크



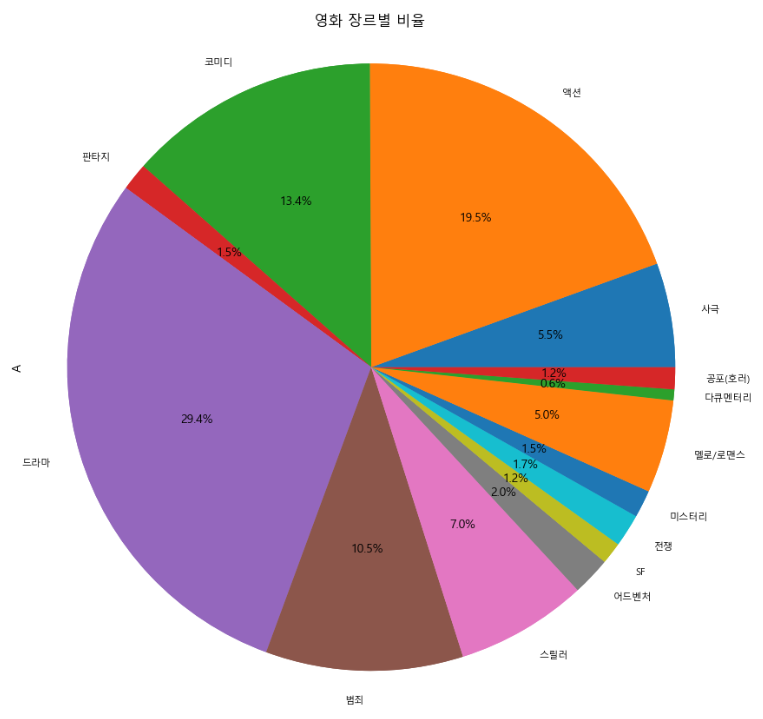
- 데이터프레임으로 변환





- 파이 차트를 이용해 시각화

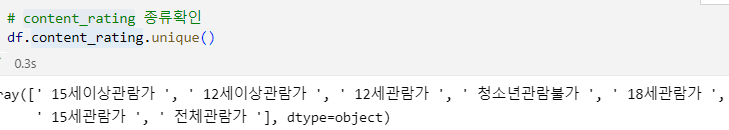




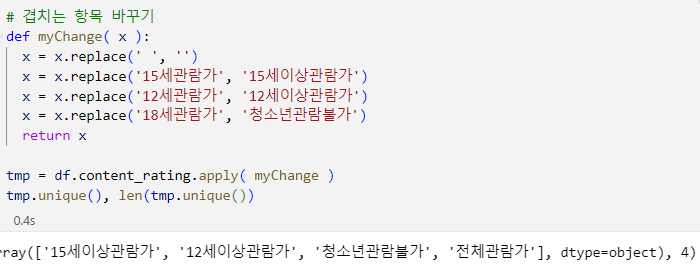
* 드라마, 액션, 코미디, 범죄, 스릴러, 사극 순으로 비율이 높고 나머지 장르들은 비교적 적은 비율이다.

3-5. 네번째 요소 : 관람 등급

- content\_rating 종류 확인



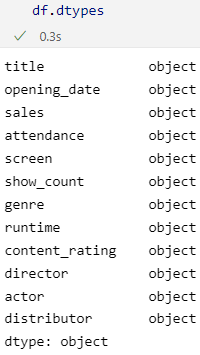
- 겹치는 항목 적절히 바꾸기



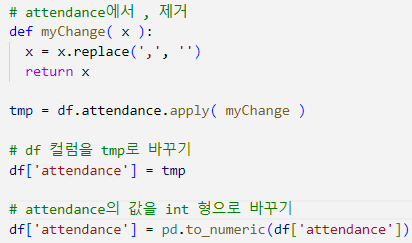
- df에 적용

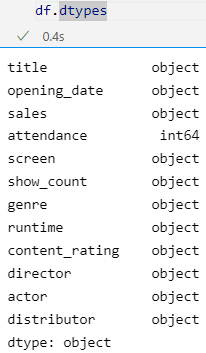


- 타입 확인(attendance와의 시각화를 할 때 대비)

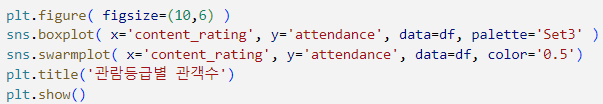


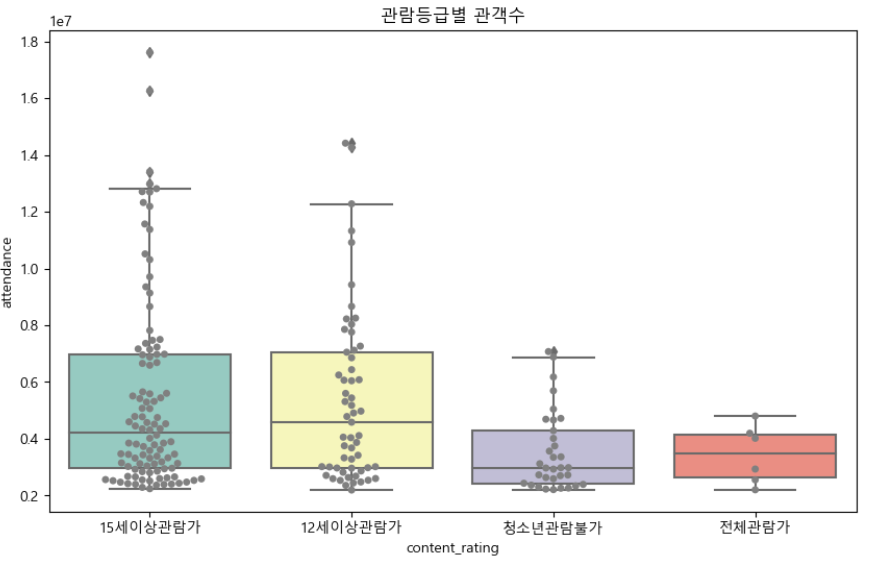
- attendance의 콤마 제거, int형으로 변환





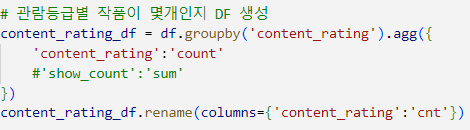
- 관람등급별 관객 수 시각화





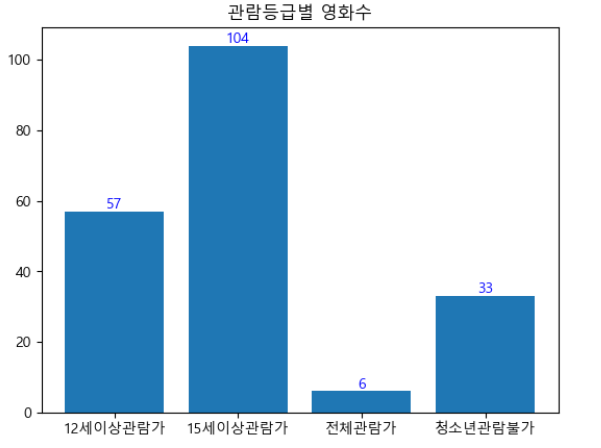
-> 15세이상에 몰려있어서 다른 방식으로 시각화 (직관적이지 않아서)

- 관람등급 별로 작품이 몇 개인지 groupby를 이용하여 그룹화



- 관람등급별 영화수 시각화

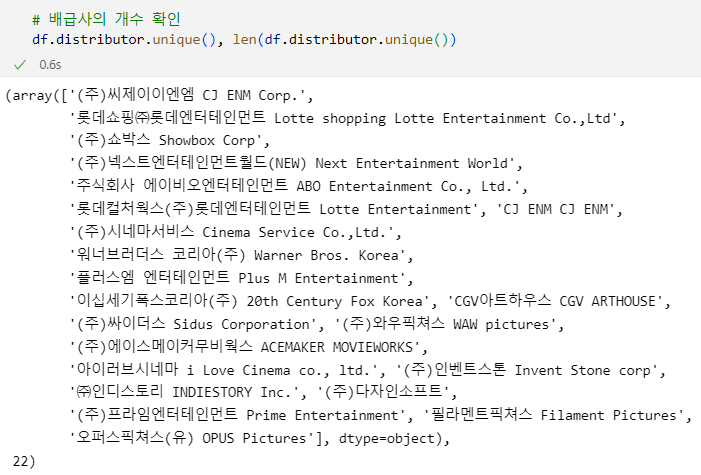




* 15세이상 관람가가 104건으로 반 이상을 차지하고 있다.

3-6. 다섯번째 요소 : 배급사

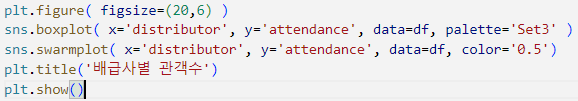
- 배급사 unique() 함수를 이용하여 확인

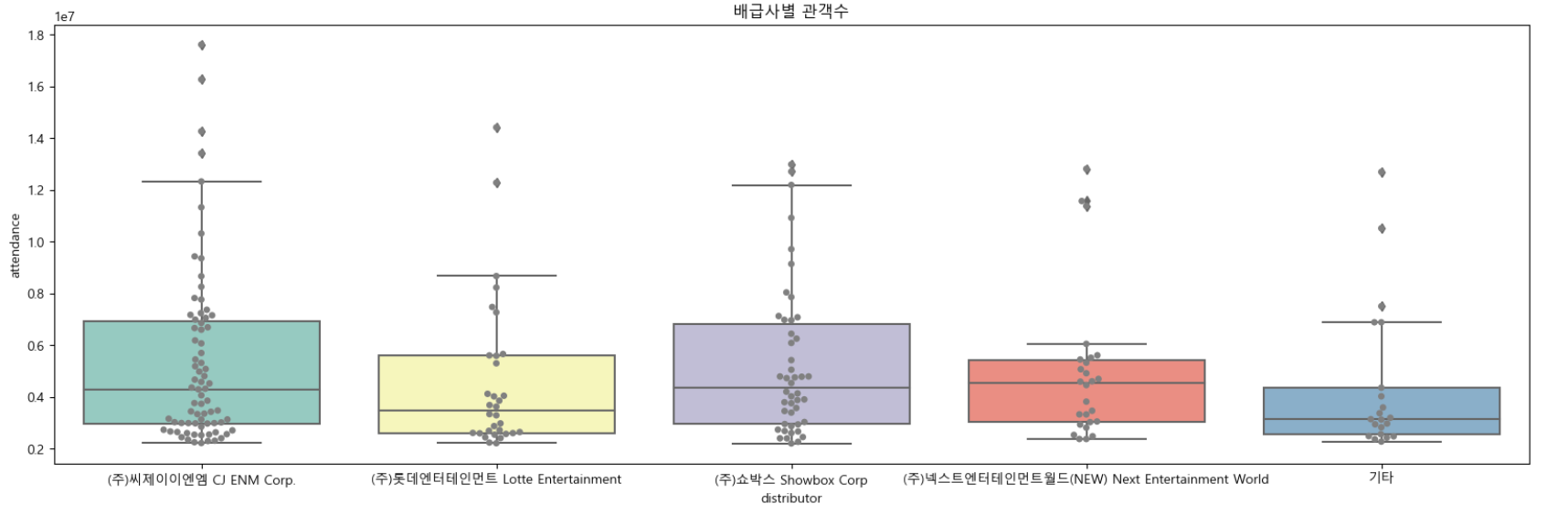


- 같은 회사는 적절히 묶기, 소수의 회사들은 ‘기타’로 변경



- 배급사별 관객 수 시각화



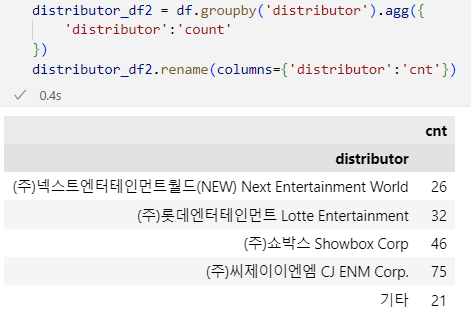


-> 배급사별 비율을 알기 위해 다른 방식으로 시각화(직관적이지 않아서)

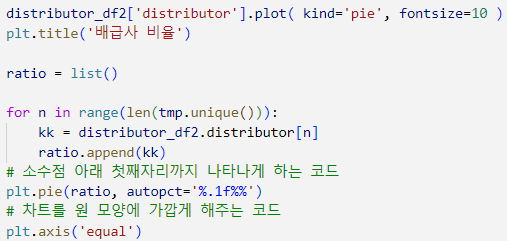
- df에 적용

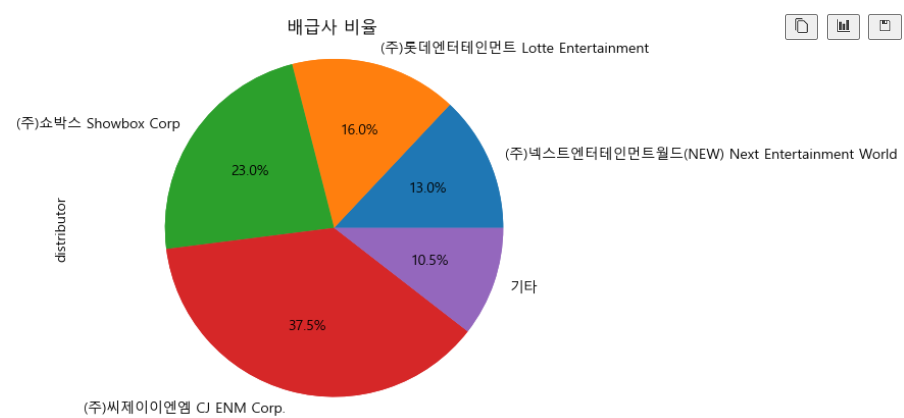


- 배급사 별로 작품이 몇 개인지 groupby를 이용하여 그룹화



- 배급사 비율 시각화

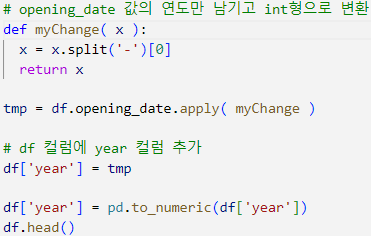


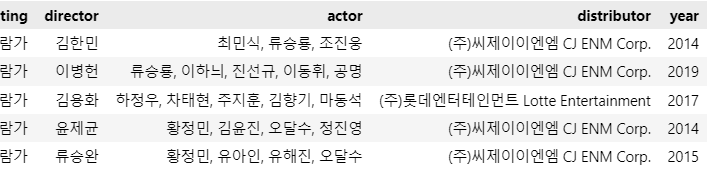


* 4대 배급사가 배급의 대부분을 담당하고 있다.

3-7. 여섯번째 요소 : 개봉시기

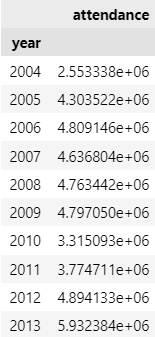
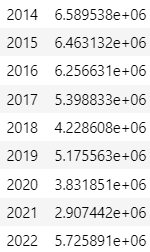
- 연도만 표시하는 year 컬럼 새로 생성



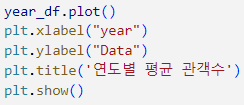


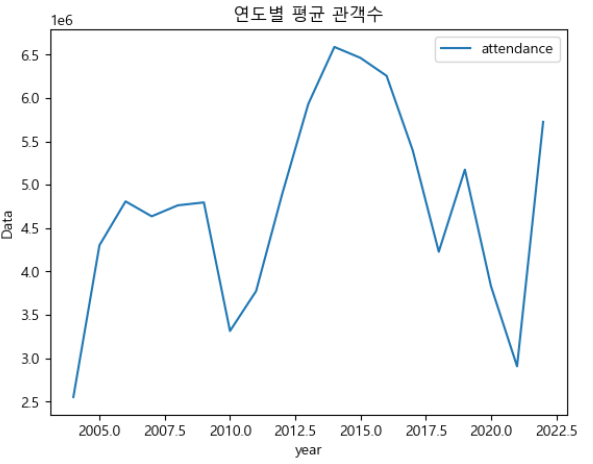
- 연도 별로 평균 관객 수를 groupby를 이용하여 그룹화



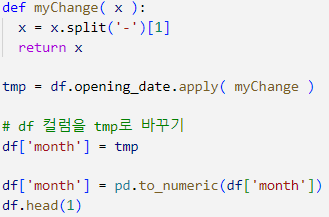
- 연도별 평균 관객 수 시각화

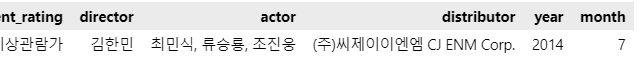




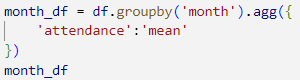
* 2014년에서 2016년 사이에 관객 수가 제일 많았으며 코로나 시기 초 중반에는 관객 수가 현저히 줄어들었지만 이후에는 다시 회복세를 보이고 있다.

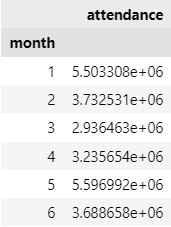
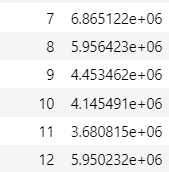
- 월만 표시하는 month 컬럼 새로 생성





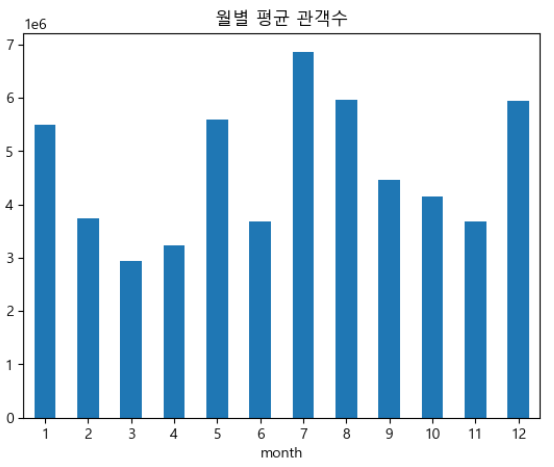
- 월별로 평균 관객 수를 groupby를 이용하여 그룹화



- 월별 평균 관객 수 시각화

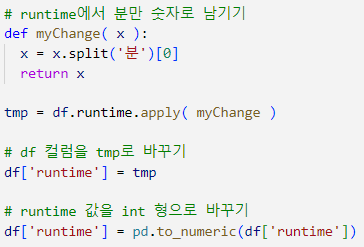




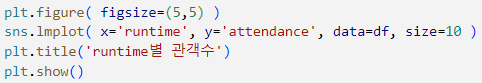
* 연말연시, 가정의 달인 5월, 여름휴가기간인 7,8월에 관객 수가 많은 편이다.

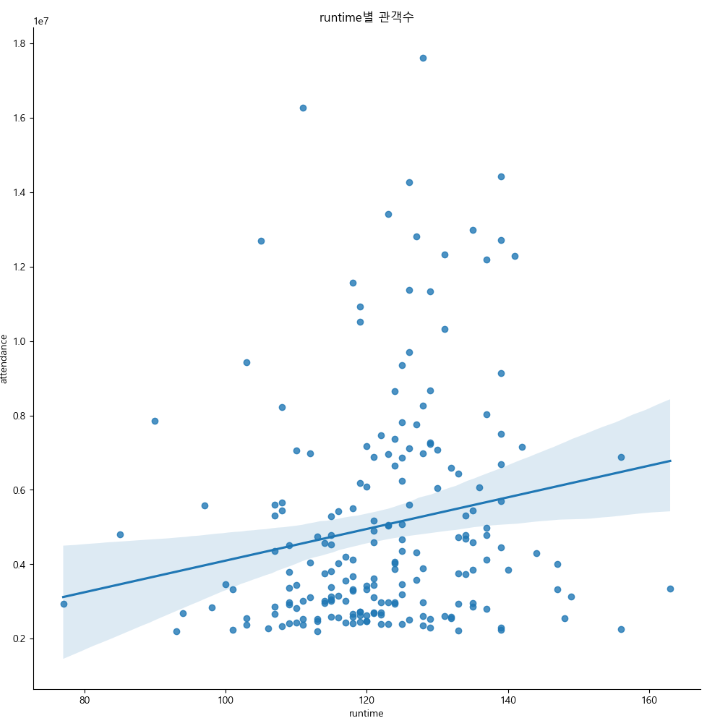
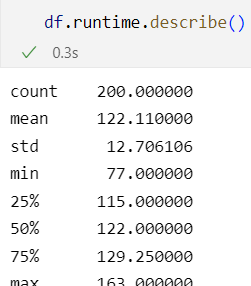
3-8. 기타 요소 : runtime

- runtime x분 x초 중 분만 남기기



- runtime별 관객 수 시각화



* 상영 길이와 관객 수 간의 유의미한 상관관계는 찾지 못했지만 평균 상영 길이는 122분으로 대부분 120분 내외에 몰려있다.

4. 결론 및 기대효과

**결론**

- 역대 흥행한 한국 영화들의 다양한 데이터(장르, 배우, 감독, 상영 시기 등)를 찾을 수 있었다.

* 배우: 흥행한 영화 200위 안에 900명의 주연배우가 출연했으며 그 중 100명의 배우만 3개 이상 영화에 출연했다.
* 감독: 흥행한 영화 200위 안에 123명의 감독이 영화를 만들었으며 그 중 21명의 감독만이 3개 이상 영화를 만들었다.
* 장르: 드라마, 액션, 코미디, 범죄, 스릴러, 사극 순으로 비율이 높고 나머지 장르들은 비교적 적은 비율이다.
* 관람 등급: 15세이상 관람가가 104건으로 반 이상을 차지하고 있으며 15세이상, 12세이상, 청소년 관람불가, 전체 관람가 순이다.
* 배급사: 4대 배급사가 배급의 대부분을 담당하고 있다.
* 연도별 관객 수: 2014년에서 2016년 사이에 관객 수가 제일 많았으며 코로나 시기 초 중반에는 관객수가 현저히 줄어들었지만 이후에는 다시 회복세를 보이고 있다.
* 월별 관객 수: 연말연시, 가정의 달인 5월, 여름휴가기간인 7,8월에 관객 수가 많은 편이다.
* 상영 길이: 상영 길이와 관객 수 간의 유의미한 상관관계는 찾지 못했지만 평균 상영 길이는 122분으로 대부분 120분 내외에 몰려있다.

- 데이터를 수집하고 정제하여 분석 및 시각화를 해보니 상대적으로 유의미한 데이터(장르, 배급사, 관람 등급 등)도 있는 반면에 의미가 있을 것이라고 예상해 추출한 데이터(상영 길이 등)가 무의미한 경우도 존재했다.

**기대효과**

- 제작사 입장에서 이 데이터를 보고 영화를 제작할 때 어떤 요소에 더 가중치를 줘야하는지 알 수 있을 것이다.

- 데이터 수집(크롤링) 및 정제, 시각화를 함으로써 수업시간에 배웠던 내용 복습 및 실력 향상