movie data analysis

KIM YE JI

Contents

데이터셋
설명
+
전처리

데이터
시각화(EDA)
+
데이터 분석

Insight

Data Set Explanation

movie_df

	title	distributor	genre	release_time	time	screening_rat	director	dir_prev_bfnum	dir_prev_num	num_staff	num_actor	box_off_num
0	개들의 전쟁	롯데엔터테인먼트	액션	2012-11-22	96	청소년 관람불가	조병옥	NaN	0	91	2	23398
1	내부자들	(주)쇼박스	느와르	2015-11-19	130	청소년 관람불가	우민호	1161602.50	2	387	3	7072501
2	은밀하게 위대하게	(주)쇼박스	액션	2013-06-05	123	15세 관람가	장철수	220775.25	4	343	4	6959083
3	나는 공무원이다	(주)NEW	코미디	2012-07-12	101	전체 관람가	구자홍	23894.00	2	20	6	217866
4	불량남녀	쇼박스(주)미디어플렉스	코미디	2010-11-04	108	15세 관람가	신근호	1.00	1	251	2	483387
...	...	...	...	...	...	...	...	...	...	...	...	...
595	해무	(주)NEW	드라마	2014-08-13	111	청소년 관람불가	심성보	3833.00	1	510	7	1475091
596	파파로티	(주)쇼박스	드라마	2013-03-14	127	15세 관람가	윤종찬	496061.00	1	286	6	1716438
597	살인의 강	(주)마운틴픽쳐스	공포	2010-09-30	99	청소년 관람불가	김대현	NaN	0	123	4	2475
598	악의 연대기	CJ 엔터테인먼트	느와르	2015-05-14	102	15세 관람가	백운학	NaN	0	431	4	2192525
599	베를린	CJ 엔터테인먼트	액션	2013-01-30	120	15세 관람가	류승완	NaN	0	363	5	7166532

600 rows × 12 columns

- 2010년대 (2010년 ~ 2015년)의 국내영화 데이터 600개를 포함
- 12개의 칼럼(영화제목, 배급사, 장르, 개봉일, 상영시간, 상영등급, 감독, 감독의 이전 참여 작품 평균 관객수, 감독의 이전 작품 개수, 스탭수, 배우수, 관객수)

Data Preprocessing

to_datetime()

#	Column	Non-Null Count	Dtype
0	title	600 non-null	object
1	distributor	600 non-null	object
2	genre	600 non-null	object
3	release_time	600 non-null	object
4	time	600 non-null	int64
5	screening_rat	600 non-null	object
6	director	600 non-null	object
7	dir_prev_bfnum	270 non-null	float64
8	dir_prev_num	600 non-null	int64
9	num_staff	600 non-null	int64
10	num_actor	600 non-null	int64
11	box_off_num	600 non-null	int64

#	Column	Non-Null Count	Dtype
0	title	600 non-null	object
1	distributor	600 non-null	object
2	genre	600 non-null	object
3	release_time	600 non-null	datetime64[ns]
4	time	600 non-null	int64
5	screening_rat	600 non-null	object
6	director	600 non-null	object
7	dir_prev_bfnum	600 non-null	float64
8	dir_prev_num	600 non-null	int64
9	num_staff	600 non-null	int64
10	num_actor	600 non-null	int64
11	box_off_num	600 non-null	int64

- 개봉일(release_time) 을 개봉연도(release_year)와 개봉달(release_month) 로 나누어 각각 칼럼으로 저장하기 위해 개봉일 (release_time) 데이터 타입을 object 에서 datetime으로 변환 : to_datetime()

Data Preprocessing

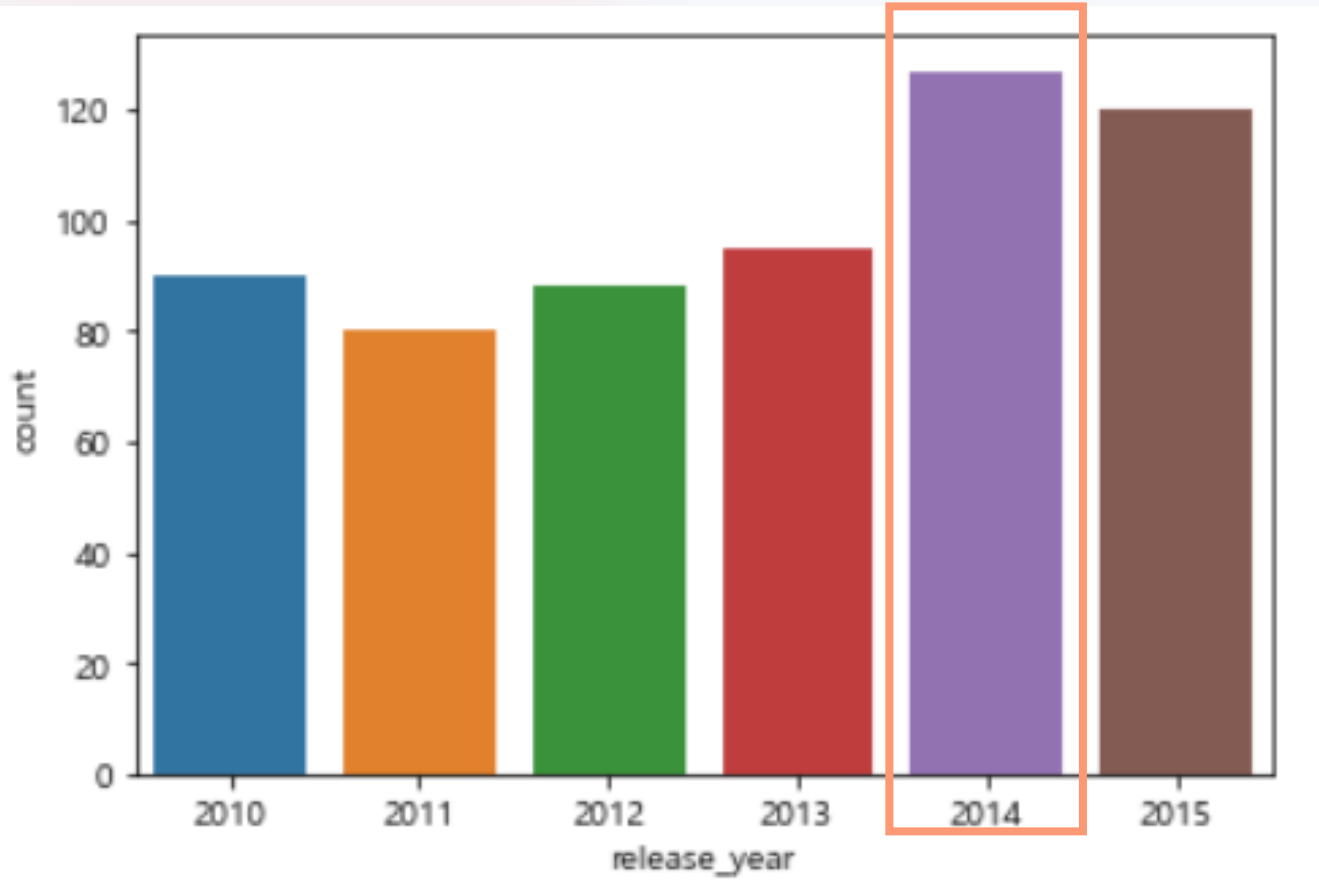
dt.year , dt.month

	title	distributor	genre	release_time	time	screening_rat	director	dir_prev_bfnum	dir_prev_num	num_staff	num_actor	box_off_num	release_year	release_month
0	개들의 전쟁	롯데엔터테인먼트	액션	2012-11-22	96	청소년 관람불가	조병옥	0.00	0	91	2	23398	2012	11
1	내부자들	(주)쇼박스	느와르	2015-11-19	130	청소년 관람불가	우민호	1161602.50	2	387	3	7072501	2015	11
2	은밀하게 위대하게	(주)쇼박스	액션	2013-06-05	123	15세 관람가	장철수	220775.25	4	343	4	6959083	2013	6
3	나는 공무원이다	(주)NEW	코미디	2012-07-12	101	전체 관람가	구자홍	23894.00	2	20	6	217866	2012	7
4	불량남녀	쇼박스(주)미디어플렉스	코미디	2010-11-04	108	15세 관람가	신근호	1.00	1	251	2	483387	2010	11
5	강철대오 : 구국의 철가방	롯데엔터테인먼트	코미디	2012-10-25	113	15세 관람가	육상효	837969.00	2	262	4	233211	2012	10
6	길위에서	백두대간	다큐멘터리	2013-05-23	104	전체 관람가	이창재	0.00	0	32	5	53526	2013	5
7	회사원	(주)쇼박스	액션	2012-10-11	96	청소년 관람불가	임상윤	739522.00	3	342	2	1110523	2012	10
8	1789, 바스티유의 연인들	유니버설픽쳐스인터내셔널코리아	뮤지컬	2014-09-18	129	전체 관람가	정성복	0.00	0	3	5	4778	2014	9
9	청춘그루브	(주)두타연	드라마	2012-03-15	94	15세 관람가	변성현	0.00	0	138	3	868	2012	3

- `movie_df['release_year'] = movie_df['release_time'].dt.year` : 개봉연도(release_year) 칼럼 추가
- `movie_df['release_month'] = movie_df['release_time'].dt.month` : 개봉달(release_month) 칼럼 추가

EDA

개봉연도(release_year) 별 데이터의 개수
: countplot()



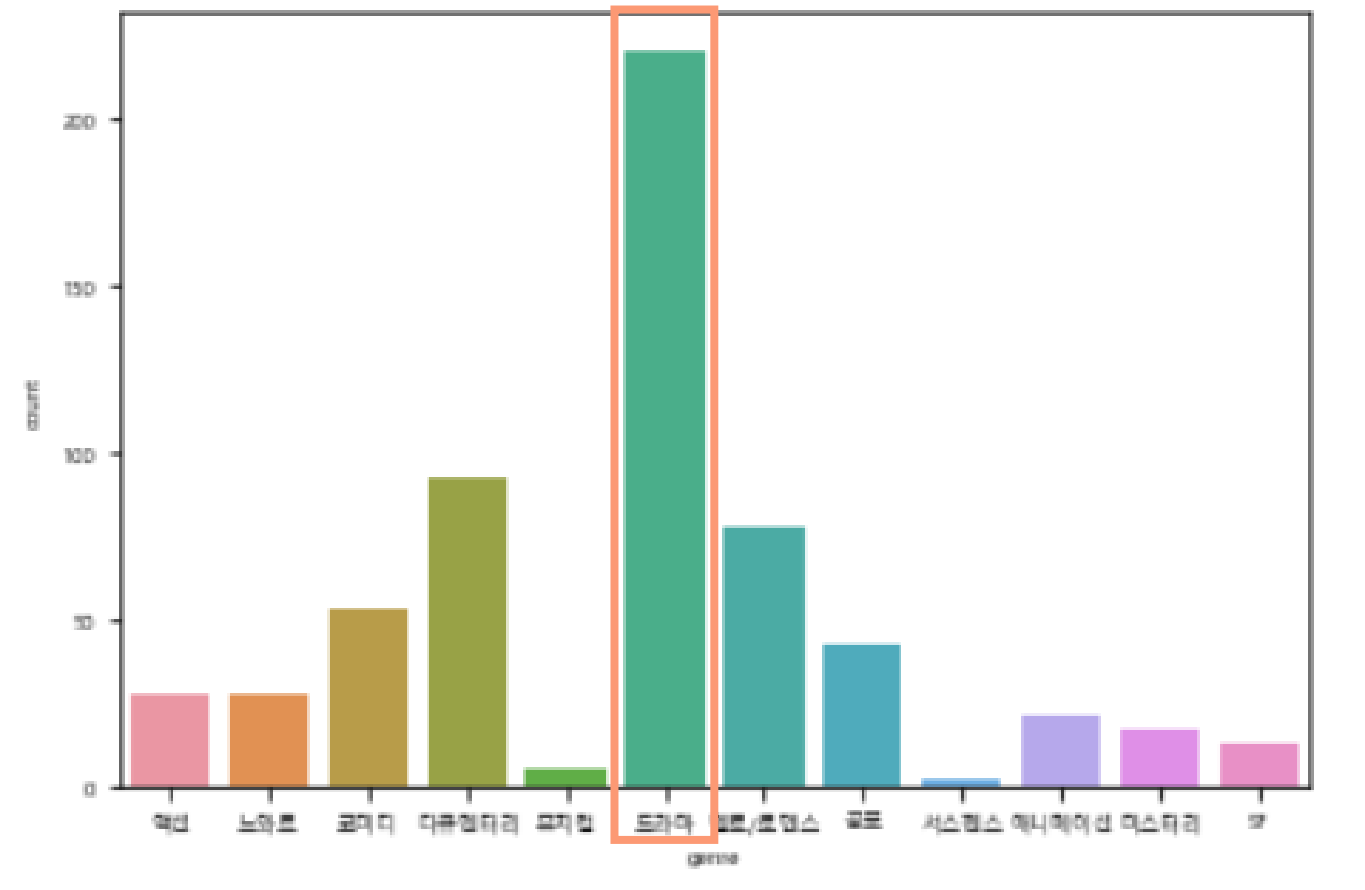
- sns.countplot(x = 'release_year',
  data = movie_df)

- 영화 데이터는 2010년부터 2015년 까지의 데이터를 포함

- 2014년도 : 한국 영화 산업의 호황기

EDA

장르(genre) 별 데이터의 개수
: countplot()



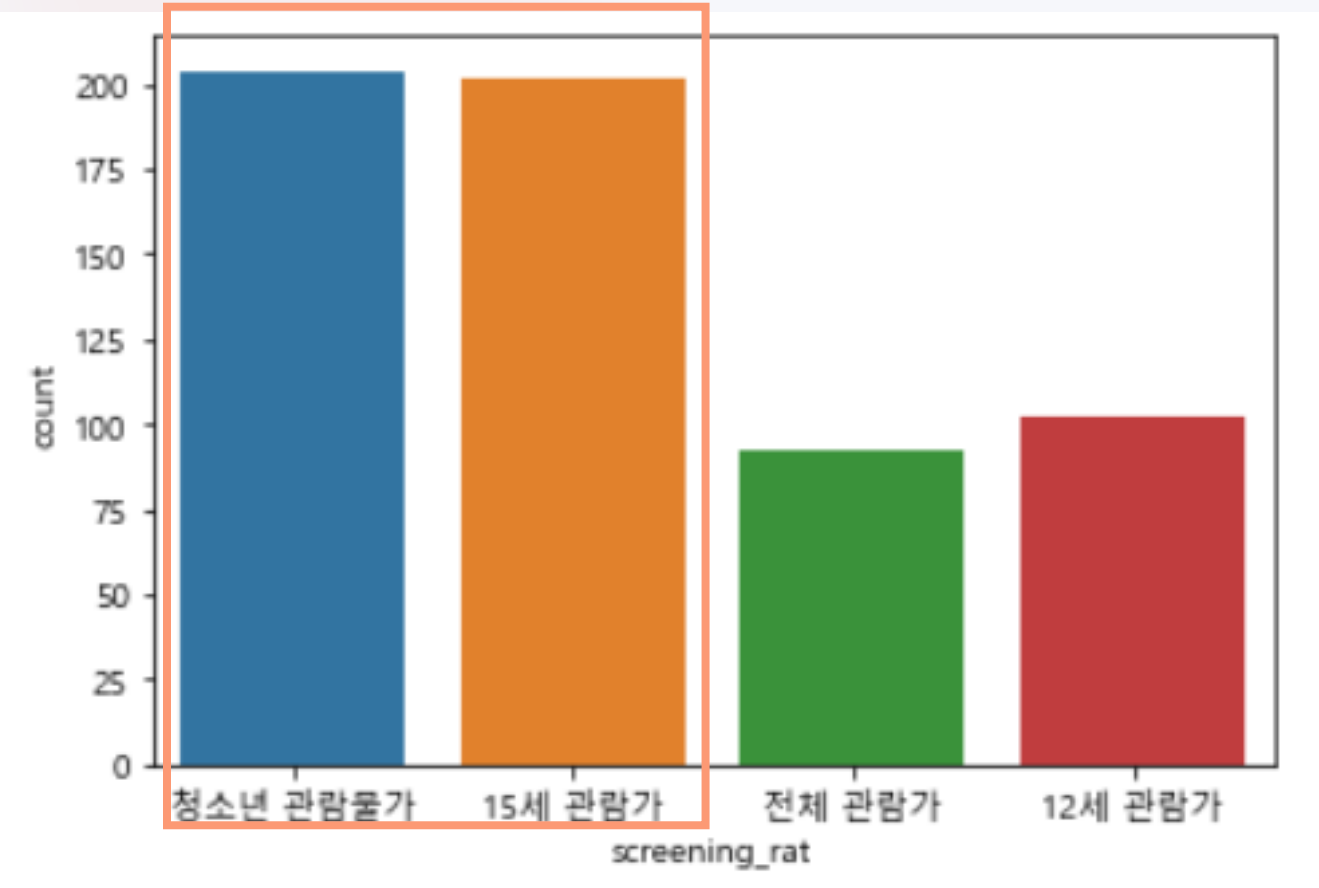
- sns.countplot(x = 'genre', data = movie_df)

- '드라마' 장르의 영화가 가장 많음

- 이외에는 다큐멘터리 > 멜로/로맨스 > 코미디 > 공포 > 액션 > 느와르 > 애니메이션 > 미스터리 > SF > 뮤지컬 > 서스펜스 순

EDA

상영등급(screening_rat) 별 데이터의 개수
: countplot()



- sns.countplot(x = 'screening_rat',
  data = movie_df)

- '청소년 관람불가 ' 와 '15세 관람가' 영화의 개수가 많음

EDA

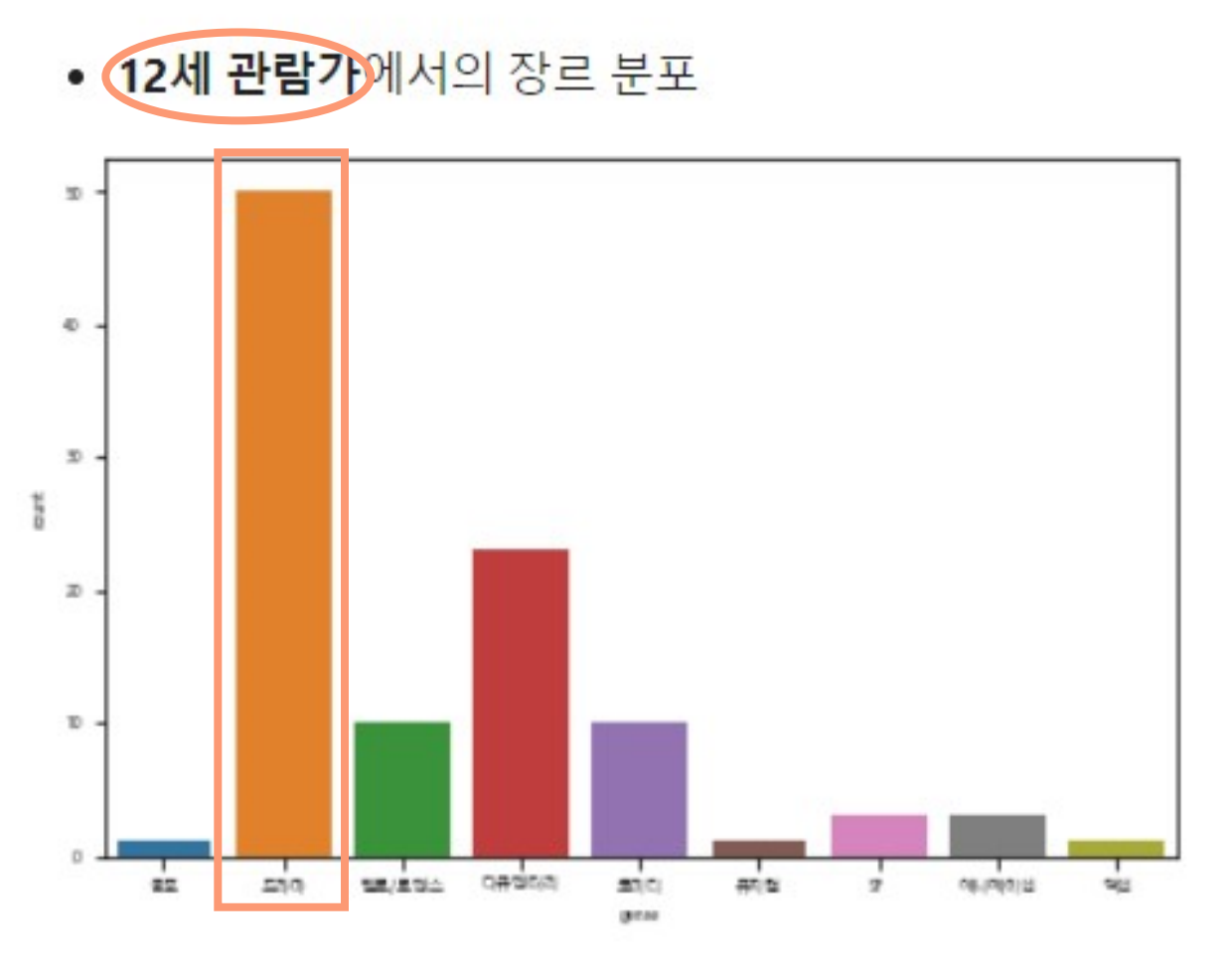
상영등급(screening_rat) 별 장르(genre) 분포
: countplot()

- movie_all = movie_df['screening_rat'] == '전체 관람가'
  movieall_df = movie_df[movie_all]
  sns.countplot(x = 'genre',
  data = movieall_df)

- '다큐멘터리' 장르가 가장 많으며, '드라마' 장르의 영화도 많음

EDA

상영등급(screening_rat) 별 장르(genre) 분포
: countplot()



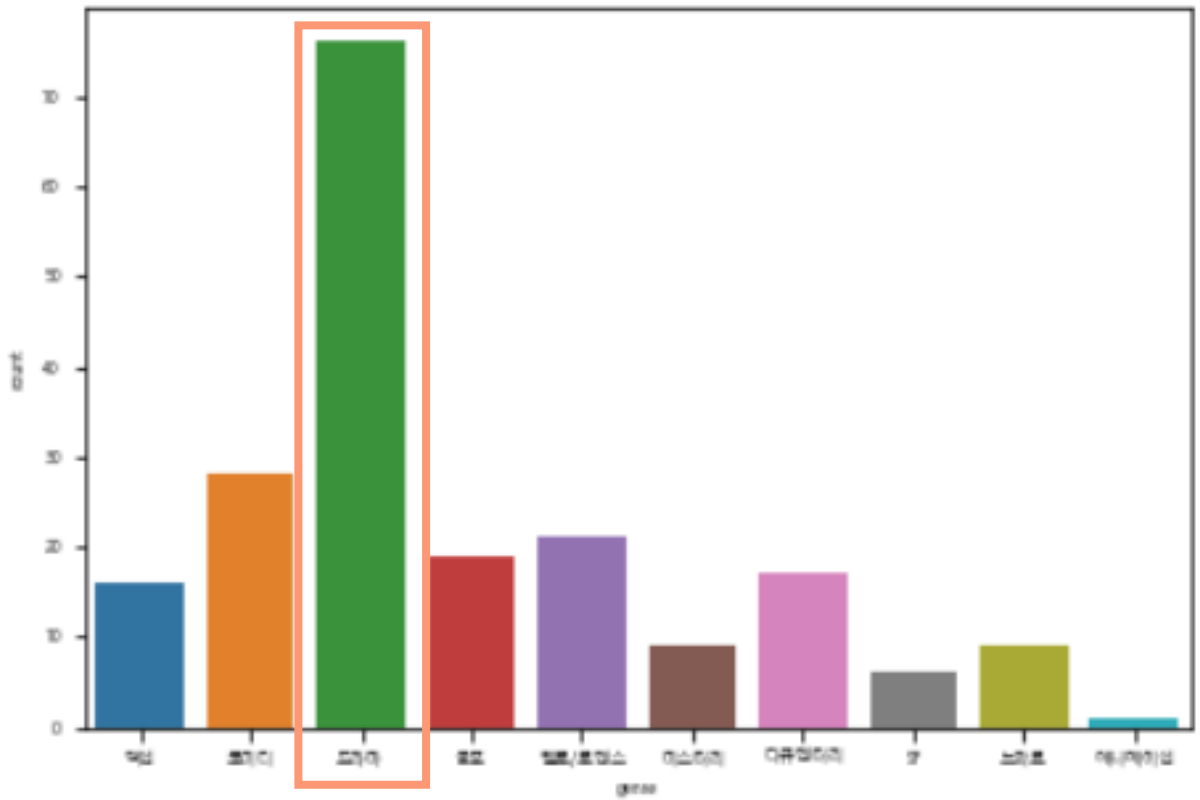
- movie_12=
  movie_df['screening_rat'] == '12세 관람가'
  movie12_df = movie_df[movie_12]
  sns.countplot(x = 'genre',
  data = movie12_df)

- '드라마' 장르가 가장 많음

EDA

상영등급(screening_rat) 별 장르(genre) 분포
: countplot()

- 15세 관람가에서의 장르 분포



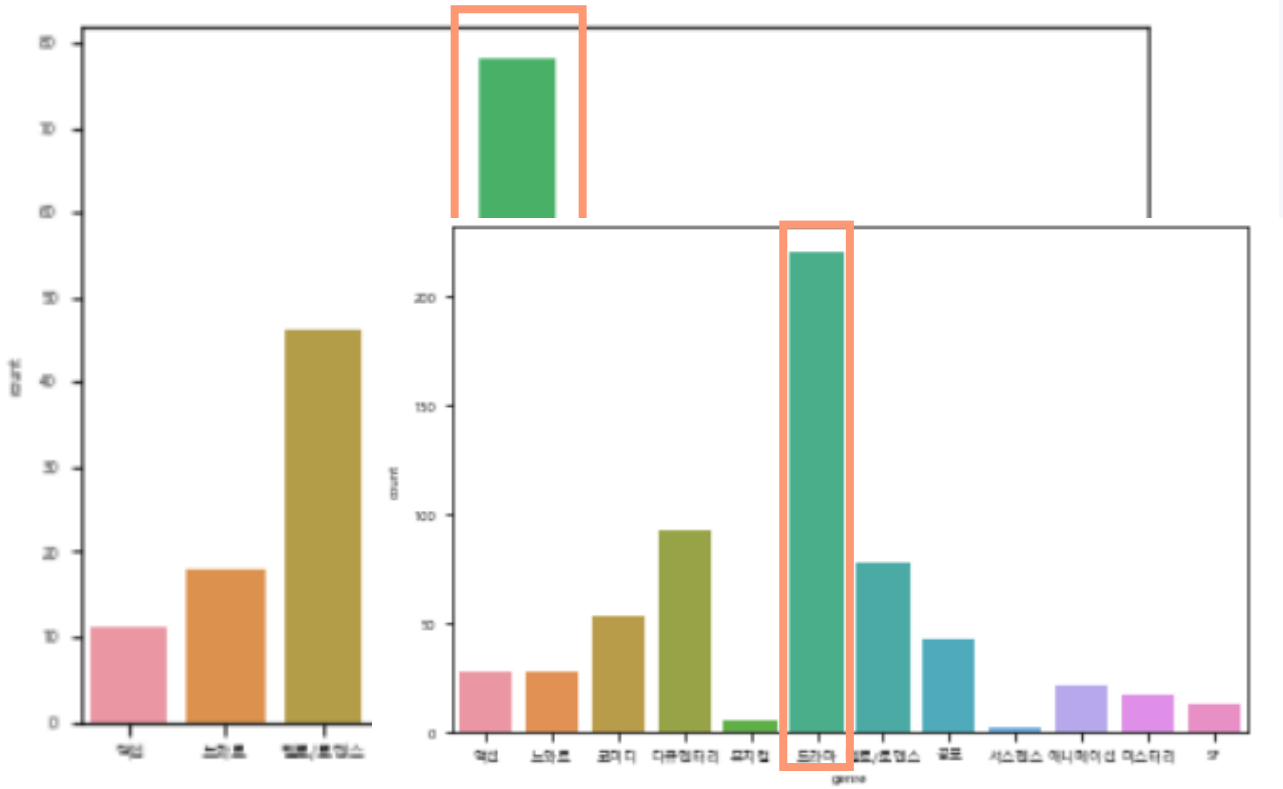
- movie_15=
  movie_df['screening_rat'] == '15세 관람가'
  movie15_df = movie_df[movie_15]
  sns.countplot(x = 'genre',
  data = movie15_df)

- '드라마' 장르가 가장 많음

EDA

상영등급(screening_rat) 별 장르(genre) 분포
: countplot()

- 청소년 관람불가에서의 장르 분포



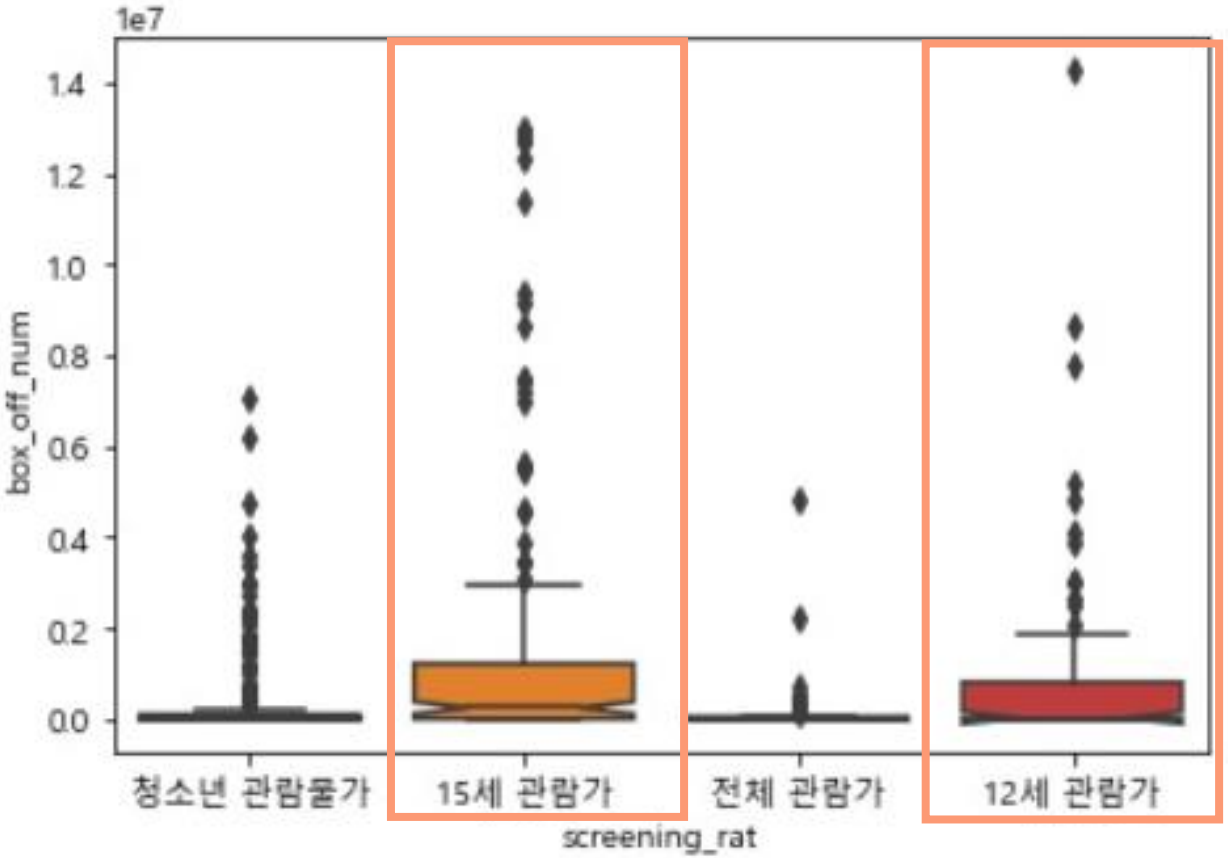
- movie_19=
  movie_df['screening_rat'] == '청소년 관람불가'
  movie19_df = movie_df[movie_19]
  sns.countplot(x = 'genre',
  data = movie19_df)

- 드라마' 장르가 가장 많음

- 대부분의 상영등급에서 '드라마' 장르가 가장 많은데, 2010년대 영화 중 '드라마' 장르의 영화 수가 가장 많기 때문

EDA

상영등급(screening_rat) 별 관객수(box_off_num) 분포 : boxplot()



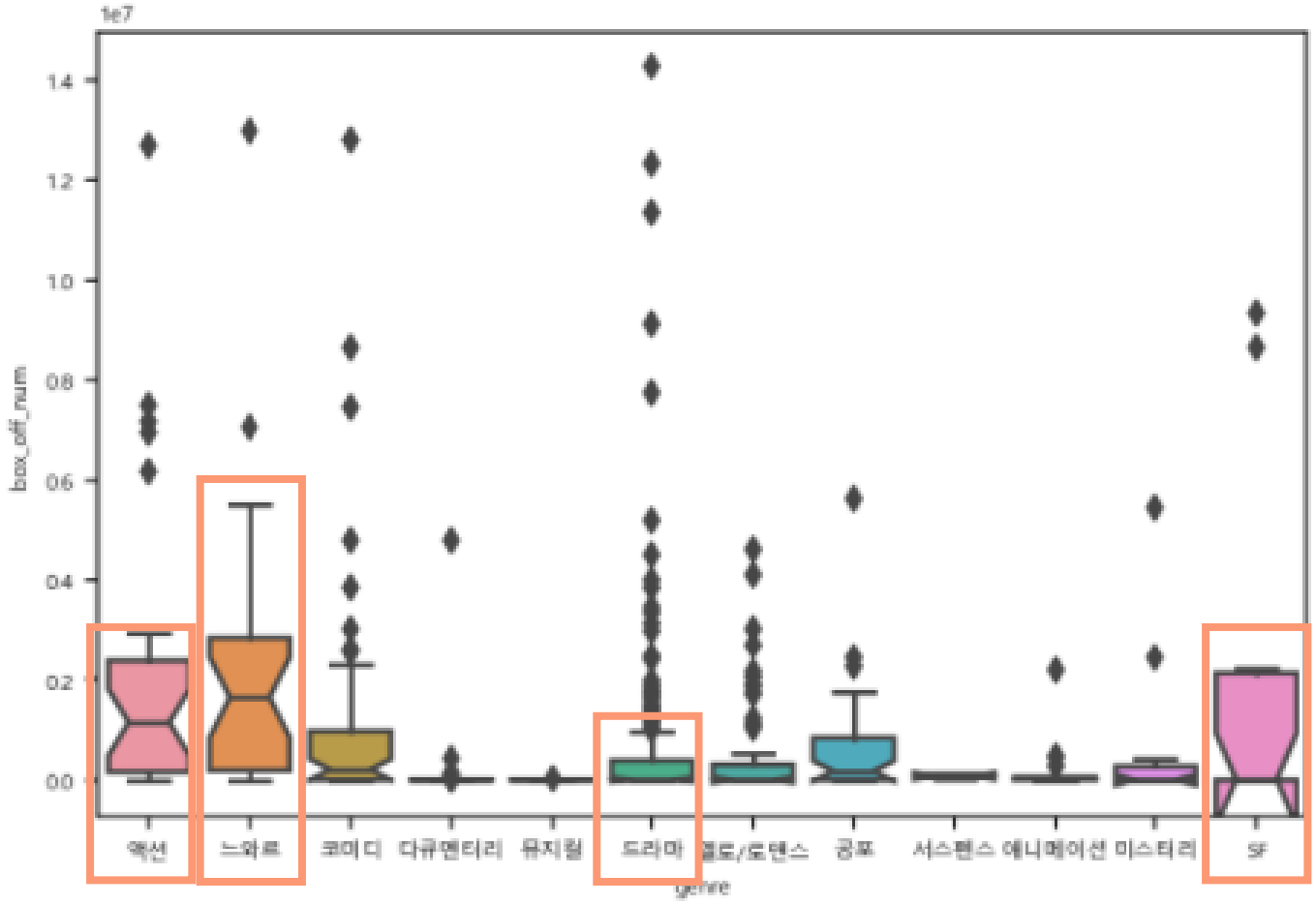
- sns.boxplot(x="screening_rat", y="box_off_num", data=movie_df, notch=True)

- '15세 관람가' 영화의 관객수 분포와 중간 관객수가 가장 큼

- '12세 관람가' 영화의 관객수 분포 또한 '청소년 관람불가'나 '전체관람가' 영화 보다 큼

EDA

장르(genre) 별 관객수(box_off_num) 분포 : boxplot()



- sns.boxplot(x="genre", y="box_off_num", data=movie_df, notch=True)

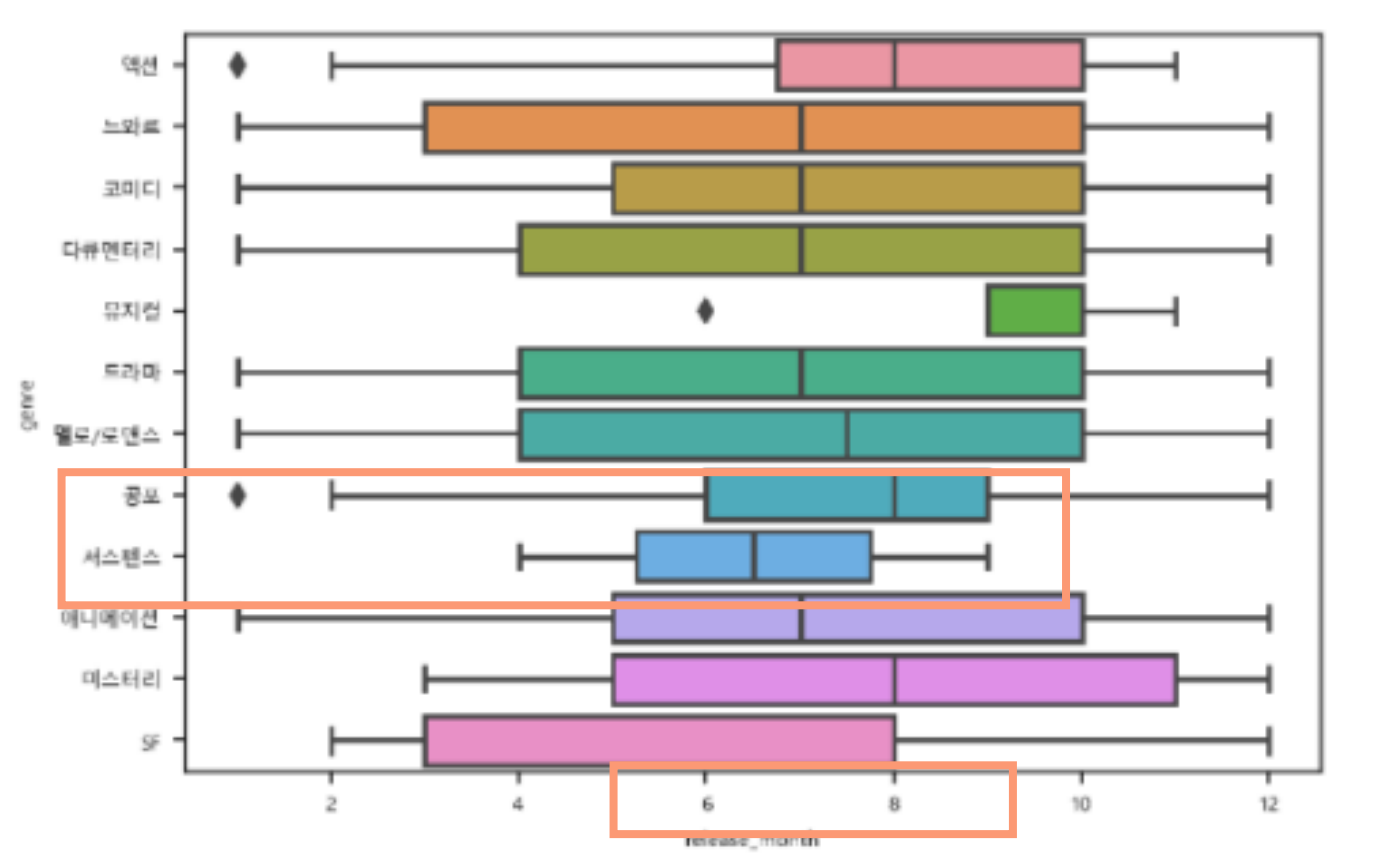
- '느와르' 장르의 영화 : 가장 광범위한 관객수의 분포와 가장 높은 중간 관객수 값을 기록

- '드라마' 장르의 영화 : 협소한 범위의 관객수 분포, 낮은 중간 관객수를 기록
  > '드라마' 장르의 특성

- '액션' 장르의 영화와 'sf' 장르의 영화 또한 광범위한 관객수의 분포
  > '액션' 장르의 영화가 'sf' 장르의 영화보다 중간 관객수가 높음

EDA

개봉월(release_month) 별 장르(genre)의 분포
: boxplot()



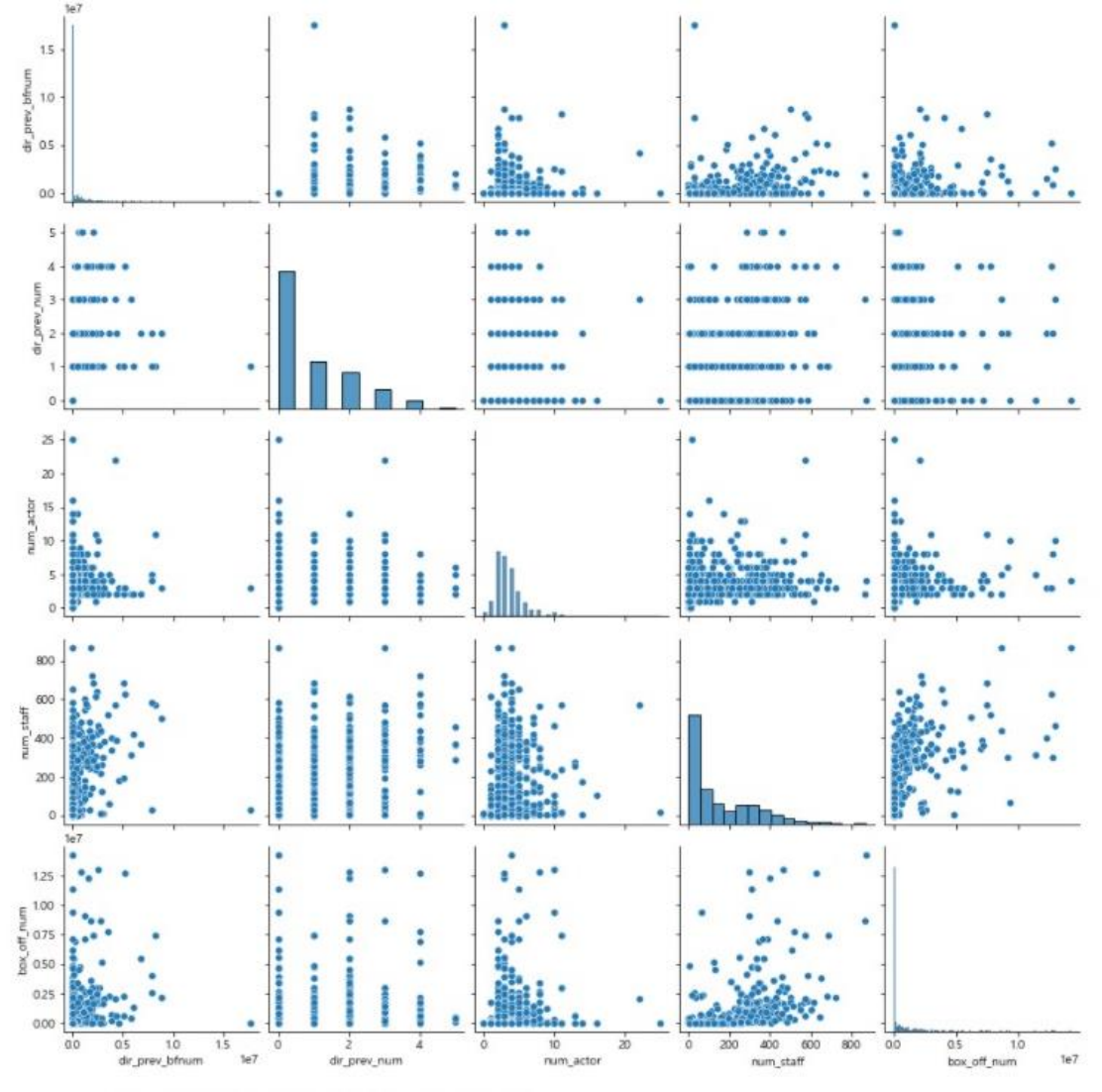
- sns.boxplot(x="release_month", y="genre", data=movie_df)

- 대부분의 장르들은 12개월 전반적으로 개봉하는 경향

- '공포' 장르와 '서스펜스 ' 장르의 영화의 개봉 달은 6월 ~ 9월에 집중
  > 대중들의 특성을 반영하여 특정 시기에 개봉

EDA

다양한 변수 간 관계를 확인
(감독의 이전작품 평균관객수, 감독의 이전 작품 수, 배우수, 스탭수, 관객수)
: pairplot()

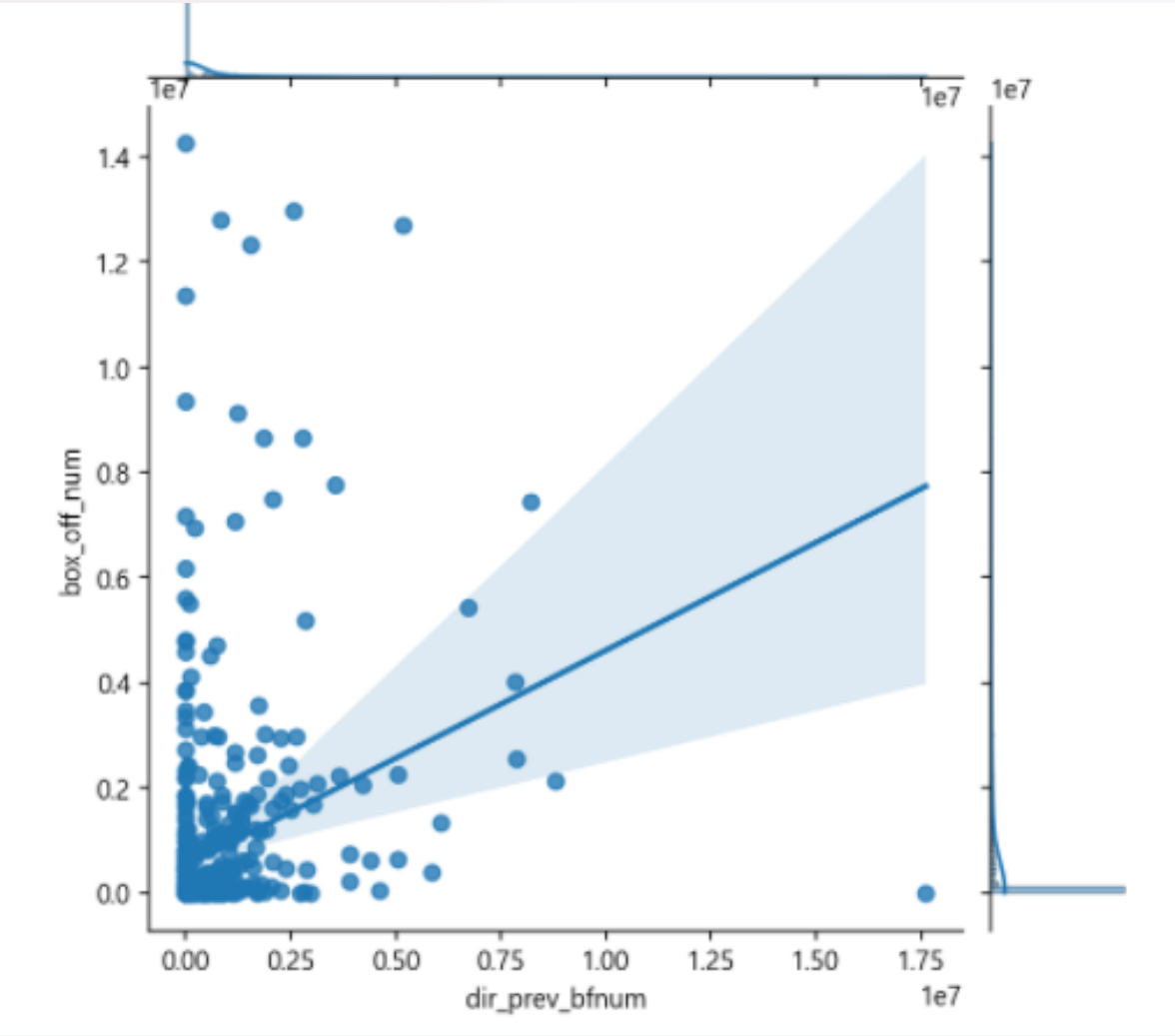


- ndf =
  movie_df[['dir_prev_bfnum','dir_prev_num',
  'num_actor','num_staff', 'box_off_num']]
  sns.pairplot(ndf)
  plt.show()
  plt.close()

- 변수 간 양의 선형관계 혹은 음의 선형관계가 나타나지 않음

EDA

감독의 이전 작품 평균 관객수(dir_prev_bfnum)와 관객수(box_off_num) 사이의 관계
: Jointplot(kind=reg)



- sns.jointplot(x='dir_prev_bfnum', y = 'box_off_num', kind = 'reg', data=movie_df)

- 오차가 심한 양의 선형관계를 보임
  > 선형관계로 보기 어려움

EDA

감독의 이전 작품 평균 관객수(dir_prev_bfnum)와 관객수(box_off_num) 사이의 관계 : LinearRegression()

- X = movie_df[['dir_prev_bfnum']]
  Y = movie_df[['box_off_num']]

- from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.5, random_state=10)

- from sklearn.linear_model import LinearRegression
  lr = LinearRegression()
  lr.fit(X_train, Y_train)
  r_square = lr.score(X_test, Y_test)
  print('회귀식 :', float(lr.coef_),'X +',lr.intercept_)
  print('결정계수(R^2) :', r_square)

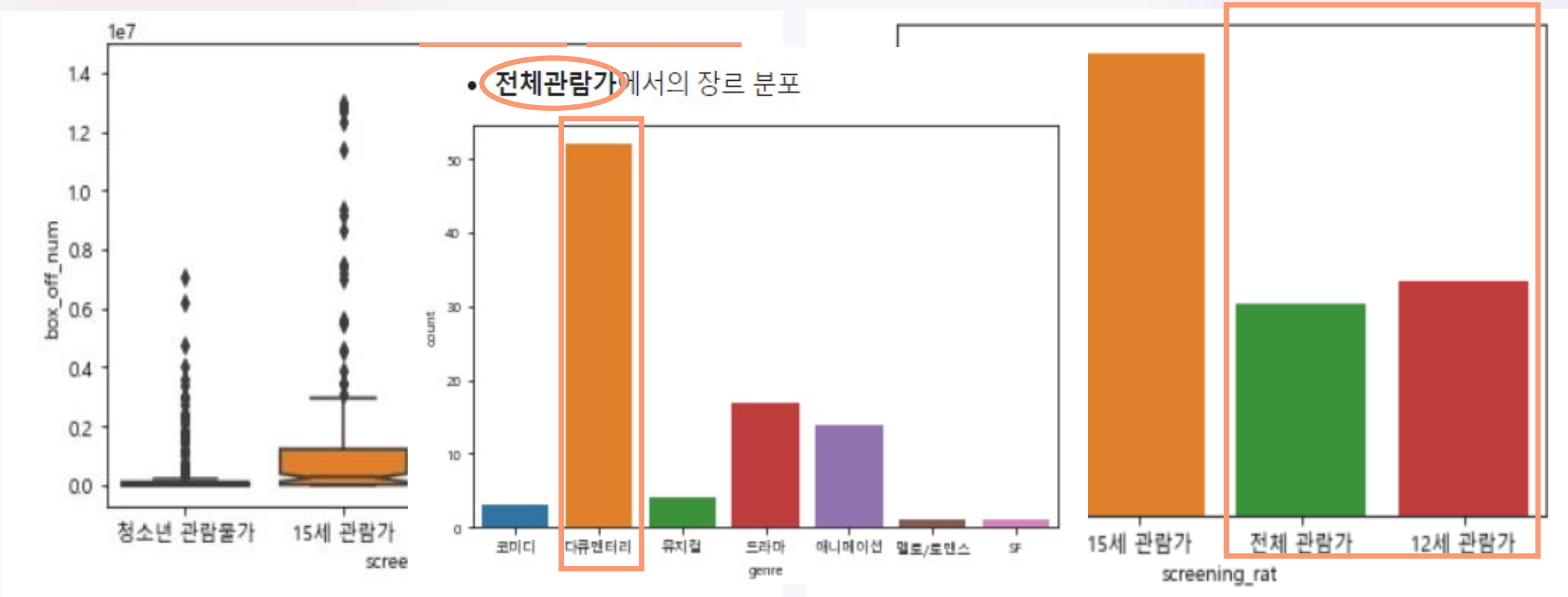
선형관계로 보기 어려움

감독의 이전 작품 평균 관객수와

관객수는 명확한 관계가 없음

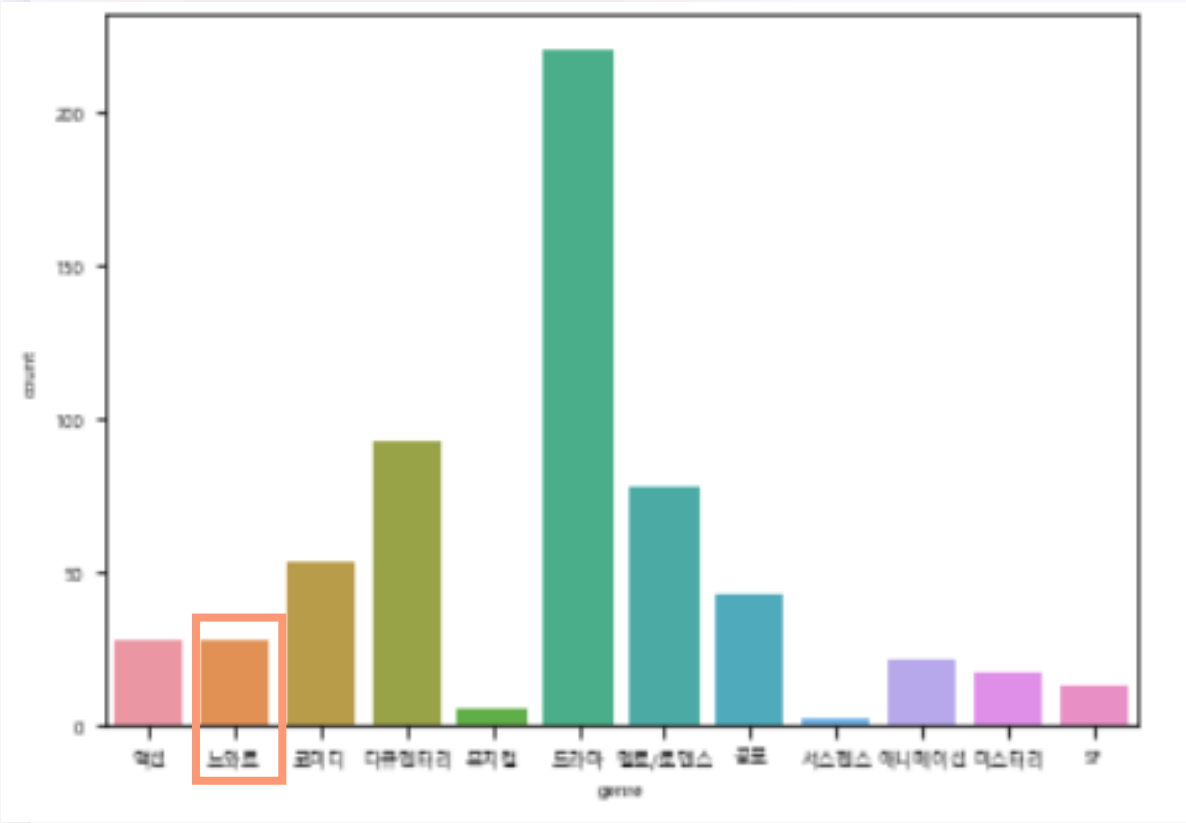
```
회귀식 : 0.31516624375491636 X + [726395.26929876]
결정계수(R^2) : 0.1299361444808902
```

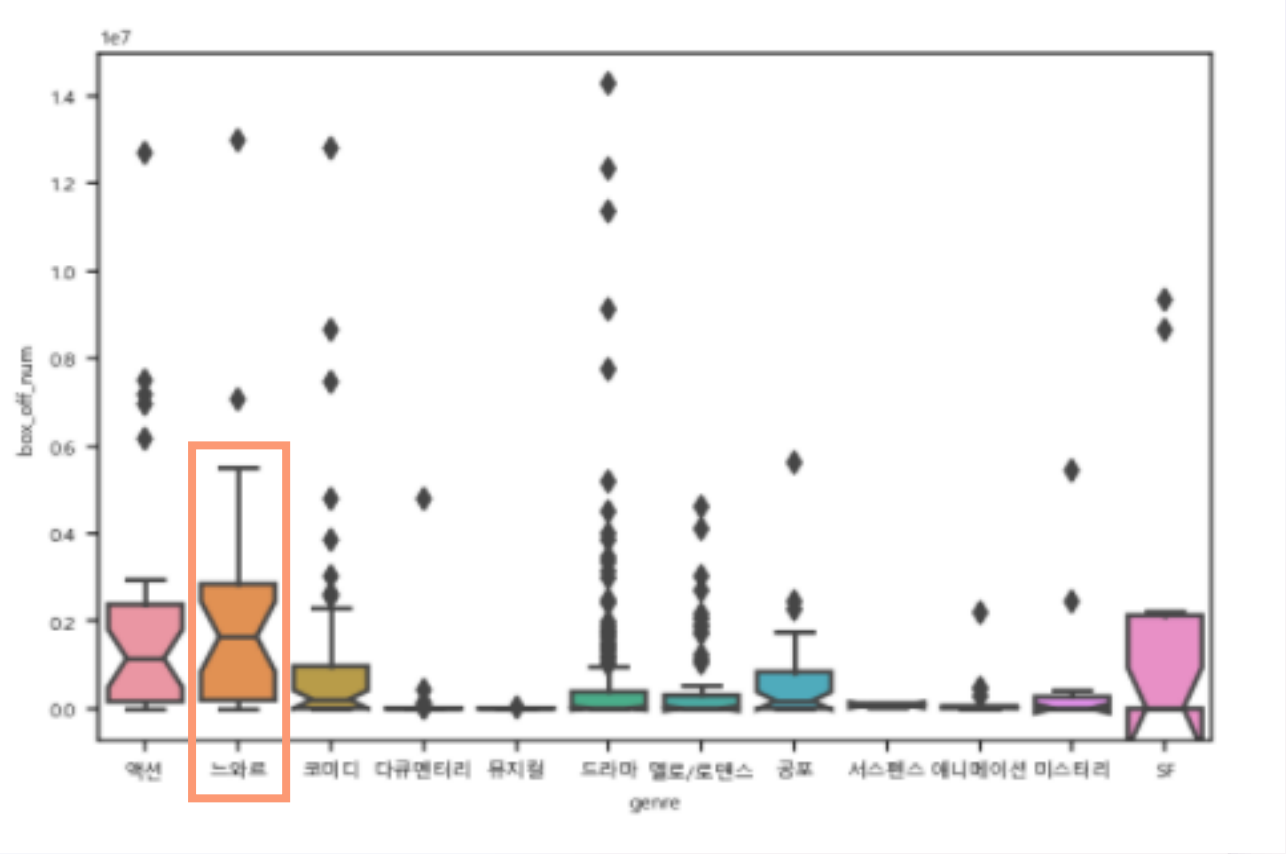
Insight - 1



- '12세 관람가' 영화의 공급과 출품을 촉진시킨다면, 더 많은 관객들을 유치 가능할 것
- 가족 단위로 즐길 수 있는 장르(코미디, 드라마 등)의 '전체관람가' 영화가 더 많이 공급되어야 함

Insight - 2





- '느와르' 장르란 범죄와 폭력세계를 다루는 영화 : 범죄와의 전쟁, 도둑들, 내부자들 (흥행작)

  > 흥행작의 장르에 따라 관객수 분포와 중간 관객수 값이 영향을 받을 것

- 흥행작이 속출한 2010년대의 사회문화적 배경 ?

  > 사회문화적 배경이 영화산업의 흥행 여부를 결정지을 수 있음

Thank you
Q & A