Analysis on IMDb Metadata

Wanhe Zhao (wanhez) Yue Yin (yuey2)

Problems

Goal: Give people insights into IMDb movie “top-ratedness” (i.e. the score IMDb assigned to the movie).

Aspects we analyze: Movie Genres, Release dates, and movie descriptions.

Can those features be the significant features that affect how people votes?

Data Collection & Processing

10264	The
10265	

about 40,000 observations

Definition of “top-ratedness”

We used the formula IMDb uses to determine top rated movies:

The formula for calculating the Top Rated 250 Titles gives a true Bayesian estimate: weighted rating (WR) = (v ÷ (v+m)) × R + (m ÷ (v+m)) × C Where:

- R = average rating for across movie = (Rating)
- v = number of votes for across movie = (votes)
- m = a good number of votes to rely on the rating based on votes (m as the average number of votes)
- C = the mean of all movies

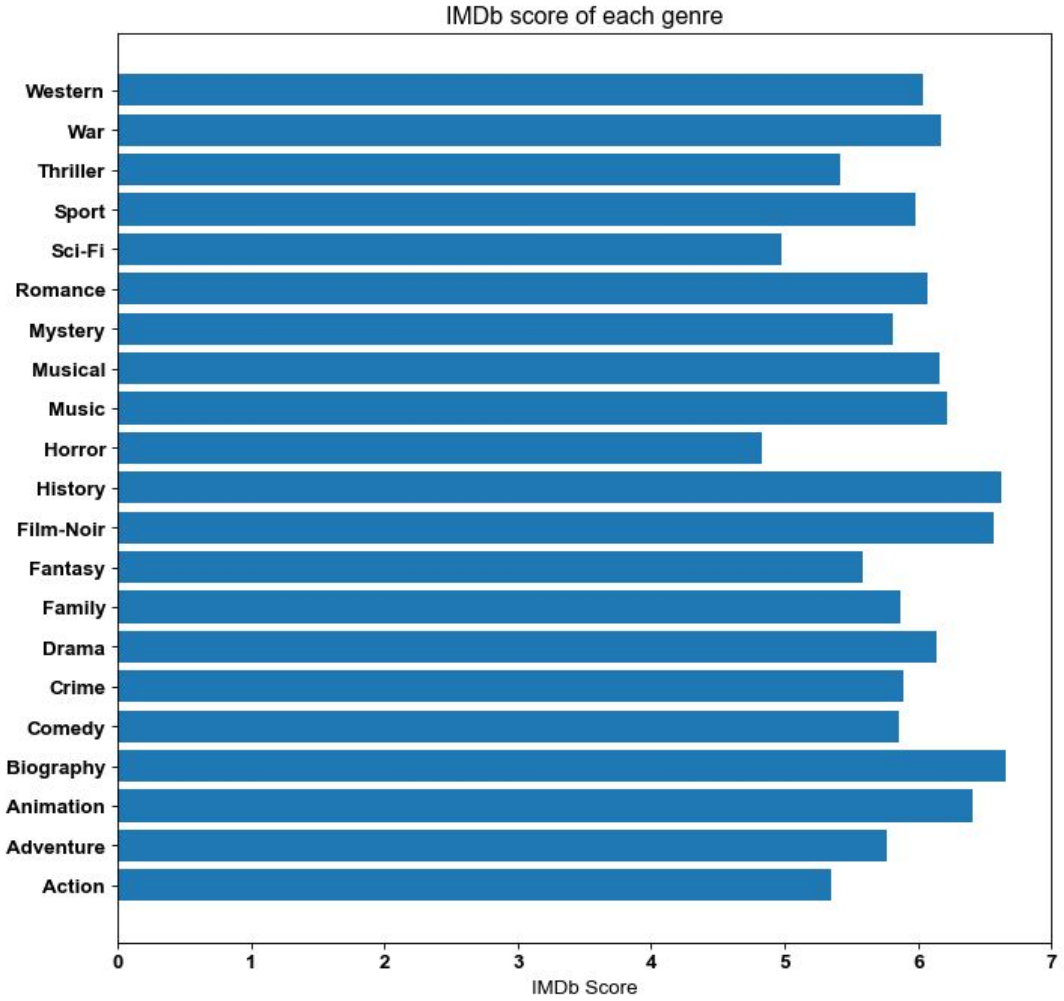
From: https://help.imdb.com/article/imdb/featured-content/why-doesn-t-a-title-with-the-average-user-vote-of-9-4-appear-in-your-top-250-movies-or-tv-list/GTU67Q5QQ8W53RJT?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=e31d89dd-322d-4646-8962-327b42fe94b1&pf_rd_r=T4PZFWW3YAJD79YH14HS&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=cons_chttp_learnmore#

GENRES

Genres

Top categories:

- Biography
- History
- Film-Noir
- Animation



Really? Let's do a hypothesis test.

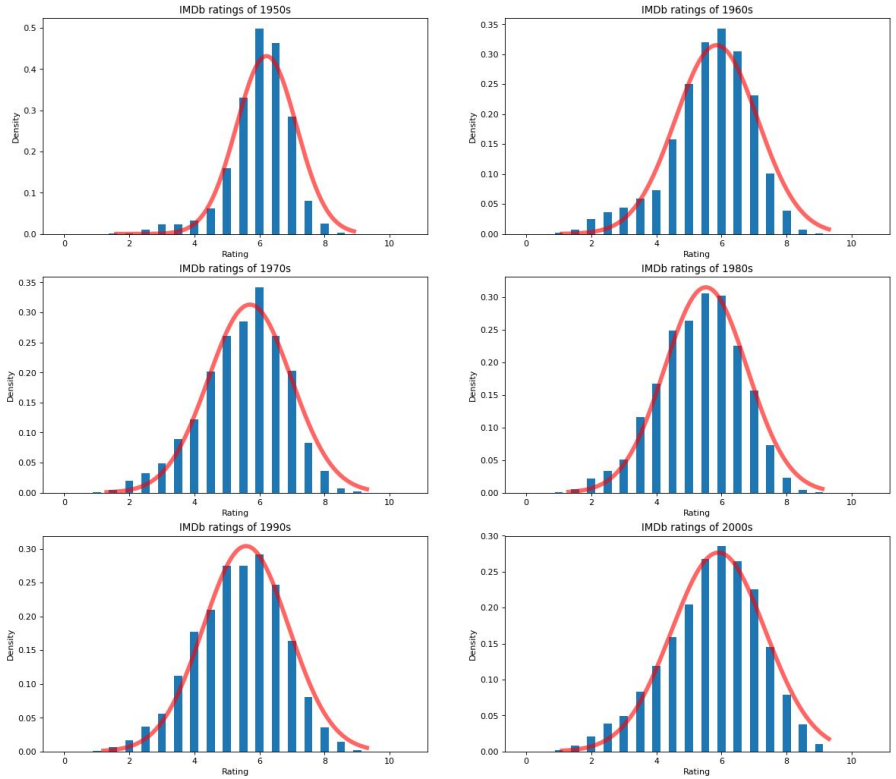
Null Hypothesis: There is no difference in scores between each one of the top genres and the other genres

Genre	p-value
Biography	1.211337525941477e-100
History	9.992623594019654e-55
Film-Noir	4.301388884286685e-46
Animation	4.358555890585483e-33

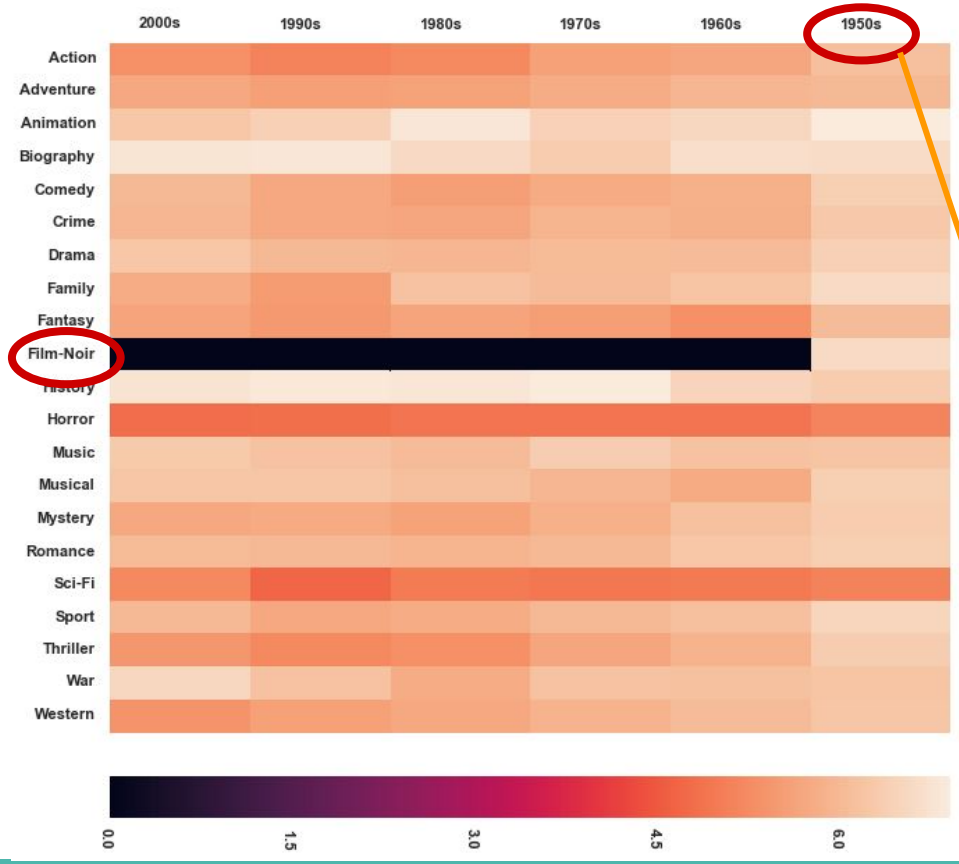
Do the ratings distribute normally over the years? Yes.

Histogram of user ratings for different decades.

Overlay the normal density curve on top of the histograms, we find that they fit well.



What genres stand out in every decade?

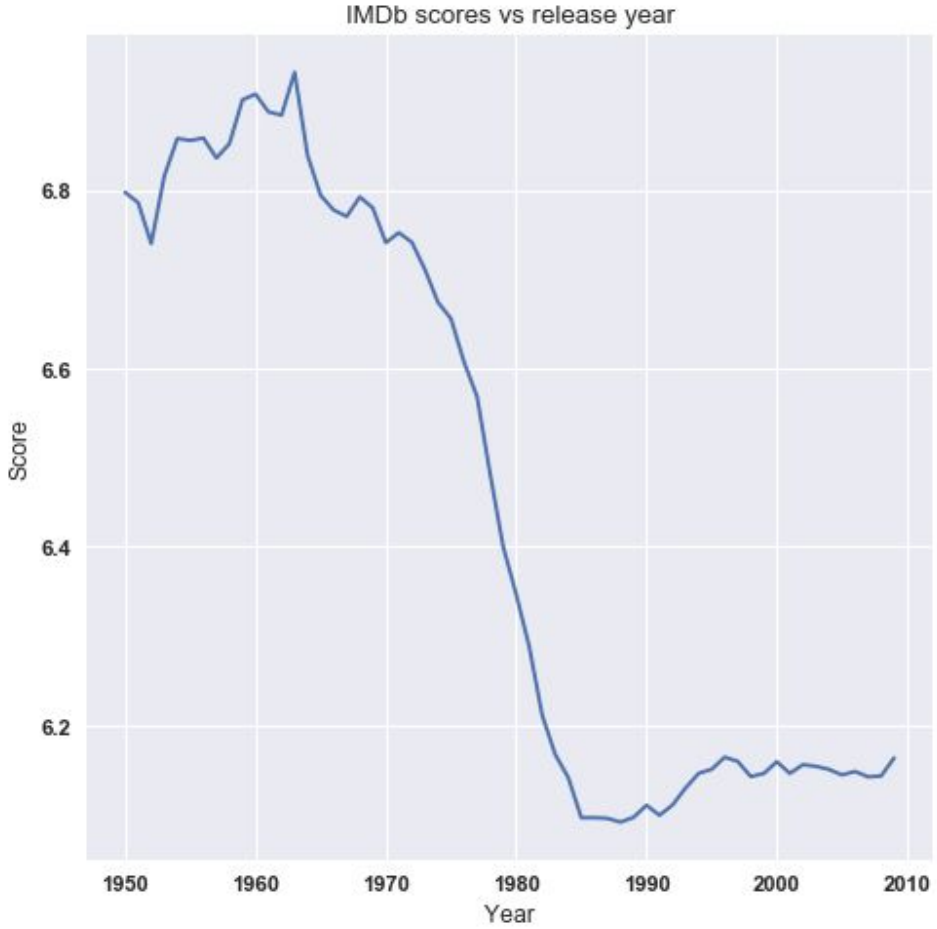


Film noir (/fɪlm nwɑːr/; French pronunciation: [film nwaʁ]) is a cinematic term used primarily to describe stylish Hollywood crime dramas, particularly those which emphasize cynical attitudes and sexual motivations. Hollywood's classical film noir period is generally regarded as extending from the early 1940s to the late 1950s. Film noir of this era is associated with a low-key, black-and-white visual

RELEASE DATE (YEARS)

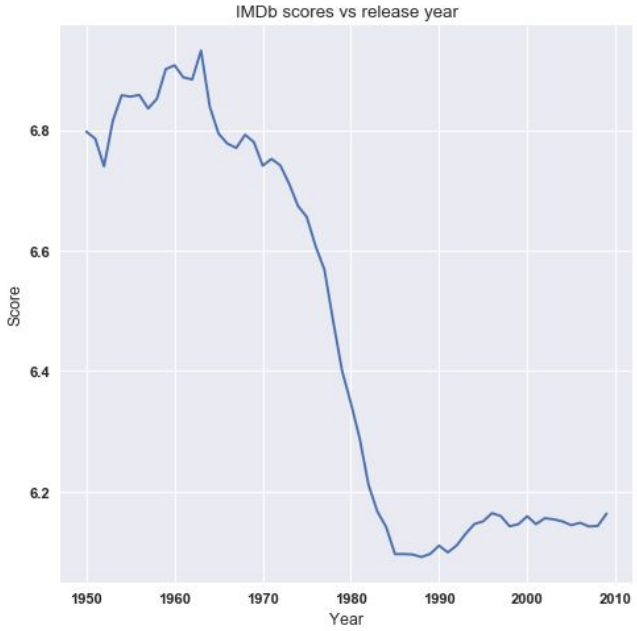
Method: Time Series analysis (sliding averages) to smooth out short-term fluctuations and find the longer-term trend.

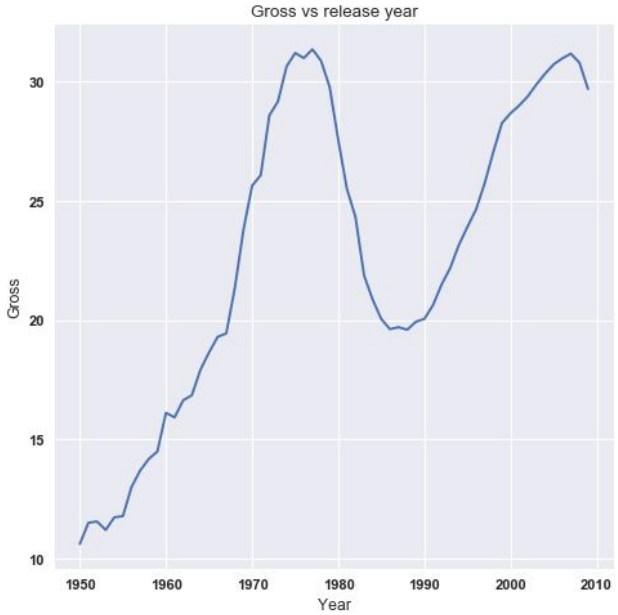
Where does the drop come from?

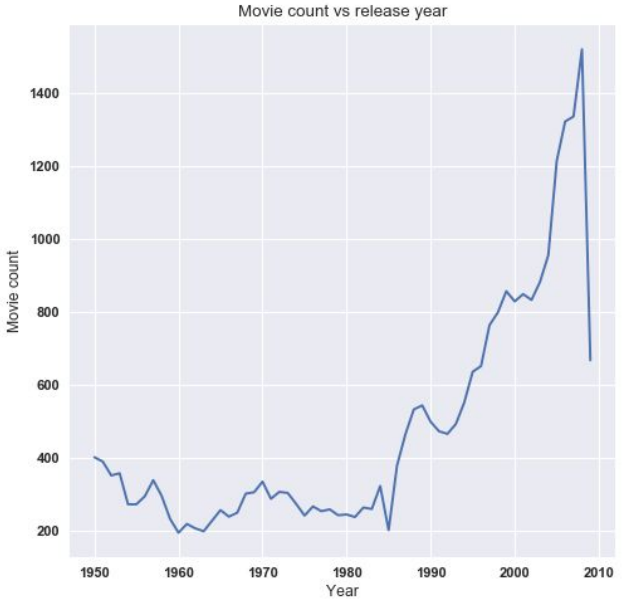


Around 1980s and 1990s there is a drop in average movie scores and gross, but an increase in the number of movie produced.

Guesses: lower production cost? Copyrights violations brought by VCRs?







GROSS

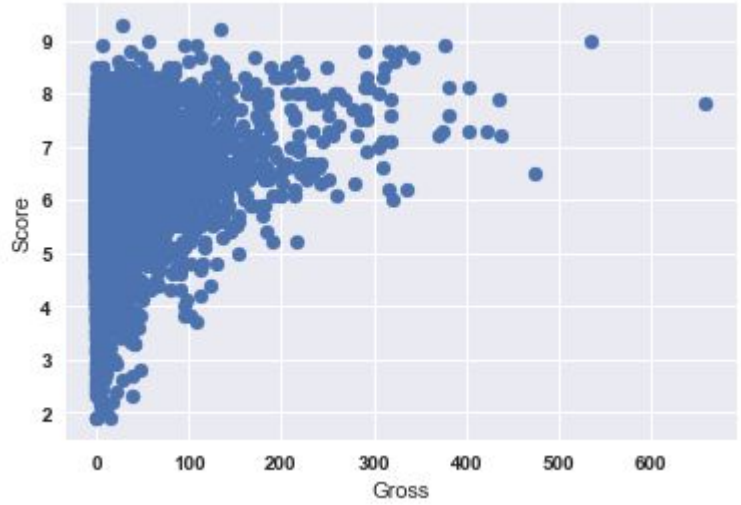
Linearity

X axis: Gross

Y axis: Score

$$\text{Score} = \beta_0 + \beta_1 \text{Gross} + \epsilon$$

Significant p-value for beta1.




```
Residuals:
Min      1Q  Median      3Q     Max
-2.3669 -0.6797 -0.3987 -0.1755  4.9582


Coefficients:
            Estimate  Std. Error   t value  p value
_intercept  6.787755    0.041151  164.9456      0.0
x1          0.005127    0.000454   11.3017      0.0
```

LINEAR REGRESSION ON GENRES, YEARS, GROSS

Linear Regression

library: scikit learn

Variables: Years, Gross, Genres

$$\text{Score} = \beta_0 + \beta_1 \text{Year} + \beta_2 \text{Action} + \beta_3 \text{Adventure} + ... + \beta_{21} \text{War} + \beta_{22} \text{Gross} + \epsilon$$

P-value < 0.05

Observation: based on beta and p-values, we find that with same year and gross, people rate action, family, horror lower, but Animation, Biography, Drama higher

	beta	p-value	variable
0	-0.012754	0.000000e+00	Year
1	-0.385830	0.000000e+00	Action
3	0.402649	5.005145e-10	Animation
4	0.422220	1.132427e-14	Biography
5	-0.199091	8.881784e-16	Comedy
6	0.084367	3.433299e-03	Crime
7	0.494219	0.000000e+00	Drama
8	-0.345096	3.330669e-15	Family
10	0.645667	2.887211e-02	Film-Noir
11	0.203504	5.309167e-03	History
12	-0.341172	0.000000e+00	Horror
20	0.165668	3.394111e-02	War
21	0.007352	0.000000e+00	Gross

Movie Description

Scores and Movie Summaries

Naive Bayes Classifier library: nltk

Most informative words:

british = True	pos	: neg	=	15.7 : 1.0
weekend = True	neg	: pos	=	14.7 : 1.0
hell = True	neg	: pos	=	13.8 : 1.0
crazed = True	neg	: pos	=	11.9 : 1.0
zombies = True	neg	: pos	=	11.7 : 1.0
beast = True	neg	: pos	=	10.9 : 1.0
widow = True	pos	: neg	=	10.5 : 1.0
english = True	pos	: neg	=	10.5 : 1.0
determined = True	pos	: neg	=	10.1 : 1.0

Maybe watch movies that has “british” and avoid movies that have “weekends”, “zombies”, “crazed”.

Conclusion

The score of movie from IMDb is a good measure of movie quality.

However, the scores varies significantly across genres.

Be aware of those factors (year, genres, ...) and don't blindly believe the scores!

A 4 star horror movie may already be the best!

Future Investigation

Explore more variables that may explain the variability in the scores.

Investigate more on some abnormal observations.

Run machine learning models for personalized recommendation.

Sources Consulted

Libraries:

Sklearn, regressors, nltk, pandas, numpy, requests, beautiful soup

Sources:

https://pythonprogramming.net/words-as-features-nltk-tutorial/?completed=/text-classification-nltk-tutorial/

https://www.quora.com/How-does-IMDbs-rating-system-work

https://stats.stackexchange.com/questions/189658/what-are-good-resources-on-bayesian-rating