What Makes a Box Office Hit?

Shanshan Yu



# Preliminary Hypotheses

The movie industry is a major business – in 2016, the global box office revenue was measured at about $38 billion dollars (Fuller). As a result, more and more studios are producing movies faster than they have in the past. Often, these movies are remakes or sequels to prior movies. Today’s moviegoer has more choices than ever before. Sometimes crowds flock to see particular movies, with crowds lining up outside theaters to watch them:



*Source:* [*http://starwarsblog.starwars.com/wp-content/uploads/2015/10/star-wars-graumans-1536x864-476674104388.jpg*](http://starwarsblog.starwars.com/wp-content/uploads/2015/10/star-wars-graumans-1536x864-476674104388.jpg)

Other movies end up being flops, in which the studios who produced them either lose money, or don’t make enough to cover the costs associated with their production. We wonder, is there a way to predict whether a movie will be a hit and bust the box office, or will it flounder and flop?

# Problem Statement

With the average budget of a major Hollywood film being $100 million or more (Mueller), production companies would benefit greatly from being able to determine if a movie will be a box office success or not before investing in that movie. Group C aims to test and quantify through predictive analysis if, given certain information pertaining to a film, if we can predict how successful it will be in terms of revenue created, as well as if it a critical hit.

# Data Source and Description

The data source Group C will be using comes from Kaggle, and includes data on over 5,000 movies. Included in this data set is a wide variety of fields, including genre, actor/actress names and Facebook likes, director name and Facebook likes, budget, IMDb scores, and many more. All in all, there are 28 variables in the original data set. Movies included in the data set range all the way from recent releases, to movies over one hundred years old. This data set is not limited to just American movies: movies from 66 countries are included.

In addition to the Kaggle data set, Group C also plans to use web scraping techniques to quantify social media sentiment from sites like Twitter. Additional data about the movies may be downloaded or scraped from websites such as FilmSite.org, The-Numbers.com, or FilmJabber.com.

# Data Preprocessing Steps

Due to the fact that our main data set came from Kaggle, it is relatively clean: major preprocessing is not necessary with this data set. However, some preprocessing is necessary to ensure our analyses do not return unwanted results.

One key preprocessing is to eliminate null value in the analysis. For example, some records are missing key fields such as director name or budget. Records missing crucial fields such as these must be excluded from our data set before beginning our analysis. Also, some movie titles, director names, and actor names contain characters that are not included in the standard English dictionary, which can cause problems. For the purpose of our analyses, we dropped any alpha characters not included in the standard English alphabet. Another important preprocessing is to correct wrong numeric records which are mainly found in the budget and gross profit data. Because some foreign movies are documented in the foreign currency and the data does not indicate the currency types. One good example of this is the famous Japanese movie “Godzilla 2000”. The budget of the movie appears to be way much higher than its peers, because the data is documented in Japanese Yen, a currency of an exchange rate over 100 yen: 1 dollar during the period of movie production.

A secondary dataset we brought in had to do with movie franchises, as our original data set did not contain any data on franchises. In this dataset, we removed any records without gross revenue numbers or release years. Additionally, we stripped non-alphanumeric characters from movie titles, as well as cleaned up blank spaces at the beginning and end of the titles. Furthermore, we releveled the categorical variables for linear regression modeling.

## Hypothesis: Movie Franchises are the Rising Box Office Stars

In order to test this hypothesis, we needed to bring in an additional data set – in this case, movie franchise data. After preprocessing this data to remove records missing important data, we then attempted to normalize gross revenue figures by accounting for inflation. To do this, we first scrapped the historical consumer price index ((CPI-U)) Data, which provides a year-to-year inflation reference of the past 100 years. With these detailed and accurate references, we scaled all the gross profit value of the past to the value that they would be worth in the most recent year 2016.

By accounting for inflation, we were able to more accurately determine which films were the biggest successes relative to all others. After adjusting revenues with inflation, we then compared the revenue of franchise movies versus non-franchise movies and total number of movies in both categories. To visualize this comparison, we created three barplot (Figure 18, Table 1-1, 1-2). it is clear that franchises on average bring in much more revenue than non-franchises, more than double in fact, although franchise movies are about one third of the other category.

After establishing that movie franchises are more valuable as a whole than single films, we moved on to performing K-means clustering on the financial data of the franchises. Before this could be done, additional data preprocessing was necessary. The preprocessing performed included removing non-numerical symbols from the pertinent financial fields, as well as removing records without financial information and writing out a detailed csv file for the franchise movies (“FranchiseMovieDetails.csv”)

The elbow plot clearly indicates that the k-means cluster performance reaches a diminishing improvement when k is over 4 (Figure 19). Therefore, we used k of 4 to perform the k-means clustering. Among all four the four clusters, the cluster 3 and 4 together accounts for 3.69% of all the franchise movies (Figure 20). The ggplot() function further characterizes more features of four clusters through side by side plots. Cluster 3 only has four franchises and it is the most profitable and successful group. They brought in total gross profit from three to seven billion dollars and three out of them brought in over five billion. All the franchises in this cluster are “alive” and new movies are lined up to 2020, regardless the fact that they already had more movies than most other franchises. The franchises in this cluster are the “Star Wars”, “James Bond”, Marvel superhero movies, and “Harry Potter” (Table 2).

The second most profitable group is cluster 4, with profit ranging between one to three billion. This group includes franchises like "Batman"，"Star Trek"，"Peter Jackson Lord of the Rings"，"Spider Man"，"X Men"，"Jurassic Park"，"Indiana Jones"，"Superman" etc. Cluster 1 and 2 are on the lower end of gross profit and the latter is the least profitable group account for 81% of all franchises. The domestic box office records seemingly well correlate with the box offices of other countries (Figure 21). Through this interesting phenomenon, not only we found these movies successfully overcame different culture barriers, but also they demonstrated their impact to the world.

The k-means clustering results make use wonder the contributing factors leading to successful franchise movies. Following k-means, we applied linear regression modeling techniques to both franchise movie data and the movie meta dataset. Prior to the regression, the first processing focused on mapping movies to their corresponding franchise titles. Movie franchises, like “Star Wars” and “Harry Potter” include a number of movies, but two thirds of movies in the movie meta dataset are not franchises. The mapping were done through the function grep(), followed by a close examination and correction to the movie titles starting with numbers.

To perform linear regression modeling, the NA value needs to be removed. This is relatively easy comparing to other processing we have to deal with. The second round of preprocessing including deep data cleaning that requires knowledge of movies. Because our movie dataset has a century time span and the MPAA rating changed multiple times towards the end of 1970s. The documented MPAA film rating mixed up the old and new rating system. The rating data is consolidated, cleaned and releveled after we studied the history of MPAA rating system. Similar preprocessing was also performed to categorical variables years and languages. Majority of films are rated at “PG”, “PG-13” and “R” (Table 3). On the other hand, a lot of franchise that has the same first year and last year will have total year value as “NA”. This is probably due to the unit (year not month) of total year leads to the data left censored. To avoid the left censored data, use 0.5 year (6 months) as an estimated average of movie running time and use 1 year for franchises having more than one movie but only lasting for a year. Most countries

Two linear regression modeling were performed. The first modeling only included numeric variables and the second one also included categorical variables with all the numeric variables. The categorical variables, “director” and “actors” each has hundreds of levels and were not considered because the number of data points were very insufficient. The initial round of regression modeling shows data was still not enough for most countries, languages and movie years. Besides, most countries with coefficient calculated showed higher P-value (over 0.4). Therefore, only “color” and “content\_rating” were used for modeling (Table 4).

To visualize the prediction, we plot two predictions of each data point against their actual value (Figure 22) The magenta dot lines in both graph are the lines when predictions equal to the actual value (y = x). Therefore, the closer data points to the magenta lines, the less errors prediction are. Both graphs show that the data points in lower profit range (less than 200 million) deviated the magenta line less than in the higher range, indicating both prediction has better performance to predict the lower range of profit. To compare the prediction errors from two models, we constructed over prediction vectors, which calculated the percentage of actual value over predicted (if positive) or under predicted (if negative). With boxplots, we further visualized the prediction error distribution from two models with boxplots statistics calculated (Figure 22). Both box plots show a number of outliers in the over prediction vectors. With the incorporation of categorical variables, half of the over prediction range shrank from [-0.32, 1.3] to [-0.17, 0.79] and the median value decreased from 0.33 to 0.018 (Tale 5). This indicated the categorical variables in the models improved the model performance.

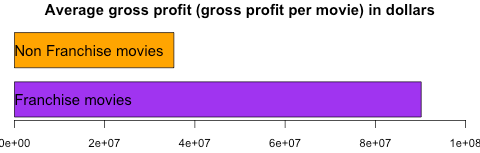
To further explore how individual variables contributing to the gross profit, we mapped the gross profit against each variable using scatter plots. The green line in each plot indicates the linear regression relationship between the individual variable and gross profit of franchise movies. Figure 23 shows variables contributing to the gross profit are “number of voted users”, “number of users for reviews”, “number of critic reviews”, “budgets” and “IMDB scores”. The first two variables imply that the words of mouths and audience’s opinions promoted the movie ticket sales. “Number of critic reviews” and “IMDB scores” reflect professional opinions on the movies, which should base on the movie quality. “Budgets” could potentially affects the movie quality during production and further affect the movie box office. As for content rating, the “Approved” rating was to the franchises made before 1978, a golden time in movie history when several famous movie franchises were made, “Star Wars”, “God Father”, “Batman” etc. The higher profit distribution (after inflation adjusted) of the “Approved” movie indicated this golden time. Other MPAA ratings are implemented after 1978 from a single “approved” to more detailed content limit. The ratings “PG”, “PG-13” have similar amount of total franchise movies comparing to rating R (Table 3), and they are more profitable. One possible explanation is that they are made for more audiences. Figure 24 shows variables that appear to be less or no regression relationship to the gross profit.

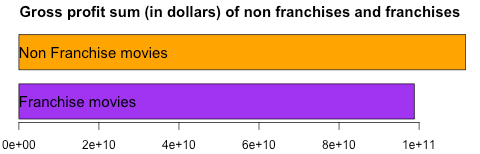
Using similar optimization strategies, we also applied linear regression models to predict profit on meta movie dataset. Although the meta movie dataset has 3 times more records than the franchise movie data, the amount of data records are still not enough for neither categorical variables “directors” and “actors”, nor “country”, “language” and “movie years”. Only “color” and “content rating” are included in the second prediction for the meta movie profit prediction. The comparison of two predictions for the movie meta data are shown in Figure 25 with a boxplot comparing the over perdition (as a measurement of errors) distribution. The incorporation of “color” and “content rating” also improved the model performance of the meta movie profit (Table 6), but the overall accuracy was better for franchise movie profit prediction.

Similarly, the “number of voted/reviewed users”, “budget”, “IMDB scores” “total FB like to cast” and “content rating” also had positive effects on the gross profit (Figure 26). “Total FB like to cast” shows impact of the social media on the movie business. This impact was trivial to franchise movie, probably due to most franchises had established their name, reputation or audience market through their earlier movies. For non-franchise movies, “total FB like” might be more important in promoting box office sales. Comparing to the “total FB like”, the “FB likes to individual actors/directors” show weaker effects on the gross profit (Figure 27).

# Appendix

Figure 18. The comparison of gross profit (average and sum) and total movie numbers between franchise and non-franchise movies.





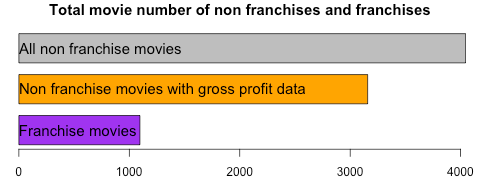


Table 1-1. Total number of franchise and non-franchise movies.

Franchise movies 1096

Non franchise movies with gross profit data 3158

All non franchise movies 4045

Table 1-2. Gross profit sum (in dollars) of franchise and non franchise movies.

Franchise movies Non Franchise movies

98819051354 111622020100

Figure 19. Elbow plot for k-means clustering.

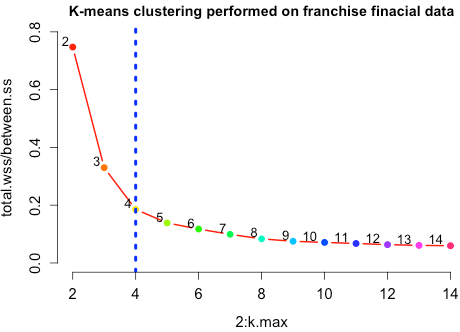


Figure 20. composition of franchise movie clusters.

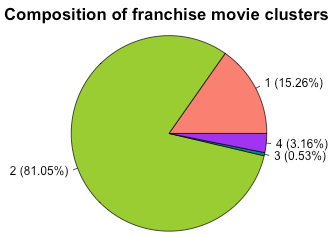
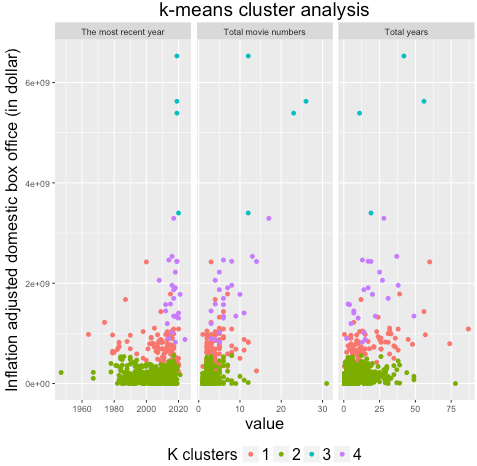


Figure 21. K-means clusters analysis.



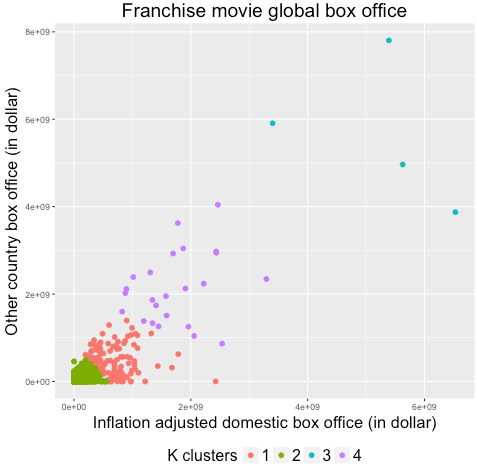


Table 2. Titles of two most profitable franchise movie clusters 3 and 4.

Cluster 3

[1] "Star Wars"

[2] "James Bond"

[3] "Marvel Cinematic Universe"

[4] "Harry Potter"

Cluster 4

[1] "Batman"

[2] "Star Trek"

[3] "Peter Jackson Lord of the Rings"

[4] "Spider Man"

[5] "X Men"

[6] "Jurassic Park"

[7] "Indiana Jones"

[8] "Superman"

[9] "Shrek"

[10] "Pirates of the Caribbean"

[11] "Fast and the Furious"

[12] "Transformers"

[13] "Hunger Games"

[14] "Twilight"

[15] "Dark Knight Trilogy"

[16] "DC Extended Universe"

[17] "Planet of the Apes"

[18] "Mission Impossible"

[19] "Despicable Me"

[20] "Iron Man"

[21] "Ice Age"

[22] "The Hobbit"

[23] "Avatar"

[24] "Madagascar"

Table 3. Movie MPAA rating summary after data cleanup.

Content rating Approved G NC-17 PG PG-13 R Unrated

Total number 21 95 16 614 1400 1856 90

Table 4. Linear regression model summary for franchise movie data.

Deviance Residuals:

Min 1Q Median 3Q Max

-280959026 -41549197 -7101628 28664207 931454904

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.723e+07 3.594e+07 1.592 0.11181

colorBlack and White -4.374e+07 2.554e+07 -1.713 0.08731 .

content\_ratingApproved 2.484e+08 9.037e+07 2.749 0.00616 \*\*

content\_ratingG 6.774e+07 1.720e+07 3.938 9.15e-05 \*\*\*

content\_ratingNC-17 1.828e+07 3.607e+07 0.507 0.61252

content\_ratingPG 7.527e+07 9.953e+06 7.562 1.45e-13 \*\*\*

content\_ratingPG-13 3.588e+07 8.767e+06 4.093 4.84e-05 \*\*\*

content\_ratingUnrated -2.405e+07 5.040e+07 -0.477 0.63336

actor\_1\_facebook\_likes -1.990e+03 7.071e+02 -2.815 0.00504 \*\*

num\_voted\_users 4.784e+02 2.756e+01 17.360 < 2e-16 \*\*\*

cast\_total\_facebook\_likes 1.346e+03 4.905e+02 2.743 0.00626 \*\*

budget 3.507e-01 6.961e-02 5.037 6.22e-07 \*\*\*

imdb\_score 5.023e+06 3.839e+06 1.308 0.19124

aspect\_ratio -4.162e+07 1.422e+07 -2.928 0.00354 \*\*

movie\_facebook\_likes -3.440e+02 1.590e+02 -2.164 0.03084 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7.42689e+15)

Null deviance: 1.0869e+19 on 628 degrees of freedom

Residual deviance: 4.5601e+18 on 614 degrees of freedom

Figure 22. Comparison of two franchise movie profit predictions and their prediction errors (over-prediction).

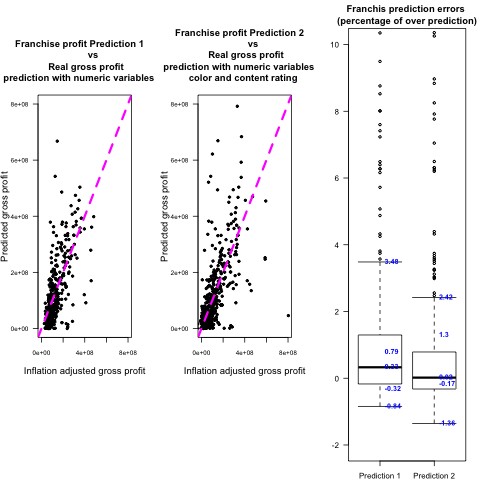


Table 5. Boxplot statistics for the prediction error comparison in figure 22.

Prediction 1 Prediction 2

Min -0.8426855 -1.35608936

2nd Q -0.1715904 -0.32311996

Median 0.3284104 0.01803949

3rd Q 1.2963696 0.78695134

Max 3.4841907 2.41642463

Figure 23. Mapping regression relationships of the individual variables to franchise gross profit.

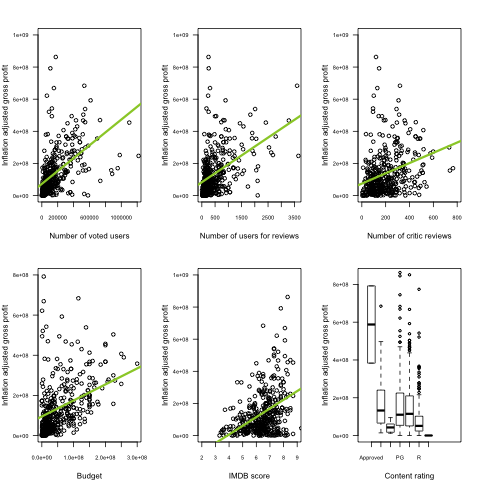


Figure 24. Numeric variables that do not show linear regression relationship to franchise gross profit and franchise profit distribution against categorical variables.

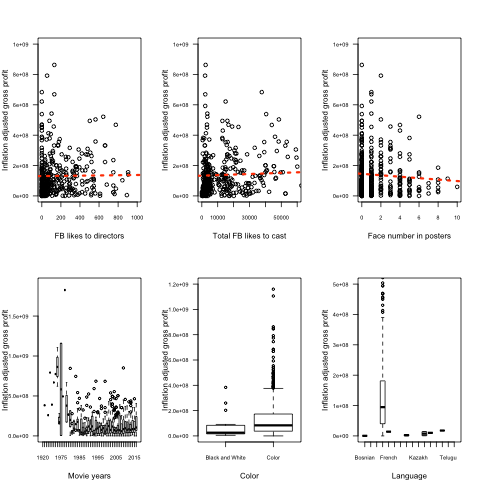


Figure 25. Comparison of two meta movie profit predictions and their prediction errors (over-prediction).

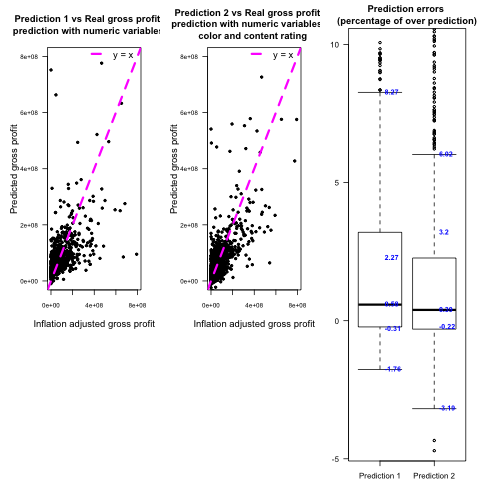


Table 6 Boxplot statistics for the prediction error comparison in figure 25.

Prediction 1 Prediction 2

Min -1.7624730 -3.1872814

2nd Q -0.2248039 -0.3058742

Median 0.5820492 0.3906960

3rd Q 3.1997884 2.2705595

Max 8.2690198 6.0223881

Figure 26. Mapping regression relationships of the individual variables to meta movie gross profit.

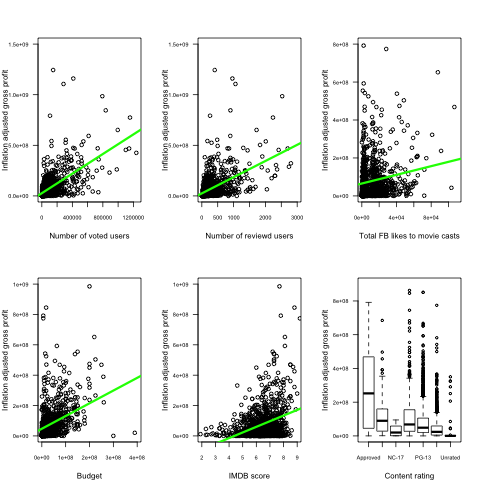
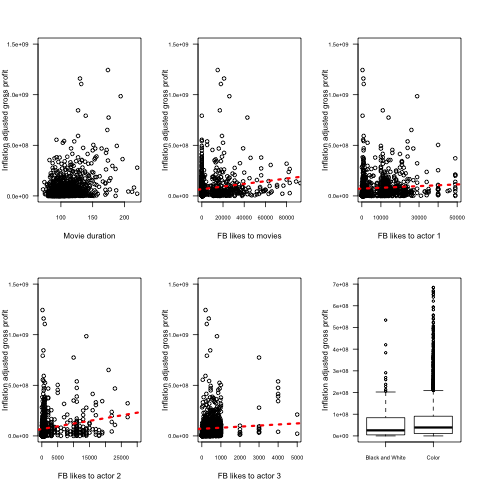


Figure 27. Numeric variables that show no or weak linear regression relationships to meta movie gross profit and gross profit distribution against movie color.



# References

Fuller, Steve. “Topic: Movie Industry.”

http://[www.statista.com/topics/964/film/](http://www.statista.com/topics/964/film/).

Mueller, Annie. “Why Movies Cost So Much To Make.” *Investopedia*, 13 June 2011, http://[www.investopedia.com/financial-edge/0611/why-movies-cost-so-much-to-make.aspx](http://www.investopedia.com/financial-edge/0611/why-movies-cost-so-much-to-make.aspx).

The Importance of a Franchise in the Movie Industry,

<http://www.yuppee.com/2013/04/16/the-importance-of-a-franchise-in-the-movie-industry>

The rise and rise of the Hollywood film franchise,

<https://www.ft.com/content/192f583e-7fa7-11e4-adff-00144feabdc0>

How Exactly Do Movies Make Money,

[http://www.investopedia.com/articles/investing/093015/how-exactly-do-moviesmake-money.asp](http://www.investopedia.com/articles/investing/093015/how-exactly-do-movies-make-money.asp)

What are the most said words in movies? Analysis of 617 scripts

<http://gettingthingsdata.com/words-in-movies/?utm_source=reddit&utm_medium=internetisinteresting&utm_campaign=est1800>

Interesting Data Sources and APIs

<http://gettingthingsdata.com/words-in-movies/?utm_source=reddit&utm_medium=internetisinteresting&utm_campaign=est1800>

Movie franchise data table

<http://www.the-numbers.com/movies/franchises/>

CPI data, <https://inflationdata.com/Inflation/Consumer_Price_Index/HistoricalCPI.aspx?reloaded=true>

Movie rating certificate

<https://help.imdb.com/article/contribution/titles/certificates/GU757M8ZJ9ZPXB39>

Motion Picture Association of America film rating system

<https://en.wikipedia.org/wiki/Motion_Picture_Association_of_America_film_rating_system>