Statpadders

Toma Shigaki-Than, CJ Frederickson, Camden Reeves, Sam Kakarla

2025-03-17

Introduction

Online review platforms have transformed how audiences evaluate and choose movies. Between professional critics and everyday viewers, these platforms host a broad spectrum of opinions that increasingly shape the trajectory of films, from box office performance to streaming recommendations. However, these two groups often evaluate films through very different lenses: critics tend to emphasize artistic merit, narrative structure, and technical execution, while audiences may prioritize entertainment, emotional resonance, and accessibility. As a result, it is commonplace for a film to receive mixed signals across platforms. A film may be highly rated by audiences but poorly reviewed by critics, or vice versa.

This divergence can be confusing for consumers, who must navigate these conflicting assessments to make viewing decisions. For example, a film might thrive on word-of-mouth and accumulate high audience ratings on IMDb but perform poorly with critics or fail to receive awards recognition.

Motivation and Importance

Understanding these differences is essential not just for filmgoers but also for industry stakeholders. Movie studios, marketers, and streaming platforms rely on both critical acclaim and mass appeal to determine promotional strategies, content investments, and recommendation algorithms. Prior research has shown that user-driven ratings have powerful word-of-mouth effects, often extending a film's relevance and commercial success (Moon, Bergey, Iacobucci, 2010). Further, the typical consumer does not evaluate films with the same lens or criteria as professional critics.

By analyzing these patterns, we aim to identify the film characteristics, such as decade of release, runtime, gross earnings, and censorship rating, that are most predictive of audiences liking a film more than critics. These insights have direct implications for consumer behavior research, content recommendation systems, and media marketing strategies in an increasingly data-driven entertainment landscape.

Given the observed divergence between critic and audience evaluations, we pose the following research question:

What factors contribute to audiences liking a movie more than critics, and how can we use these factors to predict the likelihood of a film performing better with audiences than critics?

In this study, we focus on modeling the odds that audience scores exceed critic scores. By doing so, we shift attention from understanding average film quality to identifying conditions that foster fan-favorite films and resonate with general audiences even when they fail to win over the critics.

Data

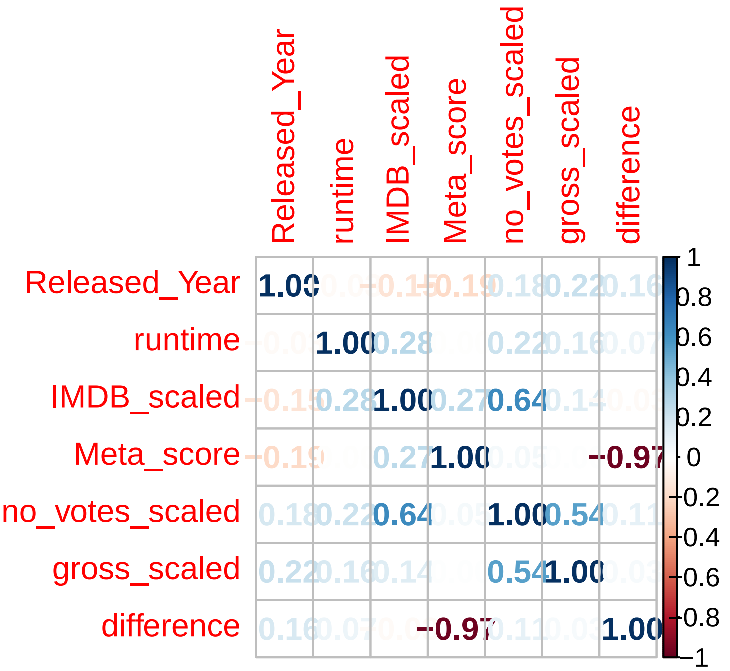
This data is taken from the top 1000 movies on IMDB, obtained through data scraping. In cleaning our data, we first had to deal with the NA values in our data. Some observations had NA values in their Gross Revenues. After examining these observations, there were no discernible patterns or connections between the NA values; they were random. As such, we were able to drop these values without compromising our data set or losing important observations. We also created the variable `difference`, whose value indicates the difference between IMDB Score and MetaScore, scaled so that they can be compared. A negative value indicates that the audience score is lower than that of critics, and positive is vice versa. Because our dataset includes older and international films with various outdated or uncommon censorship ratings, we consolidated certificates into modern, widely recognizable categories: G, PG, PG-13, R, and Other, based on their closest equivalents in the current U.S. rating system.

Predictors:

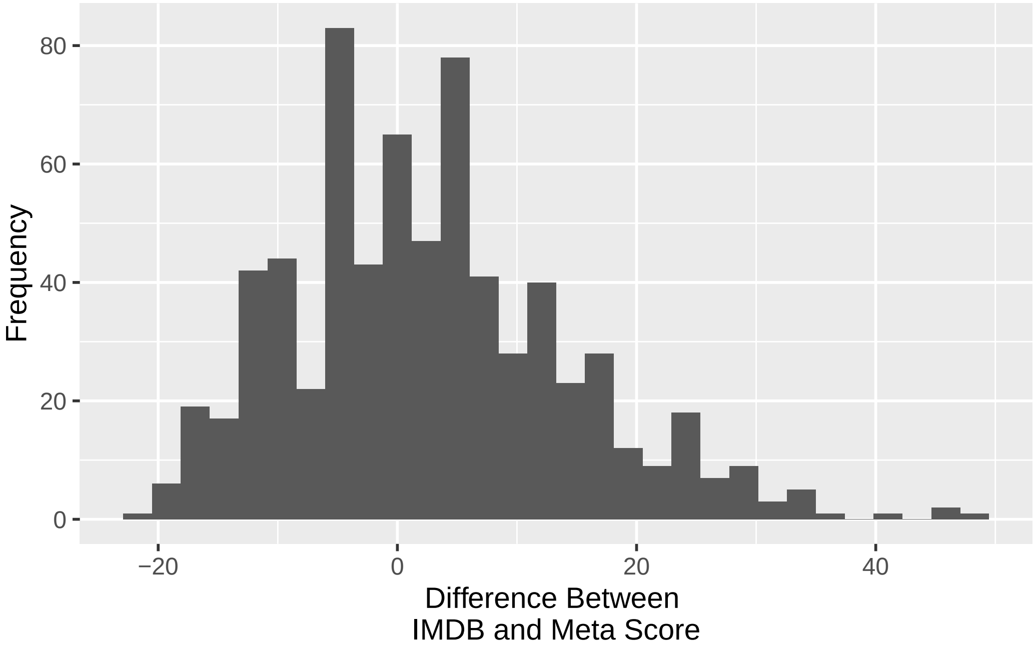
- Runtime (numerical)
- Gross Revenue (numerical)
- Censorship Certificate (categorical)
- Decade Released (categorical)
- Number of Votes (numerical)

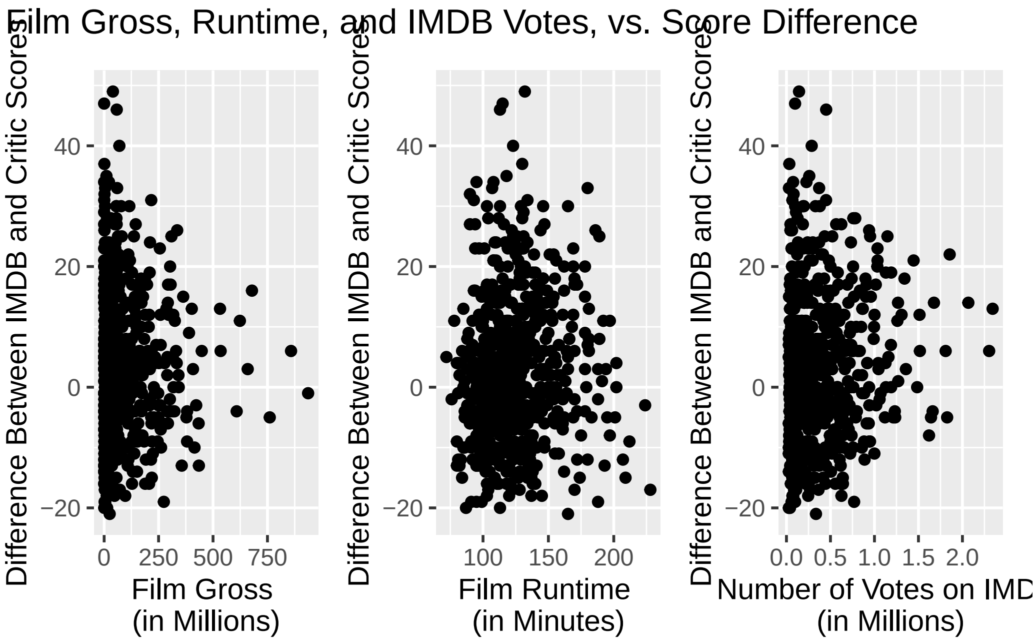
Response Variables:

- Difference: *The quantity that the IMDB score (score given by audience/fans on IMDB, scaled to match MetaScore) differs from the MetaScore (aggregate score of critics' ratings)*



We used a correlation matrix to check for multicollinearity. As expected, difference is highly correlated with `IMDB_scaled` and `meta_score`, since it's derived from them. IMDB_scaled and no_votes_scaled also show strong correlation (0.62), so we avoid including both in the same model.





In terms of the `difference` between the two scores, it seems that the values are almost normally distributed, with a slight left skew. This suggests that it is nearly equally common for a IMDB Score to be either higher or lower than the MetaScore score, though slightly more often lower.

Additionally, when plotting our numerical predictors, there appear to be no linear relationships between each predictor and our response.

Methodology

```
# A tibble: 18 x 5
   term                        estimate std.error statistic        p.value
   <chr>                          <dbl>     <dbl>     <dbl>          <dbl>
 1 (Intercept)                 -12.4      2.74       -4.51  0.00000753
 2 gross_cent                   -0.0127   0.00725    -1.75  0.0799
 3 runtime_cent                  0.0271   0.0176      1.54  0.124
 4 decade1960s                   7.61     3.21        2.37  0.0179
 5 decade1970s                   9.56     3.07        3.12  0.00191
 6 decade1980s                  14.3      2.95        4.85  0.00000151
 7 decade1990s                  16.7      2.84        5.90  0.00000000562
 8 decade2000s                  16.2      2.78        5.83  0.00000000863
 9 decade2010s                  12.2      2.82        4.34  0.0000167
```

```
10 certificateOther            -4.37    31.9        -0.137 0.891
11 certificatePG                1.55     1.31        1.18  0.237
12 certificatePG-13             1.53     2.52        0.608 0.543
13 certificateR                 1.40     1.16        1.21  0.229
14 no_votes_scaled              1.54     1.52        1.01  0.312
15 gross_cent:certificateOther  -0.141    0.454      -0.311 0.756
16 gross_cent:certificatePG     0.0122   0.00878     1.39  0.166
17 gross_cent:certificatePG-13  0.0275   0.0363      0.758 0.449
18 gross_cent:certificateR      0.0290   0.0126      2.31  0.0214
```

```
# A tibble: 18 x 5
   term                          estimate   std.error statistic p.value
   <chr>                            <dbl>       <dbl>     <dbl>   <dbl>
 1 (Intercept)                  -0.152        0.511     -0.297  0.767
 2 gross_cent                   -0.00214      0.00165   -1.30   0.194
 3 runtime_cent                  0.00948      0.00336    2.82   0.00481
 4 decade1960s                  -0.771        0.631     -1.22   0.222
 5 decade1970s                  -1.27         0.606     -2.10   0.0353
 6 decade1980s                  -0.725        0.559     -1.30   0.195
 7 decade1990s                  -0.453        0.533     -0.851  0.395
 8 decade2000s                  -0.569        0.523     -1.09   0.277
 9 decade2010s                  -0.775        0.533     -1.45   0.146
10 certificateOther           -538.       17550.        -0.0306 0.976
11 certificatePG                 0.240        0.252      0.953  0.341
12 certificatePG-13             -0.0467       0.490     -0.0952 0.924
13 certificateR                  0.0128       0.226      0.0569 0.955
14 votes_cent                    0.377        0.287      1.31   0.189
15 gross_cent:certificateOther  -7.03       231.        -0.0305 0.976
16 gross_cent:certificatePG      0.000311     0.00191    0.163  0.870
17 gross_cent:certificatePG-13   0.00299      0.00704    0.424  0.672
18 gross_cent:certificateR       0.00318      0.00252    1.26   0.207
```

```
# A tibble: 1 x 1
  adj.r.squared
          <dbl>
1        0.0881
```

We selected a logistic regression model for two primary reasons. First, none of the numerical predictors exhibited a clear linear relationship with the response variable difference. Second, our goal was to model the odds that audience scores exceed those of critics. In our initial model, most levels of the certificate variable were not statistically significant. Additionally, although we hypothesized an interaction effect between gross revenue and censorship rating, the corresponding interaction terms were not significant and thus were excluded from the next model tested.

```
# A tibble: 14 x 5
   term             estimate std.error statistic p.value
   <chr>               <dbl>     <dbl>     <dbl>   <dbl>
 1 (Intercept)      -0.118    0.501       -0.236 0.814
 2 gross_cent       -0.00158  0.000965    -1.64  0.102
 3 runtime_cent      0.00985  0.00332      2.97  0.00298
 4 decade1960s      -0.840    0.614       -1.37  0.171
 5 decade1970s      -1.28     0.596       -2.15  0.0317
 6 decade1980s      -0.754    0.550       -1.37  0.170
 7 decade1990s      -0.491    0.523       -0.940 0.347
 8 decade2000s      -0.614    0.513       -1.20  0.231
 9 decade2010s      -0.794    0.522       -1.52  0.128
10 certificateOther -1.29     1.18        -1.09  0.274
11 certificatePG     0.215    0.247        0.870 0.384
12 certificatePG-13 -0.123    0.419       -0.294 0.769
13 certificateR     -0.0648   0.219       -0.296 0.768
14 votes_cent        0.440    0.282        1.56  0.119
```

```
# A tibble: 10 x 5
   term         estimate std.error statistic p.value
   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
 1 (Intercept)  -0.308    0.450       -0.684 0.494
 2 gross_cent   -0.00130  0.000911    -1.43  0.153
 3 runtime_cent  0.00978  0.00329      2.97  0.00298
 4 decade1960s  -0.713    0.588       -1.21  0.226
 5 decade1970s  -1.09     0.563       -1.93  0.0537
 6 decade1980s  -0.550    0.516       -1.07  0.287
 7 decade1990s  -0.299    0.487       -0.614 0.539
 8 decade2000s  -0.411    0.475       -0.866 0.386
 9 decade2010s  -0.565    0.483       -1.17  0.242
10 votes_cent    0.452    0.276        1.64  0.101
```

term	df.residual	residual.deviance	df	deviance	p.value
difference_binary ~ gross_cent + runtime_cent + decade + certificate + votes_cent	681	840.351	NA	NA	NA
difference_binary ~ gross_cent + runtime_cent + decade + certificate + votes_cent + gross_cent * certificate	677	833.676	4	6.675	0.154

term	df.residual	residual.deviance	df	deviance	p.value
difference_binary ~ gross_cent + runtime_cent + decade + votes_cent	685	843.572	NA	NA	NA
difference_binary ~ gross_cent + runtime_cent + decade + certificate + votes_cent + gross_cent * certificate	677	833.676	8	9.896	0.272

A drop-in-deviance test comparing a reduced model (excluding certificate and interaction terms) to a full model (including certificate and the gross × certificate interaction) yielded a statistically significant improvement in fit ($\chi^2 = 19.648$, df = 8, p = 0.012). This suggests that certificate and its interaction with gross earnings contribute meaningfully to predicting whether audience ratings exceed those of critics. Additionally, the full model's AIC is lower, which indicates a better balance between model complexity and goodness of fit.

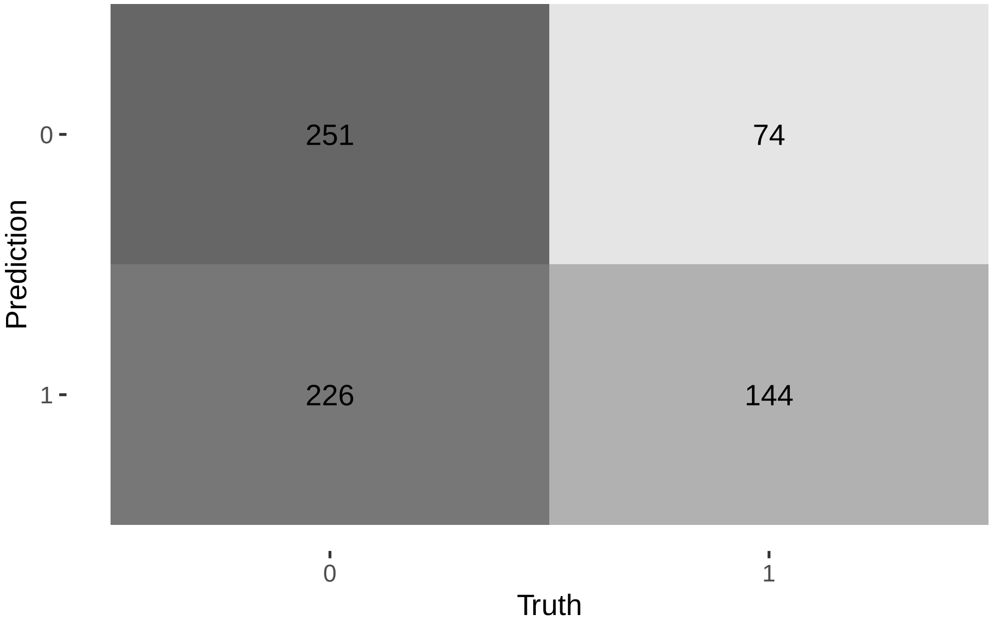
```
# A tibble: 1 x 8
  null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
          <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
1          865.     694  -417.  870.  951.     834.         677   695
```

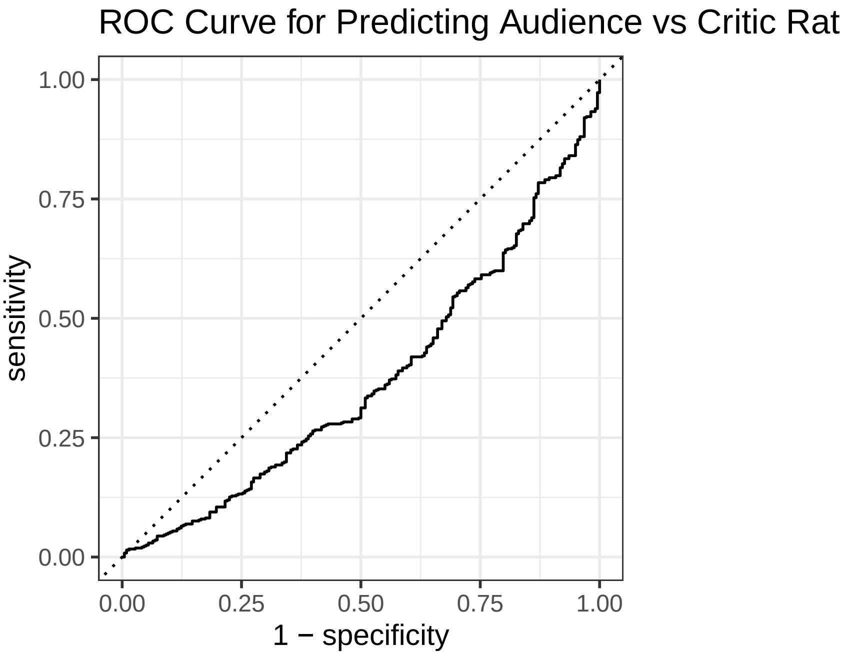
```
# A tibble: 1 x 8
  null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
          <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
1          865.     694  -422.  864.  909.     844.         685   695
```

Results

term	estimate	std.error	statistic	p.value
(Intercept)	-0.152	0.511	-0.297	0.767
gross_cent	-0.002	0.002	-1.298	0.194
runtime_cent	0.009	0.003	2.820	0.005
decade1960s	-0.771	0.631	-1.221	0.222
decade1970s	-1.275	0.606	-2.105	0.035
decade1980s	-0.725	0.559	-1.297	0.195
decade1990s	-0.453	0.533	-0.851	0.395
decade2000s	-0.569	0.523	-1.087	0.277
decade2010s	-0.775	0.533	-1.455	0.146
certificateOther	-537.772	17550.153	-0.031	0.976
certificatePG	0.240	0.252	0.953	0.341

term	estimate	std.error	statistic	p.value
certificatePG-13	-0.047	0.490	-0.095	0.924
certificateR	0.013	0.226	0.057	0.955
votes_cent	0.377	0.287	1.314	0.189
gross_cent:certificateOther	-7.030	230.694	-0.030	0.976
gross_cent:certificatePG	0.000	0.002	0.163	0.870
gross_cent:certificatePG-13	0.003	0.007	0.424	0.672
gross_cent:certificateR	0.003	0.003	1.261	0.207





```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.372
```

```
# A tibble: 5 x 3
  .metric     .estimator .estimate
  <chr>       <chr>          <dbl>
1 accuracy    binary         0.568
2 kap         binary         0.157
3 recall      binary         0.526
4 precision   binary         0.772
5 specificity binary         0.661
```

To address our research question, the final model we fit is this full logistic regression model predicting the odds that a movie's IMDb audience score exceeds its critic MetaScore. The final model included centered and scaled versions of gross revenue, runtime, and number of IMDB votes, as well as release decade, certificate rating, and an interaction term between gross revenue and certificate. A drop-in-deviance test confirmed that including certificate and its interaction with gross significantly improved model fit ($\chi^2 = 19.648$, df = 8, p = 0.012). The full model also had a lower AIC than the reduced model, indicating a more favorable balance between model complexity and fit.

Initial diagnostic checks did not reveal violations of key assumptions for logistic regression. No strong outliers or high-leverage points unduly influenced the model, and multicollinearity among predictors was low aside from expected correlations between variables involved in the interaction term.

Interpretation of model coefficients reveals several noteworthy trends. For a baseline film from the 1930s with a G rating and average values for gross, runtime, and number of votes, the predicted odds that audience ratings are higher than critic ratings is, on average, approximately 0.024, or 2.4%.

Runtime was not a statistically significant predictor (p = 0.586), though the odds ratio per unit increase is 1.002, suggesting a minimal practical effect. With every one minute increase in film runtime, the model estimates that the odds of higher audience ratings on IMDB multiply by 1.002 on average, holding all else constant.

In contrast, gross revenue had a negative effect, which was significant (p = 0.010). This suggests that for G-rated films, higher gross earnings are associated with lower odds of audiences rating a movie more favorably than critics. For every $1 million increase in gross revenue, the model estimates on average that the odds of higher audience scores is multiplied by 0.996, or 99.6%. It is important to note that while it is a decrease, it is not a huge decrease, which is an interesting relationship.

Release decade emerged as a strong predictor. Compared to films from the 1930s, those released from the 1970s through the 2010s had significantly higher odds of being rated better by audiences than critics, with odds increasing most notably in the 1990s and 2000s.

Although the main effects of certificate categories were not significant, the interaction terms reveal that the relationship between gross revenue and audience-vs-critic rating divergence depends on a film's rating. Specifically, for PG-rated and R-rated films, higher gross revenue is associated with a significantly greater chance that audiences prefer the film over critics (p = 0.004 and p = 0.006, respectively). This interaction suggests that commercially successful PG or R films are more likely to resonate with audiences than with critics.

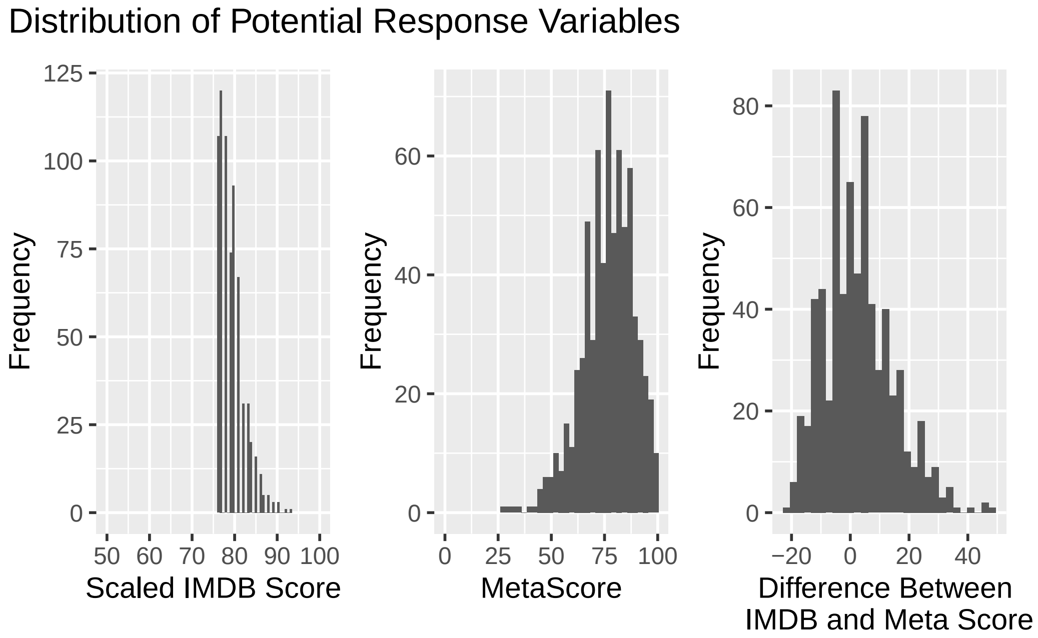
In summary, the model suggests that release decade, gross revenue, and its interaction with rating certificate are predictors in understanding when audiences are more favorable toward a film than critics. These results help contextualize the divergence in critical vs. popular reception and offer insights into the characteristics of films that perform better with general viewers.

Appendix

To begin our EDA, we first had to deal with the NA values in our data. Some observations had NA values in their Gross Revenues. After examining these observations, there were no discernible patterns or connections between the NA values; they were random. As such, we were able to drop these values without compromising our data set or losing important observations. We also created the variable `difference`, whose value indicates the difference between MetaScore and IMDB Score, scaled so that they can be compared. A negative value indicates that the MetaScore is lower than IMDB Score, and a positive value indicates that it is higher. Further, we turned our year predictor into a categorical variable by creating a new variable: `decade`. Since there is a very wide range of values in `Released_Year` for the movies selected, that variable itself is not particularly useful for our analysis. Not many observations even had the same released year, and the differences between one unit in that variable were arbitrary for some movies (for example a movie released in 1966 vs 1967 does not give much insight). For data cleaning and to improve clarity and interpretability, we changed this variable into a categorical variable `decade`, where all of the years released are grouped into decades (i.e. 1950s, 1960s, etc.). This categorical approach gives better interpretability; grouping movies into decades creates a better identifier than simply using individual years.

Additionally, the variable `Runtime` listed the runtime of each observation as a string with the number of minutes followed by the word "mins." For example, a movie 90 minutes long would be listed as the string "90 mins" instead of the number 90. As such, this made `Runtime` a categorical variable. We changed this by removing the "mins" label and refactoring it as numeric, thus making the `Runtime` into a numerical variable.

Univariate EDA

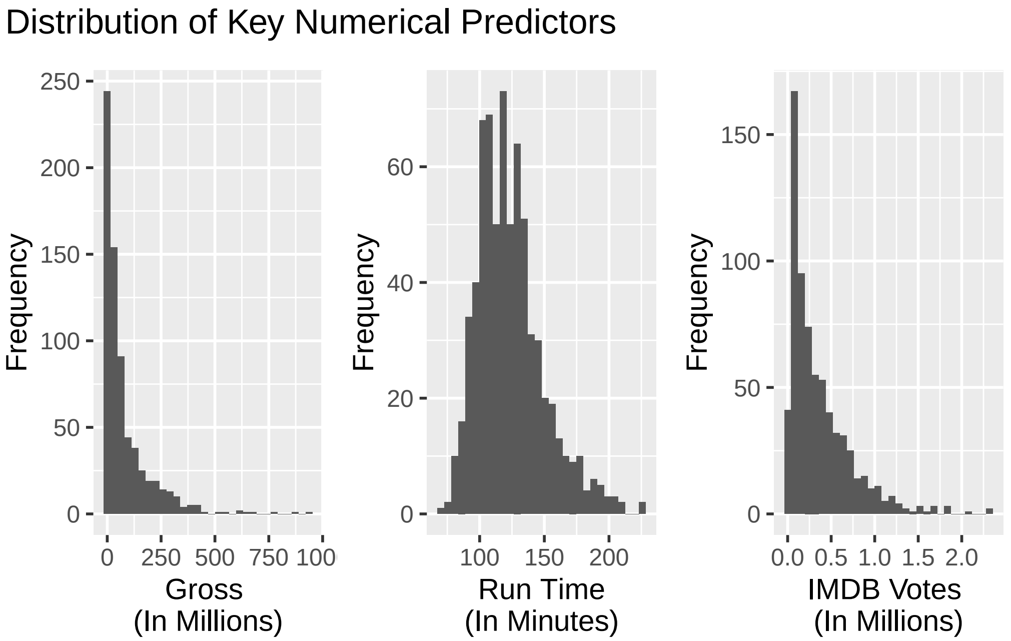


```
# A tibble: 3 x 7
  Variable    mean   med    sd   IQR   min   max
  <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 IMDB_Score 79.3     79  2.94     4    76    93
2 MetaScore  76.7     77 12.2     16    28   100
3 Difference  2.62     2 11.8     16   -21    49
```

As we can see from our response variables, scaled IMDB Score seems to be skewed right for these films, and scores tend to trend between 76 and 93. Both the mean score and median score are about 79, standard deviation of about 3, IQR of 4, and a range of 17.

The distribution for MetaScore seems to be skewed right, with a mean of about 77, a median about 78, a standard deviation of about 12, an IQR of 16 and a range of 72.

Furthermore in terms of the `difference` between the two scores, it seems that the values are almost normally distributed, with a slight left skew. This suggests that it is nearly equally common for a MetaScore to be either higher or lower than the IMDB score, though slightly more often lower. There is an outlier when MetaScore is about 49 points lower than IMDB score (-49). The mean `difference` is when about MetaScore is about two points lower than IMDB score (-2) and median at 1 point lower (-1). There is standard deviation of 12 points, IQR of about 15 points, and a range of 70 points.

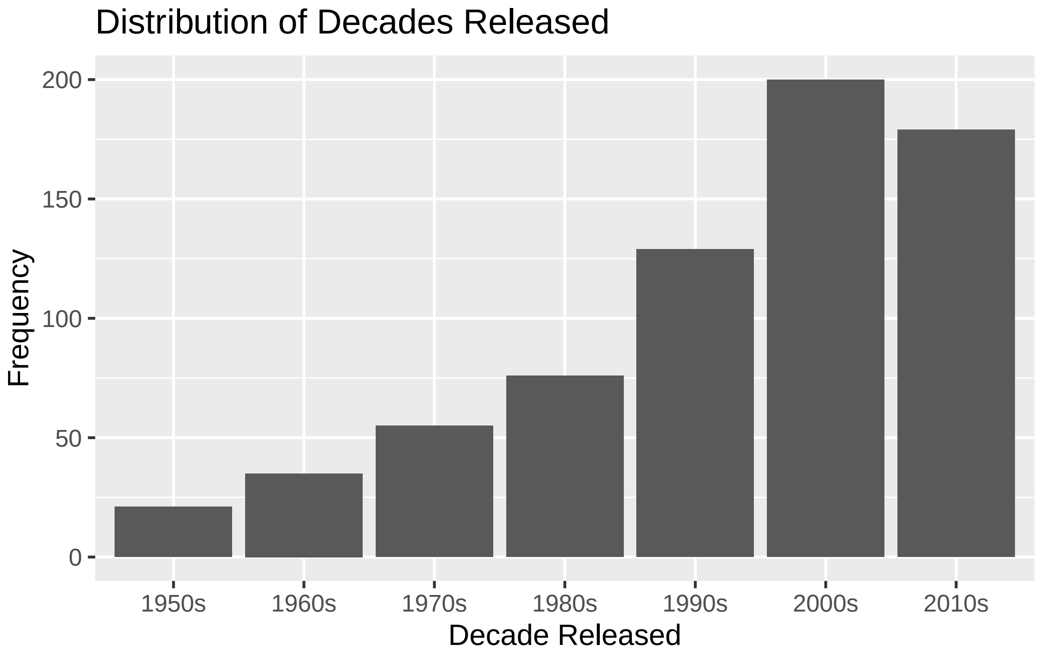


```
# A tibble: 3 x 7
  Variable    mean     med      sd    IQR       min    max
  <chr>      <dbl>   <dbl>   <dbl>  <dbl>     <dbl>  <dbl>
1 gross     80.1    35.9   116.    99.9    0.00130 937.
2 runtime  124.    120      25.6   32     72       228
3 votes      0.361   0.241   0.357  0.418  0.0252    2.34
```

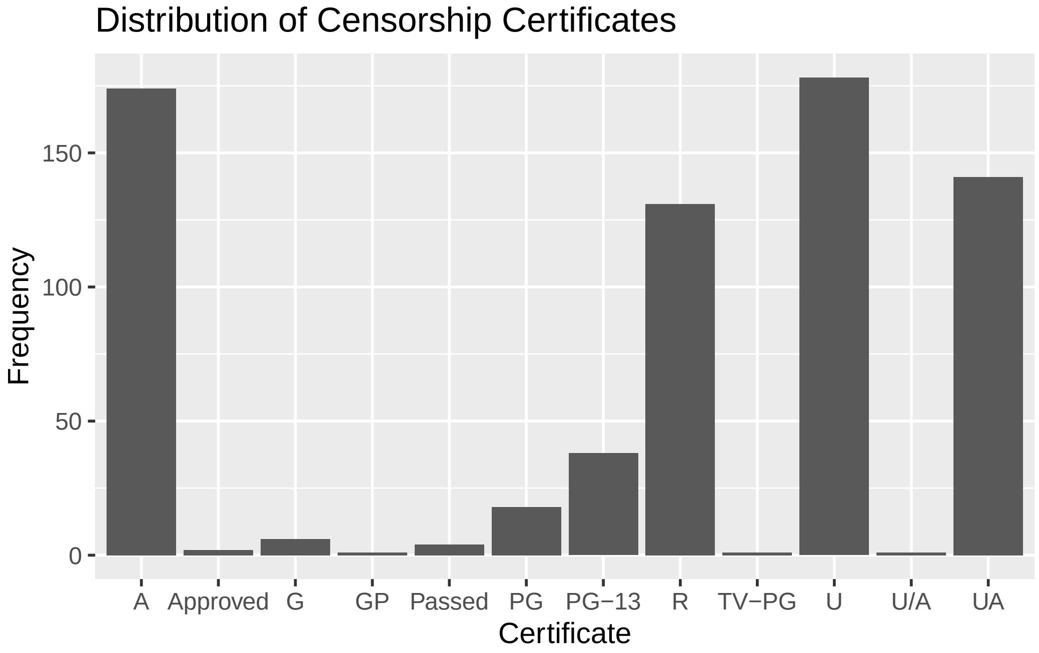
Exploring our numerical predictors, it seems that the distribution of gross revenue in millions of dollars seems to have a right skew, with most values being below $500 million. It has a mean of $78.514 million, median of $34.850 million, standard deviation of about $115 million, IQR of $96.310 million, and a range of about $937 million.

Run time seems fairly normal with a slight right skew. There is a potential outlier around 238 minutes. It has a mean of about 124 minutes, median of about 120 minutes, standard deviation of about 26 minutes, IQR of about 32 minutes, and a range of about 166 minutes.

Finally, number of votes has a right skew. With a potential outlier at about 2.34 million votes, it has a mean of about 356,000 votes, median of about 267,000 votes, standard deviation of about 354,000 votes, IQR of about 412,000 votes, and a range of about 2.32 million votes.

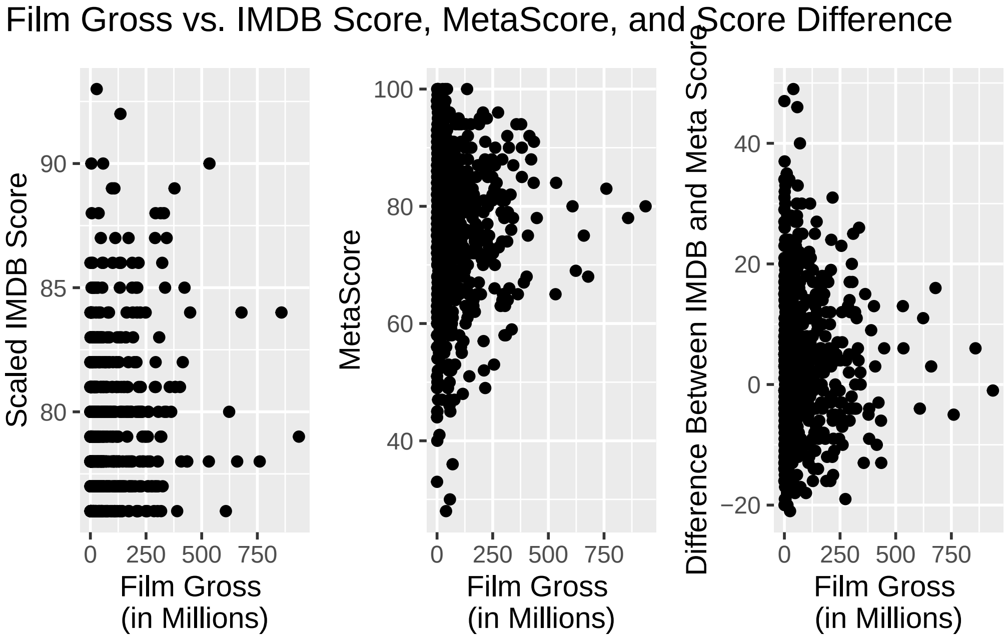


The distribution of `decade` seems to be skewed left. This is expected, as newer films are more likely to have been added to the internet in real time after release whereas older films are added retroactively.

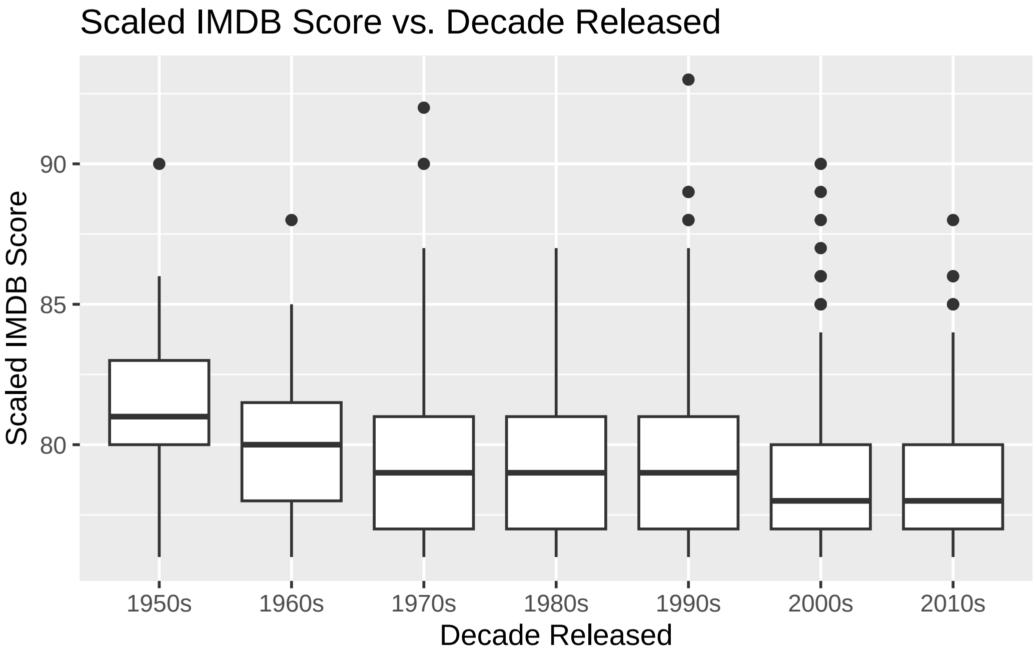


The distribution of `certificates` does not exhibit much of a normal shape, but notably the highest distribution is of "U" movies - those with unrestricted audiences.

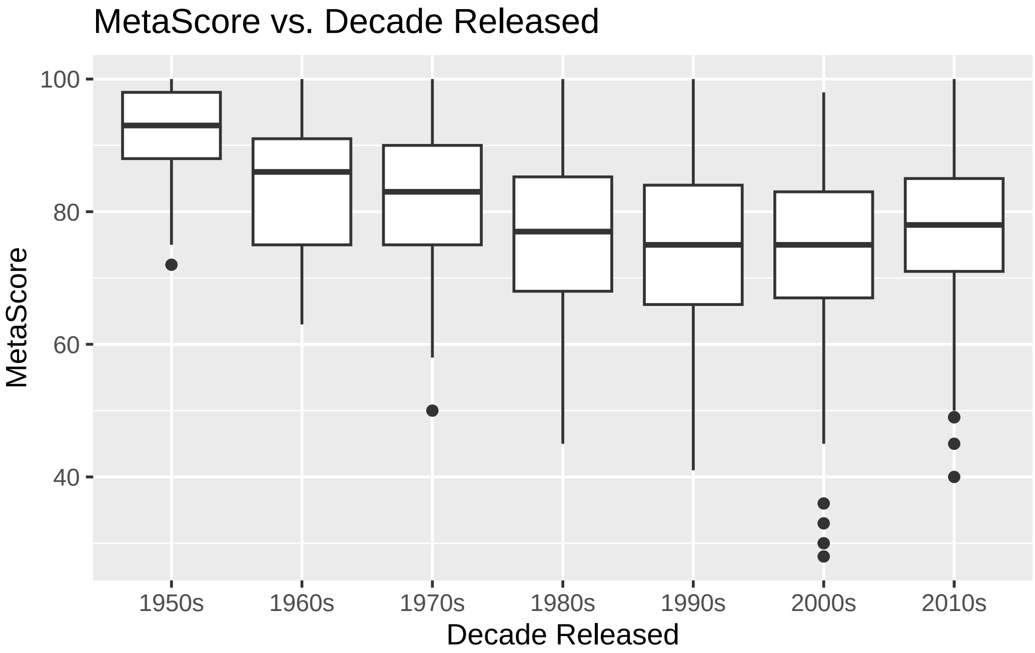
Bivariate EDA



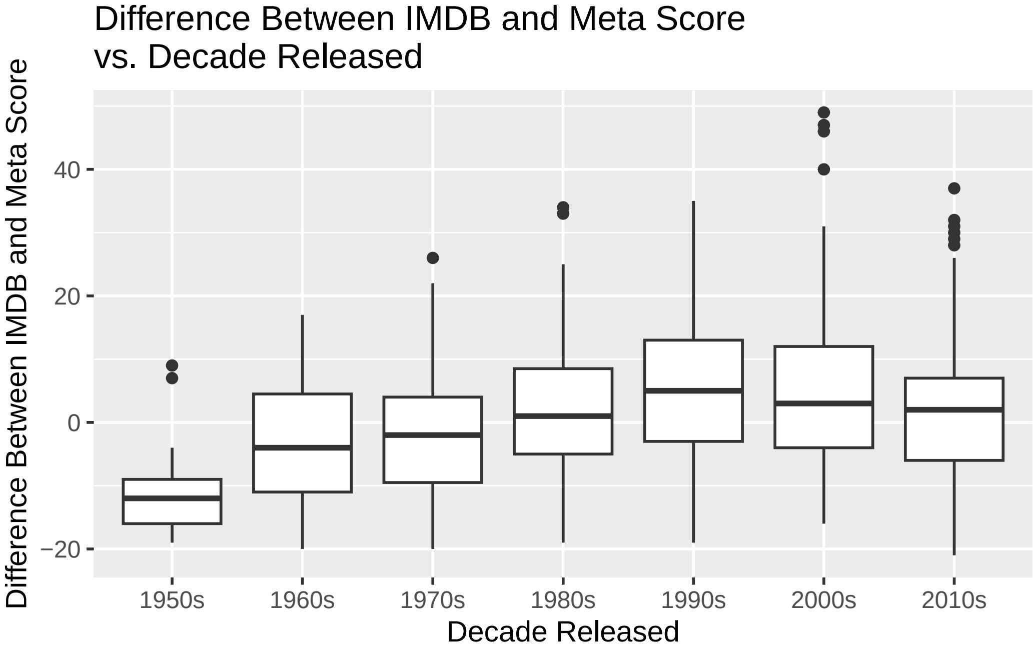
Upon initial Bivariate EDA, a clear linear relationship does not seem to appear between film's gross in millions and our three potential predictor variables. Perhaps later, to fit a model, we will need to find a variable transformation that gives us a promising model.



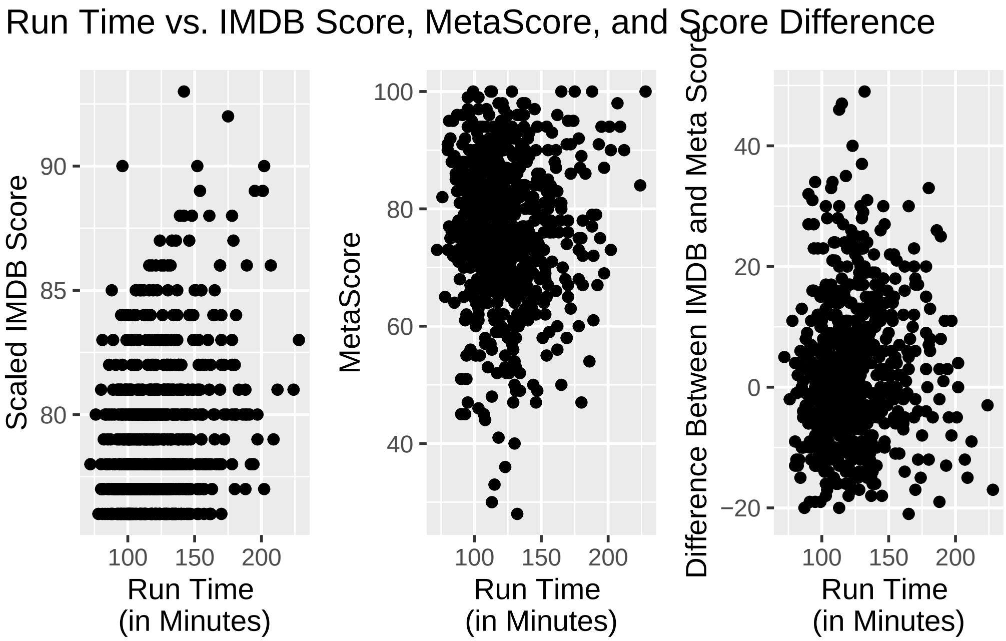
Judging from this initial bivariate EDA of decade released vs the scaled IMDB score, there seems to be a negative correlation between date and IMDB score; as movies are newer (coming out in more recent decades), the median scaled IMDB score tends to be lower.



Similarly to IMDB score, the critics' median MetaScores also seem to be lower as movies are newer. In other words, the overall aggregated critic scores for films tend to be lower for movies in more recent decades.

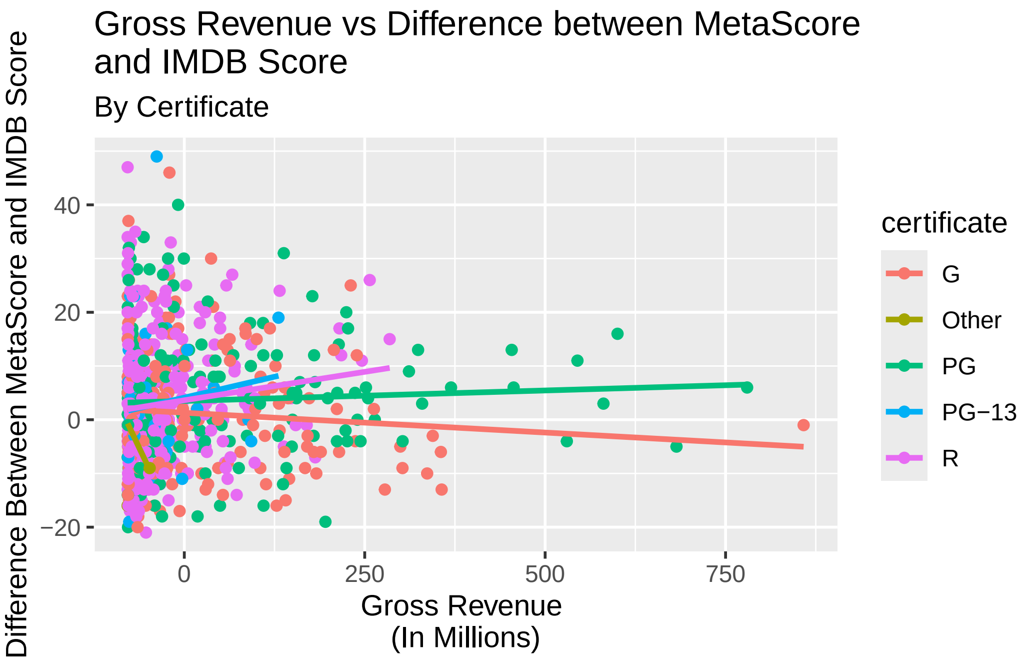


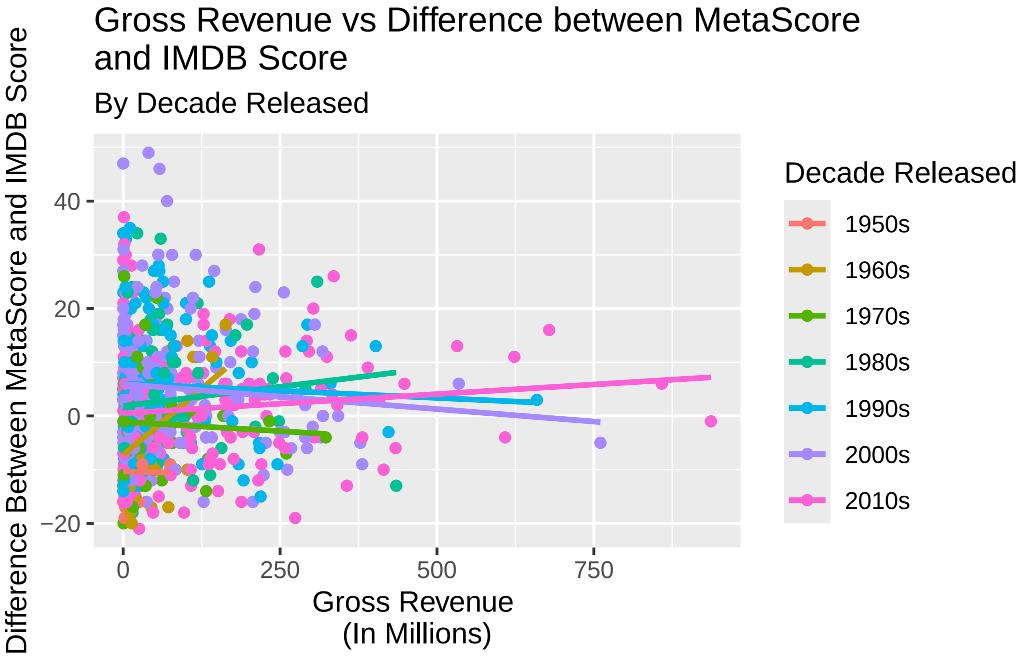
Judging from this EDA, the median difference between `Meta_score` and `IMDB_scaled` also tends to be lower as movies are newer. In other words, MetaScores tend to be lower than IMDB scores in more recent decades.



Similar to our gross predictor, a clear linear relationship does not seem to appear between film's run time in minutes and our three potential predictor variables. Perhaps later, to fit a model, we will need to find a variable transformation here as well that gives us a promising model.

Interaction Effects





Here, we plotted gross revenue against the difference in scores (MetaScore minus scaled IMDB score) to examine how these variables interact against our categorical variables. The visualizations reveal clear interaction effects in two key relationships:

Gross Revenue and Certificate

Gross Revenue and Decade Released

> ❗ Important
>
> Before you submit, make sure your code chunks are turned off with `echo: false` and there are no warnings or messages with `warning: false` and `message: false` in the YAML.

Reference: Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction. Journal of Marketing, 74(1), 108-121. https://doi.org/10.1509/jmkg.74.1.108