Analysis of Divergence Between Audience and Critic Scores

For Top IMDB Films from 1950 - 2019

Toma Shigaki-Than, CJ Frederickson, Camden Reeves

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 the differences in audience and critic scores, and how can we use these factors to predict the likelihood of significant divergence between the two?

In this study, we focus on modeling the odds that audience scores exceed critic scores by at least one standard deviation. 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. 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.Finally, we narrowed our scope by excluding films released before the 1950s, as there were relatively few observations from that period. This is understandable given that the IMDb rating system was not introduced until the 1990s—meaning older films would have had to be retroactively added and reviewed. To account for this limitation, we removed pre-1950 entries from our dataset.

Predictors:

- Runtime: *Numerical - film's duration in minutes*
- Gross Revenue: *Numerical - film's gross revenue, in millions of dollars*
- Censorship Certificate: *Categorical - censorship rating (e.g. PG, R, etc.)*
- Decade Released: *Categorical - decade of film's release*
- Number of Votes: *Numerical - number of votes film has on IMDB*

Response Variable:

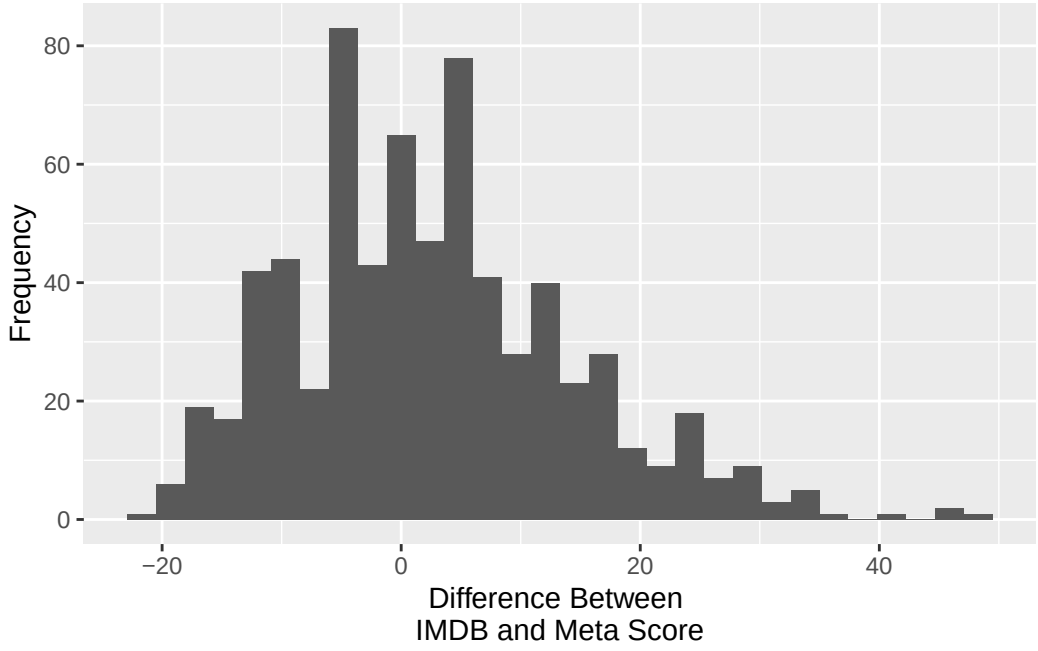
- Difference: *The quantity that the IMDB score (score given by audience/fans on IMDB, scaled to match MetaScore) differs from the MetaScore (aggregate score of critics' ratings)*

Anticipating the possibility of exploring logistic regression, we created a binary variable, `difference_binary`, based on the difference between IMDb audience scores and critic MetaScores. This variable identifies the differences that are significantly divergent versus those that are not. To capture meaningful divergence, we classified films into two categories:

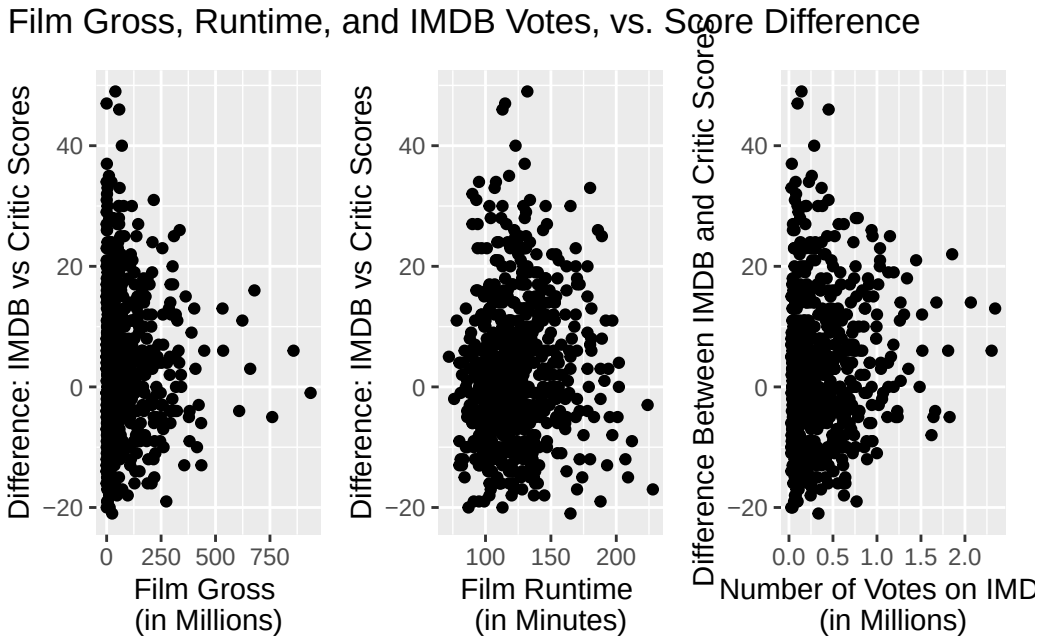
1 (divergent): if the difference was less than -13 or greater than 9, corresponding approximately to films with audience-critic score differences exceeding ±1 standard deviation from the mean.

0 (non-divergent): otherwise.

This threshold approach is grounded in the properties of the original difference distribution, which was approximately normal with a slight right skew. With a mean of approximately -2.2 and a standard deviation of about 11.9, setting thresholds at roughly one standard deviation above and below the mean allowed us to identify films with statistically meaningful deviations rather than minor fluctuations. This method balances capturing important fan-favorite or controversial films while maintaining a reasonable sample size for modeling.



We began our study with comprehensive exploratory data analysis to assess variable relationships and potential modeling challenges. A correlation matrix identified strong correlation (0.62) between the IMDB score and number of votes a film received, leading us to avoid including both in the same model to prevent multicollinearity. Additionally, initial visualizations suggested a potential interaction effect between a film's censorship certificate and its gross revenue. These findings informed the structure of our modeling approach. (*See Appendix for full EDA.*)



When plotting our numerical predictors, there appear to be no linear relationships between each predictor and our response. (*See Appendix for full EDA.*)

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

We initially explored a linear regression approach to model the difference between IMDb audience scores and critic MetaScores. However, diagnostics revealed no clear linear relationship between the response variable and the numerical predictors. Our Adjusted R^2 value of just 8.81% indicated that very little of the variation in the response was being explained by the regression model. Given this, and the bounded, binary nature of our eventual outcome (whether a film's audience score significantly diverges from critic scores), we transitioned to a logistic regression framework.

In constructing the logistic model, we first fit a full logistic regression model including:

- Gross revenue, centered for interpretability: `gross_cent`
- Runtime in minutes, centered for interpretability: `runtime_cent`
- Decade of release: `decade`
- Number of votes, centered for interpretability: `votes_cent`
- Censorship certificate: `certificate`
- An interaction term between gross revenue and certificate `gross_cent` * `certificate`

```
# 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
```

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 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
```

Upon evaluating the model, the main effects of certificate were not statistically significant, as shown by their high p-values. However, we hypothesized from our EDA that gross revenue may interact with certificate rating (*See appendix for full EDA.*). This is to assert that, for instance, commercial performance may matter differently depending on the rating of the films.

As such, we performed a drop-in-deviance test comparing the full model (with certificate and interaction) against a reduced model excluding the interaction terms. The results showed that including the gross × certificate interaction improved model fit, though modestly. Specifically, the deviance decreased with the interaction included, suggesting that there may be some moderating effect of the certificate on gross revenue's impact. Given this, we then narrowed our

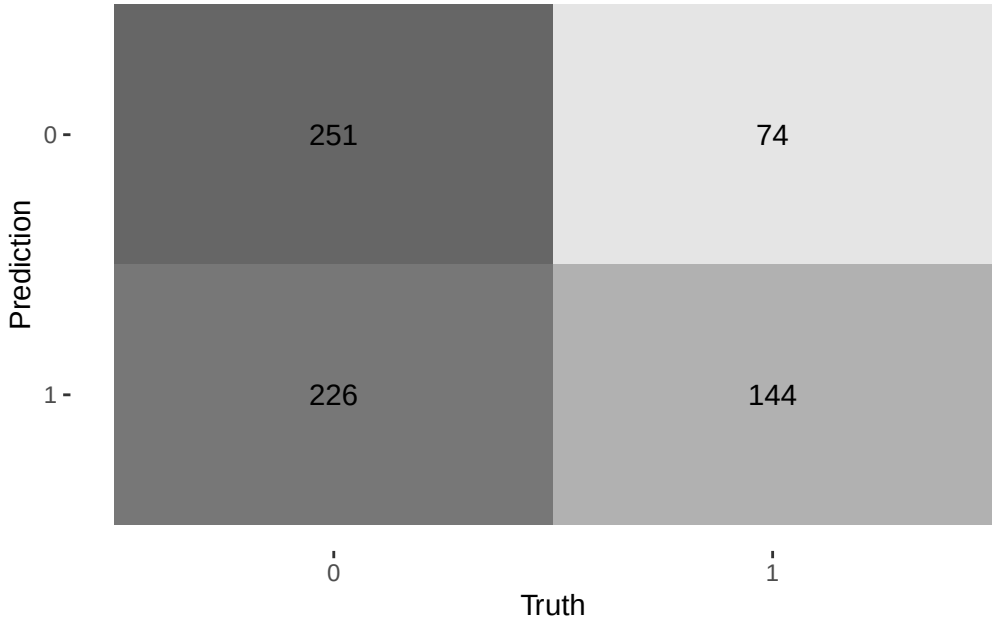
model comparison to two candidates: a model with certificate and the gross × certificate interaction terms included, and a model with no certificate variables at all (predictors: gross_cent + runtime_cent + decade + votes_cent).

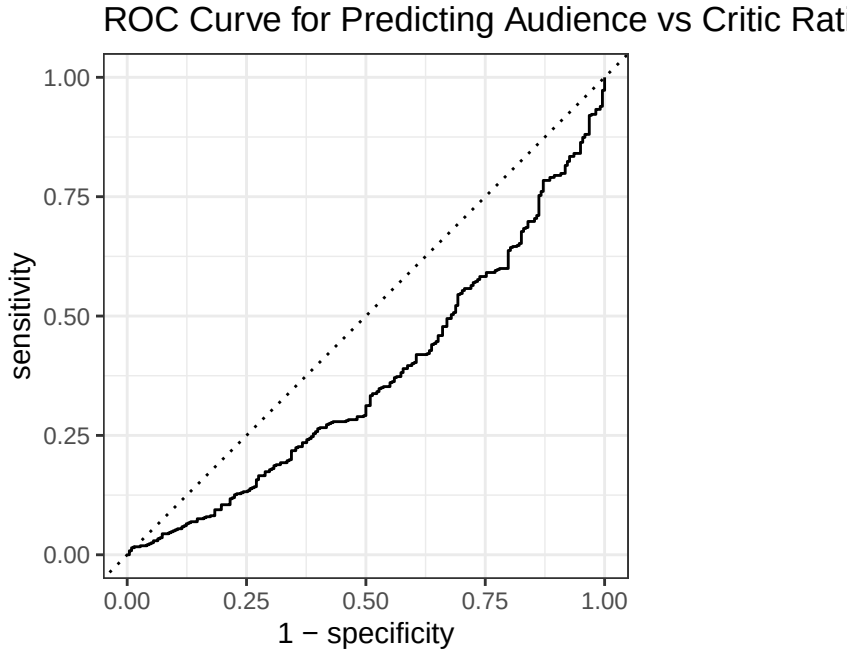
To formally select between them, we compared their AIC and BIC values. Both AIC and BIC were lower for the model without certificate variables. Since BIC penalizes model complexity more heavily than AIC, and favors simpler models, it is important to note that we would expect BIC to favor the reduced model. However, the fact that both AIC and BIC favored the simpler model presented sufficient evidence that removing `certificate` and its interaction was justified. Thus, despite the modest improvement in deviance with the interaction term, the penalized model selection criteria (AIC/BIC) led us to eliminate the censorship certificate variable and its interaction terms from the final model. The final logistic regression model, used to predict whether a movie's IMDb audience score diverged significantly from its critic MetaScore, includes centered gross revenue, centered runtime, decade of release, and centered number of votes as predictors.

Results

```
# 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
```

After fitting the final logistic regression model using `gross_cent`, `runtime_cent`, `decade`, and `votes_cent` as predictors, we initially evaluated its performance using a default classification threshold of 0.5. However, under this threshold, the model overwhelmingly predicted the majority class (0 - *non-divergent*), making it largely ineffective at identifying films where audience ratings diverged significantly from critic ratings. Very few films were classified as having a major divergence, despite a meaningful portion of the dataset meeting that condition. Recognizing that our binary outcome was based on ± 1 standard deviation from the mean difference (difference $<$ -13 or $>$ 9 being classified as 1), and thus rare by construction, we addressed this imbalance by lowering the classification threshold to 0.3. Under the adjusted 0.3 threshold, the model produced the following confusion matrix:






```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.628
```

```
# 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
```

In evaluating model performance, we considered whether to prioritize sensitivity or specificity. Sensitivity (52.6%) measures the model's ability to correctly identify films where audience scores diverge significantly from critic scores — the rare but important "fan-favorite" or "controversial" cases. Specificity, or recall, (66.1%) measures the ability to correctly classify typical films with no major divergence. Given that our research question focuses on understanding what factors drive major rating divergence, prioritizing sensitivity is more appropriate. Missing a fan-favorite or controversial film (a false negative) would be more detrimental to the purpose of our study than incorrectly flagging a typical film (a false positive).

Our goal is to capture as many truly divergent films as possible, even if it means accepting a slightly higher false positive rate. Lowering the classification threshold to 0.3 significantly improved the model's sensitivity, allowing it to identify a much greater proportion of the films where audience scores diverged meaningfully from critic scores. This came at the expected cost of increasing false positives, resulting in a moderate drop in specificity and overall precision. To assess the model's discriminative ability across all possible thresholds, we generated an ROC curve and calculated the Area Under the Curve (AUC). The resulting AUC of 0.63 suggests that the model performs just slightly better than random guessing at ranking films by their likelihood of divergence, thus still struggles. This indicates that the predictors capture some relevant patterns in the data, though their ability to sharply distinguish divergent from non-divergent films remains limited.

Among the predictors, runtime was statistically significant ($p < 0.01$), with longer films being slightly more likely to exhibit audience–critic score divergence. Specifically, for every additional minute of runtime, we estimate that the odds of a film having divergent scores are multiplied on average by a factor of 1.01, holding all other variables constant. While statistically significant, this effect is relatively small in practical terms. Decade of release also showed some influence, particularly with films from the 1970s: we estimate that a 1970s film on average has 0.33 times the odds of divergence of one from the 1950s, holding all else constanr. Again, the magnitude of this effect was modest. Gross revenue and number of votes, while included in the final model, did not emerge as strong or consistent predictors. Taken together, these results suggest that while structural film characteristics can capture some variation in audience–critic divergence, they do not provide a complete picture of what drives rating divergence.

Discussion and Conclusion

Our analysis aimed to identify film characteristics associated with significant disagreement between audience and critic evaluations. By defining divergence as a difference exceeding ± 1 standard deviation between IMDb and MetaScores, we focused on the most extreme cases: films that defy critical consensus or achieve unexpected fan acclaim. The final logistic regression model found that longer runtimes and certain decades of release were modestly associated with higher odds of divergence. However, the model's predictive performance (AUC = 0.63) reflects the limits of relying solely on structural attributes like gross revenue, runtime, decade, and number of votes. These features, while useful for identifying broad trends, do not capture the subjective or cultural factors that more directly influence how audiences and critics evaluate films. Our decision to lower the classification threshold to 0.3 was a strategic choice to prioritize sensitivity over specificity. Given the rarity of divergent films in the dataset, this adjustment allowed the model to better detect the cases central to our research question. Capturing more true divergent films—at the cost of increasing false positives—was aligned with our goal of understanding the conditions under which divergence occurs, rather than focusing only on precision. Overall, the model's limitations are understandable in the context of a rapidly evolving film industry. The relationship between critical reception and commercial success has become increasingly complex due to shifts in genre popularity, changing cultural norms, the rise of streaming platforms, and broader audience fragmentation. These are confounding variables that regression may be unable to account for. As a result, structural metadata alone is insufficient to predict the divergence in opinion between critics and general viewers. Future research would benefit from incorporating richer, content-based features such as genre, award recognition, review sentiment, and audience demographics. These variables may better capture the emotional, cultural, and narrative factors that shape divergent reception. While our model offers foundational insights, it also underscores the need for more nuanced approaches that reflect the multifaceted nature of media evaluation.

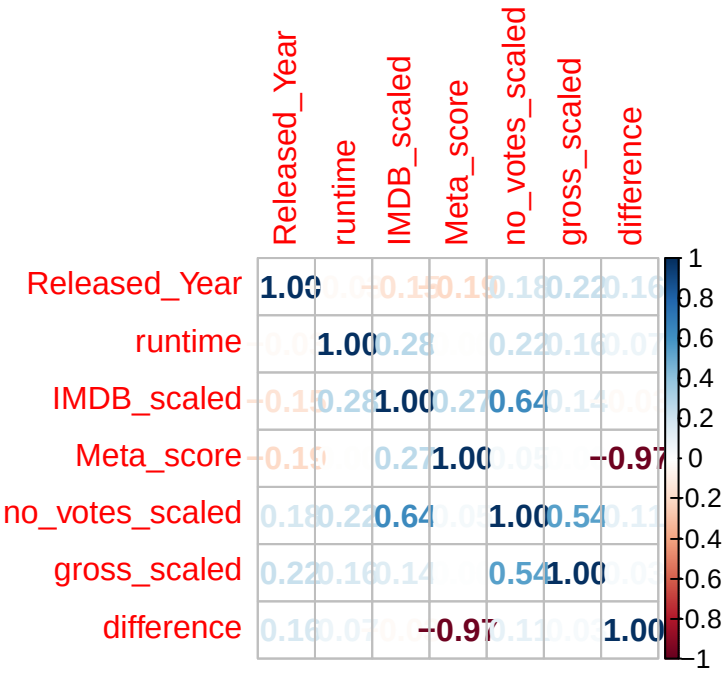
Appendix

```
# 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
```

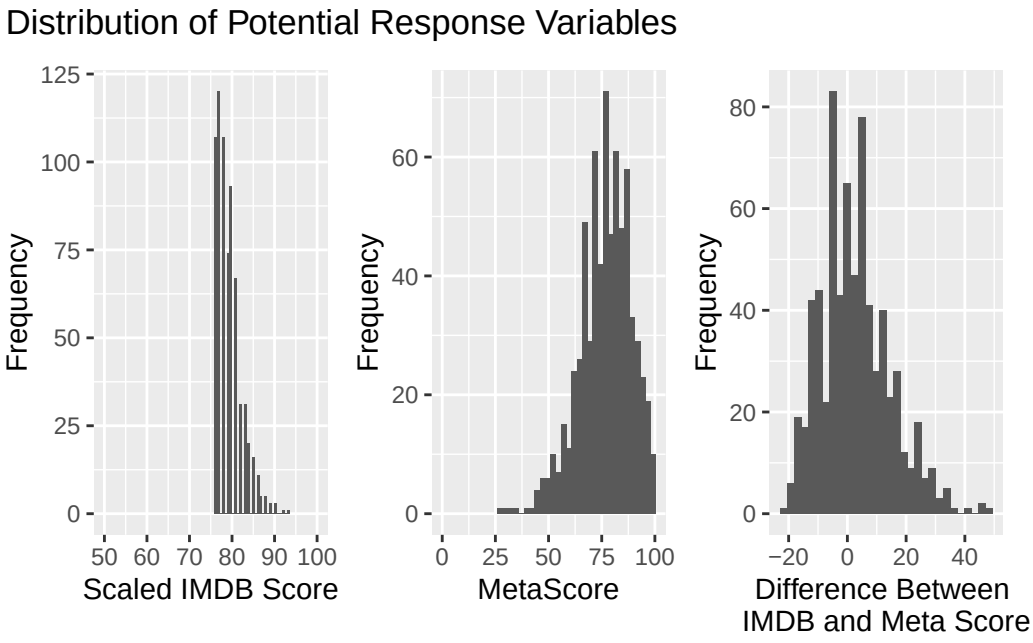
To begin our EDA, we first had to deal with the NA values in our data. Some observations had NA values in their Gross Revenues. After examining these observations, there were no discernible patterns or connections between the NA values; they were random. As such, we were able to drop these values without 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.

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.



Univariate EDA

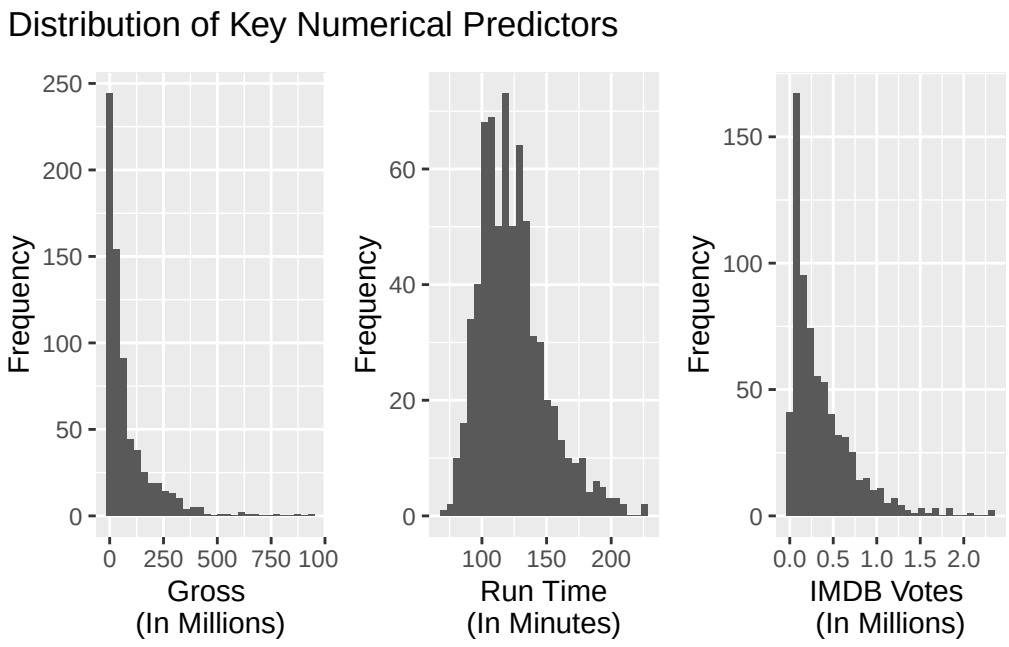


```
# A tibble: 3 x 7
  Variable    mean   med    sd   IQR   min   max
  <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 IMDB_Score 79.3     79  2.94     4    76    93
2 MetaScore  76.7     77 12.2     16    28   100
3 Difference  2.62     2 11.8     16   -21    49
```

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.



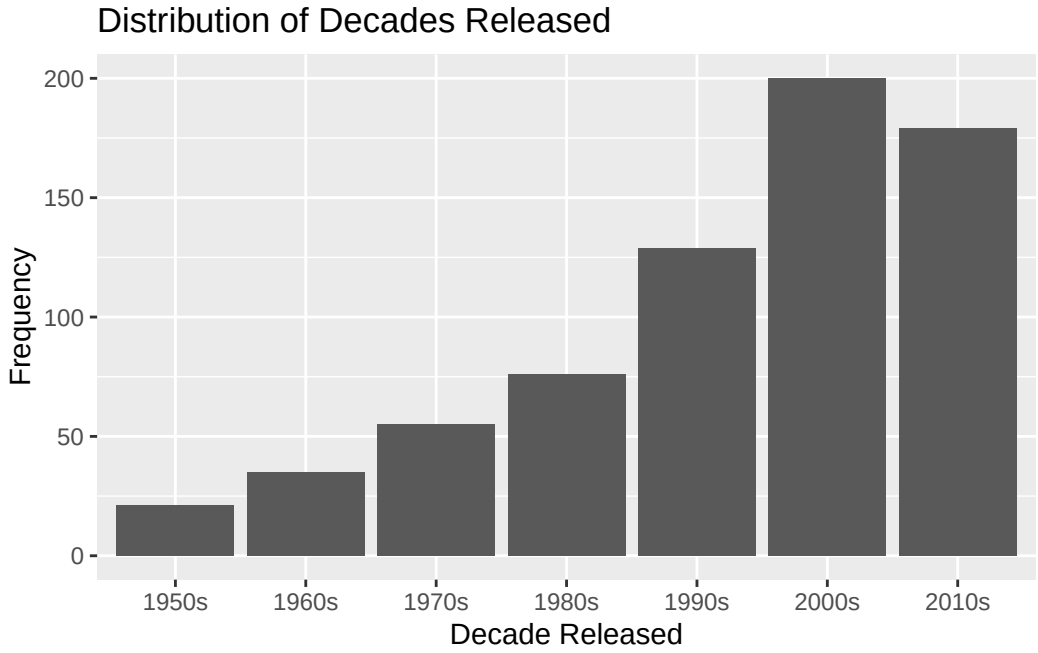

```
# A tibble: 3 x 7
  Variable    mean     med      sd    IQR      min    max
```

```
  <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
```

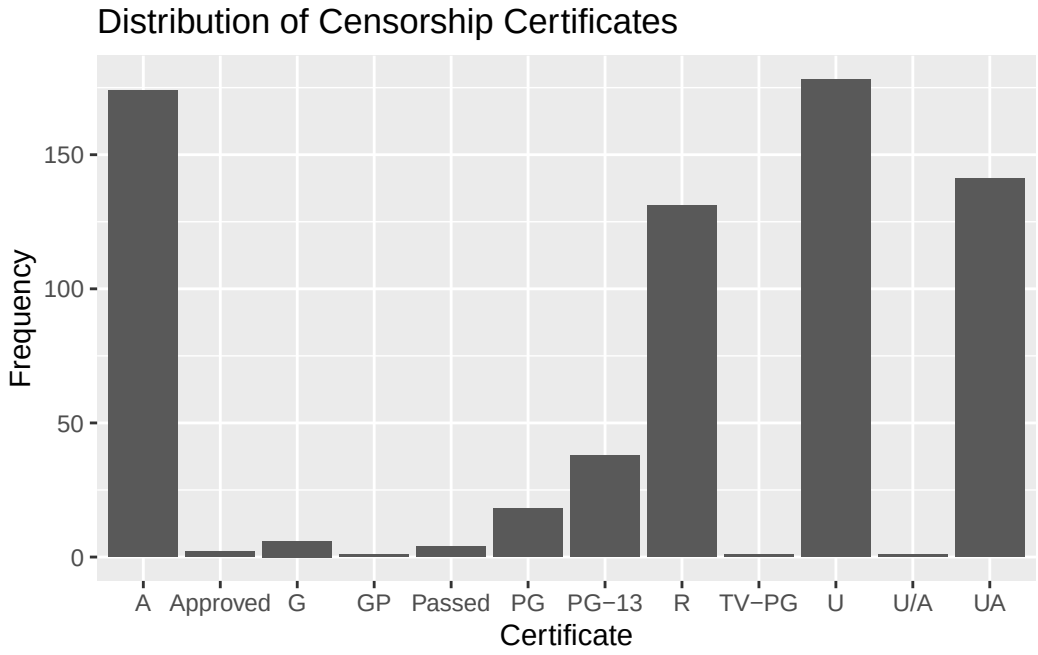
Exploring our numerical predictors, it seems that the distribution of gross revenue in millions of dollars seems to have a right skew, with most values being below $500 million. It has a mean of $78.514 million, median of $34.850 million, standard deviation of about $115 million, IQR of $96.310 million,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

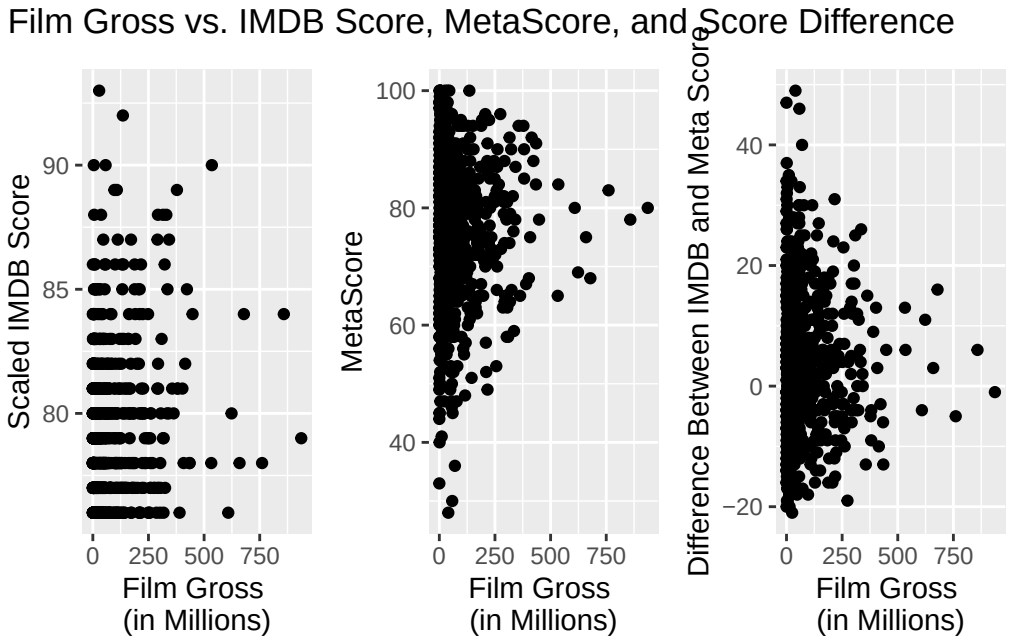


The distribution of `decade` seems to be skewed left. This is expected, as newer films are more likely to have been added to the internet in real time after release whereas older films are added retroactively.

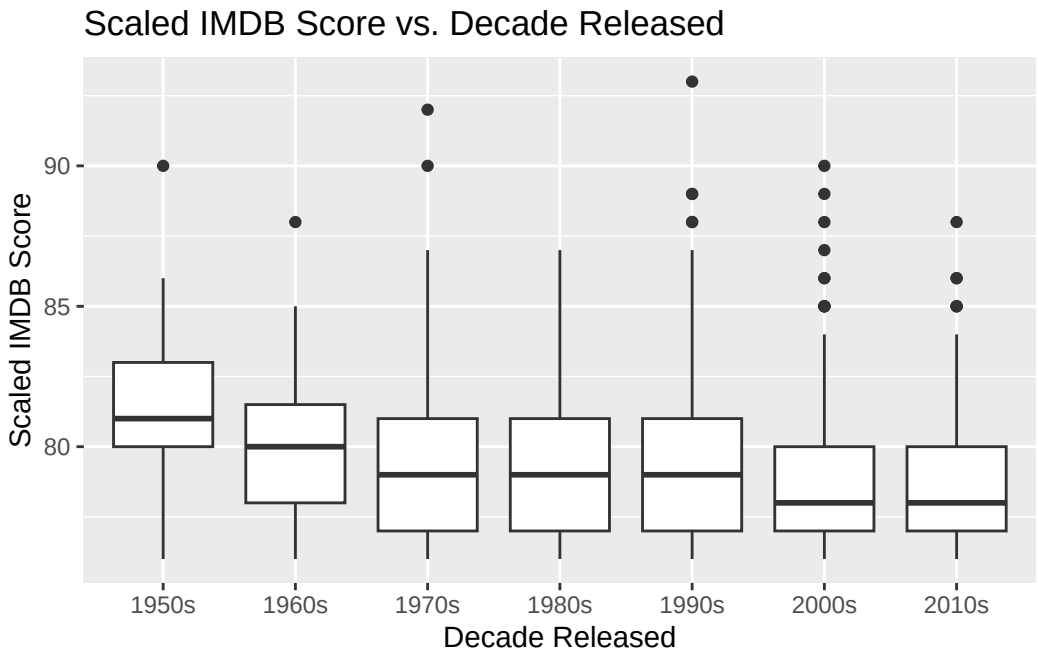


The distribution of `certificates` does not exhibit much of a normal shape, but notably the highest distribution is of "U" movies - those with unrestricted audiences.

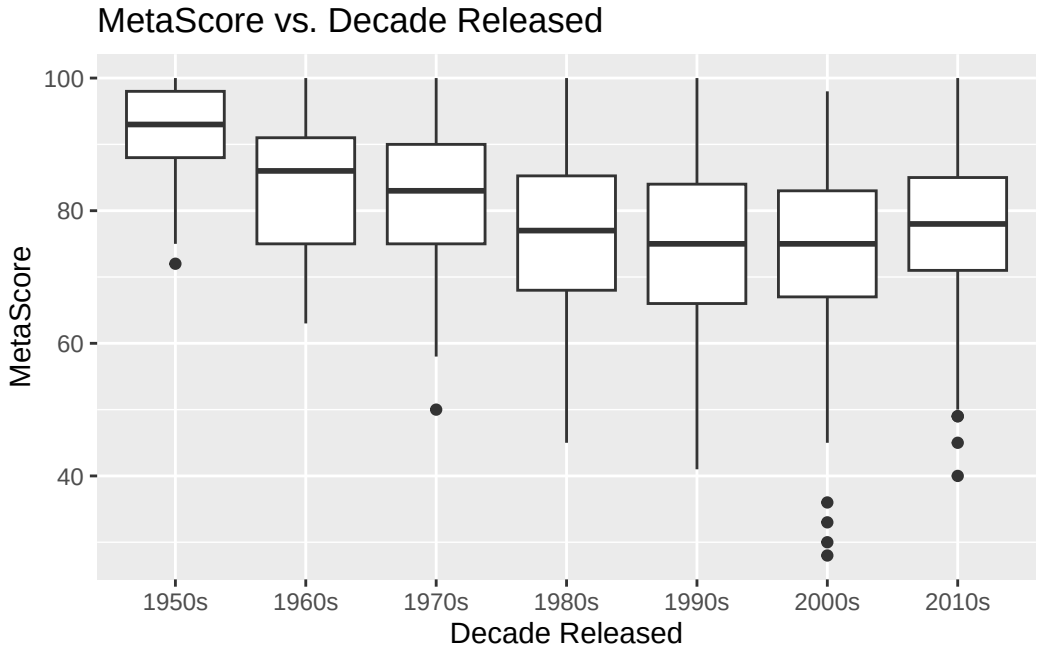
Bivariate EDA



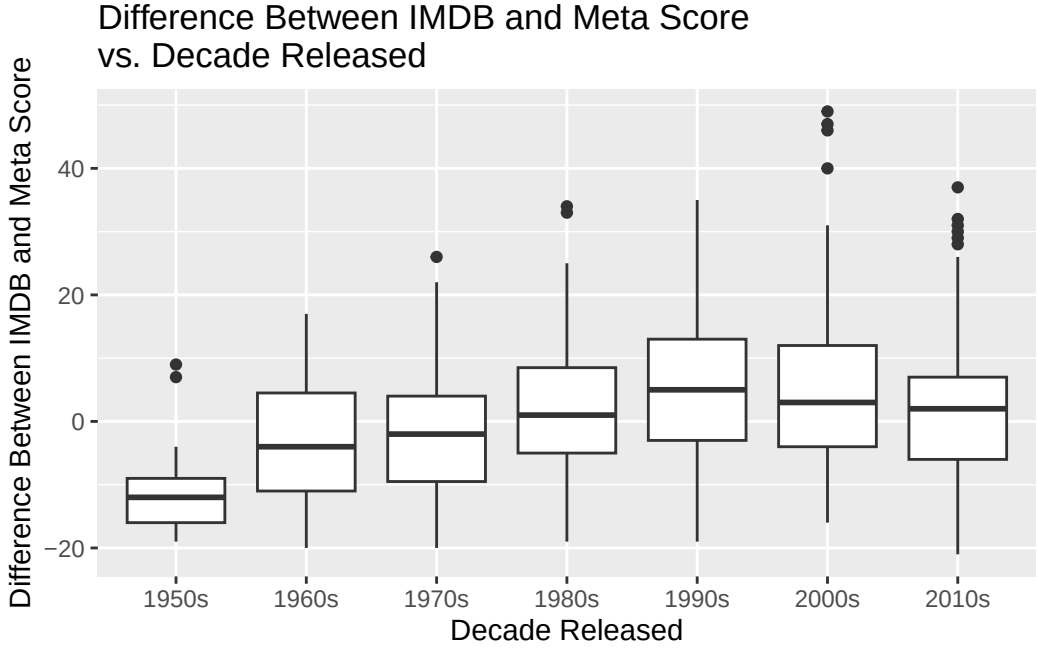
Upon initial Bivariate EDA, a clear linear relationship does not seem to appear between film's gross in millions and our three potential predictor variables. Perhaps later, to fit a model, we will need to find a variable transformation that gives us a promising model.



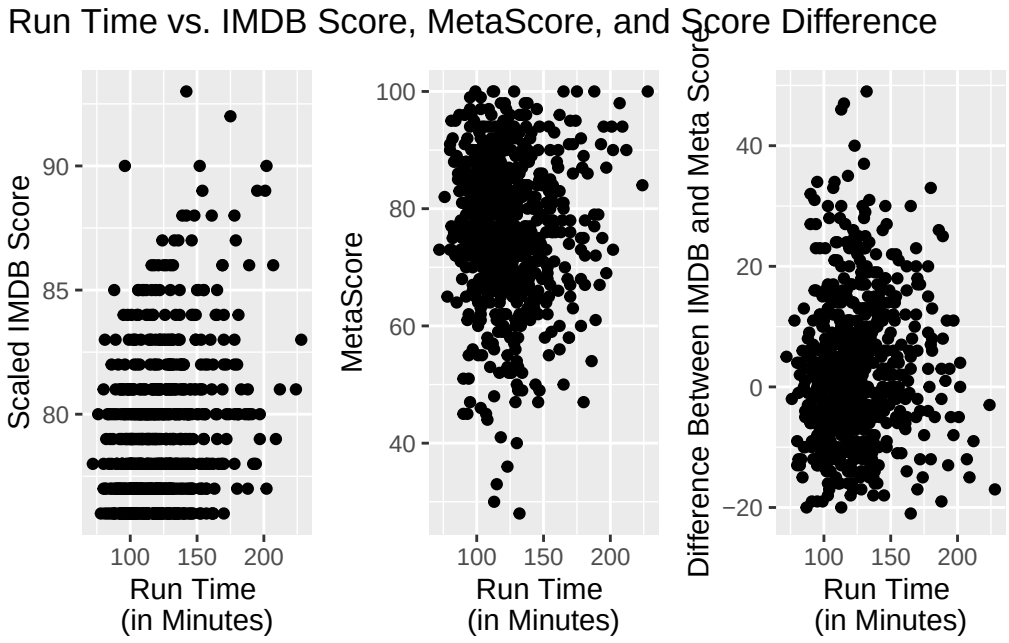
Judging from this initial bivariate EDA of decade released vs the scaled IMDB score, there seems to be a negative correlation between date and IMDB score; as movies are newer (coming out in more recent decades), the median scaled IMDB score tends to be lower.



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.

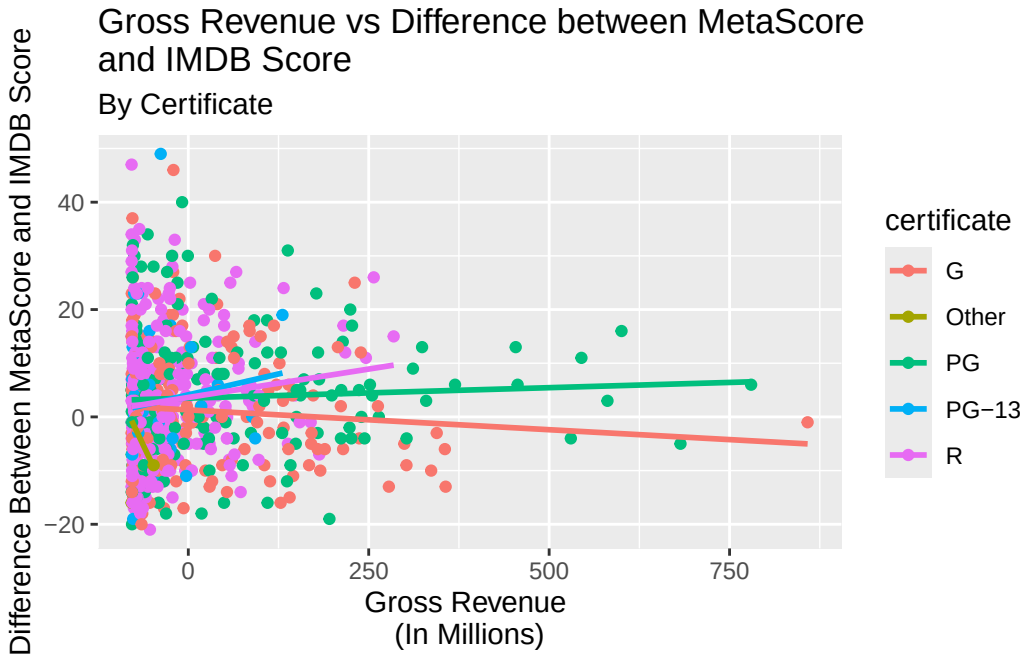


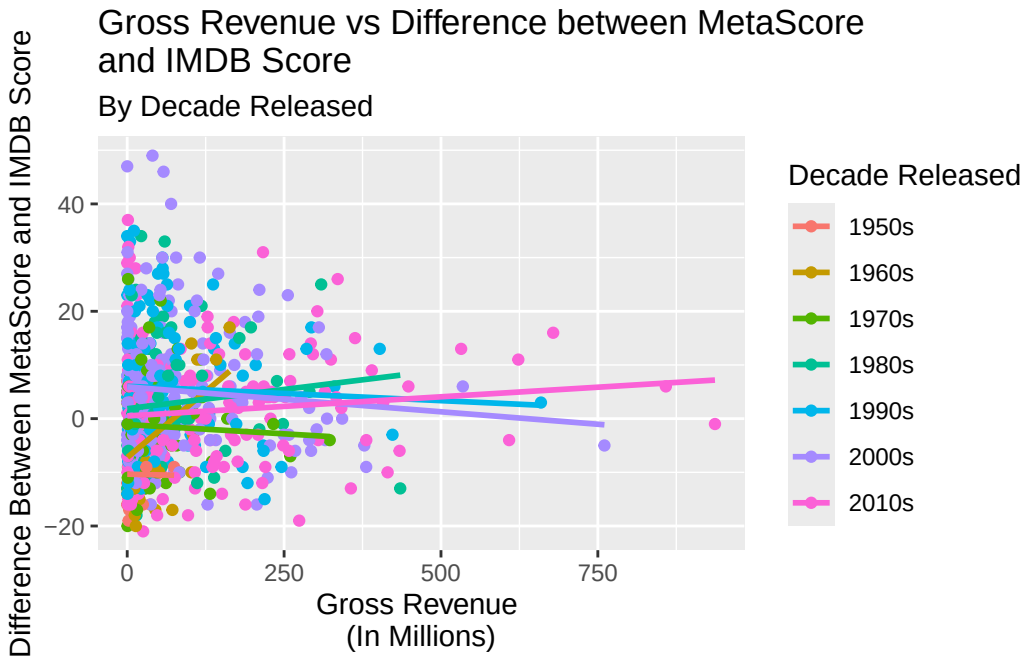
Judging from this EDA, the median difference between `Meta_score` and `IMDB_scaled` also tends to be lower as movies are newer. In other words, MetaScores tend to be lower than IMDB scores in more recent decades.



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.

References:

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

ChatGPT (*https://chatgpt.com/*) was utilized to assist in computing classification metrics. Specifically, the tool was used to generate metric outputs based on a provided confusion matrix, streamlining the calculation process and reducing the potential for manual error. The results were then reviewed and filtered to include only the evaluation metrics covered in our course.

ChatGPT. OpenAI, https://chatgpt.com/. Accessed 25 Apr. 2025.