Project Proposal

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

```
library(tidyverse)
library(tidymodels)
# add other packages as needed
#imported movie data manually (imdb_top_1000)
imdb_top_1000 <- read_csv("data/imdb_top_1000.csv")
#IGNORE IMDB_DATA AND IMDB_DATASET
```

Introduction

The proliferation of online review platforms has significantly influenced consumer decision-making across various industries, but particularly in entertainment. In the film industry, consumers rely on both professional critics and amateur audiences to gauge the quality of movies before making viewing decisions (Moon, Bergey, & Iacobucci, 2010). While professional critics evaluate films based on artistic and technical merit, amateur audiences may assess them based on personal enjoyment, accessibility, and entertainment value. This dual review system has created a dynamic where movies may receive different evaluations from critics and general audiences, raising questions about the relationship between the two and the factors influencing their divergence. Given this context, our research seeks to address the following question: What factors in a film influence IMDb user ratings and critic meta-scores; how do differences in these scores relate to movie characteristics such as gross earnings, number of votes, release year, runtime, certificate, and genre?

Motivation and Importance

Understanding the discrepancies between audience and critic ratings is critical for multiple stakeholders in the film industry. For movie studios and marketers, aligning promotional strategies with audience preferences while maintaining critical appeal can be a determinant of box office success and long-term profitability (Moon, Bergey, & Iacobucci, 2010). Platforms like Netflix and IMDb utilize rating-based recommendation systems to enhance user satisfaction and engagement, making it essential to refine these systems based on nuanced insights into rating behaviors. Moreover, prior research suggests that word-of-mouth (WOM) effects

can sustain movie revenues over time, with online user ratings playing a crucial role in influencing later viewership trends. Understanding these dynamics is essential for optimizing film production, marketing investments, and recommendation algorithms in an era where consumer feedback is widely accessible and highly influential.

Hypotheses and Theoretical Considerations

We propose several key hypotheses regarding the divergence between IMDb ratings (representing audience scores) and critic meta-scores:

The Effect of Gross Earnings and Number of Votes

Are movies with higher gross earnings and a greater number of votes more likely to have higher IMDb ratings due to their broad audience reach, (while critics may evaluate them more critically, leading to a potential divergence in scores)? Conversely, could lower-grossing films receive higher meta-scores due to stronger artistic value but may not resonate as widely with general audiences, leading to lower IMDb ratings?

Impact of Release Year and Runtime

Could older movies have higher meta-scores due to their established reputation and critical re-evaluation over time, while IMDb ratings could fluctuate based on contemporary audience preferences? What about runtime? Could longer runtime movies receive higher meta-scores as they are often associated with more complex storytelling, while audiences may rate them lower due to attention-span and pacing concerns?

Genre Differences and Their Impact on Ratings

Certain genres, such as action and comedy, may receive higher IMDb ratings due to mass mainstream appeal, but lower critic scores due to perceived lack of artistic depth. In contrast, perhaps critically favored genres such as drama and independent films may receive higher meta-scores but lower IMDb ratings due to their niche appeal.

The Role of Certificate (Censorship Rating)

R-rated movies may receive higher IMDb ratings due to mature content attracting a dedicated audience, whereas critics may assess them more rigorously depending on the execution of themes. Family-friendly movies may receive higher meta-scores due to broader accessibility but could have lower IMDb ratings if audiences find them less engaging compared to other categories. By investigating these hypotheses using this dataset that includes IMDb ratings, critic meta-scores, and various film characteristics, this study aims to provide a comprehensive analysis of the factors influencing rating discrepancies. Ultimately, our findings will offer valuable insights into how different audience segments perceive film quality, with implications for movie marketing, recommendation algorithms, and consumer behavior in the digital entertainment landscape.

In summary the following is our primary **research question:** how do different variable predictors in genre, censorship, runtime, and gross earnings predict how critics and fans will rate movies comparatively?

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

Data description

The data set is sourced from scraping all of the data of interest from IMDB into a comprehensive csv file. As stated by the individual who posted it on Kaggle, it contains "information about movies which appears on IMDB website.

Data was obtained by means of a web scraping in Python and combined with repository shared by IMDB" It was last updated in 2020. The observations are from the top 1000 rated movies from the last 50 years, up until 2020. General characteristics we are measuring include different quantitative variables pertaining to the films: runtime, year released, gross revenue, meta-score, and IMDB rating. Also, it includes categorical variables such as certificate earned by the movie (in relevance to censorship) and genre.

In terms of variable transformation, some films have more than one genre, while others have one. We could consider refactoring the variables into a factor, assigning a value to each genre. Perhaps this would allow us to associate value to films which have more than one genre.

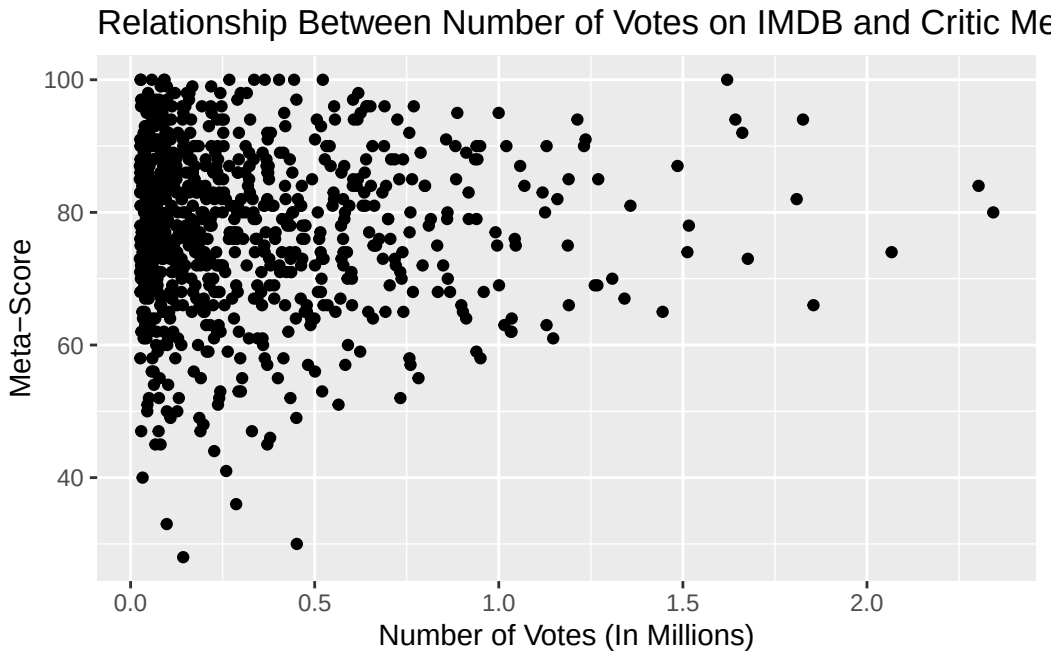
Additionally, we have the certificate variable: upon initial visualization of the distribution, we can see that there are many levels. To avoid overfitting a model, we may need to relevel these certificates to three basic categories: approved for all audiences, restricted, and other. Otherwise, if this does not pose an issue, we can keep as is.

Exploratory data analysis

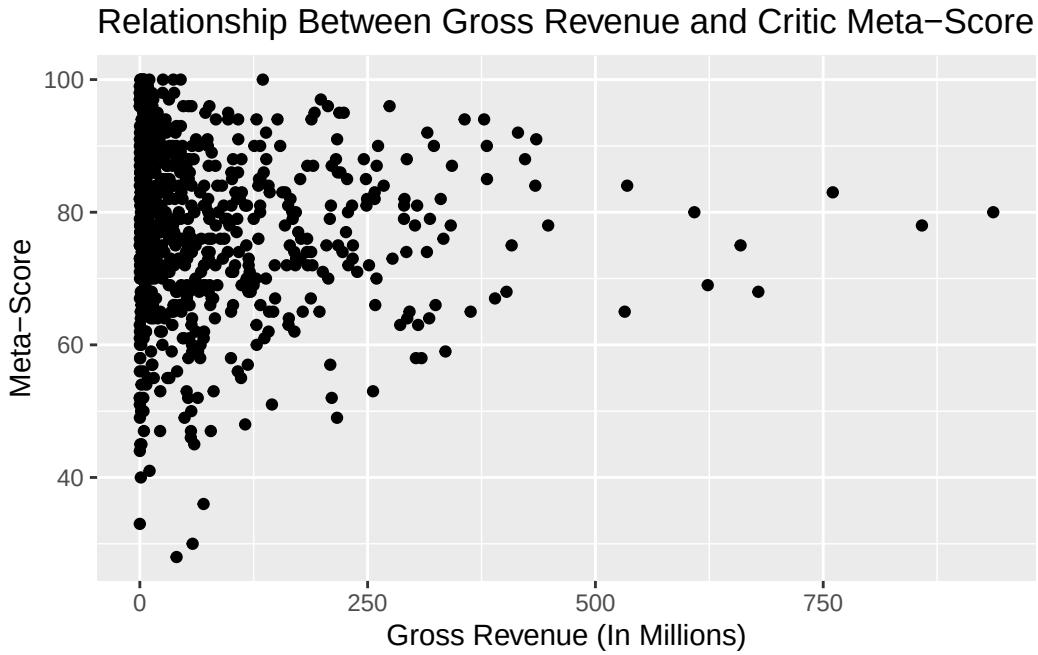
```
#mutated gross and vote predictors because they were in millions

imdb_top_1000 <- imdb_top_1000 %>%
  mutate(no_votes_scaled = No_of_Votes / 10^6,
         gross_scaled = Gross / 10^6)
         #Certificate = case_when(Certificate == "U" ~ "U",
                          #Certificate == "A" ~ "A",
                          #TRUE ~ "Other"))
```

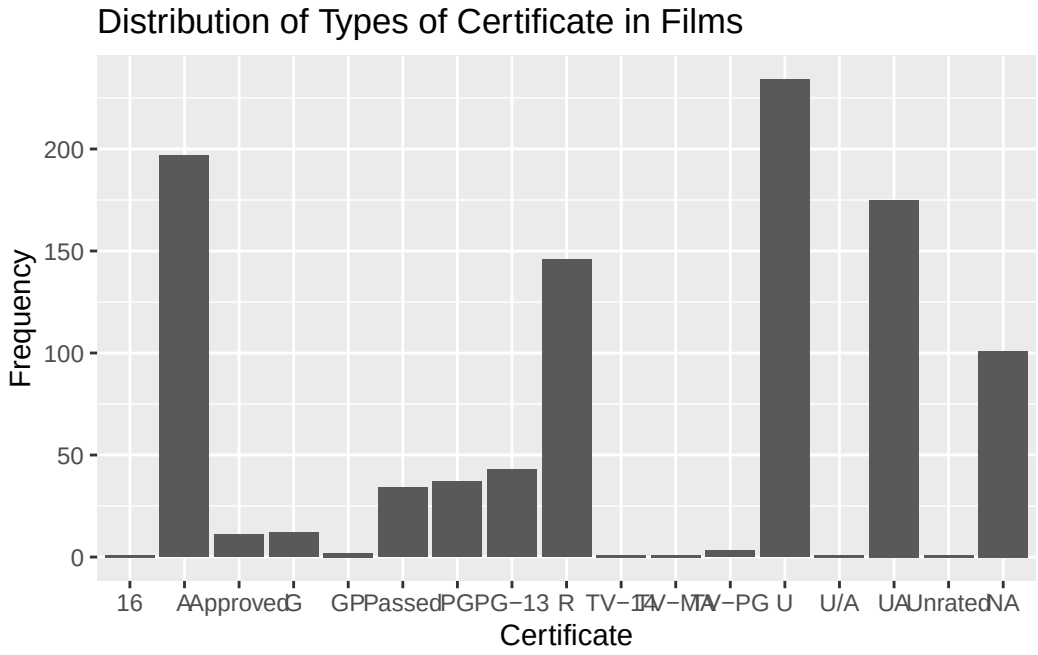
```
imdb_top_1000 %>%
  ggplot(aes(x = no_votes_scaled, y = Meta_score)) +
    geom_point() +
    labs(x = "Number of Votes (In Millions)",
         y = "Meta-Score",
         title = "Relationship Between Number of Votes on IMDB and Critic Meta-Score")
```



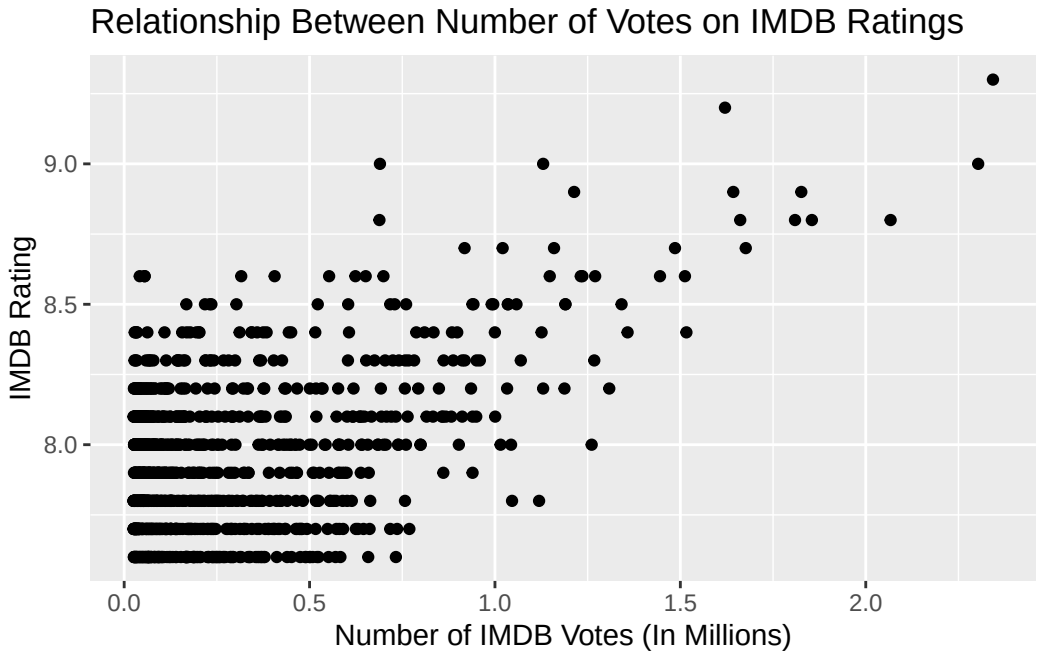

```
imdb_top_1000 %>%
  ggplot(aes(x = gross_scaled, y = Meta_score)) +
    geom_point() +
    labs(x = "Gross Revenue (In Millions)",
         y = "Meta-Score",
         title = "Relationship Between Gross Revenue and Critic Meta-Score")
```



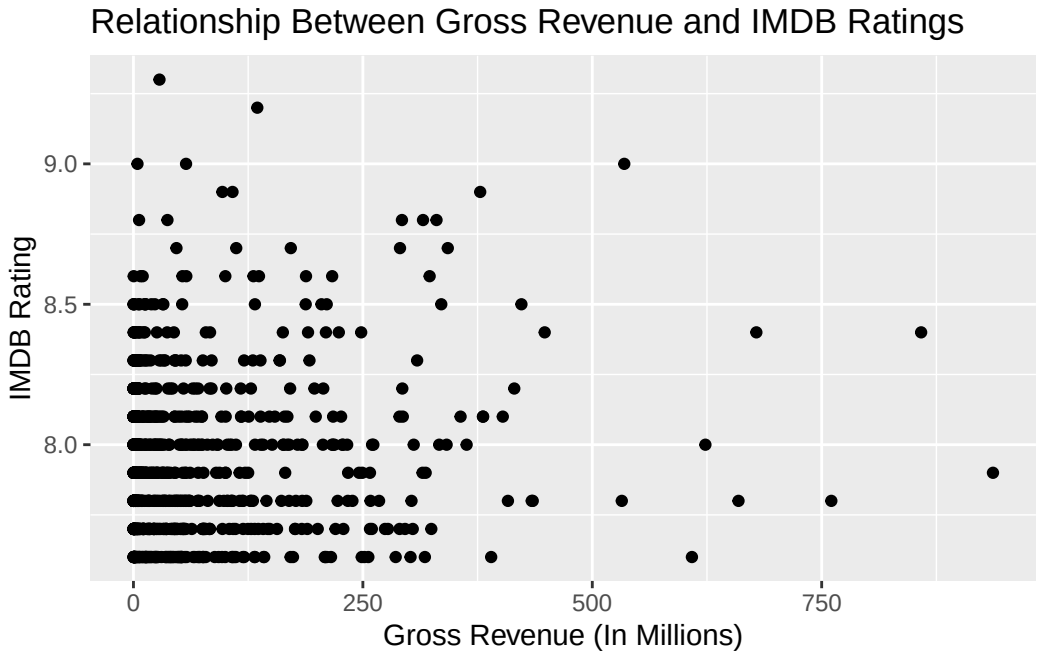

```
imdb_top_1000 %>%
  ggplot(aes(x = Certificate)) +
  geom_bar() +
  labs(x = "Certificate",
           y = "Frequency",
           title = "Distribution of Types of Certificate in Films")
```




```
imdb_top_1000 %>%
  ggplot(aes(x = no_votes_scaled, y = IMDB_Rating)) +
    geom_point()+
    labs(x = "Number of IMDB Votes (In Millions)",
         y = "IMDB Rating",
         title = "Relationship Between Number of Votes on IMDB Ratings")
```

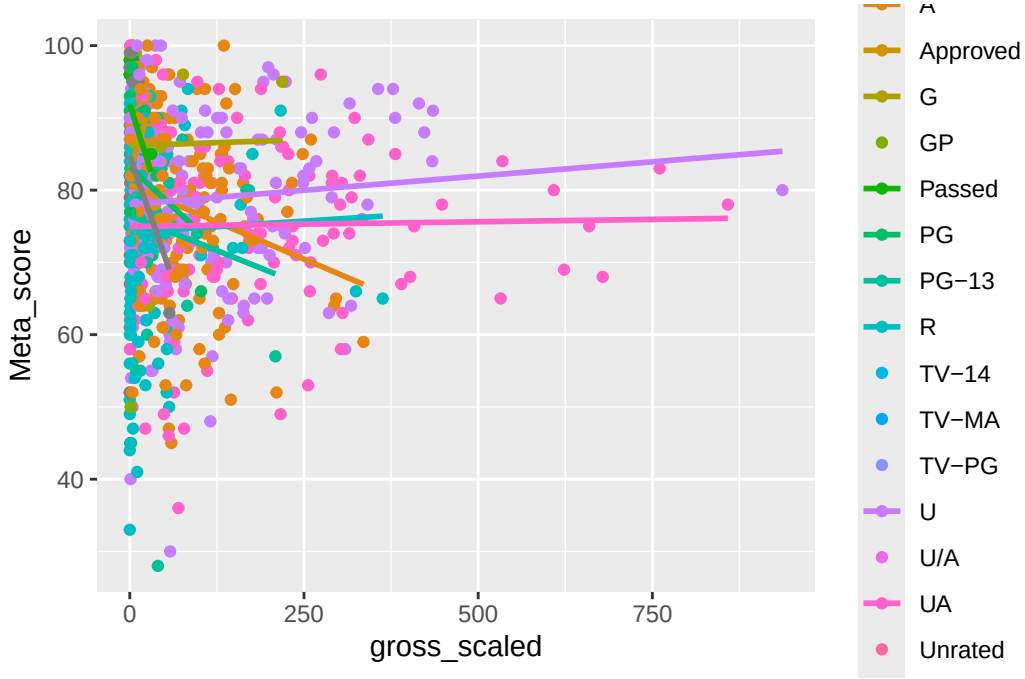



```
imdb_top_1000 %>%
  ggplot(aes(x = gross_scaled, y = IMDB_Rating)) +
    geom_point() +
    labs(x = "Gross Revenue (In Millions)",
         y = "IMDB Rating",
         title = "Relationship Between Gross Revenue and IMDB Ratings")
```



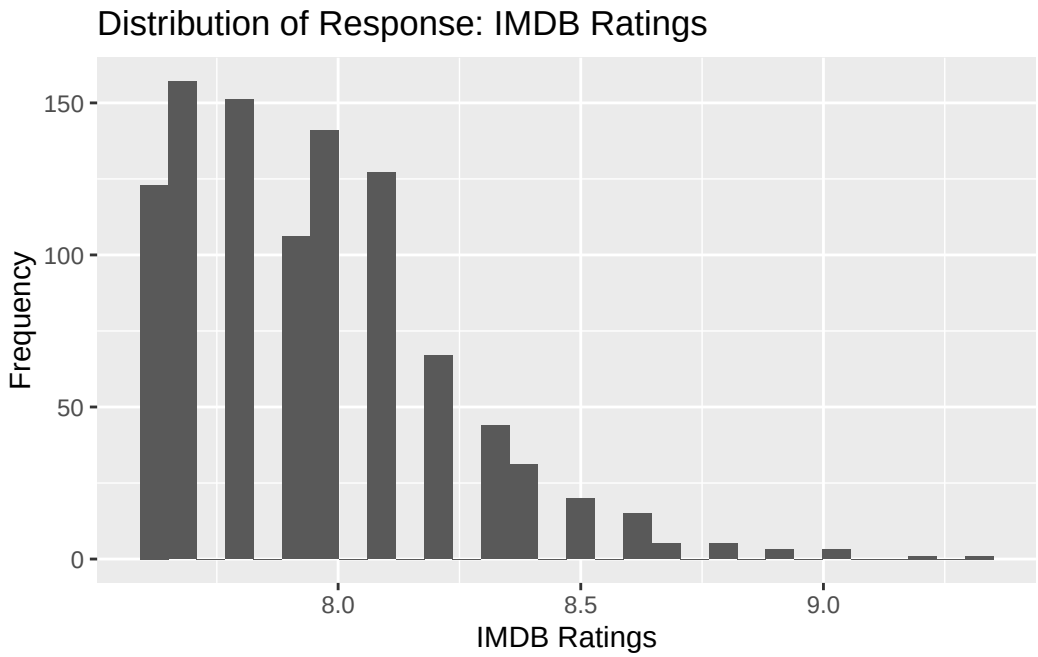
Upon initial EDA, it seems as if the relationship between gross revenue as a predictor for both IMDB and Metascore ratings are both nonlinear. Likewise, we see the same for the number of votes in relationship with the response variables. It does seem that there is a positive relationship between the number of votes on IMDB and the IMDB rating, but that is to be expected. For the other relationships, we may need to consider doing some data transformation.

```
imdb_top_1000 %>%
  ggplot(aes(x = gross_scaled, y = Meta_score, color = Certificate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



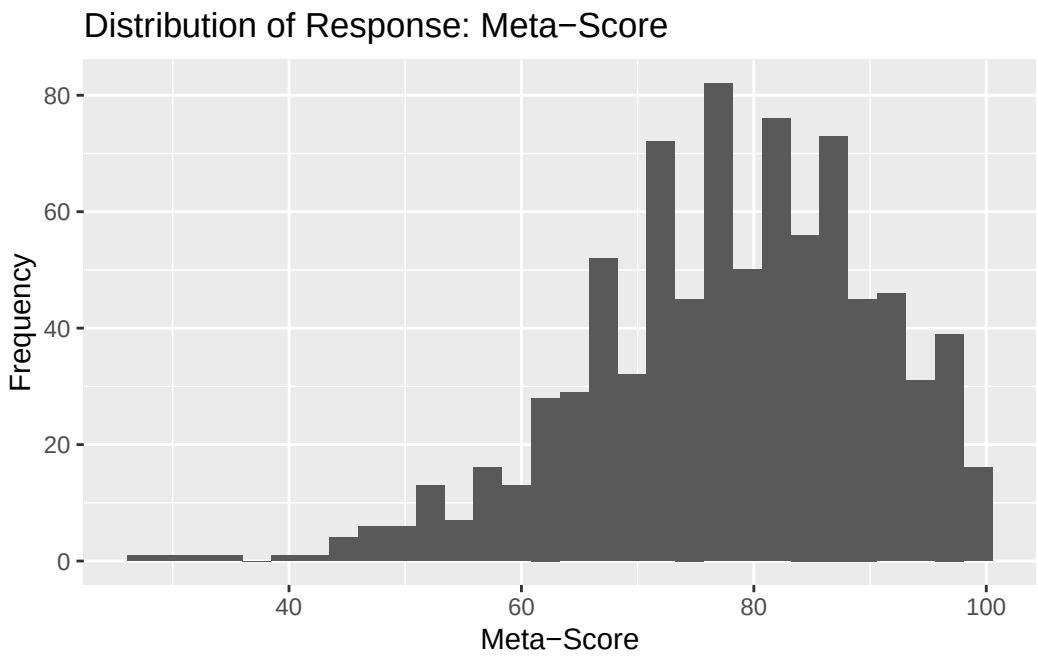
Upon initial EDA, it seems as if there is an interaction effect based on certificate of the film between films' gross revenue and their Meta score.

```
imdb_top_1000 %>%
ggplot(aes(x = IMDB_Rating)) +
  geom_histogram() +
  labs(x = "IMDB Ratings", y = "Frequency",
       title = "Distribution of Response: IMDB Ratings")
```



The distribution of IMDB Ratings (one of our response variables) seems to be skewed right.

```
imdb_top_1000 %>%
ggplot(aes(x = Meta_score)) +
  geom_histogram() +
  labs(x = "Meta-Score", y = "Frequency",
       title = "Distribution of Response: Meta-Score")
```



The distribution of meta-score (our other response variable) seems to be skewed left.

```
imdb_top_1000 %>%
  summarise(
    mean = mean(Meta_score, na.rm = TRUE),
    median = median(Meta_score, na.rm = TRUE),
    sd = sd(Meta_score, na.rm = TRUE),
    IQR = IQR(Meta_score, na.rm = TRUE),
    min = min(Meta_score, na.rm = TRUE),
    max = max(Meta_score, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 6
   mean median    sd   IQR   min   max
  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1  78.0     79  12.4    17    28   100
```

Above are the summary statistics for the response variable meta-score.

```
imdb_top_1000 %>%
  summarise(
    mean = mean(IMDB_Rating, na.rm = TRUE),
```

```
    median = median(IMDB_Rating, na.rm = TRUE),
    sd = sd(IMDB_Rating, na.rm = TRUE),
    IQR = IQR(IMDB_Rating, na.rm = TRUE),
    min = min(IMDB_Rating, na.rm = TRUE),
    max = max(IMDB_Rating, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 6
   mean median    sd   IQR   min   max
  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1  7.95    7.9 0.275 0.400   7.6   9.3
```

Above are the summary statistics for the response variable IMDB Rating.

Analysis approach

We are attempting to find how certain predictor variables will impact our response variables of meta-score and IMDB score. Potential predictor variables of interest are a film's gross revenue, the number of audience votes it receives, the year released, runtime, certificate of censorship, and genre of the movie. We plan to use multiple linear regression to see how these different factors interact with one another. For example, if we hold all else constant, does one censorship certificate (Rated R, X, etc.) have a greater effect on the score than another? And how does that differ between the two types of scores? Could we find that longer or shorter runtime presents an interaction effect?

Given our EDA with some of our primary predictor variables, we can see that not all of our visualizations are tending towards a linear form. Perhaps when we do our regression, we may need to do some transformations to make our data more fit for regression analysis.

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]