Statpadders

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

2025-03-17

```
library(tidyverse)
library(tidymodels)
library(dplyr)
library(corrplot)

imdb_top_1000 <- read_csv("data/imdb_top_1000.csv") |>
  drop_na()
```

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, decade released, 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 Decade Released 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?

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.

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

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.

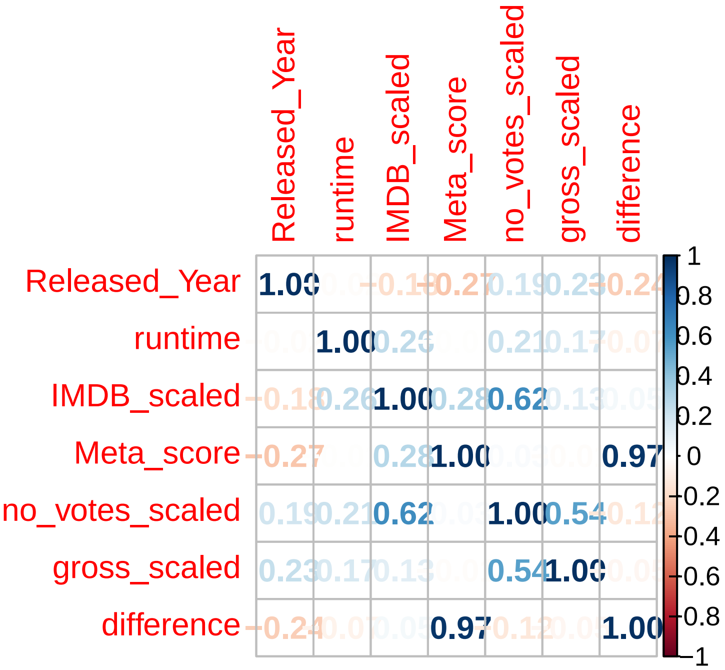
```r
imdb_top_1000 <- imdb_top_1000 |>
  mutate(no_votes_scaled = No_of_Votes / 10^6,
         gross_scaled = Gross / 10^6,
         IMDB_scaled = IMDB_Rating *10,
         Released_Year = if_else(Series_Title == "Apollo 13", "1995",
                                 Released_Year),
         difference = Meta_score - IMDB_scaled,
         runtime = as.numeric(str_remove(Runtime, " min")),
         Released_Year = as.numeric(Released_Year),
         decade = case_when(Released_Year < 1940 ~ "1930s",
                          Released_Year >= 1940 & Released_Year < 1950 ~ "1940s",
                          Released_Year >= 1950 & Released_Year < 1960 ~ "1950s",
                          Released_Year >= 1950 & Released_Year < 1960 ~ "1950s",
                          Released_Year >= 1960 & Released_Year < 1970 ~ "1960s",
                          Released_Year >= 1970 & Released_Year < 1980 ~ "1970s",
                          Released_Year >= 1980 & Released_Year < 1990 ~ "1980s",
                          Released_Year >= 1990 & Released_Year < 2000 ~ "1990s",
                          Released_Year >= 2000 & Released_Year < 2010 ~ "2000s",
                          Released_Year >= 2010 ~ "2010s",))
```

Further, we turned our year 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 productive. Very few 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 for better interpretability, we changed this variable into a categorical variable `decade`, where all of the years released are divided into decades (i.e. 1950s, 1960s, etc.) This gives better interpretability for that variable.

Additionally, the variable `Runtime` gave the runtime of each movie in the following template: "(number) mins". As such, it was a categorical variable. We changed this to remove the "mins" label, and make the runtime into a numerical variable, which only shows the number of minutes in that movie's runtime.

Potential Multicollinearity

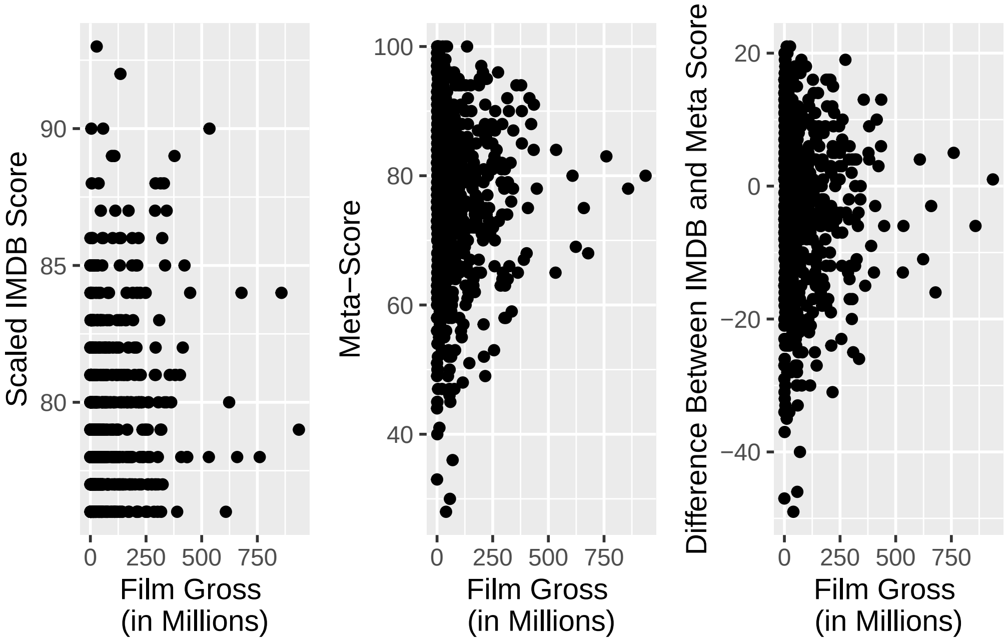
```
#geeksforgeeks.org/correlation-matrix-in-r-programming

matrix <- imdb_top_1000 %>%
  select(Released_Year, runtime, IMDB_scaled, Meta_score, no_votes_scaled,
         gross_scaled, difference)
c <- cor(matrix)
corrplot(c, method = "number")
```



Here, we created a correlation matrix to see which of our numerical predictors may be highly correlated, thus indicating potential multicollinearity. There is high correlation (0.97 correlation coefficient) between the variable `difference` and the variables `IMDB_scaled` and `meta_score`, but that is to be expected - the `difference` variable is mutated from both of those variables, to show the difference between them.`IMDB_scaled` and `no_votes_scaled` are also highly correlated with a coefficient of 0.62, but that is also to be expected - of course the votes on IMDB are correlated with an IMDB score. We would not fit a model with both of those variables.

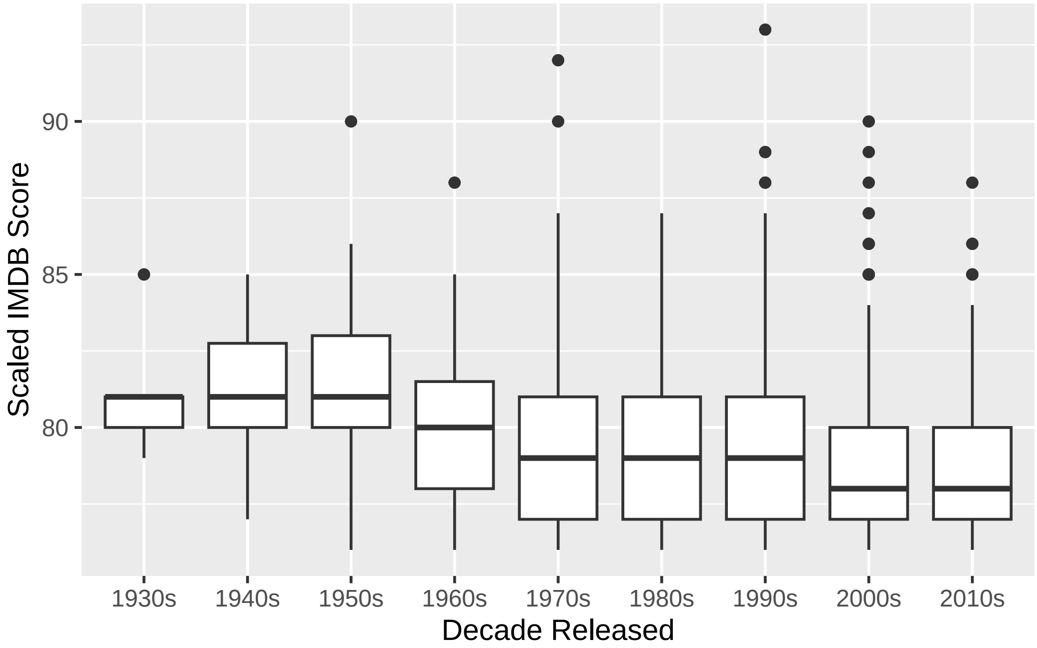
Univariate EDA

```r
library(patchwork)
p1 <- imdb_top_1000 %>%
  ggplot(aes(x = gross_scaled, y = IMDB_scaled )) +
  geom_point() + labs(
    x = "Film Gross \n(in Millions)",
    y = "Scaled IMDB Score"
  )
p2 <- imdb_top_1000 %>%
  ggplot(aes(x = gross_scaled, y = Meta_score )) +
  geom_point() +
   labs(
    x = "Film Gross \n(in Millions)",
    y = "Meta-Score"
  )
p3 <- imdb_top_1000 %>%
  ggplot(aes(x = gross_scaled, y = difference )) +
  geom_point() +
   labs(
    x = "Film Gross \n(in Millions)",
    y = "Difference Between IMDB and Meta Score"
  )

p1 + p2 + p3
```



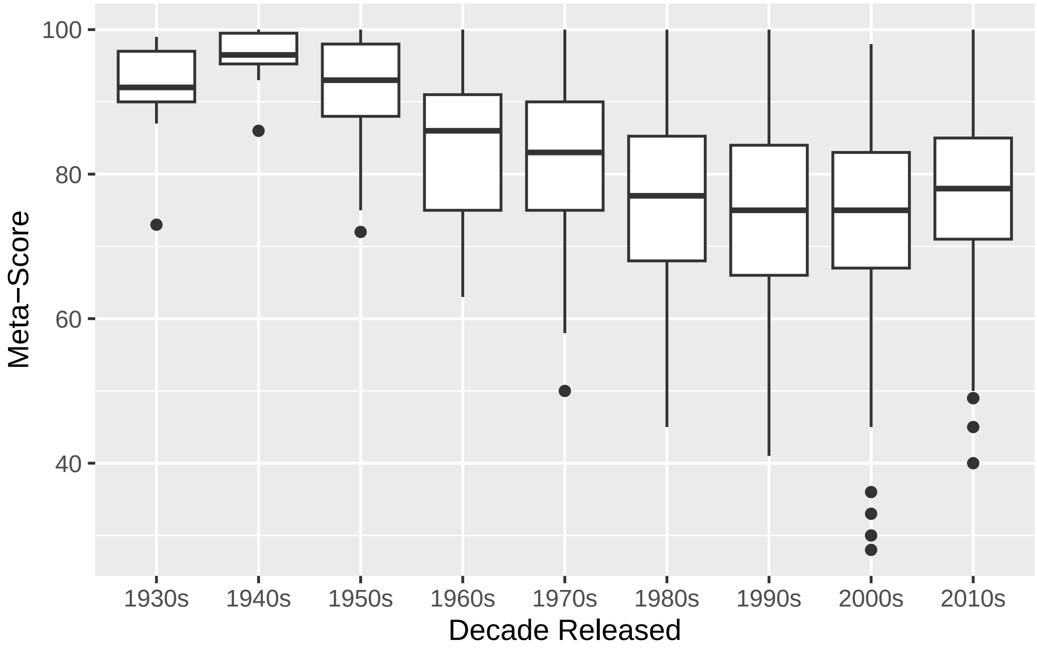
Add narrative here

```
library(patchwork)

imdb_top_1000 %>%
  ggplot(aes(x = decade, y = IMDB_scaled )) +
  geom_boxplot() + labs(
    x = "Decade Released",
    y = "Scaled IMDB Score"
  )
```



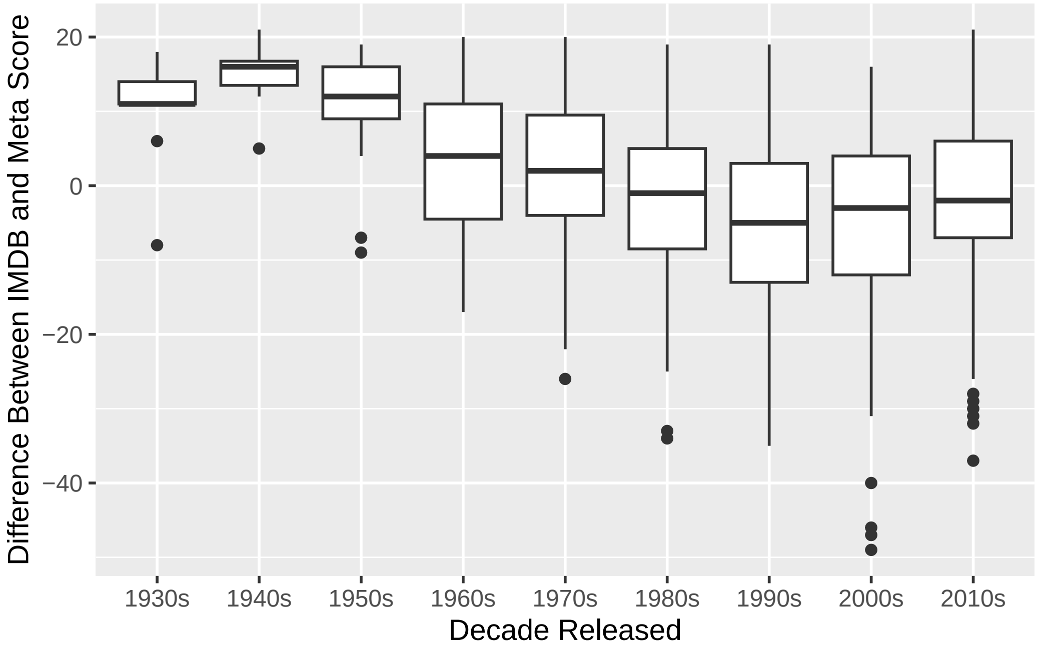
Judging from this intial univariate EDA of decade released vs the scaled IMDB score, there seems to be a negative correlation between date and IMDB score; as movies are newer (coming out in more recent decades), the median scaled IMDB score tends to be lower.

```
imdb_top_1000 %>%
  ggplot(aes(x = decade, y = Meta_score )) +
  geom_boxplot() +
   labs(
    x = "Decade Released",
    y = "Meta-Score"
  )
```



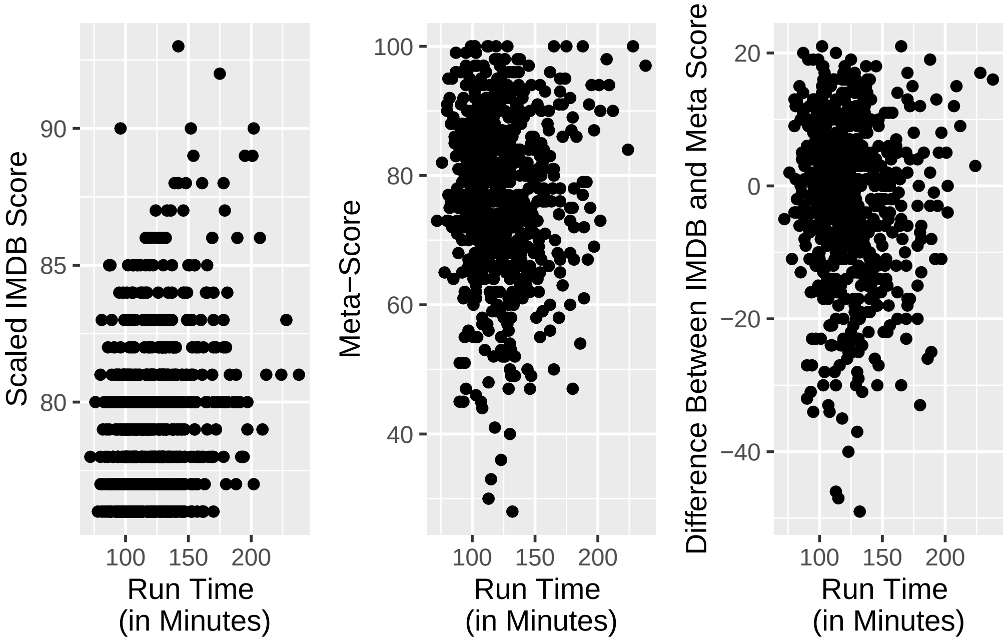
Similarly to IMDB score, the critics' median meta-scores also seem to be lower as movies are newer. In other words, the overall aggregated critic scores for films tend to be lower for movies in more recent decades.

```r
imdb_top_1000 %>%
  ggplot(aes(x = decade, y = difference )) +
  geom_boxplot() +
   labs(
    x = "Decade Released",
    y = "Difference Between IMDB and Meta Score"
  )
```



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, meta-scores tend to be lower than IMDB scores in more recent decades.

```
library(patchwork)
p1 <- imdb_top_1000 %>%
  ggplot(aes(x = runtime, y = IMDB_scaled )) +
  geom_point() + labs(
    x = "Run Time \n(in Minutes)",
    y = "Scaled IMDB Score"
  )
p2 <- imdb_top_1000 %>%
  ggplot(aes(x = runtime, y = Meta_score )) +
  geom_point() +
   labs(
    x = "Run Time \n(in Minutes)",
    y = "Meta-Score"
  )
p3 <- imdb_top_1000 %>%
  ggplot(aes(x = runtime, y = difference )) +
  geom_point() +
   labs(
    x = "Run Time \n(in Minutes)",
    y = "Difference Between IMDB and Meta Score"
  )

p1 + p2 + p3
```



Add narrative here

Bivariate EDA

do bivariate eda here

Interaction Effects

explore interaction effects here ::: callout-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. :::