Stat184_Group_Project

Do film releases from 1978 to the 2000s have a better IMDB rating on average compared to releases from the 2000s to 2022?

Madeline Hughes, Gina Seo, Chelsea Sala

Introduction

Film ratings, budgets, and genres all provide valuable insights into audience preferences and the quality of cinema. With the rapid evolution of filmmaking techniques, audience expectations, and styles, it is worth exploring whether older films have maintained their legacy in terms of viewer appreciation compared to modern releases. We are also taking into account the genre and budgets. The purpose of this study is to explore how the evolution of film production, societal tendencies, and viewing behaviors may have influenced audience reception as reflected in IMDB ratings. The main purpose of this report is to investigate whether the average IMDB ratings changed between films released from 1978 to 2000 and films released from 2000 to 2022. By analyzing these two time periods, we want to uncover any significant trends in film ratings over the decades.

Prior Investigations

As a preliminary investigation, we took a quick look at the IMDb dataset and made initial inferences about the available information before tidying it up and performing actual statistical testing. We hypothesized that there would be a stark difference in the average ratings of films released during the two periods, given the many advancements in the cinematic world. Before testing our hypothesis, we predicted that movies released between 1978 and the 2000s would exhibit a marginally bigger gap in the average ratings compared to those released from the 2000s to 2022. This difference could be attributed to variations in access to theaters, budget, advertising, screenwriting, filming technology, and other factors. We assumed that, on average, releases from 1978 to 2000 would have higher ratings, as many films from that era are considered classics, with nostalgia likely contributing to their elevated ratings.

Methodology/Data Provenance

To investigate whether film releases from 1978 – 2000 have a better IMDb rating on average compared to releases from 2000 – 2022, we will be utilizing the IMDb Top 250 Movies data set retrieved from Kaggle. The data-set contains the film releases from 1921 to 2022 along with its genre, rating, budget, box office revenue, runtime, etc. The purpose of the database is to serve as a public repository for film-related data, primarily used by audiences and researchers in the film industry to explore movie metadata and trends. With the given time range, we decided to choose a sample time frame of 44 years, from 1978 to 2022, in order to get a fair and balanced representation films released right before and after the 21st century.

With the given information, it is important that we tidy the data to filter out and extract the parameters that we need along with removing any null values and unwanted information to prevent any complications further along the line.

FAIR principles

We made sure the data-set had a clear and unique identifier to satisfy the findable aspect of the principles. To achieve accessibility the data is openly available through Kaggle and can be retrieved without any restrictive conditions. The data is interoperable, since it is stored in a widely accepted format which allows for seamless integration with analytical tools. Lastly, the data is reuseable due to its comprehensive metadata documentation which ensures reproducibility, detailing structure, variables, and collection methodologies.

CARE Principles

There is collectible benefit, since this analysis aims to benefit diverse stakeholders, including filmmakers, researchers, and audiences, by offering insights into rating trends. We have authority of control by ensuring that the data usage respects intellectual property and complies with IMDB's terms of use. We ethically handled the data to avoid misrepresentation or misuse, particularly when interpreting trends to satisfy the responsible principle. We also used transparency in methodology and acknowledgment of potential biases to follow the ethnic principle.

Results

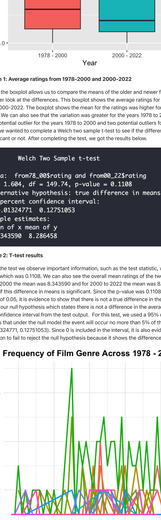

Figure 1: Average ratings from 1978–2000 and 2000–2022

Using the boxplot allows us to compare the means of the older and newer films beside each other to get a better look at the differences. This boxplot shows the average ratings for films in the years 1978-2000 and 2000-2022. The boxplot shows the mean for the ratings was higher for 1978 to 2000 than 2000 to 2022. We can also see that the variation was greater for the years 1978 to 2000. The plot shows there is one potential outlier for the years 1978 to 2000 and two potential outliers for 2000 to 2022. With this data we wanted to complete a Welch two sample t-test to see if the difference between years was significant or not. After completing the test, we got the results below.

```
        Welch Two Sample t-test

data:  from78_00$rating and from00_22$rating
t = 1.604, df = 149.74, p-value = 0.1108
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01324771  0.12751053
sample estimates:
mean of x mean of y
 8.343590  8.286458
```

Figure 2: T-test results

From the test we observe important information, such as the test statistic, which was 1.604 and the p-value which was 0.1108. We can also see the overall mean ratings of the two different time periods. For 1978-2000 the mean was 8.343590 and for 2000 to 2022 the mean was 8.286458. Our test results will show if this difference in means is significant. Since the p-value was 0.1108, which is greater than our alpha of 0.05, it is evidence to show that there is not a true difference in the average ratings. We failed to reject our null hypothesis which states there is not a difference in the average ratings. We can also obtain our confidence interval from the test output. For this test, we used a 95% confidence interval which means that under the null model the event will occur no more than 5% of the time. The interval was (-0.01324771, 0.12751053). Since 0 is included in the interval, it is also evidence that supports our decision to fail to reject the null hypothesis because it shows the difference in means is not significant.

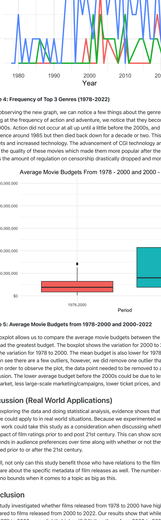

Figure 3: Frequency of Film Genre Across 1978–2022.

Using the line graph allows us to be able to see the frequency of all genres throughout the years in the same graph. This makes it easier to compare their popularity to each other. This particular line graph shows the popularity of film genres through the full 44 years that pertained to our study. Each genre is represented by a different line color, so the viewer is able to distinguish each genre. We want to take a closer look at the top 3 most common genres over the years in order to see if the frequency has changed from older to modern films. We did this by creating another line graph but only with the genres action, adventure, and drama. These 3 categories were the genres that appeared the most, so we wanted to see if more occurred before or after the 2000s to see if genre has affected ratings.

Figure 4: Frequency of Top 3 Genres (1978–2022)

After observing the new graph, we can notice a few things about the genres between the years. When looking at the frequency of action and adventure, we notice that they become much more popular after the 2000s. Action did not occur at all up until a little before the 2000s, and adventure had some prevalence around 1985 but then died back down for a decade or two. This could be due to higher budgets and increased technology. The advancement of CGI technology and budgeting could greatly affect the quality of these movies which made them more popular after the 2000s. Additionally, after the 2000s the amount of regulation on censorship drastically dropped and more societal acceptance grew.

Figure 5: Average Movie Budgets from 1978–2000 and 2000–2022

The boxplot allows us to compare the average movie budgets between the two time periods to see which year had the greatest budget. The boxplot shows the variation for 2000 to 2022 is drastically greater than the variation for 1978 to 2000. The mean budget is also lower for 1978 to 2000 than 2000 to 2022. We can see there are a few outliers, however, we did remove one outlier that was skewing the overall true data. In order to observe the plot, the data point needed to be removed to achieve the best results for our conclusion. The lower average budget before the 2000s could be due to less technology, smaller global film market, less large-scale marketing/campaigns, lower ticket prices, and inflation.

Discussion (Real World Applications)

After exploring the data and doing statistical analysis, evidence shows that this area of topic is something that we could apply to in real world situations. Because we experimented with the topic of film releases, future work could take this study as a consideration when discussing whether or not to take into account the impact of film ratings prior to and post 21st century. This can show screenwriters and producers on the trends in audience preferences over time along with whether or not they resonate more with films released prior to or after the 21st century.

Overall, not only can this study benefit those who have relations to the film industry but average viewers who care about the specific metadata of film releases as well. The number of real-world applications holds no bounds when it comes to a topic as big as this.

Conclusion

This study investigated whether films released from 1978 to 2000 have higher average IMDB ratings compared to films released from 2000 to 2022. Our results show that while the average ratings for films from 1978 to 2000 were slightly higher (8.34) than those from 2000 to 2020 (8.29), the difference is not statistically significant, as evidenced by a p-value of 0.1108 and a confidence interval (-0.0132, 0.1275) that includes zero. Therefore, we fail to reject the null hypothesis that there is no significant difference in average ratings between the two periods.

Secondary analyses revealed insights into the potential factors influencing these ratings. Genre frequency shifted significantly, with action and adventure films becoming more prominent after 2000, likely due to advancements in CGI and increased budgets. This aligns with our observation that the average budgets for films released after 2000 were notably higher, reflecting broader technological advancements and globalization in filmmaking.

While our findings suggest that the time period alone does not significantly influence average ratings, the interplay between factors such as genre popularity and budget warrants further exploration. The study highlights the evolving dynamics of the film industry and how technological advancements, societal shifts, and changes in audience preferences shape movie production and reception. Future research could extend this analysis by incorporating more nuanced metrics, such as audience demographics and streaming trends, to deepen our understanding of these patterns.

Citations

Johnson, N. (2009, October 9). Nathaniel Johnston ' imdb movie ratings over the years. Blog. https://njohnston.ca/2009/10/imdb-movie-ratings-over-the-years/

Bischoff, P. (2016, August 20). Why old movies get better ratings on Rotten Tomatoes, Metacritic, and imdb. Medium. https://medium.com/@pabischoff/why-old-movies-get-better-ratings-on-rotten-tomatoes-metacritic-and-imdb-a9503003183d

Code Appendix

```
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggplot2")
install.packages("stats")
install.packages("knitr")
library(dplyr)
library(tidyr)
library(ggplot2)
library(stats)
library(knitr)
library(readr)

IMDB_Top_250_Movies <- read_csv("~/Downloads/IMDB Top 250 Movies.csv")
View(IMDB_Top_250_Movies)

from78_00 <- IMDB_Top_250_Movies |> filter(year >= 1978 & year <= 2000)

from00_22 <- IMDB_Top_250_Movies |> filter(year >= 2000 & year <= 2022)

from78_00_avg <- mean(from78_00$rating, na.rm = TRUE)
from00_22_avg <- mean(from00_22$rating, na.rm = TRUE)

t_test_result <- t.test(from78_00$rating, from00_22$rating)
print(t_test_result)

#data:  from78_00$rating and from00_22$rating
#t = 1.604, df = 149.74, p-value = 0.1108
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
# -0.01324771  0.12751053
#sample estimates:
#mean of x mean of y
#8.343590  8.286458

IMDB_Top_250_Movies$period <- ifelse(IMDB_Top_250_Movies$year <= 1978 & IMDB_Top_250_Mo
ggplot(IMDB_Top_250_Movies, aes(x = period, y = rating, fill = period)) +
        geom_boxplot() +
        labs(title = "Average Ratings From 1978-2022 and 2000-2022",
             x = "Year",
             y = "Ratings") +
        scale_color_manual(values = c("1978-2000" = "aquamarine4", "2000-2022" = "cora
        theme_minimal()

genre78_22 <- IMDB_Top_250_Movies |>
  filter(year >= 1978 & year <= 2022)

genre_split <- genre78_22 |>
  separate_rows(genre, sep = ",") |>
  mutate(genre = trimws(genre))

genre_count <- genre_split |>
  group_by(year, genre) |>
  summarise(n = n(), .groups = "drop")

ggplot(genre_count, aes(x = year, y = n, color = genre)) +
  geom_line(linewidth = 1) +
  labs(title = "Frequency of Film Genre Across 1978 - 2022",
       x = "Year",
       y = "Frequency of Genres",
       color = "Genres") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5)
  )
```