Exploring Factors Influencing Movie Success

Stat 184 Final Project

Sara Al Riyami Layan Al Busaidi

April 28, 2025

Introduction

Movies are a huge part of our lives and a multi-billion-dollar industry that shapes and reflects society. With millions of people watching and reviewing films every year, the movie industry continues to evolve. Thousands of films are released annually, but only a few become truly successful. What factors contribute to a movie's success? Is it the genre, the budget, the actors, or something else? In this project, we set out to explore the elements that influence a movie's success by analyzing datasets of films and looking at trends in ratings, profits, and casting.

We are interested in this topic because movies are not only a major form of entertainment but also a massive industry. Understanding what leads to success can offer insights into both audience preferences and the strategies that shape film making today.

To guide our analysis, we focused on three key research questions:

1. How do audience ratings compare across the five most common movie genres?
2. Which top studios have the best return on investment, and is there a relationship between movie budget and profit?
3. Which stars appear most frequently in successful movies, and does their presence correlate with higher success?

Through these questions, we aim to uncover patterns and insights that explain what contributes to a movie's success. By comparing audience ratings across genres, we hope to identify which types of films tend to get the most viewers. Investigating studio performance and budget-profit relationships allows us to understand how financial decisions impact profitability. Finally, analyzing the presence of frequently featured stars helps us explore whether certain actors consistently appear in higher-performing films, potentially revealing a link between star presence and success. Overall, our goal is to better understand the creative and financial factors that shape successful movies.

Data Provenance

SARA

Main Dataset

SARA

Secondary Dataset

SARA

Merged and Final Dataset

SARA

FAIR and CARE Principles

FAIR SARA CARE LAYAN

EDA: Exploratory Data Analysis

In our project, we implemented **Exploratory Data Analysis** to understand the structure of our datasets and uncover any errors or patterns before diving into deeper analysis. We began by cleaning and **wrangling** the data to ensure it was accurate and usable. This included identifying and removing rows with missing values (*NAs*), as well as fixing inconsistencies.

After cleaning the data, we created **frequency** and **summary tables** to examine key statistics such as the mean, median, and standard deviation, giving us a clearer picture of the overall distribution of variables like ratings. We then used **visualizations** to uncover patterns and relationships that weren't immediately apparent from the summary statistics alone.

Through EDA, we were able to detect irregularities in our dataset, gain a better understanding of its overall structure, and identify key trends that guided our analysis. This process provided a strong foundation for answering our research questions and exploring the factors that contribute to a movie's success.

Genre and Rating

In this section, we will explore how audience ratings vary across the five most common movie genres. We begin by examining the summary statistics to better understand the distribution of ratings within each genre.

Summary Table

We created a **summary table** that highlights key statistics such as *film count, minimum, maximum, mean, median, standard deviation, and the 1st and 3rd quartiles*. This helps us better understand the distribution and variation in ratings across genres. Table 1 below showcases these statistics for the top genres.

Table 1: Summary Table of Ratings for the Top 5 Movie Genres

Genre	FilmCount	MinRating	Q1Rating	MedianRating	Q3Rating	MeanRating	MaxRating	SdRating
Comedy	208	2.1	5.7	6.3	6.900	6.263942	8.8	0.9452131
Drama	207	4.1	6.4	7.0	7.500	6.906763	9.3	0.7832919
Action	199	3.7	5.8	6.3	6.900	6.301507	8.3	0.8837750
Adventure	94	4.7	6.1	6.5	7.275	6.563830	8.6	0.8893457
Horror	77	4.0	5.3	6.1	6.600	5.974026	8.1	0.9505404

Table 1 presents an overview of the ratings for the top five movie genres: Comedy, Drama, Action, Adventure, and Horror. Among these, **Drama** clearly stands out as the highest-rated genre, with a **mean rating** of **6.91**, a **median** of **7.0**, and a **maximum rating** of **9.3**. It also shows the most consistent performance, having the **lowest standard deviation** among all genres. In contrast, **Horror** ranks the lowest, with an **average rating** of **5.97** and the **widest spread** in ratings, indicating more variability in audience reception. The other genres fall in between, though **Adventure** shows a high level of variability similar to Horror.

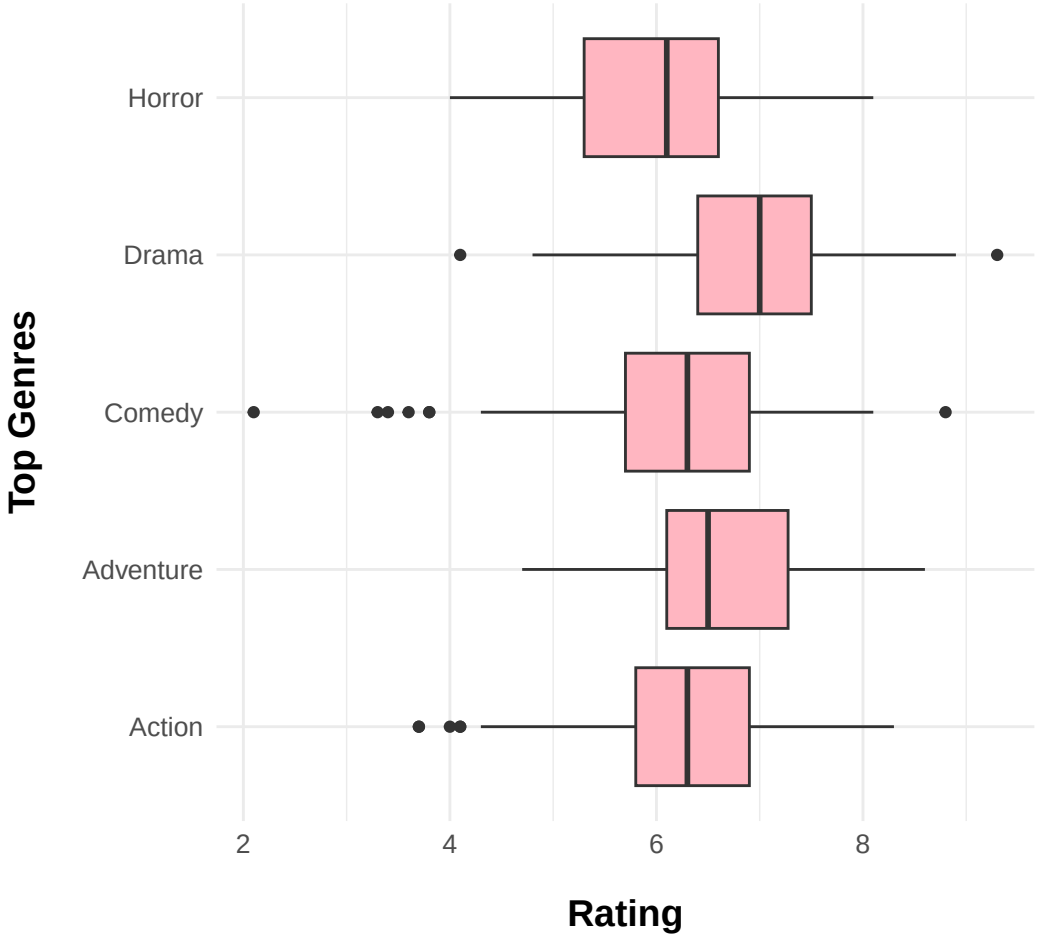
When looking at film counts, **Comedy** and **Drama** are the most represented, each with over **200 films**, while **Horror** has the lowest count with just **77 films**, which may contribute to its greater rating variability.

Overall, the data suggests that Drama is both the most consistently high-performing genre and the best-rated on average, whereas Horror tends to be more unpredictable and generally receives lower ratings.

Box Plot

Figure 1 shows a box plot of the rating distributions for the top five movie genres. This visual helps us compare the typical ratings, variation, and presence of outliers across genres. It supports and expands on the summary statistics discussed earlier.

Figure 1: Distribution of Ratings for Top 5 Genres



Companies and Movies

SARA

Q 3

Conclusion

LAYAN

Sources and References

LAYAN

Code Appendix

```
# Load all necessary packages -----
library(tidyverse)
library(rvest)
library(dplyr)
library(janitor)
library(knitr)
library(kableExtra)
library(ggplot2)

#Define global elements ----
psuPalette <- c("#1E407C", "#BC204B", "#3EA39E", "#E98300",
                "#999999", "#AC8DCE", "#F2665E", "#99CC00")

basePart <- "https://raw.githubusercontent.com/Stat184-Spring2025/"
mainPart <- "Sec4_FP_Layan_Sara/main/Data/MoviesJoined.csv"
url <- paste0(basePart,mainPart)
MoviesJoined <- read.csv(url, header = TRUE)

# Creating a summary table of Ratings by Genres ----

Genre_summary <- MoviesJoined%>%
  group_by(Genre)%>%      # Groups the data by Genre column
  summarise(           # Calculates summary statistics for each genre
    FilmCount = n(),      # Number of films in each genre
    MinRating = min(Rating, na.rm = TRUE),   #Minimum rating (ignores NA values)
    Q1Rating = quantile(Rating, 0.25, na.rm = TRUE),  # First quartile
    MedianRating = median(Rating, na.rm = TRUE),      # Median rating
    Q3Rating = quantile(Rating, 0.75, na.rm = TRUE),  # Third quartile
    MeanRating = mean(Rating, na.rm = TRUE),     # Mean (average) rating
    MaxRating = max(Rating, na.rm = TRUE),      # Maximum rating
    SdRating = sd(Rating, na.rm = TRUE)        # Standard deviation of ratings
  ) %>%
  arrange(desc(FilmCount))%>%          # Sorts the genres by film count
  slice_head(n=5)       # Selects the top 5 movie genres with the most films

# Displaying the summary table ----
Genre_summary%>%
  kable(
    booktabs = TRUE,
    align = c("l", rep("c",8)), # Left-aligns the first column, centers the rest
    format = "latex"
  )%>%
  kableExtra::kable_styling(
    font_size = 16,          # Sets font size of the table
```

```
    latex_options = c("striped","scale_down"),
  )%>%
  row_spec(0, bold = TRUE, background = "pink")%>%  # Styles the header
  column_spec(1, italic = TRUE) # Styles the 1 column

# Wrangling Data ----
## Get Top 5 Genres
TopGenres <- MoviesJoined %>%
  count(Genre, sort = TRUE) %>%  # Counts num of movies per genre and sorts them
  slice_max(order_by = n, n = 5) %>%  # Selects top 5 genres w most movies
  pull(Genre)

## Show data for only the Top 5 genres
MovieGenre <- MoviesJoined %>%
  filter(Genre %in% TopGenres)  # Filter movies of only the top 5 genres


# Create the box plot for Genre and Ratings----
ggplot(
  data = MovieGenre,
  mapping = aes(
    x = Rating,      # Set the x-axis to represent Rating
    y = Genre    # Set the y-axis to represent Genre
  )
) +
geom_boxplot(fill = "lightpink") +   # Creates box plot with pink boxes
labs(                #labels the x and y axis
  y = "Top Genres",
  x = "Rating"
) +
theme_minimal()+
theme(
  text = element_text(size = 12),
  axis.title.x = element_text(face = "bold",   # Make the x-axis title bold
                              size = 14,    # Set font size to 14
                              margin = margin(t = 15)
                              ),
  axis.title.y = element_text(face = "bold",
                              size = 14,
                              margin = margin(r = 15)
                              ) # margin pushes titles away from axis
)
```