# 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

To explore the factors contributing to movie success, we used two well-sourced datasets that each provided different but complementary information from major online movie databases. We cleaned and combined these datasets through a detailed wrangling and merging process to create a final dataset. This merged version gave us the foundation needed to conduct a thorough Exploratory Data Analysis (EDA) and effectively address our research questions.

Before settling on these data sources, we evaluated multiple datasets. An earlier dataset **link the source of old?**, included many relevant attributes, but the newer ones offered more complete and consistent data, especially in financials and studio identifiers, making them a better fit for our analysis.

### Main Dataset

Our main dataset was sourced from a **Github** repository created by the authors Tran Hieu Le, Totyana Hill, Fahim Ishrak and Zhilin Wang. It was originally scraped from TMDb (The Movie Database) this dataset contains information on approximately 5,000 movies released between the 1930s and 2016. Each *case* in the dataset represents a single movie, and includes essential attributes related to its financial performance and basic descriptors. The key *attributes* include production budget, worldwide revenue, title, production company, runtime, and genre. The *purpose* of collecting this data was to offer a structured overview of film financials, enabling analysis of profitability and return on investment.

### Secondary Dataset

Our secondary dataset was obtained from **Kaggle** and published by Daniel Grijalva. It was originally scraped from IMDb and contains metadata on approximately 7,600 movies spanning from 1980 to 2020. Each *case* corresponds to an individual movie and captures its descriptive, creative, and audience-related *attributes*. These include age rating, director, writer, lead actor (star), release year, rating count, and rating score. The *purpose* of this dataset was to supplement our main data by providing richer descriptive and audience-related information, which helped us better address our research questions.

### Merged and Final Dataset

To answer our research questions, we combined our two datasets into a single, cleaned, and well-structured table. The datasets were merged based on matching *movie titles* and *production companies* to ensure accuracy and consistency. To maintain data quality, we excluded records with missing or invalid values and treated zero values in financial fields as missing. We also created a new variable, *Profit*, by subtracting budget from revenue to better assess the financial success in our analysis.

The final dataset includes movies released from 1980 to 2019, which aligns with the expected coverage based on the original time spans of the two source datasets. It contains 15 key attributes across essential categories: content (title, genre, runtime, rating), financials (budget, revenue,

profit), production details (studio, director, star), and audience reception (rating score and count). This comprehensive dataset provided a strong foundation to explore our research questions related to genre-based ratings, profit of companies, and the influence of star actors on a film's success.

## FAIR and CARE Principles

The **FAIR** principles are essential for promoting high standards in data sharing, transparency, and longterm usability which our datasets do align to these guidelines. **Findable**: Our primary dataset was sourced from Kaggle and the secondary from GitHub, both of which are widely recognized platforms that provide basic metadata and descriptive tagging, making the data easy to locate. **Accessible**: Both datasets are publicly available and can be retrieved without login or payment barriers through standard web protocols. However, because the long-term availability of external repositories is not guaranteed, we also uploaded both original datasets as well as our cleaned and merged version to our own GitHub repository to ensure continued access and reproducibility. **Interoperable**: The datasets are provided in CSV format, a widely accepted and accessible format that enables easy integration into R and other data analysis tools, which insures smooth integration and analysis across various platforms. **Reusable**: The data includes meaningful attributes that provide context for interpretation. Through our cleaning and documentation process, we enhanced its structure, added derived variables like Profit, and clearly described the data's provenance, making it suitable for future research.

The **CARE** principles, focus on the ethical use of data for all people, especially for Indigenous or marginalized communities. While our datasets do not directly involve sensitive or community specific data, it's still important to reflect on these principles to ensure our work is thoughtful, fair, and respectful. **Collective Benefit**: Our project aims to uncover trends in the film industry that can help others better understand what drives movie success, whether creatively or financially. We hope our analysis supports open learning and sparks interest in data-driven storytelling. **Authority to Control**: The datasets we used weren't officially released by the platforms that host the data (IMDb, TMDb). That being said, they do algin with their terms and conditions of public usage. **Responsibility**: We made sure to treat the data with care. Since the dataset is non-sensitive and publicly accessible, we mainly focused on cleaning, organizing, and documenting it properly so our analysis would be accurate, transparent, and respectful. **Ethics**: We focused on minimizing harm by ensuring that our dataset did not include private or sensitive information and we avoided making claims that could misrepresent people or groups. We used the data strictly for educational purposes, and we made sure our work promoted fairness.

## 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.

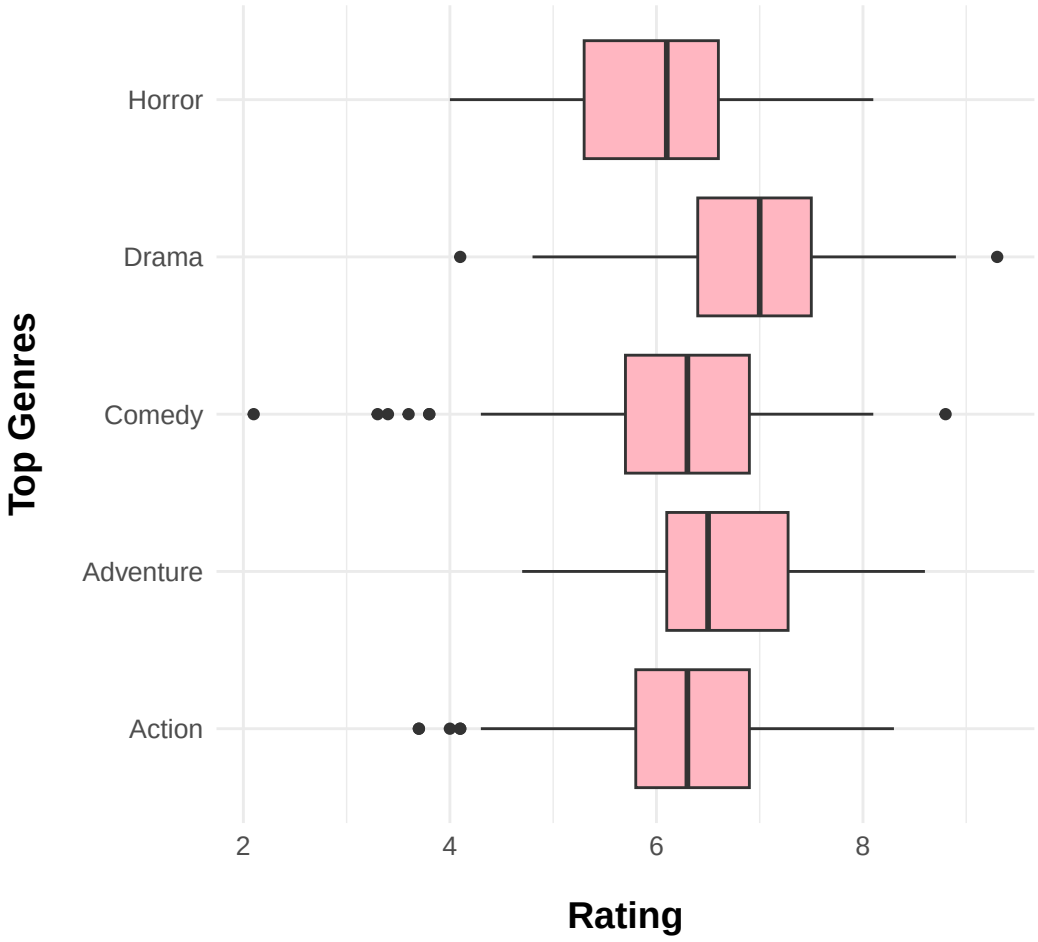
When looking at film counts, **Comedy** and **Drama** are the most represented, each with over **200 films**, while **Horror** has the lowest count with just **77 films**, which may contribute to its greater rating variability.

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")


# Tidy and Merging the two datasets: ----

# Read Data ----
url1 <- "https://raw.githubusercontent.com/Stat184-Spring2025/Sec4_FP_Layan_Sara/main/Data/Budg
MoviesSubRaw <- read.csv(url1, header = TRUE)
url2 <- "https://raw.githubusercontent.com/Stat184-Spring2025/Sec4_FP_Layan_Sara/main/Data/Movi
MoviesMainRaw <- read.csv(url2, header = TRUE)

# Tidying Secondary Movies data ----
# Filter out unwanted columns
MoviesSubTidy <- MoviesSubRaw %>%
  select(-c("released","budget","gross","genre","runtime"))

#Tidy the Main Movie data----
# Filter out unwanted columns
MoviesMainTidy <- MoviesMainRaw%>%
  select(-c("X","popularity", "release_date", "vote_average",
            "vote_count","Number_Genres"))

#Merging other two datasets----
MoviesJoined <- MoviesMainTidy%>%
  # Join by Title and Company
  inner_join(MoviesSubTidy, by = c("title"="name",
                                   "production_companies"="company") )%>%
  # Rename the columns
  rename(
    Title = title,
    Genre = genres,
    Company = `production_companies`,
    AgeRating= rating,
```

```
    Year = year,
    Rating = score,
    RatingCount = votes,
    Director = director,
    Writer = writer,
    Star = star,
    Budget = budget,
    Revenue = revenue,
    RunTime= runtime,
    Country = country
  ) %>%
  # Re-order the columns
  select(Title, Genre, Company, AgeRating, Year,
         Rating, RatingCount, Director, Writer, Star, Budget, Revenue,
         RunTime, Country) %>%
  # Found two movies with revenues that are not in millions
  # Changed the two values
  mutate(
    Revenue = ifelse(Title == "Chasing Liberty", 12000000, Revenue),
    Revenue = ifelse(Title == "Death at a Funeral", 46000000, Revenue)
  ) %>%
  # Replace 0 with NA in Budget and Revenue
  mutate(
    Budget = ifelse(Budget == 0, NA, Budget),
    Revenue = ifelse(Revenue == 0, NA, Revenue)
  ) %>%
  # Make a Profit column
  mutate(Profit = Revenue-Budget) %>%
  # Replace "Not Rated" and "Unrated" in AgeRating with NA
  mutate(
    AgeRating = ifelse(AgeRating %in% c("Not Rated", "Unrated"),
                       NA, AgeRating)
  ) %>%
  # Drop rows with missing Budget, Revenue, or AgeRating
  drop_na(Budget, Revenue, AgeRating)

# Create a csv file and save it
#write.csv(
#  MoviesJoined,
#  file = "MoviesJoined.csv",
#  row.names = FALSE
#)


# Read the joined dataset ----
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----
```

```r
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
)
```