Exploring Factors Influencing Movie Success

Stat 184 Final Project

Sara Al Riyami Layan Al Busaidi

April 28, 2025

1 Introduction

2 Hypotheses and 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. Who are the most frequently featured stars among in the movie dataset?

3 Data Provenance

3.1 Main Dataset

3.2 Secondary Dataset

3.3 Merged and Final Dataset

4 FAIR and CARE Principles

5 EDA: Exploratory Data Analysis

5.1 Genre and Rating

5.1.1 Summary

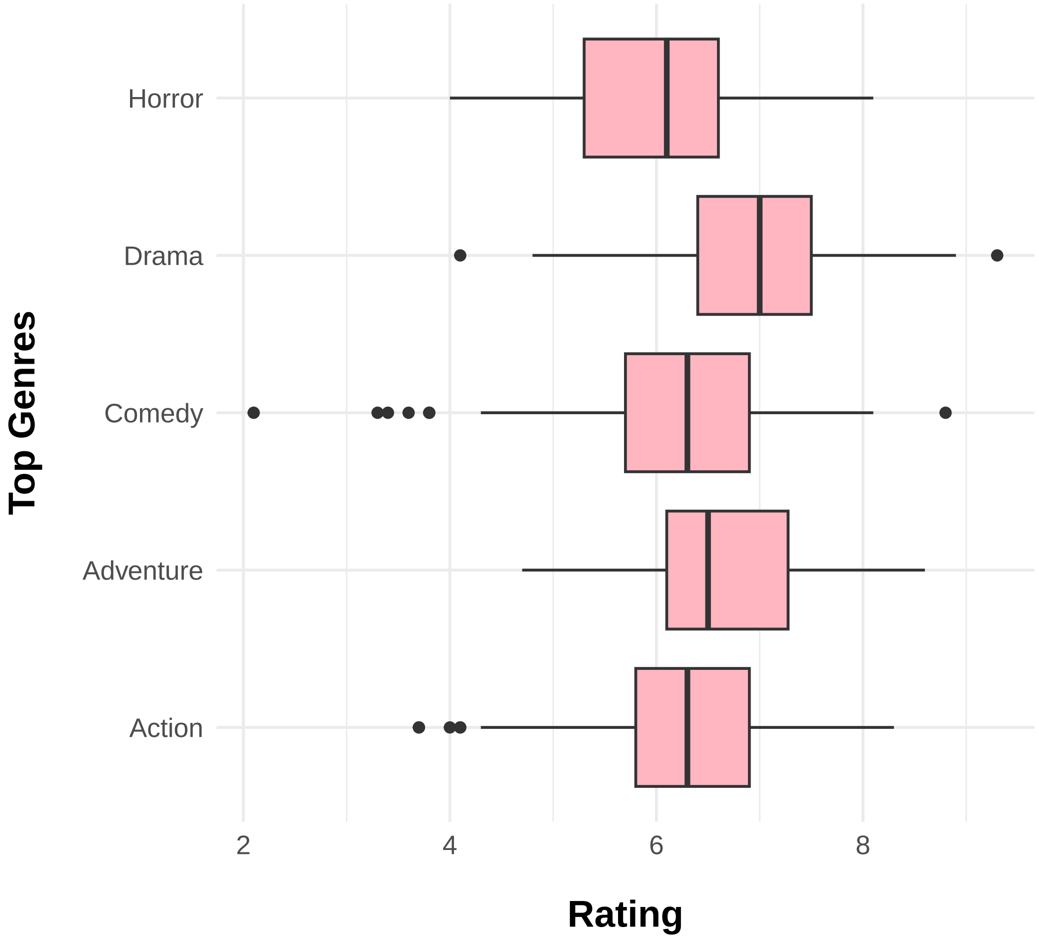
Table 1: Rating Summary by Top 5 Movie Genres

Genre	FilmCount	MinRating	Q1Rating	MedianRating	Q3Rating	Mea

Comedy	208	2.1	5.7	6.3	6.900	6
Drama	207	4.1	6.4	7.0	7.500	6
Action	199	3.7	5.8	6.3	6.900	6
Adventure	94	4.7	6.1	6.5	7.275	6
Horror	77	4.0	5.3	6.1	6.600	5

5.1.2 Box Plot

Figure 1: Distribution of Ratings for Top 5 Genres



6 Conclusion

7 Sources and References

# Code Appendix

```r
# 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
  )%>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped","hover"),
    font_size = 14,          # Sets font size of the table
    full_width = FALSE
  )%>%
  row_spec(0, bold = TRUE, background = "#AC8DCE")%>%  # 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
)
```