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 the presence of specific stars tend to be linked to higher movie 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 them through an extensive data wrangling and merging process to create the 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 (Chhikara, 2024), 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 (Hieu2695, 2019). 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 (Grijalva, 2019). Each *case* corresponds to an individual movie and includes details about its content, creators, and audience ratings. 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 long-term usability which our datasets do align to these guidelines.

- **Findable**: Our primary dataset was sourced from Kaggle (Hieu2695, 2019) and the secondary from GitHub (Grijalva, 2019), both of which are widely recognized platforms that provide basic metadata and descriptive tagging, making the data easy to locate. To support findability, we also included clear instructions in the repository's *README* file to help users locate and understand the datasets.
- **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 are not guaranteed, we uploaded both original datasets as well as our cleaned and merged version to our own GitHub repository and made it public to ensure continued access and reproducibility.
- **Interoperable**: The datasets are provided in CSV format, a widely accepted and accessible format that enables easy use in 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 and transparency.

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.

Rating Distributions in the Five Most Common Genres

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.

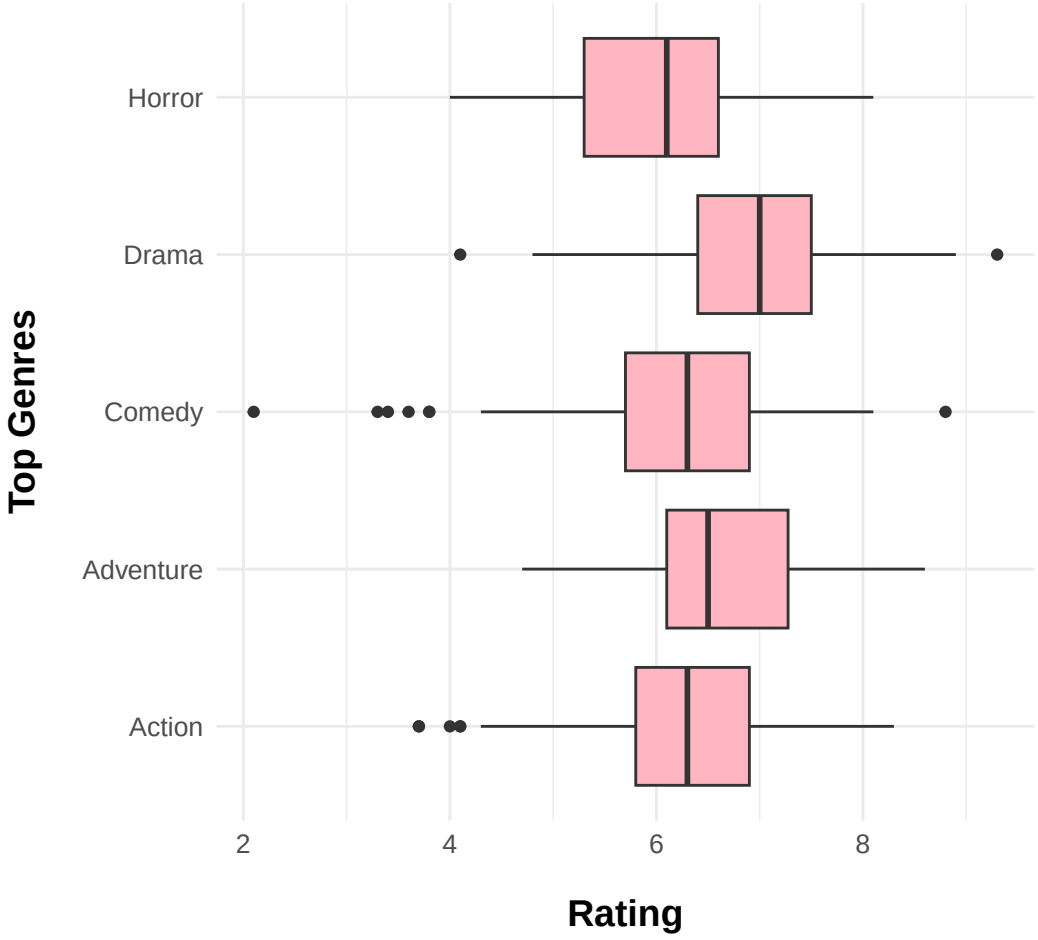
When looking at film counts, **Comedy** and **Drama** are the most represented, each with over **200 films**, while **Horror** has the lowest count with just **77 films**, which may contribute to its greater rating variability.

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



The box plot in Figure 1 shows the distribution of movie ratings across the top five genres and helps bring the numbers in Table 1 to life through a visual format. One thing that really stands out is how **Drama**'s entire box is shifted higher than the others, meaning that not only is its average rating high, but most of its movies consistently land in that higher range. It also has a **tighter**

interquartile range, meaning the ratings are tightly clustered, further reinforcing that Drama is the most consistent genre overall. In contrast, **Horror** displays a much wider spread, this suggests more variability in audience reception.

Comedy shows several low-rated *outliers* in the box plot, indicating that while many comedy films fall around the average, a few are rated significantly lower. According to the Summary Table 1, Comedy has a **minimum rating** of **2.1**, and the box plot confirms this value is a outlier with a point far below the lower whisker.

This highlights how a few poorly received films can drag down the overall average, even if most of the genre performs reasonably well. It's a reminder that summary statistics alone don't always tell the full story, without a visual, we wouldn't easily notice how rare or extreme those low ratings are. The visual helps us see that while these outliers exist, they are not representative of the genre as a whole.

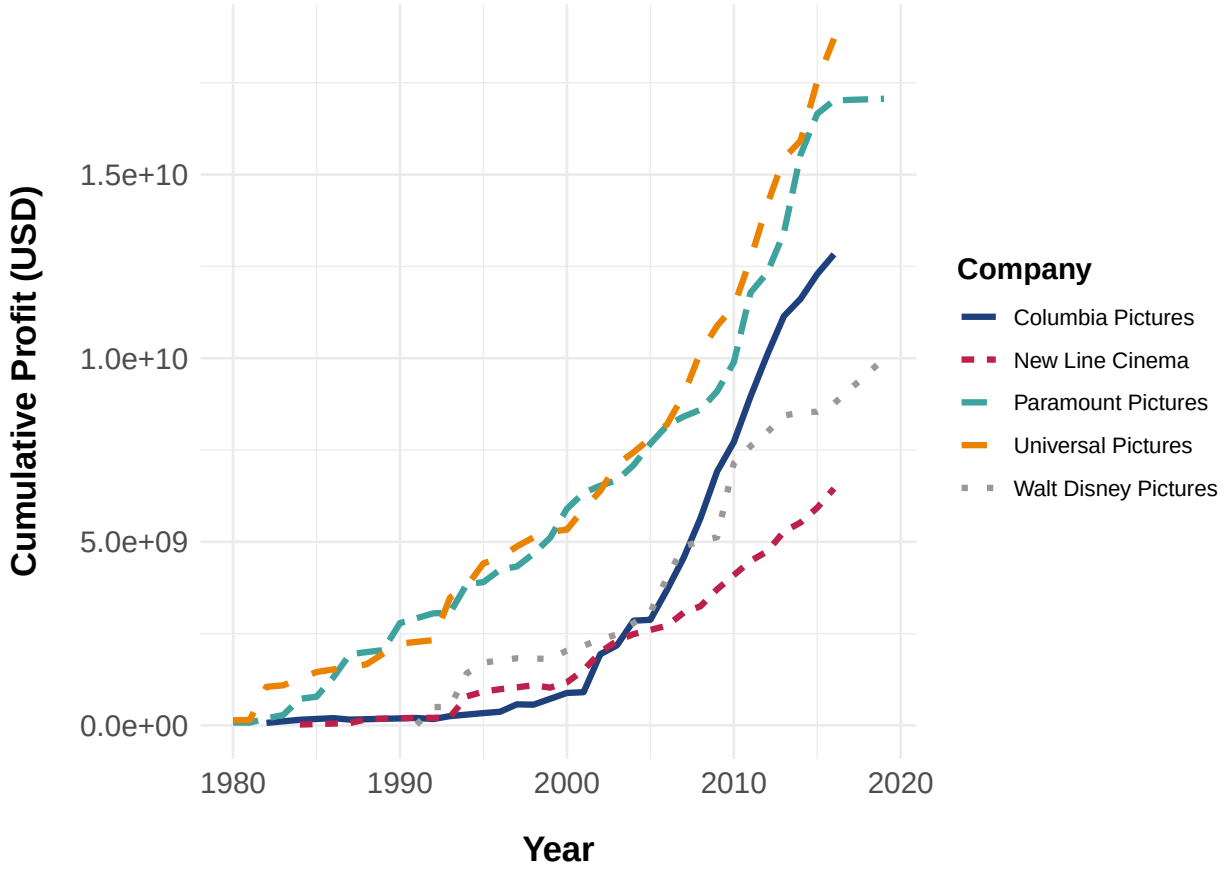
Company Profitability and Financial Patterns

In this section, we will explore which production studios have the best return on investment (ROI), by analyzing both the **budget** and the **profit** of the most active companies in our data, over time. We first identified the *top five production companies* based on the *number of films* they released. Focusing on these studios allowed us to base our comparisons on a solid number of films, making the patterns we observed more reliable.

Profit Line Graph

We began by calculating the average profit for each of the top 5 companies. To visualize this, we created a line graph (Figure 2) that shows the *Year* on the y-axis and its *cumulative profit* earnings in USD on the y-axis. Each line represents a different company, distinguished by color and line type, making it easy to compare their financial performance over time.

Figure 2: Cumulative Profit Over Time for Top 5 Production Companies



Overall, Figure 2 shows that all *five* studios experienced growth in profit over time, though at different rates. Notably, **Universal Pictures** and **Paramount Pictures** lead in cumulative profit, especially after 2000, where their lines show a sharp upward trend. **Universal Pictures**, in particular, reaches the highest total profit by the end of the time range with, 18 billion dollars in cumulative profit over the years. **Columbia Pictures** experiences substantial growth in later years, eventually overtaking **Walt Disney Pictures**, whose curve begins to grow less rapidly around 2015. This may suggest a decrease in the number of Disney films represented in our dataset during those years, rather than a decline in actual profitability. Meanwhile, **New Line Cinema** trails behind in cumulative earnings, showing it had a smaller impact at the box office overall.

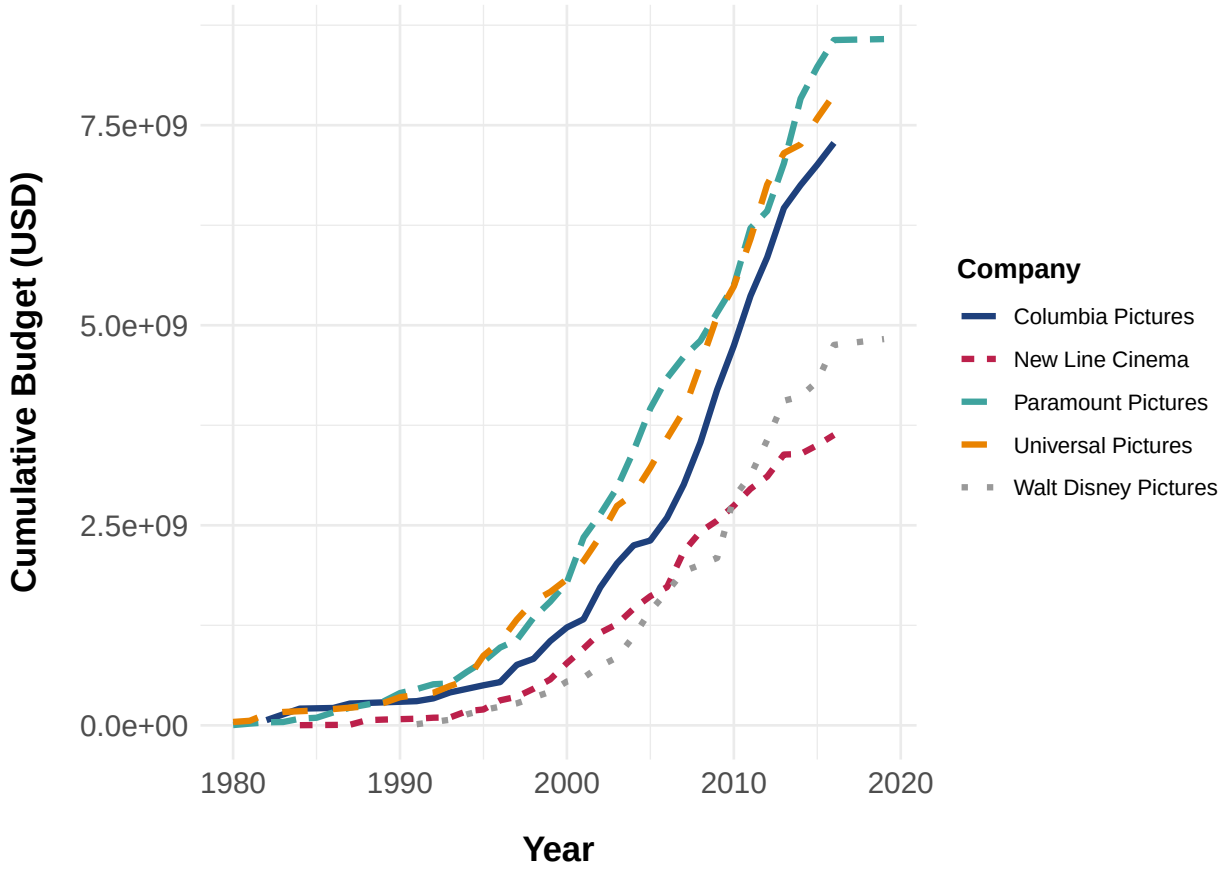
It's important to note that while these trends offer valuable insight, our dataset may not fully represent every studio's complete catalog. For example, Columbia Pictures probability stops after 2015, not because they stopped releasing but because our data doesn't include their movies after that year. That being said, this visualization still provides us with an idea of how studio profits have evolved and highlights which companies have consistently *delivered financial success over time.*

Budget Line Graph

Next, we created a similar line graph but changed the cumulative profit to cumulative budget to compare total spending by each company over the same period of time (Figure 3). Like the profit

graph, it shows cumulative amounts over time, helping us see which companies invest the most in production.

Figure 3: Cumulative Budget Over Time for Top 5 Production Companies



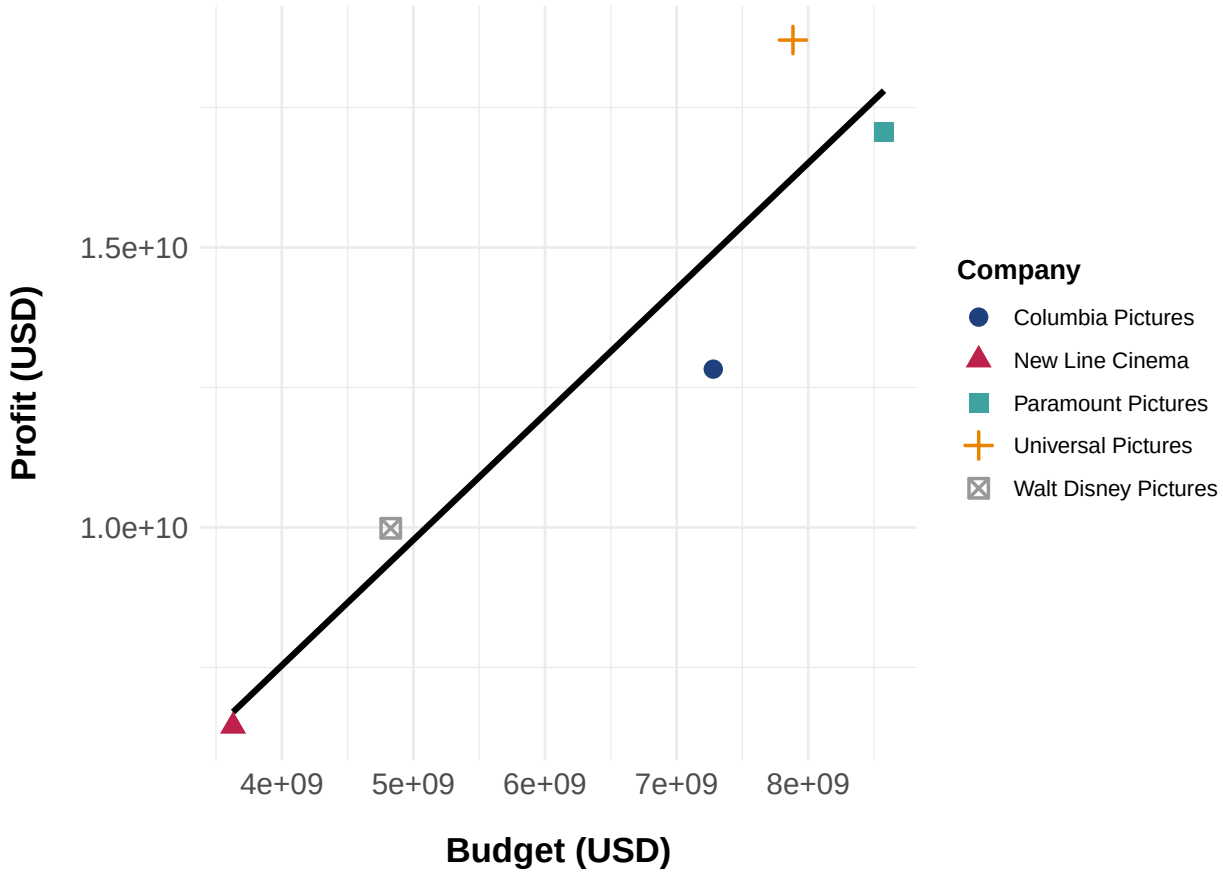
In Figure 3, we observe that just like in the profit line graph, **Paramount Pictures** and **Universal Pictures** had the highest cumulative budgets, reflecting on their large scale investments across their film portfolios. Where Paramount reached its highest commulative budget of about 8 billion dollars. Walt Disney Pictures and New Line Cinema, on the other hand, show more moderate cumulative spending, with noticeably slower growth in budget accumulation throughout the observed period, just like their profit.

An especially interesting case is **Columbia Pictures**. Despite having one of the highest cumulative budgets, its total profit in Figure 2 is only middle-ranking compared to the other top studios. This implies that Columbia's return on investment is relatively lower, spending more over time but not yielding profits at the same scale as studios like Universal or Paramount Pictures. In contrast, Universal appears more efficient, achieving strong cumulative profits with about the same slightly overall budget. This comparison highlights an important insight: *greater spending* does not always lead to proportionally *higher returns*.

Budget-Profit Relationship

Now, to examine the direct relationship between investment and return, we plotted each company's total budget against total profit in a scatter plot (Figure 4). Each point represents one of the top five companies, as seen from the legend on the right, and a black regression line shows the trend. The x-axis is the total budget in USD and the y-axis is the total profit also in USD.

Figure 4: Relationship Between Budget and Profit of the Top 5 Companies



As expected from our line graphs, Figure 4 reveals a positive linear relationship: studios that spend more generally earn more. However, the position of each studio relative to the trend line provides valuable insights. For example, **Universal Pictures** sits above the line, suggesting it outperforms expectations based on its spending. In contrast, **Columbia Pictures**, while having a high budget, falls below the line—indicating a lower return relative to its total investment, which we also noticed from Figure 2 and Figure 3.

Interpretation

Together, these three visuals (Figure 4, Figure 2 and Figure 3) provide a comprehensive view of company performances over time. The line graphs illustrate which studios consistently grew their profit and spending, while the scatter plot highlights differences in return of investments efficiency. **Universal and Paramount Pictures** emerge as *top performers* in both absolute profitability

and investment effectiveness. Meanwhile, Columbia, despite high spending, show slightly lower relative returns, and New Line Cinema operates with lower budgets and profits overall. This analysis highlights that although large budgets can contribute to high profits, a studio's return on investment is ultimately determined by how effectively those funds are used, through efficient operations and smart production decisions.

Star Impact on Movie Success

To explore our third research question, examining which actors appear most frequently in movies and whether their presence contributes to a film's success. We defined success of movies in two distinct ways to deepen our analysis. The first definition focuses on **financial success** (Figure 5), measured by a profit exceeding $50 million. The second centers on **critical success** (Figure 6), evaluated through audience ratings. By analyzing how often the top five most frequent stars appear in successful films under each definition, we were able to identify consistent patterns and better understand the impact of star power on movie performance.

Summary Table

This Summary table presents statistics for the five most frequently appearing stars in the dataset. The columns include: **MovieCount** (the number of films each star appeared in), **AvgRating** (the average audience rating for those films), **RatingSD** (the standard deviation of ratings), and **AvgProfit** (the average profit for the star's films). This table helps us begin exploring whether the most active stars tend to be in more successful films. In the following sections, we'll also visualize this data to examine whether their presence is linked to higher ratings or greater profitability.

Table 2: Summary Statistics for Top Stars

Star	MovieCount	AvgRating	RatingSD	AvgProfit
Nicolas Cage	13	6.07	1.24	94028034
Adam Sandler	12	5.89	1.01	146358425
Dwayne Johnson	11	5.99	0.38	110519841
Tom Cruise	11	6.97	0.41	322109979
Tom Hanks	11	7.33	0.95	313617993

Table 2 shows that **Nicolas Cage** appears in the most films, with a total of **13**, but also has the *highest variability* in ratings, with a *standard deviation* of **1.24**, suggesting mixed audience responses. In contrast, **Dwayne Johnson** is the most consistent performer, with a *low standard deviation* of **0.38**, though his *average rating* is relatively normal at **5.99**. **Tom Hanks** stands out with the **highest average rating** of **7.33**, indicating that his films are well-received by viewers, along with a *high average profit*. **Tom Cruise** offers the strongest balance, combining a *high average rating* of **6.97** with the *highest average profit* in the group.

While these numbers reveal interesting patterns, we cannot confidently conclude that a star's presence directly affects a film's success. There may be other underlying factors at play that

influence these outcomes. In fact, one possible influence could be **genre**. As seen in the box plot in Figure 1, **drama** films tend to receive the **highest ratings**, which could help explain *Tom Hanks*' strong performance. **Comedy** showed **greater variability** and several low-rated outliers, possibly contributing to *Adam Sandler's* lower ratings. **Action** films fall somewhere in the middle in terms of *rating consistency*, which may account for the **steadier results** seen with *Tom Cruise and Dwayne Johnson*. Meanwhile, *Nicolas Cage* has appeared across a wide range of genres, likely contributing to the **broader spread** in his ratings. However, further exploration will be needed to determine whether these patterns hold across a larger set of stars and whether other factors like genre play a more significant role.

Bar Chart

To dig deeper into financial success, we classified each movie as either "Successful" where the profit is more than $50 million or "Unsuccessful". Then created a bar chart showing the count of each category for the top five stars (Figure 5). The bars are color-coded: *teal* for successful films and *orange* for unsuccessful films.

Figure 5: Number of Successful Movies of Top 5 Most Frequent Stars

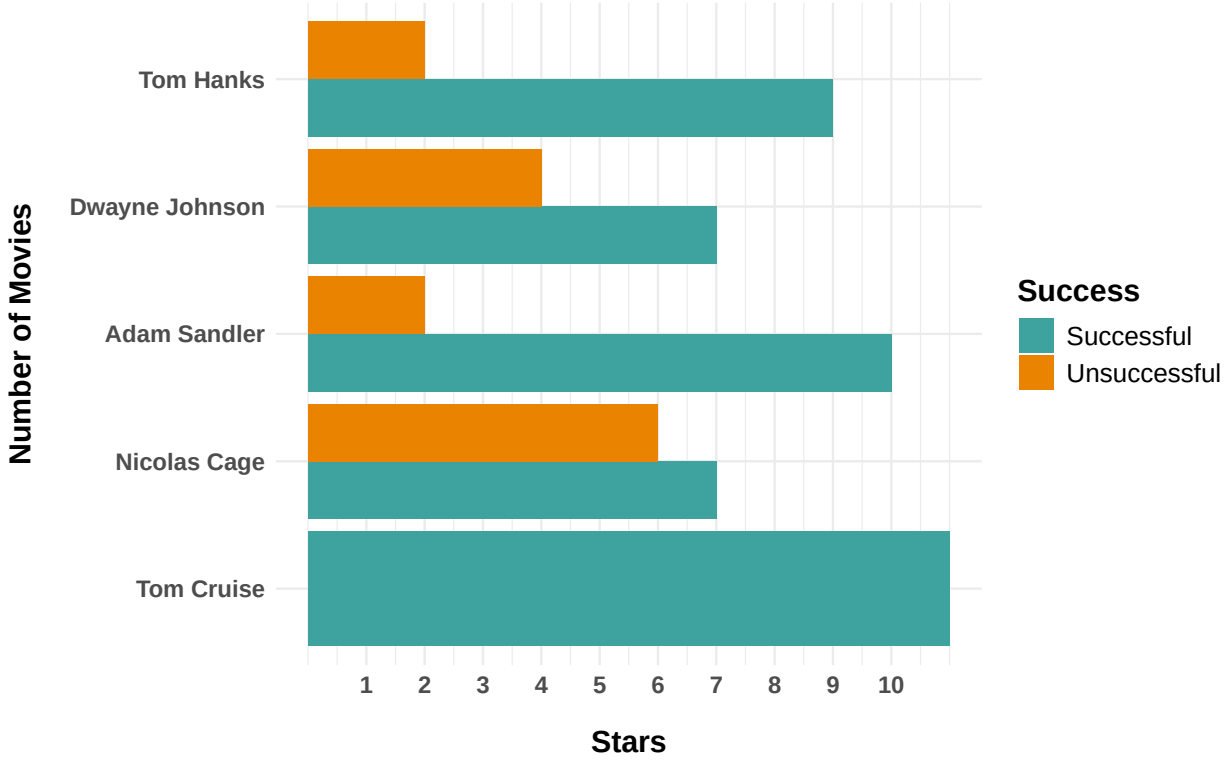


Figure 5 complements the earlier summary table (Table 2). We can see that all *13* of **Tom Cruise's** movies are considered successful under our definition, aligning with his position in the summary table, Table 2, as the star with the *highest average profit*. This consistency supports the idea that his presence in a film is a strong indicator of financial success for that film.

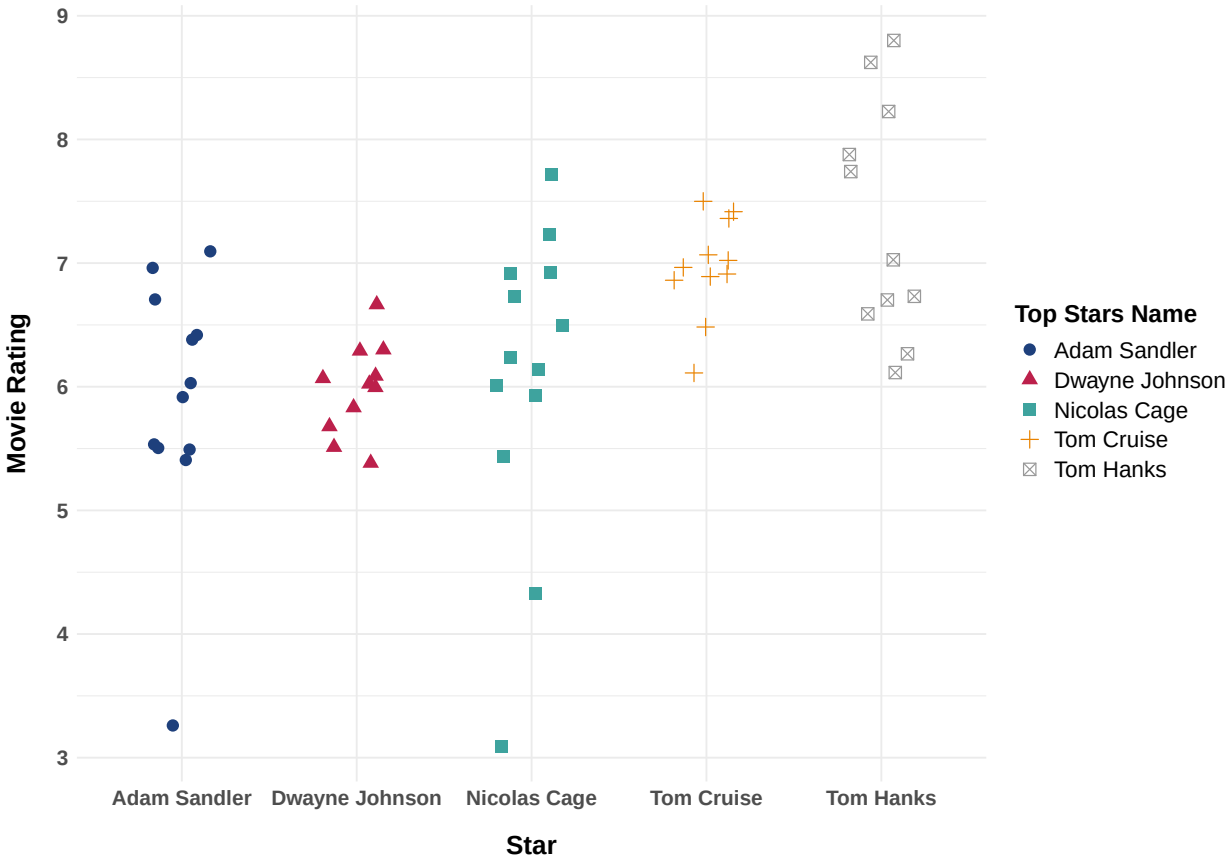
An interesting comparison arises between **Tom Hanks** and **Adam Sandler**, who both have two unsuccessful films, and Adam Sandler has just one more successful film (*10*) than Tom Hanks (*9*). Yet despite this, Table 2 shows that Tom Hanks has a significantly higher average profit, suggesting that the magnitude of financial success in Hanks's films is much greater—likely due to higher box office returns per film—even if he has fewer successes by count. On the other hand, **Nicolas Cage**, who has the *lowest average profit*, also appears in the highest number of unsuccessful films (*6*), compared to *7* successful ones. This aligns with his broader performance variability. **Dwayne Johnson** sits in the middle, with a moderate average profit and a fairly balanced split: *7* successful and *4* unsuccessful films.

Together, Table 2 and Figure 5 provide a clearer understanding of the consistency of each star's success, not just how many movies they've made, but how often those movies perform well financially.

Scatter Plot

To better understand how audiences respond to the films of the most frequently featured stars, we created a scatterplot that maps the individual movie ratings for each of the top five actors (Figure 6). Each point represents a single film, with its vertical position showing the audience rating and its color and shape identifying the star. This visualization allows us to look beyond averages and see the full range of how each star's movies are received, helping us explore whether certain actors consistently appear in higher-rated films or if their performance doesn't have much of an impact.

Figure 6: Movie Ratings of Top 5 Most Frequent Stars



In Figure 6, *Tom Hanks* clearly stands out, with all of his movies receiving **ratings above 6.0**, and several reaching as high as 8.5. This pattern reflects what we saw in the Summary Table 2, where he had the **highest average rating** overall. The cluster of tightly grouped, high ratings also suggests that his performances tend to be well-received.

In contrast, *Nicolas Cage*'s points are much more scattered, ranging from just over **3 to above 7**. This **wide spread** supports the **high standard deviation** shown in the Table 2 and illustrates the inconsistency in how his movies are rated by audiences. While he's been in some well-rated movies, he's also been in several that scored poorly.

Adam Sandler's ratings mostly fall in the mid range, with one movie below 4. This matches his **lower average rating** from Table 2 and shows that even though he appears in many films, they don't always perform well with audiences.

Tom Cruise has no ratings below 6, and most of his films land in a narrow range between **6.5 and 7.5**. This confirms what we saw earlier, he not only scores well but does so consistently, which could be part of why his movies tend to be financially successful too. Similarly, *Dwayne Johnson*'s ratings are mostly **clustered around 6** with little spread, reflecting consistent but average performance.

Conclusion

To wrap up, our project investigated three main questions about what makes a movie successful. First, we examined how audience ratings differ across genres using Table 1 and Figure 1, and found that genres vary widely in performance. **Drama** had the *highest* and *most consistent ratings*, while **Horror** showed more *variability* and *lower scores*, with **Comedy** showing notable *outliers*. Second, we explored the relationship between production budgets and profits through Figure 2 (Profit by Company), Figure 3 (Budget by Company), and the Budget vs. Profit Scatter Plot (Figure 4). Together, these visuals suggest a *positive relationship between budget and profit*, where **higher investment** is often linked to **higher returns**. Lastly, in analyzing the impact of movie stars using Table 2, Figure 5 (Success Count), and Figure 6 (Ratings Scatterplot), we found that frequently appearing stars do not necessarily guarantee higher-rated or more profitable movies. While some stars like **Tom Cruise** and **Tom Hanks** show more consistent success, overall, our analysis **did not find strong evidence** that star presence alone drives movie success, more investigation would be needed to draw a firm conclusion. Overall, exploring these patterns helps deepen our understanding of what drives both critical and commercial success in the movie industry.

Sources and References

Main Dataset:

Hieu2695. (2019, November 5). Movie Industry [Data set]. GitHub. Retrieved May 6, 2025, from https://github.com/hieu2695/Movie-Industry/blob/master/Movie.csv

Supplemantry Dataset:

Grijalva, D. (2019, October). Movie Industry [Data set]. Retrieved May 6, 2025, from https://www.kaggle.com/datasets/danielgrijalvas/movies

Old Dataset:

Deepanshu Chhikara. (2024). IMDb Top250 Movies [Data set]. Kaggle.

https://doi.org/10.34740/KAGGLE/DSV/7990386

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 ----
mainChunk <- "https://raw.githubusercontent.com/Stat184-Spring2025/"
ExtraChunk <- "Sec4_FP_Layan_Sara/main/Data/Budget_Revnue.csv"
url1 <- paste0(mainChunk,ExtraChunk)
MoviesSubRaw <- read.csv(url1, header = TRUE)
ExtraChunk2 <- "Sec4_FP_Layan_Sara/main/Data/Movie.csv"
url2 <- paste0(mainChunk,ExtraChunk2)
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
#)
```

```r
# 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 Box Plot 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
```

```r
# 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(
  y = "Top Genres",   #labels the x axis
  x = "Rating"      #labels the y axis
) +
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
)

# Creating visualizations for 2nd Question ----

# Get top 5 companies by number of movies ----
TopCompanies5 <- MoviesJoined %>%
  count(Company, sort = TRUE) %>% # Count number of movies per company
  slice_max(n, n = 5) %>%  # Select the top 5 companies
  pull(Company)

# Create the Profit Line Graph -----
MoviesJoined %>%
  filter(Company %in% TopCompanies5) %>%  # Filter data to get top 5 companies
  group_by(Company, Year) %>%       # Group by company and year
  summarize(YearlyProfit = sum(Profit, na.rm = TRUE), .groups = "drop") %>%

  arrange(Company, Year) %>% # Sorts data so cumulative sum works correctly
  group_by(Company) %>%
  mutate(CumulativeProfit = cumsum(as.numeric(YearlyProfit))) %>%
  # Gets cumulative sum so line graph is smooth
  ggplot(mapping = aes(x = Year,            # Create the line chart
                       y = CumulativeProfit,
```

```
                        color = Company,
                        linetype = Company
                        )
          ) +
  geom_line(size = 1.2) +    # Draw the lines with thicker width
  scale_color_manual(values = psuPalette) +  # Apply custom colors
  labs(           #Add labels and formatting
    x = "Year",
    y = "Cumulative Profit (USD)",
    color = "Company",
    linetype = "Company"
  ) +
  theme_minimal() +
  theme(
    legend.title = element_text(face = "bold", size = 12),
    axis.title = element_text(size = 14, face = "bold"),   # Style axis titles
    axis.text = element_text(size = 12), # Adjust axis text size
    axis.title.x = element_text(margin = margin(t = 15)
                                ),
  axis.title.y = element_text(
    margin = margin(r = 15)
                                )
  )

# Calculate Max Profit  ----
CumulativeProfitData <- MoviesJoined %>%
  filter(Company %in% TopCompanies5) %>%  # Filter for the top 5 companies
  group_by(Company, Year) %>%   # Group by company and year
  summarize(YearlyProfit = sum(Profit, na.rm = TRUE), .groups = "drop") %>%
  arrange(Company, Year) %>%
  group_by(Company) %>%
  mutate(CumulativeProfit = cumsum(YearlyProfit)) # Compute cumulative profit per company

# Calculate Max Budget ----
CumulativeBudgetData <- MoviesJoined %>%
  filter(Company %in% TopCompanies5) %>%
  group_by(Company, Year) %>%
  summarize(YearlyBudget = sum(Budget, na.rm = TRUE), .groups = "drop") %>%
  arrange(Company, Year) %>%
  group_by(Company) %>%
  mutate(CumulativeBudget = cumsum(as.numeric(YearlyBudget))
         ) # Compute cumulative budget per company


# Create the Budget Line Graph -----
MoviesJoined %>%
  filter(Company %in% TopCompanies5) %>%
```

```
  group_by(Company, Year) %>%
  summarize(YearlyBudget = sum(Budget, na.rm = TRUE), .groups = "drop") %>%
  arrange(Company, Year) %>%
  group_by(Company) %>%
  #Get cumulative sum so that line graph is smooth
  mutate(CumulativeBudget = cumsum(as.numeric(YearlyBudget))) %>%
  ggplot(           # Create the line chart
    mapping = aes(
      x = Year,
      y = CumulativeBudget,
      color = Company,
      linetype = Company)
    ) +
  geom_line(size = 1.2) + # Draw the lines with thicker width
  scale_color_manual(values = psuPalette) +  # Apply custom colors
  labs(             # Add labels
    x = "Year",
    y = "Cumulative Budget (USD)",
    color = "Company",
    linetype = "Company"
  ) +
  theme_minimal() +
  theme(                  # Customizing and formatting
    legend.title = element_text(face = "bold"),
    axis.title = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    axis.title.x = element_text(margin = margin(t = 15)
                                ),
  axis.title.y = element_text(
    margin = margin(r = 15)
                                )
  )

# Create the budget-profit Scatter Plot -----
# Get the Total profit and budget ----
CompanySummary <- MoviesJoined %>%
  filter(Company %in% TopCompanies5) %>%
  group_by(Company)%>%
  summarize(TotalProfit = sum(Profit, na.rm = TRUE),
            TotalBudget = sum(Budget, na.rm = TRUE))

# Create and map the plots ----
CompanySummary %>%
  ggplot(mapping = aes(x = TotalBudget,
              y = TotalProfit,
              color = Company,
              shape = Company
```

```
            )
        ) +
  geom_point(size = 3, stroke = 1) +
  geom_smooth(aes(group = 1), method = "lm", se = FALSE, color = "black", size = 1.2) +
  scale_color_manual(values = psuPalette) + #Same palette as the lines
  labs(
    x = "Budget (USD)",
    y = "Profit (USD)",
    color = "Company",
    shape = "Company"
  ) +
# Customize axes and legend title to be more accessible ----
  theme_minimal() +
  theme(
    legend.title = element_text(face = "bold"),
    axis.title = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    axis.title.x = element_text(margin = margin(t = 15)),
    axis.title.y = element_text( margin = margin(r = 15))
  )

# Wrangling Star Summary Table Data ----
star_summary <- MoviesJoined %>%
  group_by(Star) %>%  # Group data by each unique star
  summarise(
    Movie_Count = n(),    # Count number of movies each star appeared in
    Avg_Rating = mean(Rating, na.rm = TRUE), # Calculate the average rating
    SD_Rating = sd(Rating, na.rm = TRUE), # Standard deviation of ratings
    Avg_Profit = mean(Profit, na.rm = TRUE) # Average profit per movie
  ) %>%
  arrange(desc(Movie_Count))%>%      # Sort stars by number of movies
  slice_head(n=5)%>%    # Keeps only the top 5 most frequent stars
  mutate(across(where(is.numeric), round, 2)) # Rounds to 2 decimal places

# Displaying the Summary Table ----
star_summary %>%
  kable(booktabs = TRUE,
        col.names = c("Star", "MovieCount", "AvgRating",
                      "RatingSD", "AvgProfit"), # Names columns
        align = c("l", rep("c", 4)), # Align first column left, others center
        format = "latex" # Output the table in LaTeX format (for PDF rendering)
  )%>%
  kableExtra::kable_styling(
    font_size = 16,            # Sets font size of the table
    latex_options = c("striped","scale_down") # Apply striped rows
  )%>%
  row_spec(0, bold = TRUE)  # Bold the header row
```

```r
# Wrangling ScatterPlot of the Rating for Top Stars ----

# Identify Top 5 Stars by number of movies
top5_stars <- MoviesJoined %>%
  count(Star, sort = TRUE) %>%  # Count how many times each star appears
  slice_head(n = 5) %>%  # Keep the top 5 most frequent stars
  pull(Star)

# Filter movies for those stars
top_star_movies <- MoviesJoined %>%
  filter(Star %in% top5_stars)

# Create Bar Chart of Successful movies -----
MoviesJoined %>%
  # Define Success in movies as Profit > $50M
  mutate(Success = ifelse(Profit > 50000000, "Successful", "Unsuccessful")) %>%
  # Take the Top 5 Stars
  filter(Star %in% top5_stars) %>%
  # Get their counts
  count(Star, Success) %>%
  ggplot(aes(x = n,
             y = reorder(Star, -n),
             fill = Success)) +
  geom_bar(stat = "identity",
           position = "dodge") + # Set the bars side to side
  scale_x_continuous(breaks = 1:10) + # Number the x-axis
  scale_fill_manual(values = c("Successful" = "#3EA39E",
                               "Unsuccessful" = "#E98300")) +
  labs(
    x = "Stars",
    y = "Number of Movies") +
  theme_minimal() +
  theme(
    legend.title = element_text(face = "bold", size = 14),
    legend.text = element_text(size = 12),
    axis.title = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 11, face = "bold"),
    axis.title.x = element_text(margin = margin(t = 15)),
    axis.title.y = element_text( margin = margin(r = 15))
  )



# Plot the Scatterplot of Stars/Rating ----

ggplot(
  data = top_star_movies, # Use the filtered dataset
```

```
  mapping = aes(x = Star,    # X-axis shows the star's name
                y = Rating,   # Y-axis shows the movie rating
                color = Star,   # Color points by star for distinction
                shape = Star # Shape for star distinction
                )
  ) +
  geom_jitter(width = 0.2, size = 3) +
  scale_color_manual(values = psuPalette) +  # Apply custom colors
  labs(
    x = "Star",   # X-axis label
    y = "Movie Rating",  # Y-axis label
    color = "Top Stars Name",
    shape = "Top Stars Name"
  ) +
  theme_minimal() +     # Uses a clean, minimal theme
  theme(
    legend.position = "right",  # Place the legend on the right
    legend.title = element_text(size = 14, face ="bold"),  # Set legend text size and bolds
    legend.text = element_text(size = 13),
    axis.title.x = element_text(face = "bold",   # Make the x-axis title bold
                                size = 15,    # Set font size to 14
                                margin = margin(t = 15)
                                ),
    axis.title.y = element_text(face = "bold", # Customize y-axis title
                                size = 15,
                                margin = margin(r = 15)
                                ),
    axis.text = element_text(size = 12, face = "bold")
  ) # margin pushes title away from axis
```