Mini Project Report on TMDB 5000 Dataset

TAMILARASAN

i. Introduction

The global film industry is a powerful blend of art, culture, and commerce, with significant economic influence. The TMDB 5000 Movies Dataset contains metadata for 5,000 films released between 1960 and 2017, covering essential details like titles, release dates, genres, budgets, revenues, cast, crew, and user ratings. This dataset offers a unique opportunity to explore trends in cinema, such as the profitability of different genres, the role of directors and production companies, and the evolution of audience preferences over time. By analyzing this data, we can uncover key factors that influence a film's financial success, critical reception, and cultural impact. This report leverages the dataset to answer key analytical questions through data preprocessing, statistical testing, and visualization.

ii. Description of the Data

The TMDB 5000 Movies dataset contains detailed information on approximately 5,000 films across two CSV files:

- tmdb_5000_movies.csv: Contains metadata about each film, including title, score, release date, genre, revenue, budget, language, and production details.
- tmdb_5000_credits.csv: Includes cast and crew information, specifying roles and gender for each member.

The dataset consists of 24 variables, which are categorized into 17 qualitative and 7 quantitative attributes.

Qualitative Variables

These are categorical and non-numeric descriptors. Examples from the movies.csv file include:

- Film title and ID
- Genre (e.g., Sci-Fi, Comedy, Horror)
- Language of the film

- Production status (Released or Post-production)
- Production companies and countries
- Tagline, homepage, and keywords

The credits.csv file adds:

- Cast details: actor names and gender (1 for male, 2 for female, 0 if undefined)
- Crew details: name, gender, role, and department

Quantitative Variables

These numeric values are only found in the movies.csv file and are further split into:

- Discrete Variables: Fixed numeric values such as number of votes, release year, and film duration.
- Continuous Variables: Metrics like revenue, budget, popularity score, and average rating, which can take on a wide range of values.

iii. Data Cleaning

Before analysis, several preprocessing steps were necessary to prepare the dataset:

- **Parsing JSON Strings:** Many columns (e.g., genres, keywords, production compa- nies, cast, crew) store data in JSON format. These were parsed and converted into usable string formats using the json.loads() function.
- **Handling Null Values:** Several columns contained missing data. Fields like homepage and tagline were removed due to excessive null values and low analytical relevance. Columns such as runtime, overview, and release date had only a few nulls. For these:
  - Missing runtimes were replaced with the average duration.
  - Missing overviews were labeled as "unspecified."
  - Films with missing release dates were excluded.
- **Zero Values:** Variables like budget and revenue sometimes had values of zero, which can skew results. All entries with a budget or revenue under $1,000 were dropped. Zeroes in other metrics (e.g., vote count, rating) were also discarded.
- **Duplicate Titles:** To avoid ambiguity, duplicate movie titles were made unique by appending the year of release.
- **Date Formatting:** The release date column was split into year, month, and day to facilitate temporal analysis.

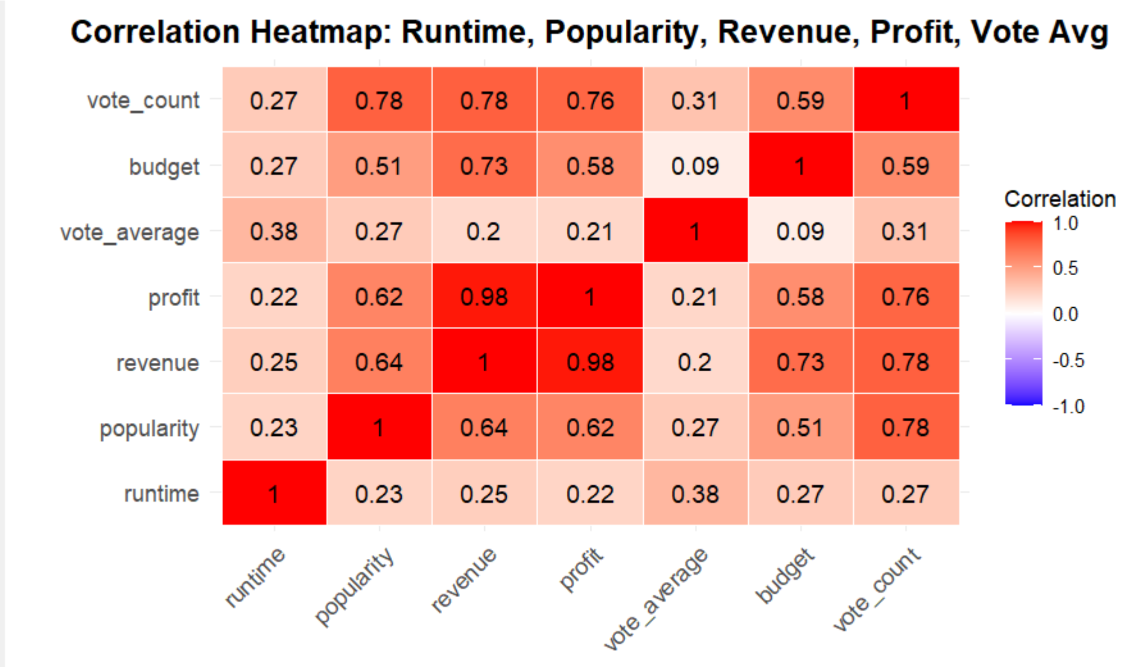
After cleaning, the dataset was refined and ready for deeper statistical exploration.

Q1 :Exploring the Correlation, Relationship Between Budget, Popularity, and Movie Revenue

I aimed to explore the relationships between key movie attributes such as budget, revenue, popularity, runtime, and voting metrics. By understanding how these variables interact, I hoped to uncover patterns that might explain financial success in films.

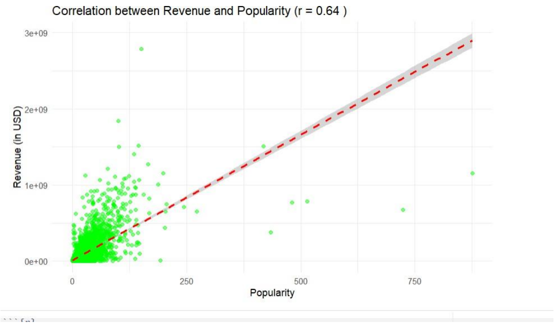
Methodology: I computed the Pearson correlation coefficients among all selected variables. To visually communicate these relationships, I plotted a correlation heatmap. The heatmap used a diverging color scale from blue (negative correlation) to red (positive correlation), with annotations to display exact correlation values.Also, Pearson's correlation was used to find correlation between Revenue and Popularity and between Budget and Revenue. The results were plotted in scatter plots.

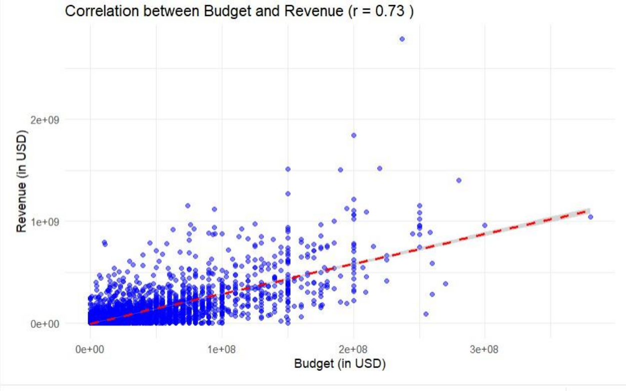
Visualization:



Correlation Heatmap: Runtime, Popularity, Revenue, Profit, Vote Avg

	runtime	popularity	revenue	profit	vote_average	budget	vote_count
vote_count	0.27	0.78	0.78	0.76	0.31	0.59	1
budget	0.27	0.51	0.73	0.58	0.09	1	0.59
vote_average	0.38	0.27	0.2	0.21	1	0.09	0.31
profit	0.22	0.62	0.98	1	0.21	0.58	0.76
revenue	0.25	0.64	1	0.98	0.2	0.73	0.78
popularity	0.23	1	0.64	0.62	0.27	0.51	0.78
runtime	1	0.23	0.25	0.22	0.38	0.27	0.27





Inference

Based on the correlation heatmap and scatter plots, we can infer that a movie's budget, popularity, and vote count are strong indicators of its financial success, particularly in terms of revenue and profit. Higher-budget movies tend to bring in more revenue, likely due to better production quality, marketing, and distribution. Popularity and vote count, which reflect how much attention a movie receives from audiences, also show a strong positive relationship with revenue and profit. This suggests that movies that are widely talked about or viewed are more likely to perform well at the box office. In contrast, vote average and runtime have weak correlations with financial outcomes, indicating that critical reception and movie length play a much smaller role in determining profitability. Overall, visibility and investment appear to be more influential than reviews or duration when it comes to a movie's commercial success.

Q2: Testing Profit Differences Across Companies

Hypotheses

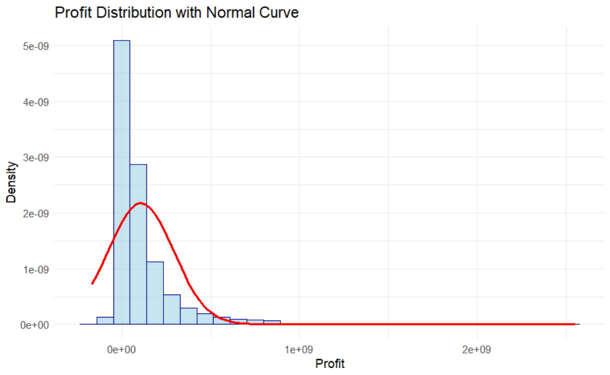
- **Null Hypothesis** (H_0): There is no difference in the distribution of profits among the production companies.
- **Alternative Hypothesis** (H_a):At least one production company's profit distribution differs from the others.

Assumption Checking

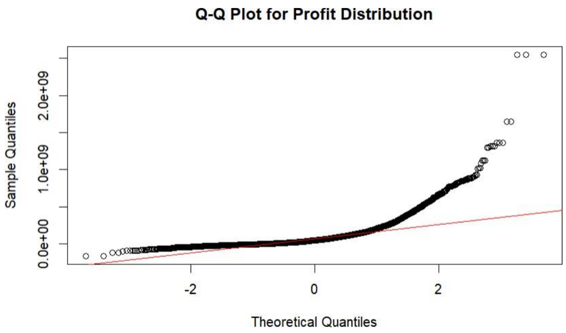
Before choosing a statistical test, I first checked whether the distribution of profits met the assumptions required for parametric testing. Specifically:

- I plotted a histogram of the profit variable, which indicated a right-skewed distribution.
- I also generated a Q-Q plot to assess normality and observed significant deviation from the normal line.

Since the assumption of normality was clearly violated, I decided to opt for a non-parametric test rather than the traditional one-way ANOVA.



(a) Profit Distribution with Normal curve



(b) Q-Q Plot of Profit

Statistical Method

I chose the **Kruskal-Wallis test**, a non-parametric alternative to one-way ANOVA, as it does not assume normality and is suitable for comparing the medians of more than two independent groups.

Test Result

The Kruskal-Wallis rank sum test yielded the following result:

$$\text{Kruskal-Wallis chi-squared: } \chi^2 = 740.76, \; df = 146, \quad p\text{-value} < 2.2 \times 10^{-16}$$

Inference

Since the p-value is significantly less than common significance levels such as 0.05 or 0.01, I rejected the null hypothesis H_0. **Therefore, I concluded that there is strong statistical evidence to suggest that not all production companies have the same profit distribution.** At least one company's profit distribution is significantly different from the others.

Q3 :Which production company generates the highest profit?

My goal in this part of the analysis was to determine which production company consistently generates the highest profit from its films. Understanding this can offer valuable insights into which companies are most effective in managing budgets and producing commercially successful movies.

Approach and Methodology: I began by cleaning the dataset to ensure that only movies with valid, non-zero budget and revenue values were included. From this, I computed the profit for each film using:

$$\text{Profit} = \text{Revenue} - \text{Budget}$$

Each movie in the dataset could be associated with multiple production companies. To accurately account for this, I expanded the nested JSON structure so that every production company involved in a film was given its own row linked to that film's profit.

To compare companies fairly, I grouped the data by production company and calculated the average profit across all their films:

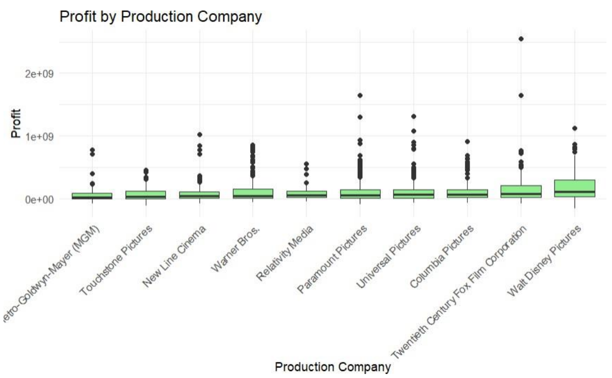
$$\text{Average Profit}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\text{Revenue}_j - \text{Budget}_j)$$

where n_i is the number of movies credited to company i.

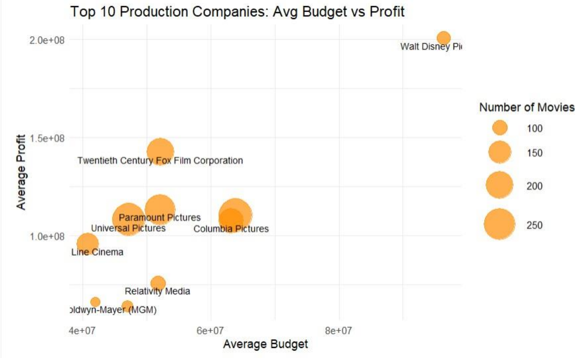
Box Plot Analysis: To visualize how profits are distributed across companies, I created a box plot for the top 10 production companies. This helped me identify variations in median profit, interquartile ranges, and the presence of outliers. The box plot was particularly useful in showing that some companies have a more consistent profit margin, while others exhibit high variability due to blockbusters or flops.

Why This Method is used: I chose to use the average profit rather than total profit to avoid bias toward companies that simply produce more movies. This allows for a more meaningful comparison of financial efficiency. The method is descriptive in nature and doesn't assume normality or require complex statistical assumptions — it's ideal for ranking performance.

Visual Representation: To bring this comparison to life, I created a bubble chart showing the average budget versus average profit for the top 10 production companies. Each bubble's size reflected the number of films produced, allowing for a quick visual assessment of scale and profitability.



(a) Profit of Production Company



(b) Profit of Production Company based on number of produced

Inference

This aligns with the insights drawn from the boxplot, where both the spread (interquartile range) and central tendency (median) of profits differ significantly across production companies. For instance, companies like *Walt Disney Pictures* show higher median profits and greater variability, while others like *Metro-Goldwyn-Mayer (MGM)* display a narrower profit range.

Additionally, the bubble plot further supports these findings by visually emphasizing the magnitude and concentration of profits among companies. Larger bubbles for certain companies indicate a stronger impact or dominance in terms of profit, reinforcing the result that profit distributions are not uniform across production houses.

Q4: Analysis of Annual Movie Counts

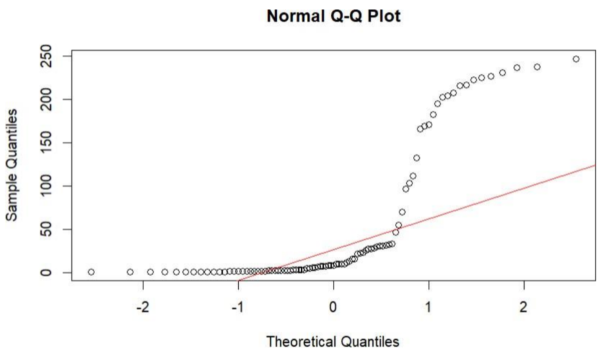
In this analysis, I explore various statistical aspects of movie release counts over time, with a focus on normality, variance across different periods, and potential trends in movie production.

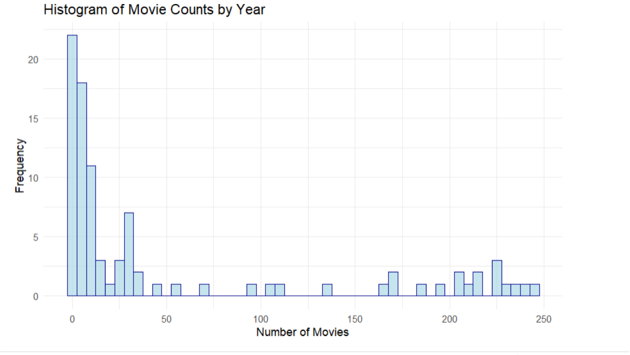
Q4.1: Test for Normality of Annual Movie Counts

To determine whether the number of movies released each year follows a normal distribution, I applied the **Shapiro-Wilk test**. This test evaluates the null hypothesis that the sample is drawn from a normal distribution.

Hypotheses

- H_0: The annual movie counts are normally distributed.
- H_1: The annual movie counts are not normally distributed.





The **Shapiro-Wilk test** is highly effective for small to moderate sample sizes and is sensitive to deviations from normality, including skewness and tailedness. Given the time-series nature of the data, which often exhibits skewed distributions, the Shapiro-Wilk test provides a robust assessment of the normality assumption. I also used Q-Q plots and histograms to corroborate the results visually.

Test Result

$W = 0.65968, \quad \text{p-value} = 3.278 \times 10^{-13}$

Since the **p-value** is significantly less than 0.05, I rejected the null hypothesis. This provides strong evidence that the distribution of annual movie counts significantly deviates from normality. Therefore, the assumption of normality is violated.

Inference

Given the extremely low p-value, I can confidently conclude that the annual movie counts do not follow a normal distribution.

—

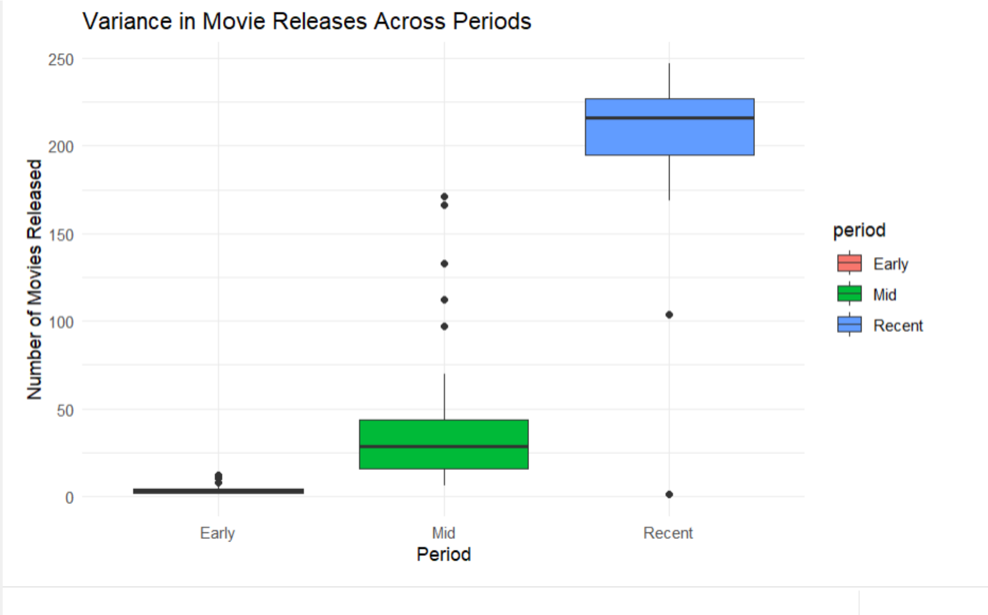
Q4.2: Variance Across Time Periods

Next, I compared the variability in movie release frequency across three historical periods: **Early**, **Mid**, and **Recent**. To assess the homogeneity of variances, I applied **Levene's Test**, which is particularly robust when normality assumptions are violated.

Since the data did not meet normality (as shown earlier), Levene's Test is a more appropriate method for assessing the equality of variances across groups defined by non-numeric categories like time periods.

Hypotheses

- H_0: The variances in movie release counts are equal across the three periods.
- H_1: At least one period has a significantly different variance.



The result of **Levene's Test** for homogeneity of variance yielded a test statistic of $F = 8.5353$ with a corresponding **p-value of 0.0004124**. Since the p-value is well below the conventional significance level of 0.05, I rejected the null hypothesis.

Inference

This result provides strong evidence that the variability in the number of movies released per year differs significantly across the three time periods. In other words, at least one of the periods has a substantially different spread in movie counts compared to the others.

—

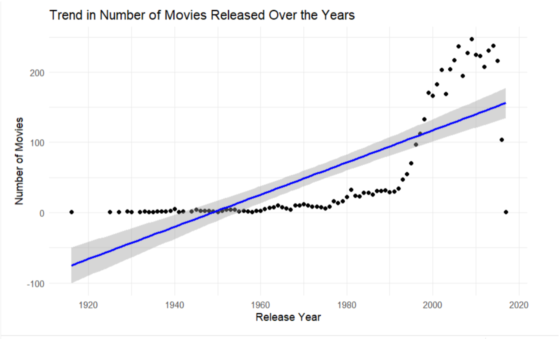
Q4.3: Test for Linear Trend Over Time

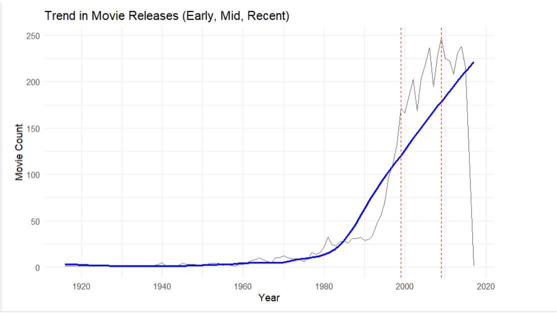
To explore whether the number of movie releases has changed significantly over time, I fitted a **linear regression model**, with the release year as the predictor and the annual movie count as the response variable.

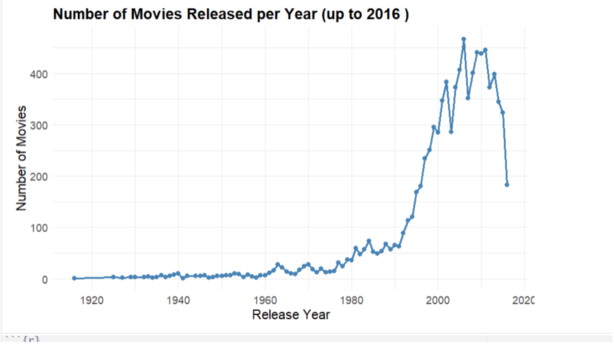
Hypotheses

- H_0: There is no linear trend in the number of movies released per year.
- H_1: There is a linear trend in the number of movies released per year.

Linear regression is appropriate when assessing whether a continuous response variable, like movie count, exhibits a trend over a continuous predictor, such as time. This model estimates the strength and significance of the trend, helping determine whether movie pro- duction has increased or decreased over the years.







The estimated regression coefficient for release_year is **2.2885**, with a **p-value** $< 2 \times 10^{-16}$. This indicates that, on average, the number of movies released increases by approximately **2.29** movies per year. The extremely small p-value suggests that this relationship is highly statistically significant and unlikely to have occurred by chance.

The **coefficient of determination** (R^2) is **0.582**, meaning that approximately **58.2%** of the variability in the number of movies released per year can be explained by the model with release year as the predictor.

The **adjusted R²** is **0.577**, which slightly adjusts for the number of predictors (in this case, only one) and provides a more realistic estimate of model fit. This still suggests a moderately strong relationship between release year and movie count.

The **F-statistic** is **122.5**, with a **p-value** $< 2.2 \times 10^{-16}$ confirming that the overall regression model is statistically significant.

Inference

The **positive linear trend** is highly statistically significant, and I can conclude that the number of movies released per year has **increased significantly** over time. This suggests a growing trend in movie production, with an average annual increase of about **2.29 movies**.

The Shapiro-Wilk test assessed normality in small samples, linear regression revealed trends over time, and Levene's Test examined variance homogeneity without assuming normality. Together, these methods provided a comprehensive understanding of how movie production patterns have evolved over the years.

Q5: Comparison Between Genre Groups

Hypotheses

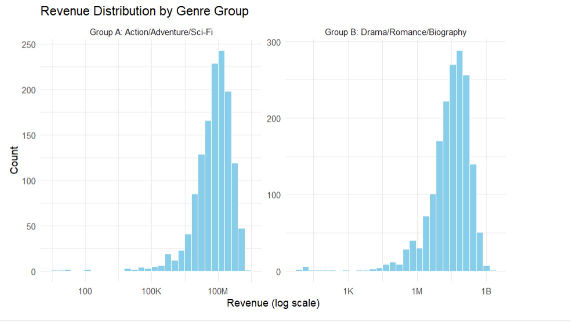
In this analysis, I am testing whether the distribution of movie revenues differs between two genre groups: Group A (Action/Adventure/Science Fiction) and Group B (Drama/Ro- mance/Biography).

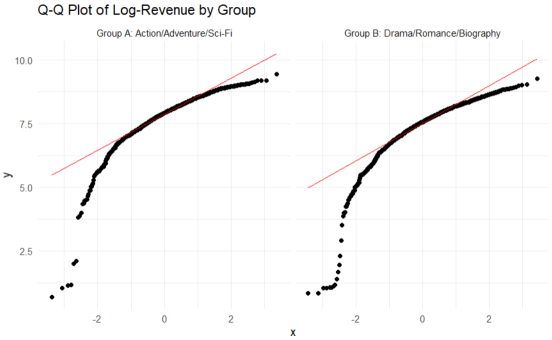
- **Null Hypothesis** (H_0): The distributions of movie revenues are identical for Group A (Action/Adventure/Science Fiction) and Group B (Drama/Romance/Biography).
- **Alternative Hypothesis** (H_a): The distribution of movie revenues differs between the two groups (i.e., the true location shift is not equal to zero).

Table 1: Normality Assessment using Shapiro-Wilk Test

Genre Group	Shapiro-Wilk p-value	Normality Condition
Group A: Action/Adventure/Sci-Fi	1.05×10^{-34}	Not Normally Distributed
Group B: Drama/Romance/Biography	2.46×10^{-39}	Not Normally Distributed

Test Results



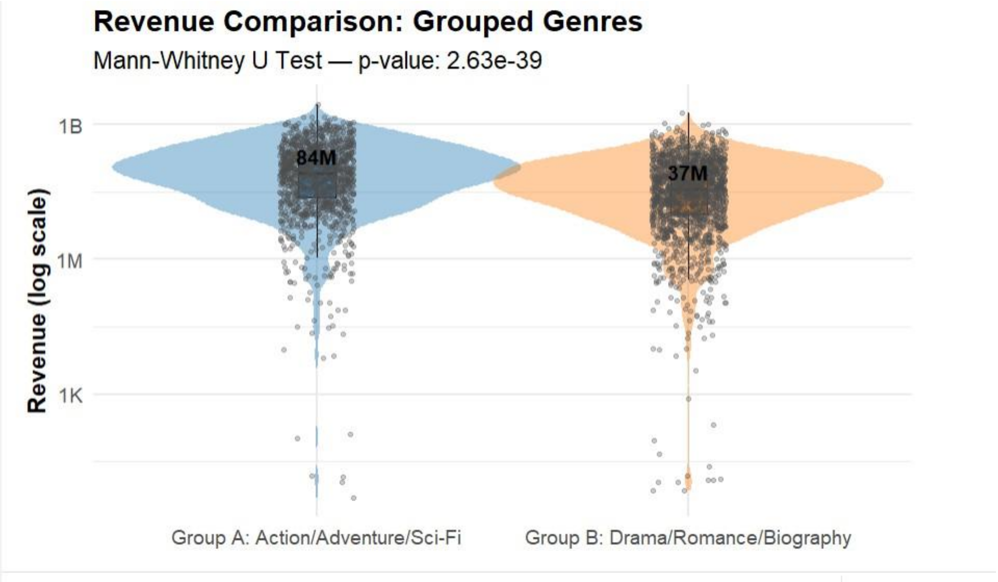


Mann-Whitney U test
W = 1,478,813
p-value $< 2.2 \times 10^{-16}$

Both groups significantly deviate from normality (with *p*-values less than 0.05), confirming that the use of the Mann-Whitney U test is appropriate for comparing the revenue distributions between the groups.

Inference

Since the p-value is significantly less than 0.05, I reject the null hypothesis. This provides strong evidence that there is a statistically significant difference in the distribution of movie revenues between the two genre groups.



Descriptive Statistics

Genre Group	Median Revenue	IQR	Sample Size (*n*)
Group A: Action/Adventure/Sci-Fi	$83,615,414	$191,315,940	1,341
Group B: Drama/Romance/Biography	$36,733,909	$90,932,962	1,729

From the descriptive statistics, I observe that Group A (Action/Adventure/Sci-Fi) movies tend to have higher median revenues and greater revenue spread compared to Group B (Drama/Romance/Biography). This observation aligns with the result of the Mann-Whitney U test, which indicated a significant difference in the distribution of movie revenues between the two groups. Based on this, I selected the alternative hypothesis (H_a), which suggests that there is a significant difference in the distribution of revenues between the two groups. Therefore, I conclude that there is strong statistical evidence supporting the claim that Action/Adventure/Science Fiction films tend to earn more than Drama/Romance/Biography films, both in terms of central tendency and distribution spread.

Q6 :How Movie Genres Evolve: Trends in Ratings, Rev- enue, Releases, and Budgets

I wanted to explore four key questions:

1. How have user ratings changed over time for different movie genres?
2. Which movie genres generate the highest average revenue?
3. How has the number of movies released per year varied by genre?
4. How does the budget distribution vary across different genres?

Understanding these trends helps to analyze whether certain genres have become more appreciated by audiences over time and which genres have seen growth or decline in production.

What and Why were these methods chosen? I calculated the average user rating per year per genre and visualized the results using a loess (locally estimated scatterplot smoothing) curve with a span of 0.3 to highlight long-term trends. I chose loess smoothing for visualizing average ratings because it is well-suited for identifying underlying patterns in noisy data without assuming a strict linear relationship. It helps highlight general trends while preserving important fluctuations. For the **number of movies released**, I counted how many movies were released each year per genre and plotted this using line graphs to show volume over time. For the number of movies, simple line plots were sufficient because I wanted to show raw trends in frequency without the need for statistical estimation.

The bar chart is a straightforward and effective way to compare average values across categories, making it perfect for visualizing average revenue by genre.

The box plot was chosen because it provides a deeper statistical view of the budget distribution. It shows the median, interquartile range, and outliers — which are important for financial variables like budget that can vary widely between movies.

Visualizations

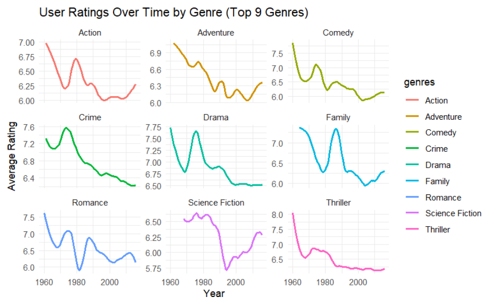


Figure 7: Movie Ratings Over Time

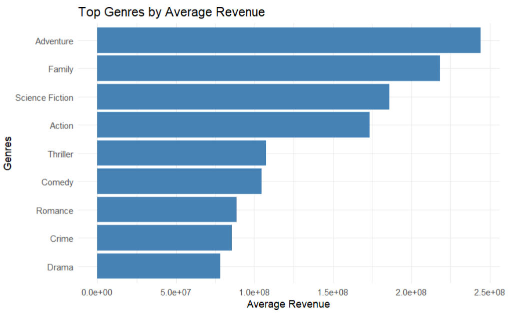


Figure 8: Average Revenue by Genre

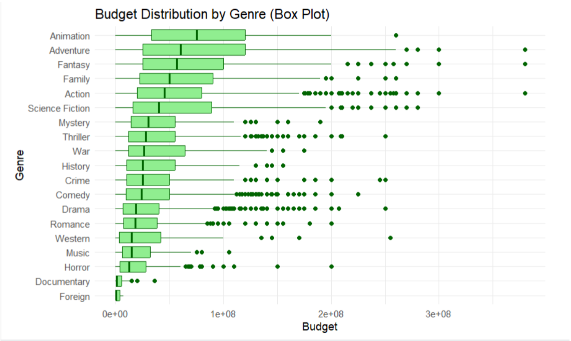


Figure 9: Budget Distribution by Genre

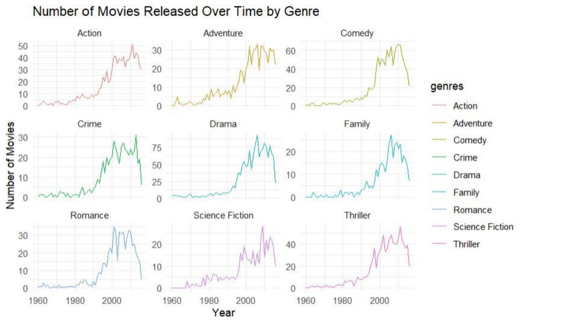


Figure 10: Movies Released Per Year by Genre Over Time

Inferences

From the **user ratings over time** plot: User ratings across genres reveal shifting audience preferences. Drama and Crime have generally maintained higher average ratings, suggesting strong storytelling and critical appeal. On the other hand, genres like Action, Comedy, and Romance have seen a gradual decline in user ratings over time. Interestingly, Science Fiction and Adventure, despite their commercial success, tend to have moderate to lower average ratings, hinting at a possible gap between financial performance and critical reception.

From the **Top Genres by Average Revenue** plot: From the bar chart, we can infer that Adventure, Family, and Science Fiction genres generate the highest average revenue. These genres are often associated with big-budget productions and appeal to broad audiences, which likely contributes to their commercial success. Genres like Drama and Crime, while commonly produced, tend to earn less revenue on average, possibly due to their niche appeal or lower production scales.

From the **Budget Distribution by Genre** plot: The chart shows that Animation, Adventure, and Fantasy films receive the highest median budgets. Action, Family, and Sci-Fi also have relatively high budgets with many outliers. In contrast, Documentary, Foreign, and Horror genres have much lower budgets. This indicates a clear preference for funding commercially promising genres. Studios invest more in genres with broader audience appeal and higher returns.

From the **number of movies released** plot: The line charts show how movie production trends have evolved across genres over the years. Notably, there has been a consistent rise in the number of movies released in genres such as Action, Comedy, and Drama, peaking around the 2000s. Family and Adventure films also saw significant growth, especially in the 1990s. This indicates that studios increasingly invested in genres that were gaining popularity and delivering better financial returns.

Q7 :Analysis of Revenue by Director

Q7.1 :Testing Revenue Distribution Across Directors

Hypotheses

To investigate whether the revenue distributions differ by director, I tested the following hypotheses using the Kruskal-Wallis rank sum test:

- **Null Hypothesis** (H_0): There is no difference in the distribution of revenues among the directors.

- **Alternative Hypothesis** (H_a): At least one director has a revenue distribution that differs significantly from the others.

Test Results

Kruskal-Wallis rank sum test:
Chi-squared $= 39.482$, **df** $= 2$, **p-value** $= 2.67 \times 10^{-9}$

The very small p-value suggests a statistically significant difference in revenue distributions among the directors.

Before conducting the Kruskal-Wallis test, I performed diagnostic tests to assess the data's suitability:

- **Shapiro-Wilk test for normality:**
  $W = 0.66636, \quad p < 2.2 \times 10^{-16}$
  This indicates that the residuals deviate significantly from a normal distribution.

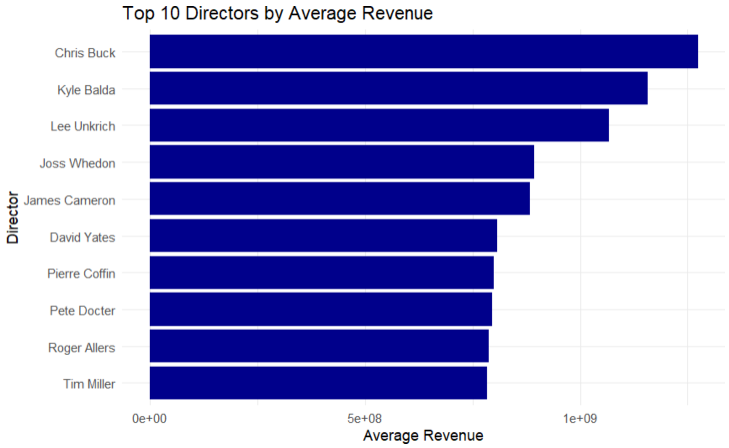
- **Levene's Test for homogeneity of variances:**
  $F(2, 65) = 8.8121, \ p = 0.0004109$
  This result suggests unequal variances across groups (heteroscedasticity).

Since both the assumptions of normality and homogeneity of variances were violated, a non-parametric alternative was necessary. The Kruskal-Wallis test was adopted in this scenario because it does not assume normality or equal variances and is robust for comparing more than two independent groups based on ranks.

Inference

Given the extremely small p-value of 2.67×10^{-9}, I reject the null hypothesis. This provides strong evidence that at least one director has a significantly different revenue distribution compared to others. The Kruskal-Wallis test has shown that there are indeed significant differences in revenue distributions among the directors analyzed. This suggests that the directors' revenues are not all from the same population distribution.

Q7.2 :Revenue Comparison Among Film Directors



From the chart, it is evident that Chris Buck has the highest average revenue among the directors, exceeding $1 billion. Animated film directors such as Kyle Balda, Lee Unkrich, and Pierre Coffin dominate the top ranks, emphasizing the strong box office performance of animated movies. In general, while the top three directors stand out significantly, the remaining directors have average revenues clustered between $700 million and $900 million.

Q8 :The Influence of Star Power: Analyzing Actor Fre- quency and Movie Profits

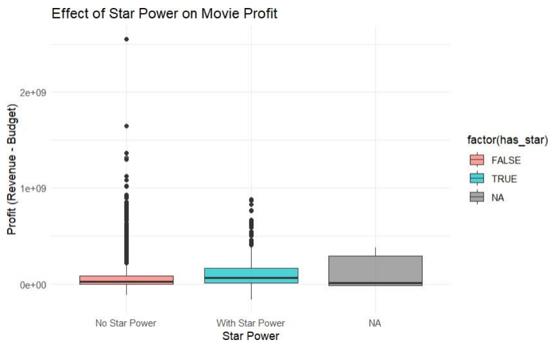
What is the question? In this section, I explored two connected questions:

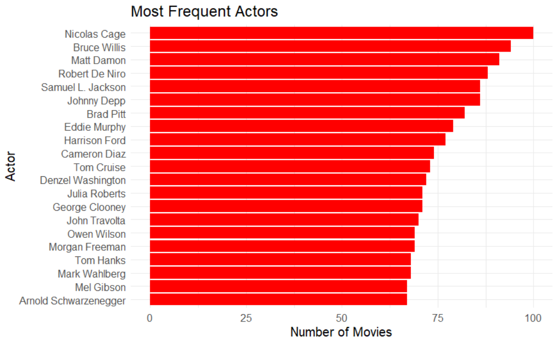
1. Who are the most frequently appearing actors in the dataset?
2. Does the presence of these top actors (i.e., "star power") have a measurable impact on a movie's profitability?

Why were these methods chosen Counting actor appearances allowed me to quantitatively identify the most prominent actors in the dataset. A bar chart was a natural fit for clearly visualizing frequency rankings. To identify the most frequent actors, I separated the cast list for each movie and counted how often each actor appeared. I then visualized the top 20 most frequent actors using a horizontal bar chart.

For analyzing star power, the binary classification approach made it easy to group movies and assess profitability differences. A box plot was ideal here because it shows not only the central tendency (median) but also the distribution and potential outliers, providing more insight than just average values. To analyze the effect of star power, I created a binary indicator (‘has star‘) showing whether a movie featured any of the top 20 actors. I calculated the profit for each movie and compared average profits between movies with and without star power. A box plot was used to visualize the distribution of profits across these two categories.

Visualization





Inference

From the **Star Power and Profit** box plot: The chart compares movie profits with and without star power. Movies with star power generally have higher median profits than those without. However, some no-star movies achieve extreme profits, as seen in the outliers. The NA group shows a wider spread but limited data. Overall, star power can boost profits but doesn't guarantee success.

From the **Most Frequent Actors** bar chart: I found that actors like Nicolas Cage, Bruce Willis and Matt Damon appeared in the most movies.These actors are known for being prolific, which aligns well with the data.

Conclusion

This report reveals that high-budget, visually engaging genres like Action and Animation earn significantly more revenue. Star power boosts profitability, but some low-profile films still succeed. Directors and genre choices greatly influence earnings, while budget and popularity strongly correlate with success. Overall, strategic production choices drive financial outcomes in filmmaking.