Mangira Project

Anestis,Juljan,Nikos,Dimitris

November 2024

Abstract

This analysis is intended to explore the movie dataset in maximum possible extent and uncover patterns and trends that may result in a success of next movie. The purpose of this report is to examine a variety of influential variables such as movie budget, revenues, genres, production company, language and others to guide decision making for a movie to the success of box office. In this report we go over the steps taken for the dataset's analysis, we made relevant visualizations and we extract key insights for future movie production.

Contents

1 Data Processing **4**

2 Log-Transformed Distribution of Budget and Revenue **5**

2.1 Key insights 5

3 Statistical Analysis of Variables **6**

3.1 Key Observations 6

4 Relationships Between Budget and Revenue **7**

4.1 Key Finding 8

5 Relationship Between Runtime and Popularity **9**

5.1 Key Findings 9

6 Correlation Heatmap of Key Variables **11**

6.1 Key Findings 11

7 Top 5 genres by Profit,Revenue and Popularity **14**

7.1 Key Insights 15

8 Seasonal Trends **16**

9 Production Companies and Language Analysis **18**

9.1 Key Observations 18

9.2 Language Analysis 19

10 Revenue Trends Over Time **19**

11 Ideas for Future Exploration **20**

12 Conclusions **21**

13 Limitations and Future Work **22**

List of Figures

1 HistPlot for the Distribution of Budget and Revenue 5
2 Satistical Summary of Key Varuables 6
3 ScatterPlot Between Budget and Revenue 7
4 ScatterPlot Between Runtime and Popularity 9
5 Heatmap of Key Variables 11
6 Top 5 Genres based on Profit, Revenue and Popularity 14
7 Top 5 Most Profitable Genres 16
8 Detailed Genre Revenue Analysis 16
9 Seasonal Trends in Movie Success 17
10 Top 10 Production Companies by Total Revenue 18
11 Top 5 Most Profitable Spoken Languages for Movies 20
12 Average Revenue Over the Years 20
13 Average Revenue Trend Over Decades 21

1 Data Processing

Before conducting any analysis, data preprocessing is a crucial step to clean and prepare the dataset. We began by handling missing values, particularly for essential columns like revenue, budget, runtime, and release date. Missing values were dropped to maintain data quality, and release dates were converted to datetime format to extract useful information, such as release years and months. We filtered out movies where both budget and revenue were zero, as these entries did not contribute meaningfully to the analysis. Additionally, profit and profit margin were calculated to help understand movie profitability.

Listing 1: Filtering/Loading the Dataset

```
1
2  # Load dataset
3  movies_df = pd.read_csv('movies.csv')
4
5  # Preprocessing: Handle missing values
6  movies_df = movies_df.dropna(subset=['revenue', 'budget', 'runtime',
       'release_date'])
7  movies_df['release_date'] = pd.to_datetime(movies_df['release_date'],
       dayfirst=True, errors='coerce')
8  movies_df = movies_df.dropna(subset=['release_date'])
9
10 # Extract year from release_date for analysis
11 movies_df['year'] = movies_df['release_date'].dt.year
12
13 # Filter out rows where both budget and revenue are zero
14 movies_df = movies_df[~((movies_df['revenue'] == 0) &
15 (movies_df['budget'] == 0))]
16
17 # Calculate profit and profit margin
18 movies_df['profit'] = movies_df['revenue'] - movies_df['budget']
       movies_df['profit_margin'] = (movies_df['profit'] /
19 movies_df['budget']) * 100
20 movies_df['profit_margin'].replace([float('inf'), -float('inf')],
       float('nan'), inplace=True)
```

2 Log-Transformed Distribution of Budget and Revenue

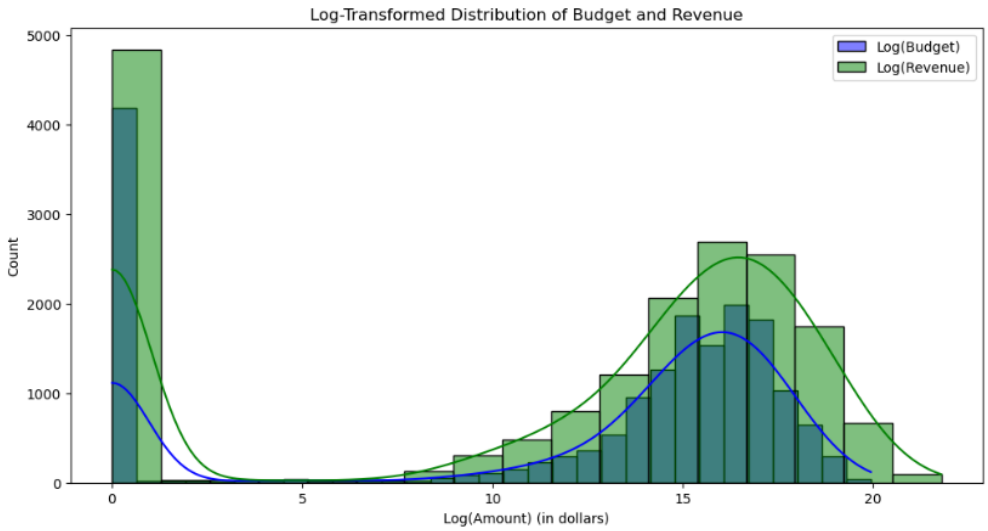

Figure 1: HistPlot for the Distribution of Budget and Revenue

For easier interpretation and to reduce skewness, we applied a logarithmic transformation to the budget and revenue variables, as visualized in a log-transformed histogram.

As a result from the graph a general pattern emerged, indicating that movies with higher budgets often achieve higher revenues through not strictly proportionate. The overlap in the distribution suggests that while many movies achieve moderate success, only a subset of movies with high budgets stands out in terms of revenue.

2.1 Key insights

1. Hight Variability: Revenue exhibits higher variability than budget, as seen in the broader spread of the green histogram

2. Cost-Efficiency: Some low-budget movies achieve relatively hight revenue, indicated by the KDE peak in the budget distribution and it overlaps with a portion of the revenue distribution.

3 Statistical Analysis of Variables

A statistical summary of key variables such as budget, revenue, profit, profit margin and popularity was computed. This summary provided insights into average values, distribution and variability of these variables. The profit margin, in particular served as a useful metric for understanding the financial success of the movies relative to their production costs.

Statistical Summary of Key Variables:

	budget	revenue	profit	profit_margin	popularity
count	17,708.00	17,708.00	17,708.00	17,708.00	17,708.00
mean	14,814,312.54	39,712,186.15	24,897,873.61	-	15.78
std	30,896,814.78	122,162,609.29	101,548,530.44	-	49.00
min	0.00	0.00	-199,545,977.00	-100.00	0.60
25%	3,000.00	0.00	-2,342,693.25	-100.00	5.58
50%	3,000,000.00	2,500,000.00	218,347.50	65.24	10.33
75%	15,000,000.00	23,000,000.00	11,688,851.75	2,400.00	16.60
max	460,000,000.00	2,923,706,026.00	2,686,706,026.00	-	2,994.36

Figure 2: Satistical Summary of Key Varuables

3.1 Key Observations

1. Both variables(Budget and Revenue) showed significant variability with mean values being much higher than medians, indicating the presence of high-value outliers.

2. Profit and Profit Margin:

   (a) Movies with a high profit margin tended to have moderate budhets and carefully controlled costs.

   (b) There were notable outliers where movies achieve extraordinary profits,

suggesting rare but impactful successes.

(c) Popularity and Runtime

i. Popular Movies generally exhibited an optimal runtime range, with diminishing return in popularity for excessively long or short films.

4 Relationships Between Budget and Revenue

To explore the relationships between various variables, several analyses were performed:

- Budget and Revenue: A scatter plot of budget vs revenue showed a generally positive relationship, indicating that higher-budget movies tend to generate higher revenues. However there were some outliers, suggesting that spending a lot of money does not always guarantee success.

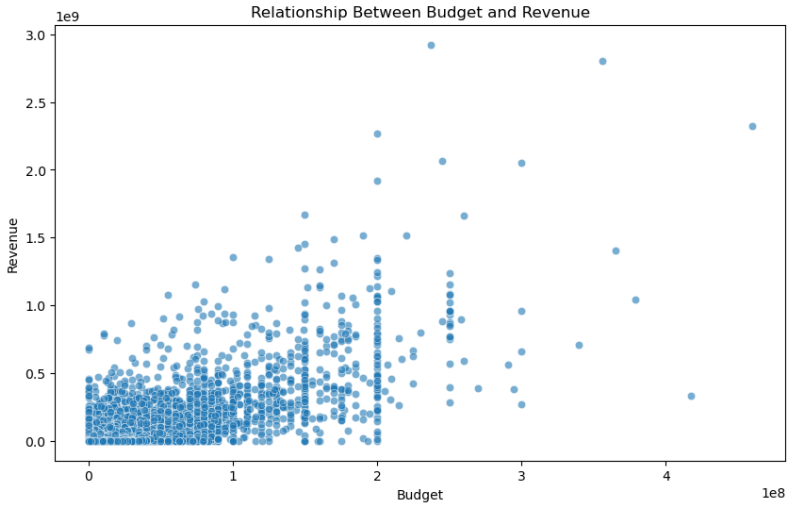

Figure 3: ScatterPlot Between Budget and Revenue

4.1 Key Finding

- Positive Correlation:
  - There is a general positive trend, where higher budgets tend to be associated with higher revenues. However, this relationship is not perfectly linear.
- High Variability:
  - Movies with smaller budgets exhibit significant variability in revenue. While some low-budget movies achieve high revenues, many generate only modest earnings.
  - Higher-Budget movies generally show less variability, with most achieving higher revenues.
- Saturation Point:
  - Beyond a certain budget threshold the incremental revenue increase diminishes. This could imply that additional spending doesn't guarantee financial returns.
- Outliers:
  - Several high-budget movies have extremely high revenues. These could represent blockbuster films or highly successful franchise.
  - A few high-budget movies fail to generate revenues, suggesting factors like poor marketing, bad reviews or limited audience.
- Clusters of Success:
  - A significant number of movies with a budget between 100m$ have revenues below 500m$.
  - Outliers exist where movies with budgets below 100m$ achieve revenues exceeding 1b$ indicating rare but highly successful low-budget films.

5 Relationship Between Runtime and Popularity

This is a scatter plot indicating the relationship between runtime and Popularity.

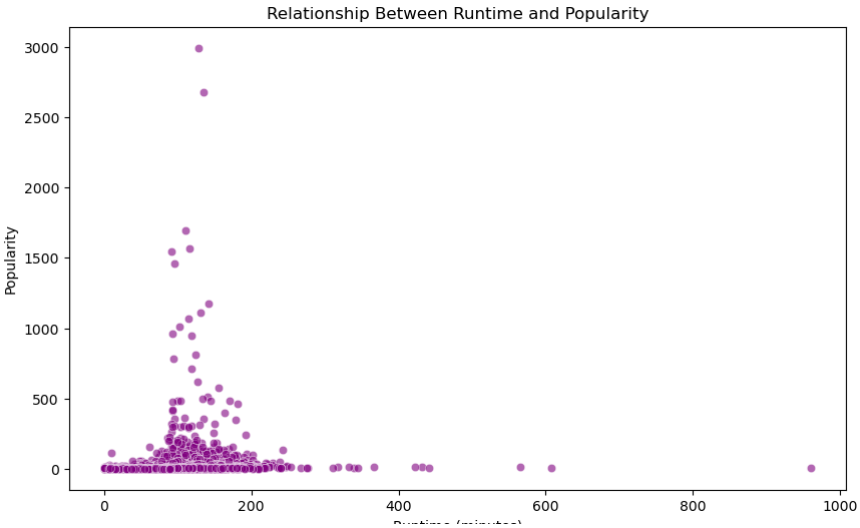

Figure 4: ScatterPlot Between Runtime and Popularity

5.1 Key Findings

- Optimal Runtime Range:
  - Most popular movies appear to have runtime between 90 and 150 minutes, with a noticeable cluster of higher popularity scores in this range.
  - This range likely represents the ”sweet spot” for engaging audiences without causing fatigue or being too short to fully developed the story.
- Diminishing Returns for Long Movies:
  - Movies with runtime beyond 150 minutes (especially over 200 minutes) tend to have lower popularity scores on average.
  - This suggests that excessively long runtimes may reduce audience engagement, possibly due to viewer fatigue.

- Short Movies (Under 90 Minutes):
  - Movies with runtimes under 90 minutes generally show lower popularity, indicating that very short movies might struggle to provide enough content or storytelling to satisfy audiences.
- Outliers:
  - A few movies with extreme popularity scores (above 1500) and runtimes within the optimal range (90-150 minutes) likely represent blockbuster hits to highly anticipated films.
  - There are a few unusual movies with runtimes exceeding 400 minutes, but they are not popular likely reflecting niche or experimental content.
- Weak Correlation:
  - There is no strong linear relationship between runtime and popularity. Instead the data suggests a moderate bell-shaped trend with the highest popularity concentrated within a specific runtime range.

6 Correlation Heatmap of Key Variables

A correlation heatmap of key variables was generated to identify relationships among budget, revenue, profit, popularity, runtime and audience ratings. This analysis highlighted the strongest correlations and indicated where relationships might exist.

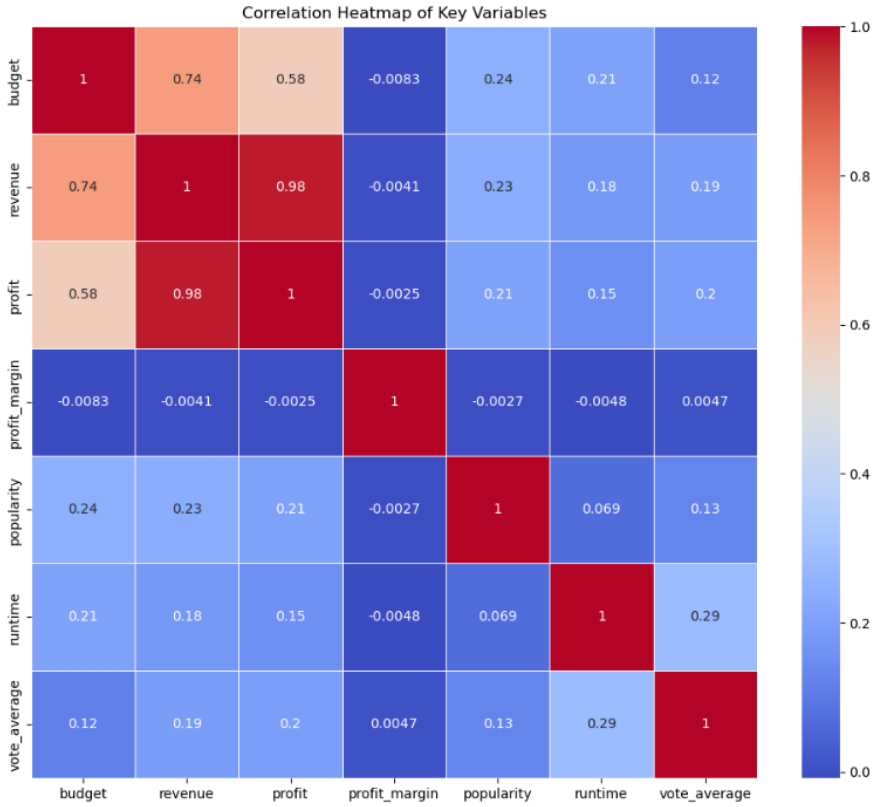

Figure 5: Heatmap of Key Variables

6.1 Key Findings

- Strong Positive Correlation Between Budget and Revenue (0,74):
  - Higher budgets are strongly associated with higher revenues, indicating that investing more in production often leads to greater financial returns.

  - However the correlation is not perfect suggesting that factors other than budget(e.g., marketing,genre,audience reception) also play a role in determining revenue.

- Strong Positive Correlation Between Revenue and Profit (0.98):
  - Revenue and profit are almost perfectly correlated, as profit is directly derived from revenue. This result is expected.

- Moderate Positive Correlation Between Budget and Profit (0.58):
  - Higher budgets tend to generate higher profits this relationship is weaker than the budget-revenue correlation. This suggests that while large budgets may lead to high revenues, they don't always translate into proportional profits due to higher costs.

- Weak Correlation Between Budget and Profit Margin (-0.0083):
  - There is essentially no relationships between budget and profit margin. This implies that smaller-budget films can achieve high profit margins, whereas higher-budget films may have lower margins despite generating large profits.

- Popularity has Weak Positive Correaltions:
  - With budget (0.24), revenue (0.23) and profit (0.21):
    * More popular movies tend to have higher budgets, revenues and profits, but the relationships are not strong. Popularity is influenced by the factors such as marketing critical acclaim or franchise appeal.
    * Weak correlation suggests that popularity alone is not a strong predictor of financial performance.

- Runtime has Weak Positive and Correaltions:
  - With vote average (0.29) and budget (0.21):

    - Longer runtimes are weakly associated with higher audience ratings and larger budgets. However the correlation is not strong enough to generalize that longer movies are always better received or higher-budget productions.

- Vote Average(Audience Rating) Correlation:
  - With runtime (0.29) and profit (0.2):
    - Higher-rated movies tend to have slightly longer runtimes and higher profits. This suggests that well-received movies often have sufficient runtime to develop engaging stories and deliver better received financial results.

- Profit Margin has Almost No Correlation with Any Variable:
  - The lack of significant correlations indicates that profit margins are not directly tied to budget revenue or popularity. Smaller-budget films can achieve exceptional margins though cost efficiency, while large-budget films may struggle due to high productions costs.

7 Top 5 genres by Profit,Revenue and Popularity

Genre Analysis play a significant role in determining a movies success. We split the genres column into individual genres to analyze their impact separately. The top genres were identified based on average profit, average revenue and popularity.

```
Top 5 Genres by Average Profit:

genres
Family, Fantasy, Romance                          1.106116e+09
Action, Animation, Comedy, Family, Adventure      9.515526e+08
Adventure, Comedy, Animation, Family              8.004579e+08
Science Fiction, Adventure, Family, Fantasy       7.824655e+08
Animation, Family, Comedy, Fantasy, Adventure     7.787608e+08
Name: profit, dtype: float64

Top 5 Genres by Average Revenue:

genres
Family, Fantasy, Romance                          1.266116e+09
Action, Animation, Comedy, Family, Adventure      1.031553e+09
Animation, Family, Comedy, Fantasy, Adventure     9.287608e+08
Adventure, Comedy, Animation, Family              8.754579e+08
Animation, Family, Adventure, Drama, Comedy       8.576112e+08
Name: revenue, dtype: float64

Top 5 Genres by Average Popularity:

genres
Action, Mystery, Thriller, Crime                 1547.2200
Animation, Comedy, Family, Fantasy, Romance      1008.9420
Action, Drama, Adventure                          559.1372
Adventure, Family, Fantasy, Romance               353.5430
Comedy, Adventure, Fantasy                        283.9435
Name: popularity, dtype: float64
```

Figure 6: Top 5 Genres based on Profit, Revenue and Popularity

7.1 Key Insights

- Top 5 Genres by Average Profit:
  - Family, Fantasy, Romance ranks the highest in profitability, generating over 1.1 billion in average profit.
  - Genres combining Action, Animations, Comedy, Family and Adventure follow closely with nearly 951 million in average profit.
  - The consistent inclusion of Family,Animation and Adventure genres highlights their strong profitability.
- Top 5 Genres by Average Revenue:
  - Family,Fantasy and Romance once again leads with over 1.26 billion in average revenue.
  - Similar to the profit ranking genres involving Action,Animation,Comedy,Family and Adventure rank high in revenue generation.
  - The overlap between high-profit and high-revenue genres suggests that these genres not only generate substantial revenues but also remain controlled costs, leading to profitability.
- Top 5 Genres by Average Popularity:
  - Actions, Mystery, Thriller, Crime stands out with a significantly higher average popularity (1547.22), far exceeding other genres.
  - Animation,Comedy,Family,Fantasy and Romance comes second showing that animated family movies are not only profitable and revenue-generating but also popular among audiences.
  - Popularity rankings include genres like Aciton,Drama and Adventure indicating that audiences are drawn to diverse genres.

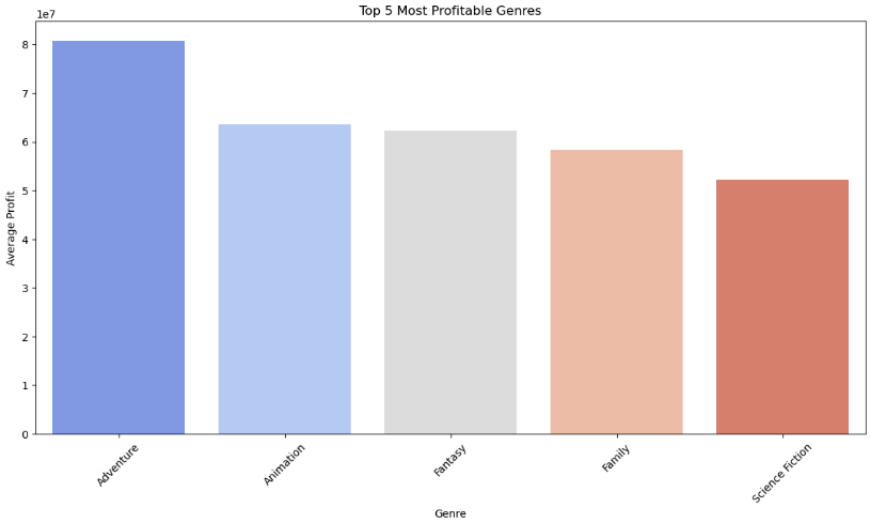

Figure 7: Top 5 Most Profitable Genres

```
Detailed Genre Revenue Analysis:

Adventure: Average Revenue = $120,099,473.68, Total Revenue = $288,719,134,734.00
Fantasy: Average Revenue = $94,260,520.24, Total Revenue = $143,370,251,278.00
Animation: Average Revenue = $90,115,249.37, Total Revenue = $94,530,896,585.00
Family: Average Revenue = $88,487,148.95, Total Revenue = $134,323,492,113.00
Science Fiction: Average Revenue = $81,082,124.84, Total Revenue = $142,299,129,090.00
Action: Average Revenue = $76,505,338.57, Total Revenue = $294,010,016,109.00
Comedy: Average Revenue = $40,097,993.99, Total Revenue = $237,500,418,379.00
War: Average Revenue = $38,403,259.61, Total Revenue = $23,694,811,180.00
Thriller: Average Revenue = $37,308,897.03, Total Revenue = $156,286,969,672.00
Mystery: Average Revenue = $35,345,165.34, Total Revenue = $47,291,831,223.00
```

Figure 8: Detailed Genre Revenue Analysis

8 Seasonal Trends

To understand seasonal trends, we extracted the release month from the release date and analyzed the average popularity and revenue of movies each month. It was observed that certain months tend to have higher average popularity and revenue, indicating the significance of seasonal timing for movie releases. A line plot was used to illustrate these trends, highlighting the months when movie release might be most profitable.

- Average Revenue Peaks in Specific Months:
  - High-Revenue Months:

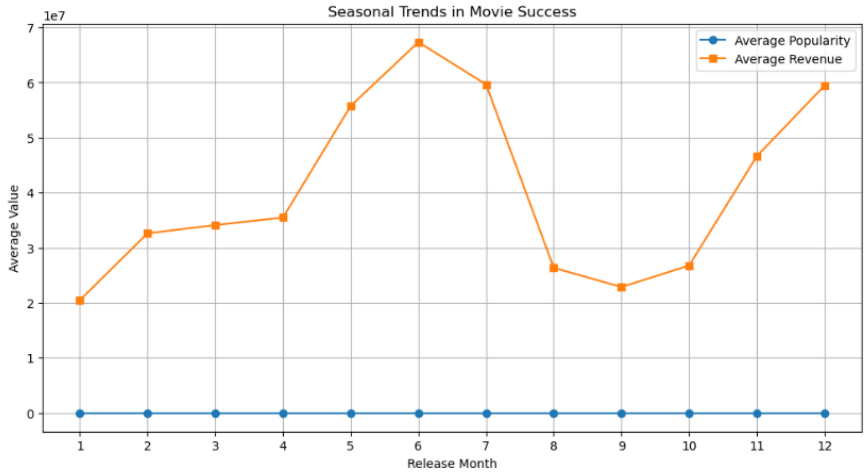

Figure 9: Seasonal Trends in Movie Success

  - ∗ June (6) and July (7) show the highest average revenue suggesting these months are prime for movie releases. These peaks are likely driven by summer blockbusters, vacation periods and high audience.

  - ∗ December (12) is another revenue peak, likely attributed to holiday seasons, family gathering and festive releases.

  - Low-Revenue Months:

    - ∗ September and January have significantly lower revenues, possibly due to fewer releases or less audience engagement during these months.

- Average Popularity Remains Relatively Constant:

  - Populairty values are consistently low across all months showing minimal variation compared to revenue trends.

  - This suggests that popularity alone may not be influences by the release month but by other factors like marketing genre or star power.

- Revenue and Popularity Do not Always Align:

- While revenue spikes in specific months, popularity shows no significant change, indicating that revenue generation more heavily influenced by external factors like seasonal timing and audience availability.

9 Production Companies and Language Analysis

We identified the most profitable production companies by exploring the ”production_companies” column and grouping by each company. This analysis provided insights into which companies are leading in terms of movie profitability helping us identify potential partenships for future productions.

Top 10 Production Companies by Total Revenue:

	production_companies	revenue
0	Warner Bros. Pictures	80214538856
1	Universal Pictures	77295116145
2	20th Century Fox	64019637767
3	Paramount	63335237082
4	Columbia Pictures	59944912191
5	Walt Disney Pictures	53931217272
6	Marvel Studios	30961973635
7	New Line Cinema	27787064203
8	Metro-Goldwyn-Mayer	21413009368
9	DreamWorks Pictures	20196810541

Figure 10: Top 10 Production Companies by Total Revenue

9.1 Key Observations

- Dominance of Major Studios:
  - Warner Bros.Pictures leads the list with a total revenue of over 802 billion, followed closely by Universal Pictures (772 billion). Both companies have a significant lead over others, emphasizing their dominant position in the global film market.
  - Other High-ranking studios like 20th Century Fox, Pramount and Columbia

Pictures also show substantial revenues, reflecting their historical success and strong market presence.

- Disney and Marvel Studios:
  - Walt Disney Pictures ranks 6th with over 539 billion and Marvel studios a subsidiary of Disney ranks 7th with 309 billion.
  - Combines, Disney and Marvel represent a considerable share of the market, driven by franchises like the Marvel Cinematic Universe animated classics and live action adaptations,
- Specialized and Iconic Studios:
  - New Line Cinema and Metro Goldwyn Mayer (MGM) show impressive revenues highlighting their historical contributions and focus on specific types of films such as franchises and blockbusters.
- Dreamworks Pictures:
  - Ranking 10th,DreamWorks has generated over 201 billion showcasing its success in the animation secotr and family oriented films.

9.2 Language Analysis

We analyzed both original languages and spoken languages in terms of total profit. English was by far the most profitable language, but other languages like French,Spanish and Mandarin also showed considerable profitability.

10 Revenue Trends Over Time

To further understand trends in Revenue, we analyzed average revenue over different periods. Average Revenue Over the Years: A line plot showed how average revenue evolved over time, with significant spikes in recent decades.

Top 5 Most Profitable Spoken Languages for Movies:

	spoken_languages	profit
0	English	407953406689
1	French	44731125299
2	Spanish	44669886799
3	Mandarin	31384547043
4	German	26204471986

Figure 11: Top 5 Most Profitable Spoken Languages for Movies

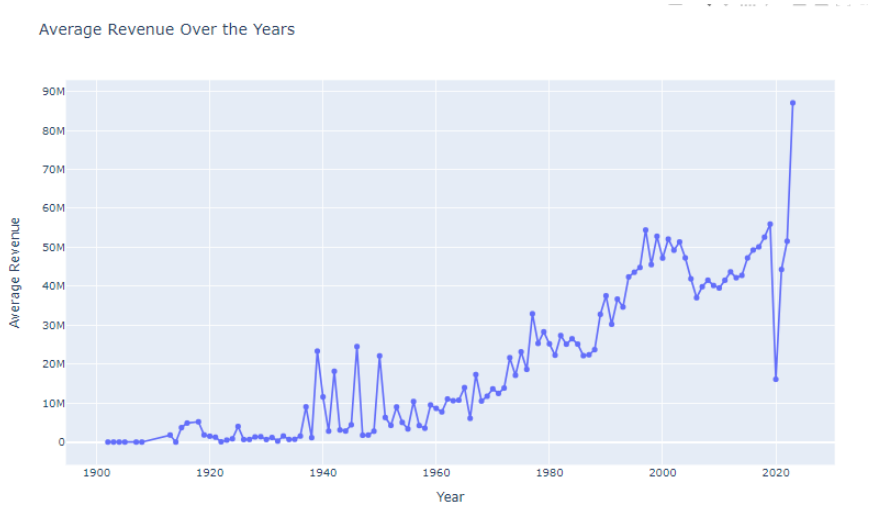

Figure 12: Average Revenue Over the Years

Average Revenue Trend Over Decades: We analyzed average revenue grouped by decade, highlighting that the recent decades have shown a substantial increase in revenue.

11 Ideas for Future Exploration

Based on the analysis, several recommendations and ideas could help improve future movie production strategies:

- Target Optimal Release Timing: Plan releases during months when movies historically perform well to maximize audience reach and revenue.

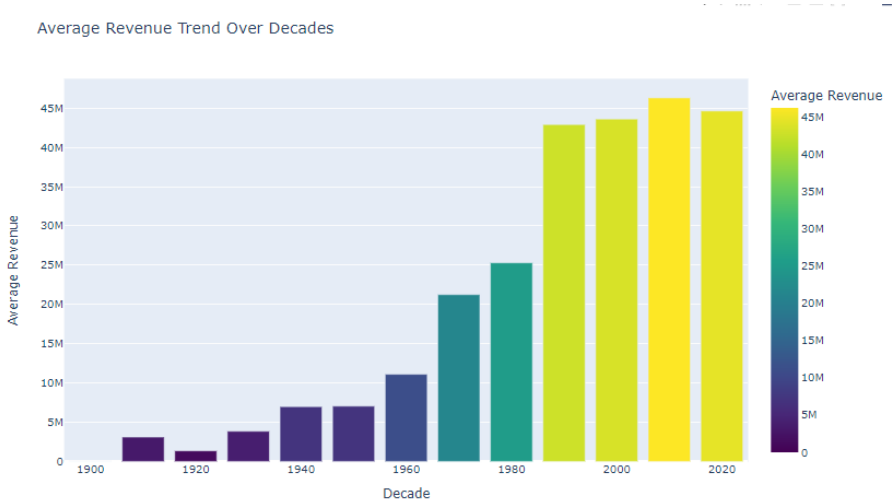

Figure 13: Average Revenue Trend Over Decades

- Genre Focus: Focus on genres that have proven profitability such as Adventure and Fantasy while also exploring genre combinations to attract diverse audience segments.
- Cost Efficiency: Develop movies with careful budget control aiming for cost efficient genres to maximize profit margins.
- International Language Films: Conisder

12 Conclusions

The findings from the analysis offer several important insights:

- Higher budgets generally lead to higher revenues but not always.Some low-budget movies achieved remarkable profitability as evidenced by the profit margin analysis
- The choice of genre has a significant impact on a movies financial success. Adventure, Animation and Fantasy consistently ranked among the most successful

genres in terms of profit revenue and popularity.

- Release timing plays an essential role in a movies success. Certain months showed higher average revenues suggesting that releasing during these months might maximize returns.
- English-language movies dominate in terms of profitability. However there is potential for growth in movies produced in languages like French, Spanish and Mandarin.

13 Limitations and Future Work