Behavioral Research in Statistical Methods
Mini Project 4
Report

Mayank Mittal (2022101094)

March 2, 2025

1 Introduction

This project involves performing descriptive statistics, exploration, and visualizations on the TMDB 5000 Movies Dataset. The dataset provides comprehensive information about movies, including budget, revenue, ratings, genres, and more. We answer twelve key questions by exploring trends in revenue, budget, ratings, directors' success, user preferences, and other movie-related factors.
The dataset can be accessed from the following link:
Dataset Link: https://www.kaggle.com/datasets/akshaydattatraykhare/movies-dataset/data

2 EDA and Preprocessing

We imported necessary libraries and loaded the `movies.csv` and `credits.csv` datasets, containing information about 4803 movies with 20 columns and 4 columns respectively.

2.1 Data Preprocessing

The dataset undergoes the following preprocessing steps making new columns:

- Convert JSON strings into lists or dictionaries for columns: `genres`, `keywords`, `production_companies`, `production_countries`, and `spoken_languages`.
- Replace infinite values with NaN.
- Extract genre names into a new column `genre_list`.
- Fill missing values in `tagline` with an empty string.
- Convert `release_date` to datetime format and extract and made new columns:
  - `release_year`
  - `release_month`
- Compute the number of spoken languages.
- Convert revenue to millions (`revenue_millions`).
- Create a binary indicator for English movies (`is_english`).
- Compute Return on Investment (ROI) percentage and made a column 'roi'.
- Calculate the length of the `tagline` and made a new column 'tagline length'.

2.2 Handling Missing Values and Feature Selection

We performed Exploratory Data Analysis (EDA) and preprocessing to handle missing values, clean the data, and select relevant features.

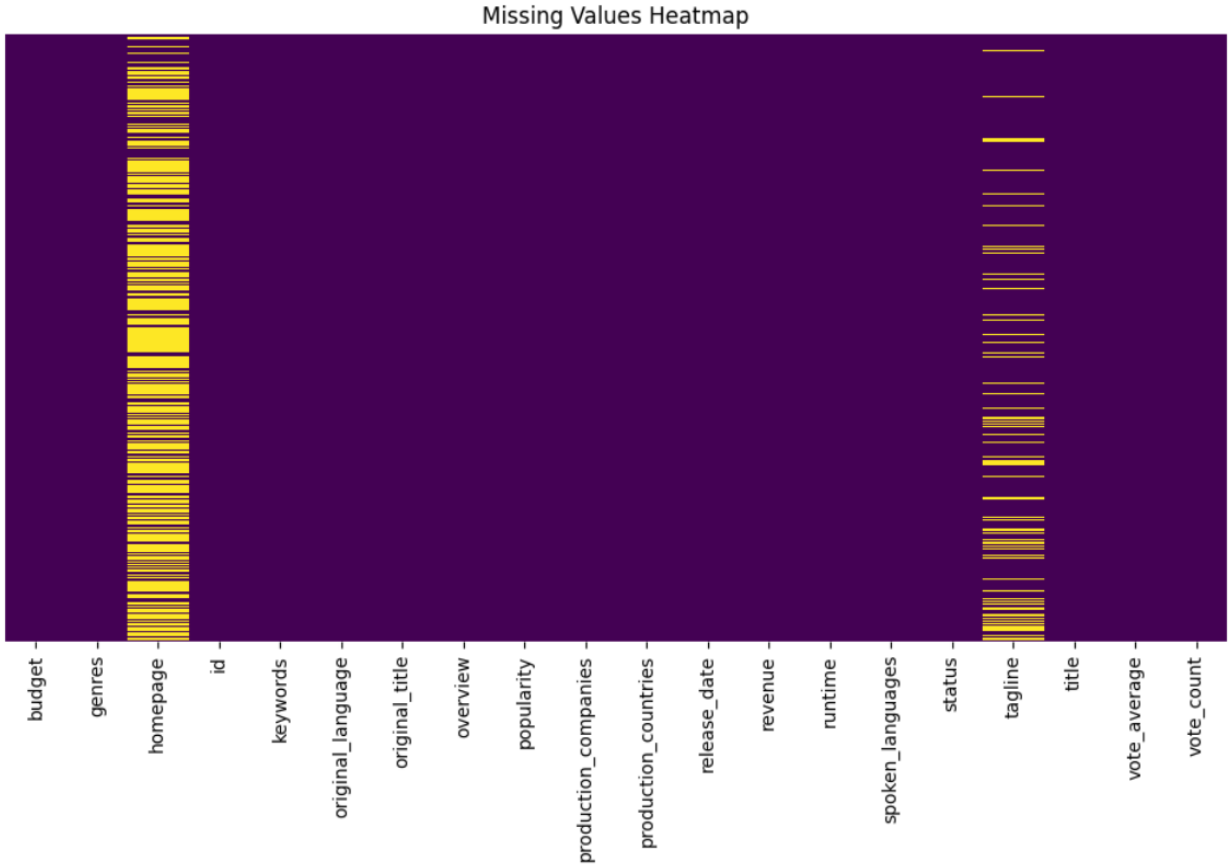


Figure 1: Heatmap of Missing Values

As shown in Figure 1, only two columns had significant missing values. Due to the high number of missing values, imputation or removal of these rows was not feasible. Instead, we removed the `homepage` column, as it was deemed less important for our analysis. For `tagline` and `ROI`, we proceeded with the analysis using only non-null values.

2.3 Dropping Unnecessary Features

We dropped the `homepage` column due to its low relevance.

2.4 Handling Release Date and Overview

We removed the row with missing `release_date`, `month`, and `year` values, as it contained mostly missing data, using the `movie_id` column. We imputed the three missing `overview` values with empty strings, as removing these rows was not feasible due to other non-missing data.

	budget	id	popularity	release_date	revenue	runtime	vote_average	vote_count	release_year	release_month	num_languages	revenue_millions	roi	tagline_
count	4.802000e+03	4802.000000	4802.000000	4802	4.802000e+03	4802.000000	4802.000000	4802.000000	4802.000000	4802.000000	4802.000000	4802.000000	4.802000e+03	4802.
mean	2.905109e+07	57098.234902	21.496776	2002-12-27 23:45:54.352353280	8.227777e+07	106.853603	6.093440	690.361724	2002.468763	6.795918	1.444606	82.277769	1.986122e+05	34.
min	0.000000e+00	5.000000	0.000372	1916-09-04 00:00:00	0.000000e+00	0.000000	0.000000	0.000000	1916.000000	1.000000	0.000000	0.000000	-1.000000e+02	0.
25%	8.000000e+05	9013.750000	4.671734	1999-07-14 00:00:00	0.000000e+00	94.000000	5.600000	54.000000	1999.000000	4.000000	1.000000	0.000000	-1.632024e+01	18.
50%	1.500000e+07	14626.500000	12.924931	2005-10-03 00:00:00	1.917498e+07	103.000000	6.200000	235.500000	2005.000000	7.000000	1.000000	19.174985	4.741262e+00	32.
75%	4.000000e+07	58589.750000	28.332017	2011-02-16 00:00:00	9.291920e+07	117.750000	6.800000	737.000000	2011.000000	10.000000	2.000000	92.919195	2.157480e+02	47.
max	3.800000e+08	459488.000000	875.581305	2017-02-03 00:00:00	2.787965e+09	338.000000	10.000000	13752.000000	2017.000000	12.000000	9.000000	2787.965087	8.499999e+08	252.
std	4.072447e+07	88581.302370	31.818451	NaN	1.628697e+08	22.662122	1.191496	1234.674268	12.414354	3.424187	0.926790	162.869733	1.235044e+07	27.

Figure 2: Descriptive Statistics

2.5 Descriptive Statistics

Figure 2 displays the descriptive statistics of the dataset.

2.6 Pearson's Correlation Heatmap

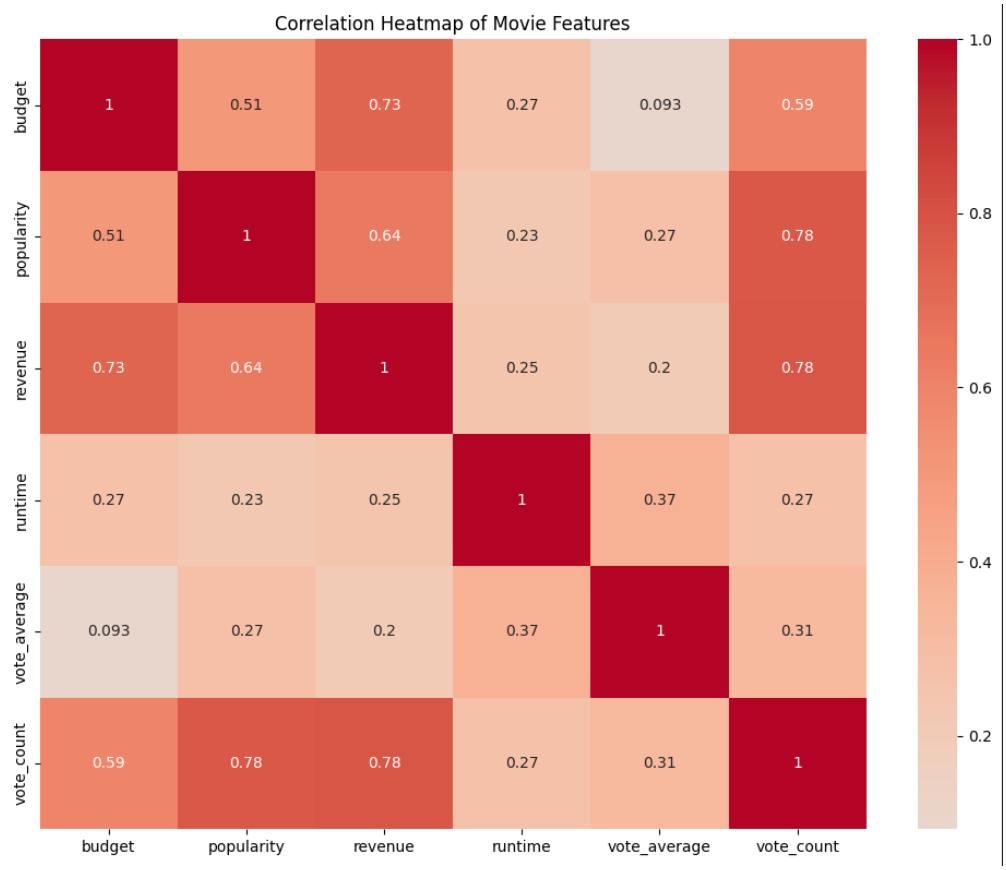


Figure 3: Correlation Heatmap

Figure 3 shows the correlation heatmap for numerical columns. Observations from the heatmap:

- Popularity and revenue have a high positive correlation (0.78) with vote count as obvious that more votes on the movie will likely results in more popularity and revenue..
- Revenue and budget have a high positive correlation (0.73) as more budget invested for the movie will likely results in more revenue until the movie becomes flop due to poor storyline.

- Revenue and popularity have a good positive correlation (0.64) as more revenue is only when movie popularity is high in the market.
- Budget and vote count have a good positive correlation (0.59) as more budget likely results in good movie and more votes.
- Budget and popularity have a good positive correlation (0.51) as more budget results in good movie with high popularity.

2.7 Spearman Correlation Heatmap

Figure 4 shows the spearman correlation heatmap for numerical columns. Observations from the heatmap:

- Revenue and budget have a strong positive correlation (0.76), suggesting higher budgets often lead to higher revenue.
- Popularity and vote count show an extremely strong correlation (0.96), indicating popular movies get more votes.
- All financial metrics (budget, revenue, popularity, vote count) are strongly correlated with each other.

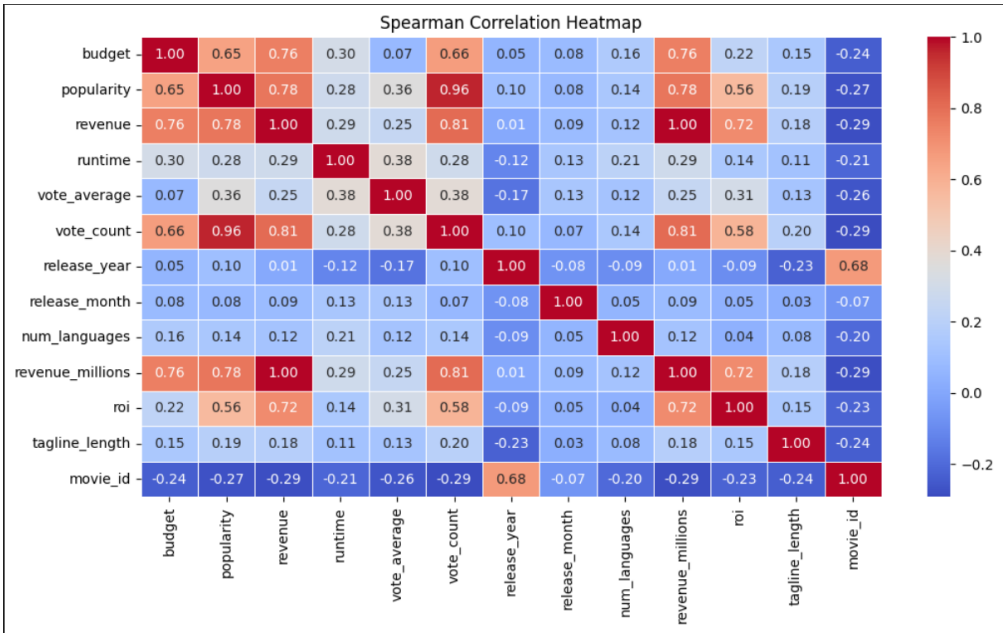


Figure 4: Spearman Correlation Heatmap

2.8 Partial, Semi-Partial Correlation and Normality Checks

2.8.1 Partial Correlation: Vote Average and Revenue Controlling for Budget

To examine the relationship between *vote_average* and *revenue* while controlling for *budget*, we performed a partial correlation analysis. The resulting correlation coefficient was found to be $r = 0.293$.

Interpretation:

- Since $r = 0.293$ is positive, it suggests that higher *vote_average* is associated with higher *revenue*, even after accounting for *budget*.

- However, the correlation is moderate, indicating only a partial dependency.
- The confidence interval is narrow, making the estimate reliable.
- The p-value is extremely low, indicating high statistical significance and ruling out randomness.

2.8.2 Semi-Partial Correlation

A semi-partial correlation analysis was also conducted, yielding a correlation coefficient of $r = 0.2466$. This suggests that when only the unique contribution of *budget* is removed from *vote_average*, the correlation with *revenue* slightly decreases but remains statistically significant.

2.8.3 Normality Checks

To assess the normality of the dataset, we conducted the Kolmogorov-Smirnov (KS) test and Anderson-Darling test. Both tests indicated that the data is **not normally distributed**.

Additionally, we visualized normality using Q-Q plots for all numerical features, revealing that:

- Most features exhibit **strong right or left skewness**.
- Some features follow an **S-shaped distribution**, deviating significantly from normality.

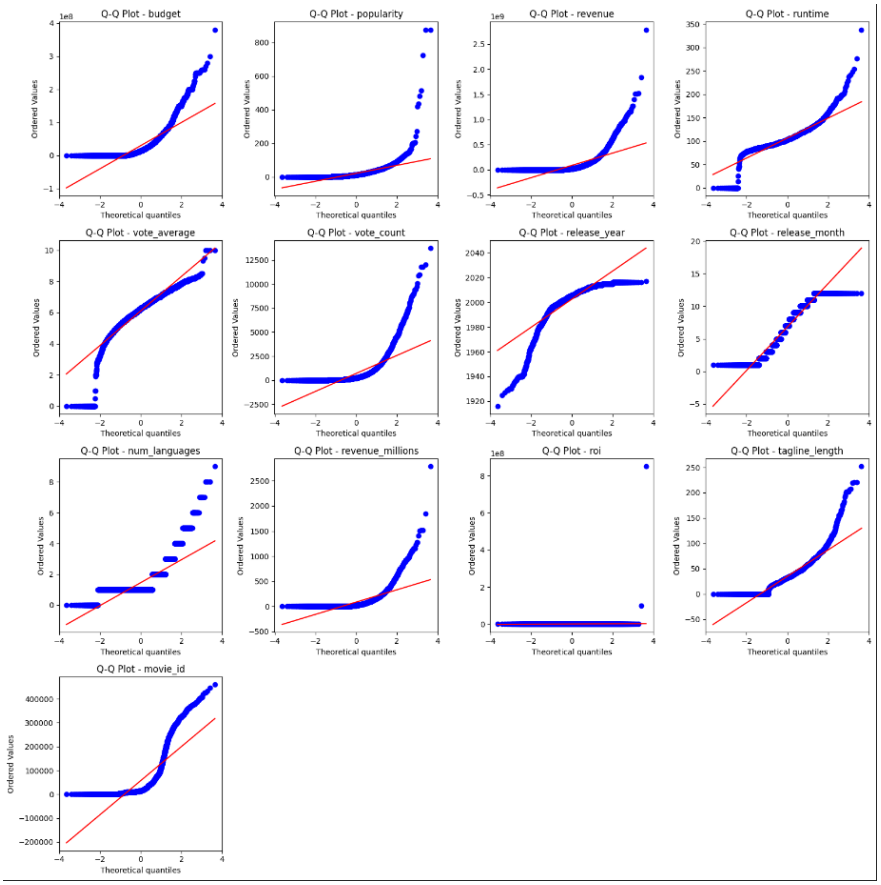


Figure 5: Q-Q plot of Numerical features

These findings suggest that transformations or non-parametric methods may be required for further statistical modeling.

2.9 Distribution of Numerical Features

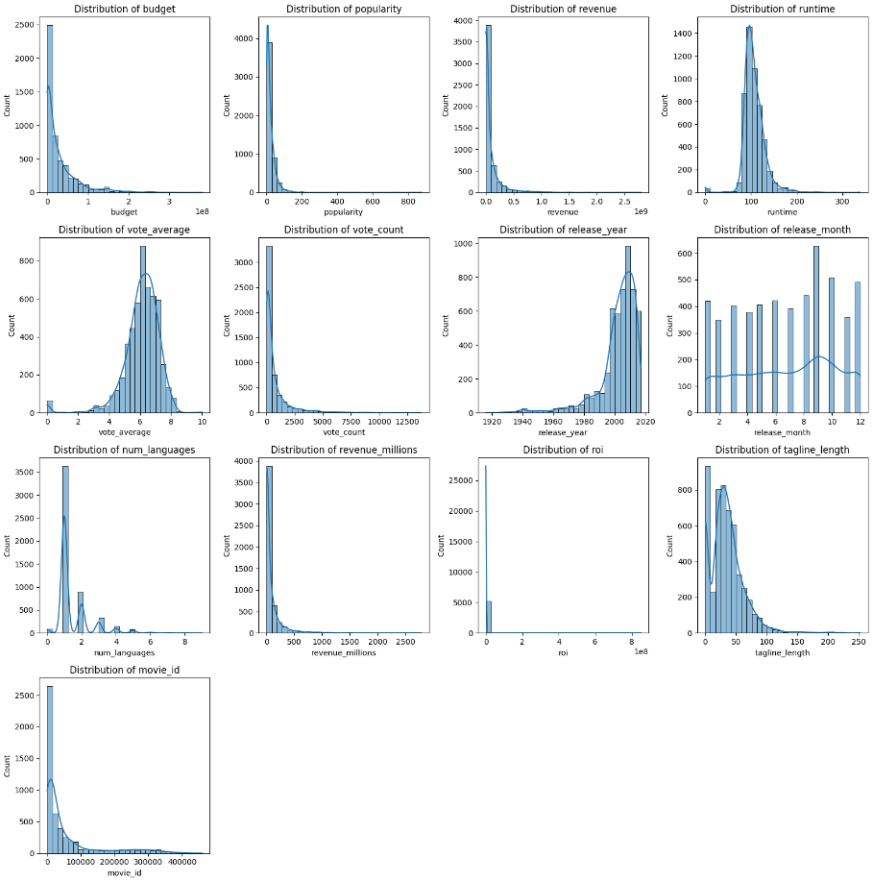


Figure 6: Distribution of Numerical features

2.10 Revenue vs. Budget Analysis

Figure 7 shows the relationship between revenue and budget.
Observations:

- Most high-budget movies have high revenue.
- Some movies like 5 in number with high budget but very less revenue like flopped movies
- Many movies have both low budget and low revenue.
- Can see at top right corner some movies with big circles like too much popular and also giving high revenue although of high budget

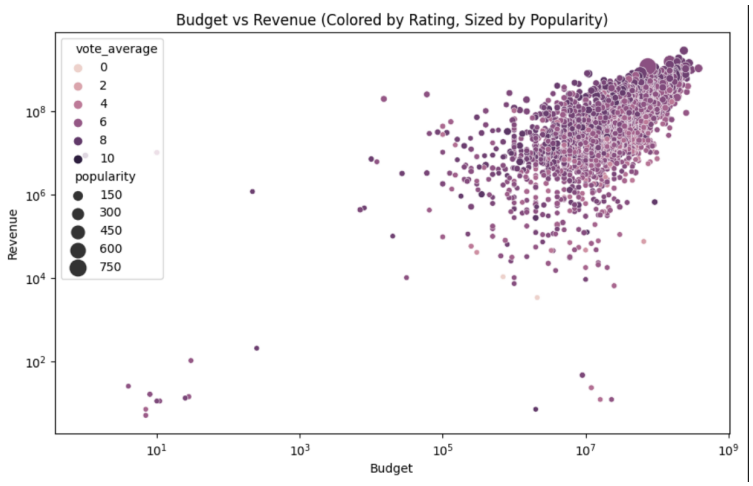
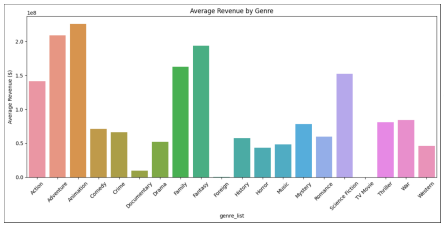


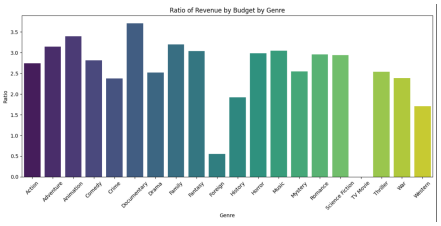
Figure 7: Revenue vs. Budget Analysis

3 Questions

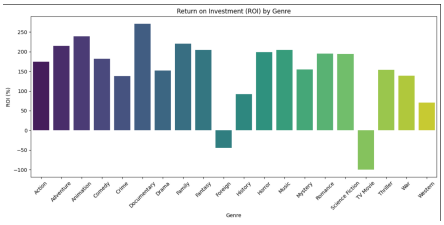
3.1 Q1: What genres of movies have the highest average revenue?



(a) Average revenue by Genre



(b) Ratio of Revenue to budget by Genre



(c) ROI on Genres

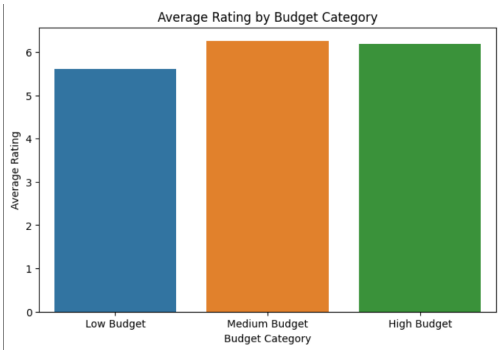
Figure 8: Plots of Question 1

3.1.1 Observations:

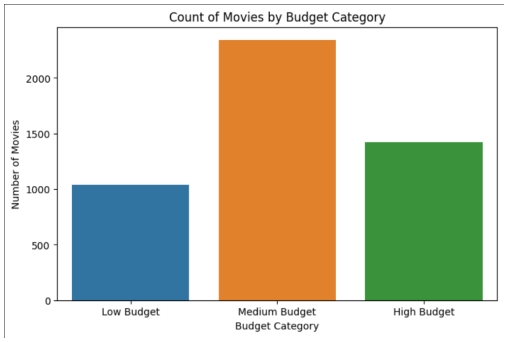
- Animation, Adventure, Fantasy, Family, and Science Fiction have the highest average revenue.
- Drama is the most occurring genre in all movies being in 2297 movies.
- Animation is only in 234 movies like very less movies but still having highest average revenue.
- Foreign and TV Movie are lowest revenue making genres.

- Documentary has highest revenue to budget ratio of more than 3.5 followed by Animation, Adventure and Family and all have atleast ratio more than 2 except Foreign which has ratio near to 0.5.
- On ROI plot, Documentary has highest among all more than 250% followed by Animation. Foreign and TV Movie has negative ROI thus movies will be in loss.

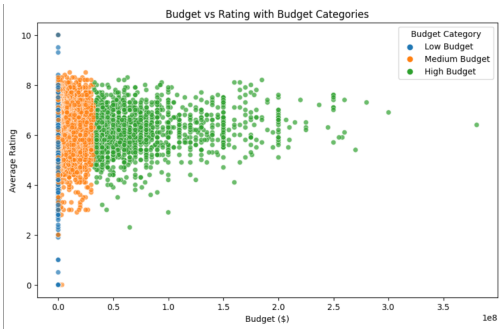
3.2 Q2: Is there a correlation between movie budgets and Ratings?



(a) Average rating by budget category



(b) Count of Movies by category



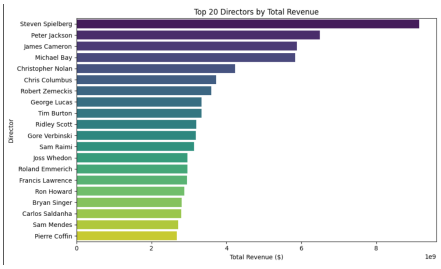
(c) Budget vs Rating by Category

Figure 9: Plots of Question 2

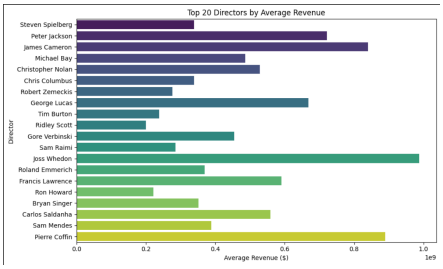
3.2.1 Observations:

- No strong correlation exists between budget and ratings.
- Very less positive correlation between them suggesting no correlation as it might be there that some high budget movies goes flop while some low budget movies like animation becomes record breaking with good rating

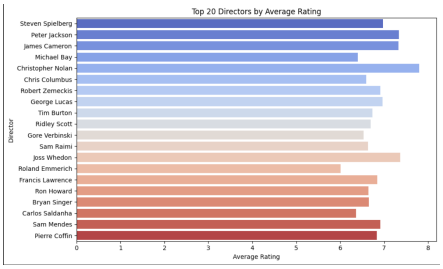
3.3 Q3: Which directors have consistently high box office success?



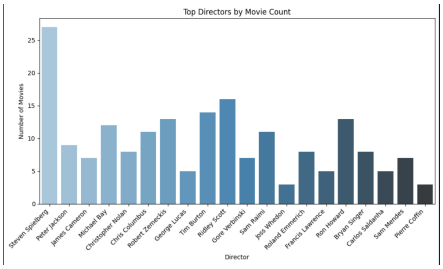
(a) Top 20 Directors by Total Revenue



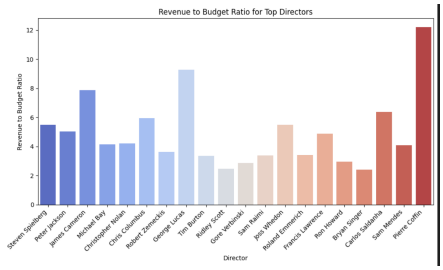
(b) Top 20 Directors by Average Revenue



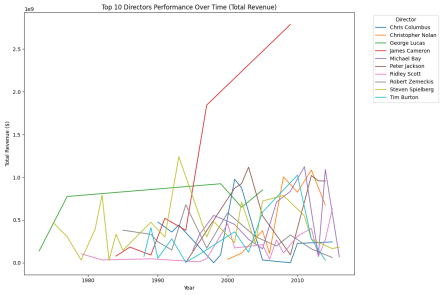
(c) Top 20 Directors by Average Rating



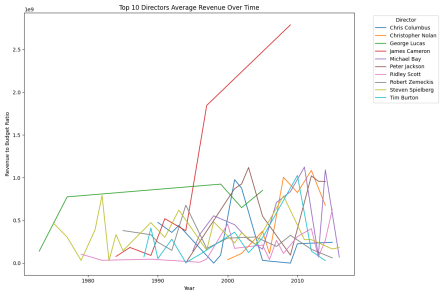
(d) Top 20 Directors by Movie Count



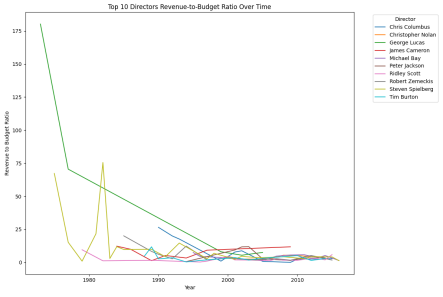
(a) Top 20 Directors by Revenue to Budget ratio



(b) Top 10 Directors Performance over Years (Total revenue)



(c) Top 10 Directors Performance over Years (Average revenue)



(d) Top 10 Directors Performance over Years (Revenue to Budget ratio)

Figure 10: Plots of Question 3

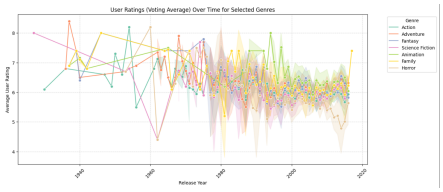
3.3.1 Observations:

- Steven Spielberg, Peter Jackson, and James Cameron have the highest total revenue.
- Joss Whedon, Pierre Coffin, James Cameron, Peter Jackson, and George Lucas have the highest average revenue.
- Christopher Nolan, Joss Whedon, Peter Jackson, and James Cameron have the highest average ratings as people loved their movies the most.
- Steven Spielberg has the highest movie count of more than 25 movies.
- Pierre Coffin, George Lucas, and James Cameron have the highest average revenue/budget ratio varying between 8-12.
- In terms of total revenue from 1970s to 2020s, James Cameron showed the greatest growth till now in his career and same with Christopher Nolan. Steven Spielberg peaked two times in 1982 and 1992 and Peter Jackson in 2003 , Chris Columbus in 2002 and Michael Bay in 2010.
- In terms of average revenue with time, James Cameron ,George Lucas, Tim Burton and Christopher Nolan showed consistency in growth
- In terms of Revenue-Budget ratio, all directors showed decline in the ratio due to coming inflation and more money to make movies nowdays like more budget as more advanced technologies, VFX and high salaries taking casting but James Cameron showed consistency in its ratio which is increasing gradually.
- So overall James Cameron , Christopher Nolan , Ridley Scott, Peter Jackson and George Lucas are top directors having consistently high box office success

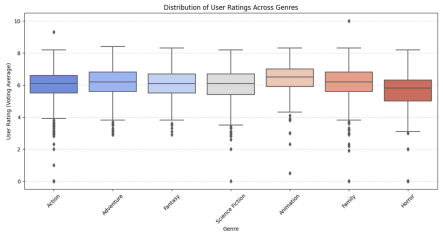
3.4 Q4: How do user ratings compare between movies in the same genre but different production years?

3.4.1 Observations:

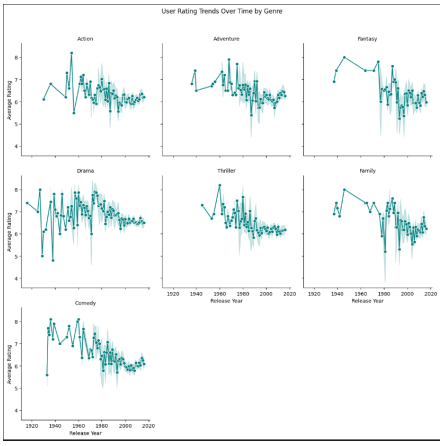
- Overall rating decline from 1920 to 2020.
- Films from 1930-1950 have higher ratings.
- Ratings converged since the 1980s.
- Animation has the highest median rating and least volatility.
- Horror has the lowest median rating and highest variability.
- Modern films have increased rating variance.
- Thriller shows the steepest decline over time.
- Drama has consistent ratings.
- Fantasy has dramatic fluctuations.
- Modern films frequently receive low ratings.
- Perfect 10 ratings appear more frequently in recent decades.



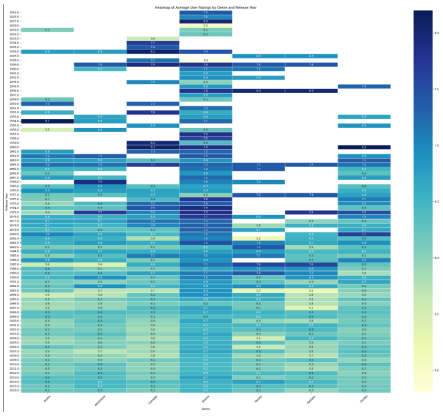
(a) User Ratings Over Years For Genres (Voting Average)



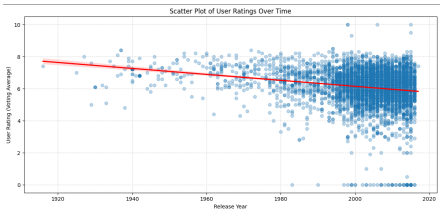
(b) Distribution of User Ratings Across Genres



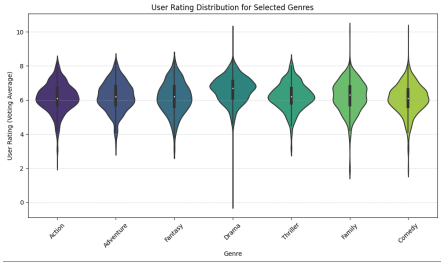
(c) User Rating Trend Over Time by Genre



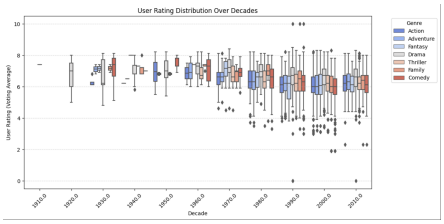
(d) Average user rating Over Genre and Year Heatmap



(a) Scatter Plot of User ratings Over Time



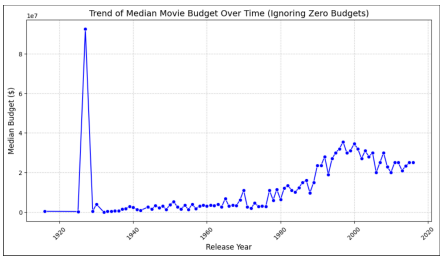
(b) BoxPlot of User Rating for Selected Genres



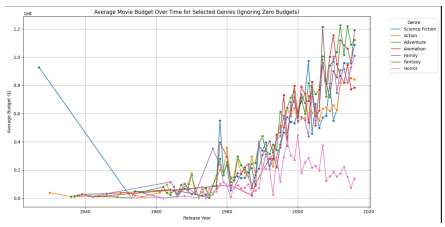
(c) User Rating Distribution over Decades

Figure 11: Plots of Question 4

3.5 Q5: How has the average movie budget changed over time?



(a) Trend of Median Movie Budget over Release Years



(b) Average Movie Budget over Time for selected Genres

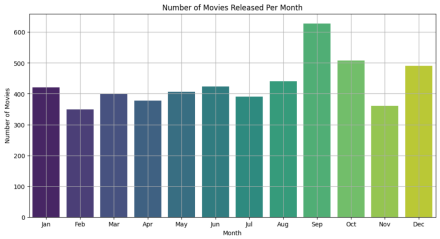
3.5.1 Observations:

- There was a dramatic spike in median movie budgets around 1925, reaching nearly $90 million (adjusted for inflation). This could be due to the transition from silent films to "talkies" requiring significant technological investment.
- From 1930 to 1970, median budgets remained relatively stable and low, hovering around $5-10 million.
- Starting from 1980, there's been a steady upward trend in median budgets, rising from about $10 million to stabilizing around $20-30 million in the 2000s-2010s.
- Science Fiction had high budgets in the 1930s around $90 million that declined sharply through the 1940s-50s, before rising again from the 1980s onward.
- Fantasy, Adventure, Animation, and Science Fiction have high budgets since 2000 frequently exceeding $100 million.
- Horror has consistently low budgets rarely exceeding $40 million even in recent years.
- Fantasy and Adventure films have become the most expensive to produce in recent years, with budgets often reaching $120 million.
- Technological advancements, competition, marketing costs, globalization, and star power increase budgets.

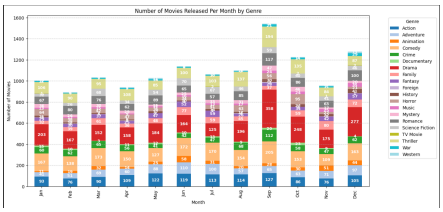
3.6 Q6: Which months are most popular for movie releases?

3.6.1 Observations:

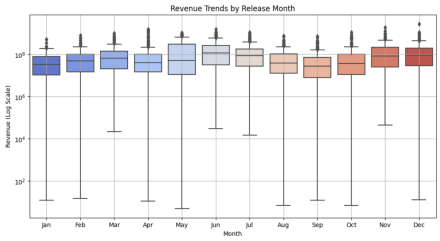
- September has the highest number of releases, with approximately 1,500 movies. This coincides with the fall film festival season (Venice, Toronto, etc.) and the beginning of awards season.
- December shows the second-highest number of releases (around 1,250 movies), which aligns with the holiday season and Oscar-qualifying releases.
- June and July also show strong numbers (approximately 1,100-1,150 movies each), corresponding to the summer blockbuster season.
- Drama dominates releases across most months.
- Action and Comedy show higher numbers in summer.
- Release numbers have increased substantially since the 1970s.



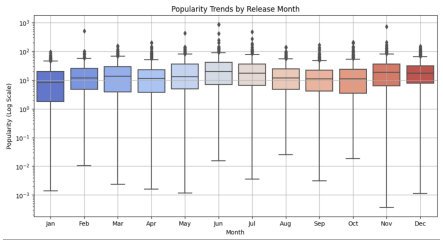
(a) Number of Movies Released Per Month



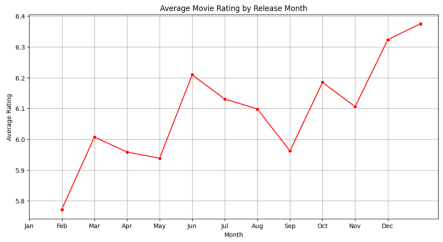
(b) Number of Movies Released Per Month by Genre



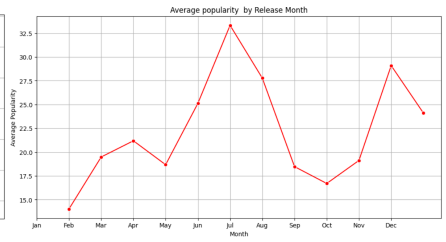
(c) Revenue Trends by Release Month



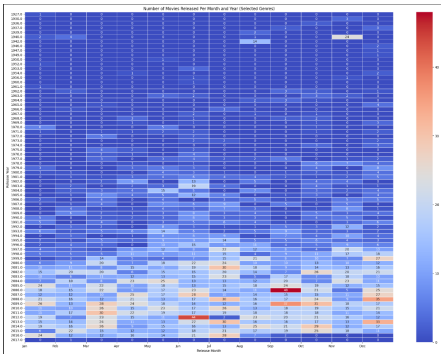
(d) Popularity Trends by Release Months



(e) Average Movie Rating by Release Month



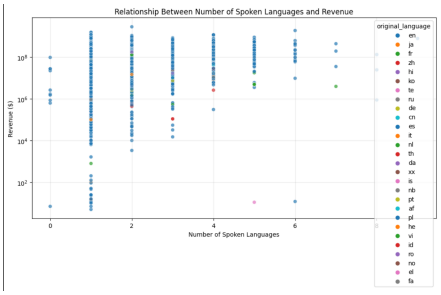
(f) Average Popularity by Release Month



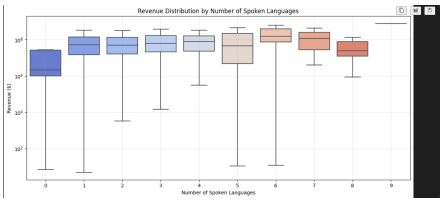
(g) Number of Movies Released Per Month and Year for Selected Genres Heatmap

- The number of monthly releases has increased substantially from the 1920s-1970s (0-5 movies per month) to the 2000s-2010s (15-40 movies per month).
- Strategic scheduling based on audience availability, award timing, and viewing habits.

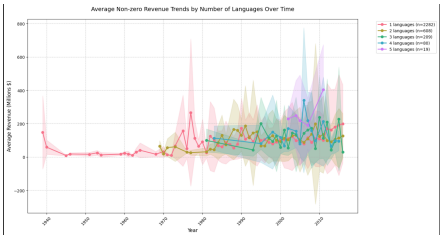
3.7 Q7: How does the number of spoken languages correlate with international revenue?



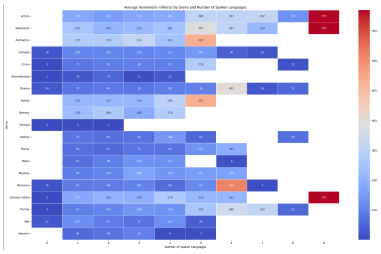
(a) Number of spoken languages vs Revenue



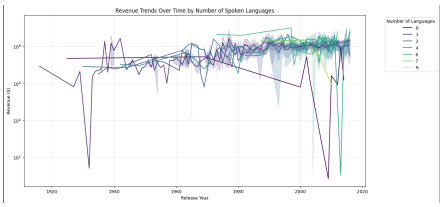
(b) Boxplot of Revenue vs No. of Spoken languages



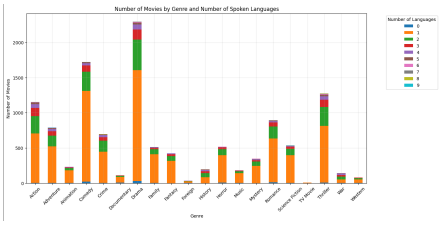
(c) Average Revenue vs Number of Languages over Time



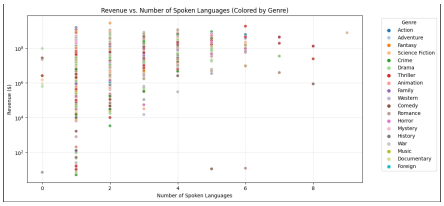
(d) Average Revenue by Number of Spoken Languages and Genres



(e) Revenue Trends Over Time by Number of Spoken Languages



(f) Stacked Bar chart for No. of Movies by Genre and Spoken languages

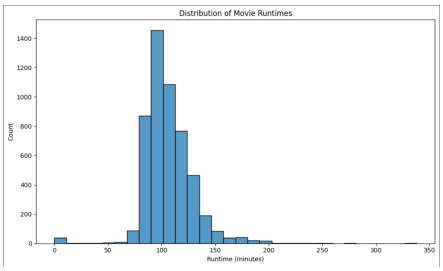


(g) Revenue vs Spoken languages by Genre

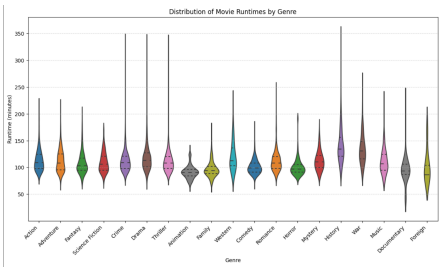
3.7.1 Observations:

- Movies with 4-6 languages have higher median revenues (around $100M-$200M).
- Action, Adventure, and Sci-Fi films with 9 languages show the highest average revenue ($770M)
- English dominates as the original language.
- Most movies are made in 1-2 languages.
- Since 1970, movies with 3-5 languages have shown increasingly stable revenue performance.
- Multilingual films have increased revenue since 1990.
- 4-6 languages appear to be the "sweet spot" for maximizing revenue potential."
- Big-budget genres (Action, Adventure, Sci-Fi) benefit most from multiple languages.
- Risk vs. reward: adding languages increases revenue but also variability.

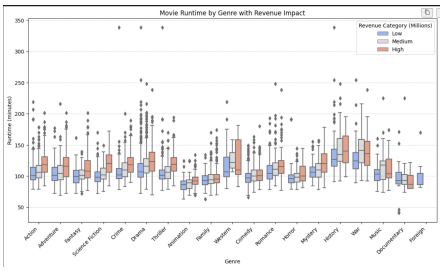
3.8 Q8: What's the distribution of movie runtimes by genre?



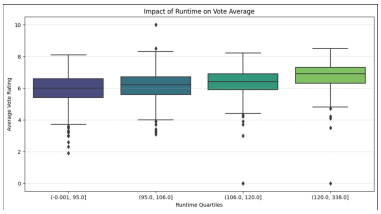
(a) Number of spoken languages vs Revenue



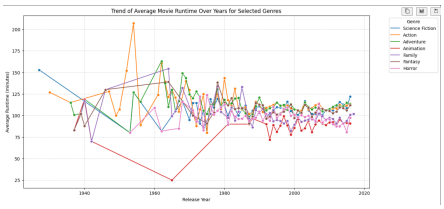
(b) Boxplot of Revenue vs No. of Spoken languages



(c) Average Revenue vs Number of Languages over Time



(d) Average Revenue by Number of Spoken Languages and Genres

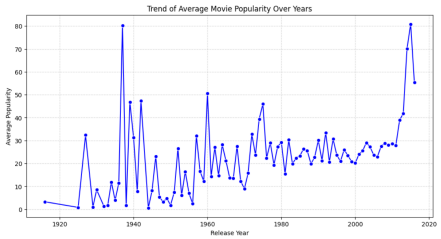


(e) Revenue Trends Over Time by Number of Spoken Languages

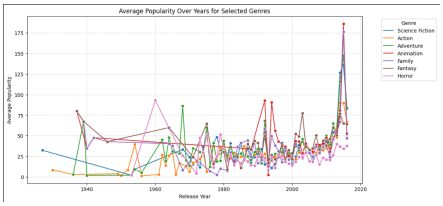
3.8.1 Observations:

- Most movies are between 90-120 minutes, with the peak around 100 minutes.
- Animation and Family films tend to be the shortest (medians around 90 minutes).
- History, War, and Crime genres genres have the longest typical runtimes.
- Higher-revenue films tend to have longer runtimes.
- Longer films receive higher average ratings.
- Animation shows the most dramatic change, with very short runtimes in the 1960s.
- Runtimes have generally decreased since the 1960s.
- Runtimes have stabilized since the 1980s between 90-110 minutes.

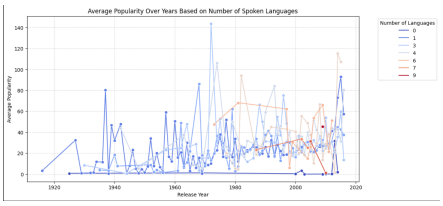
3.9 Q9: Is there a trend in movie popularity over the years?



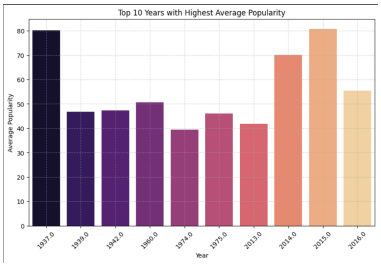
(a) Line Plot of Average Popularity over Time



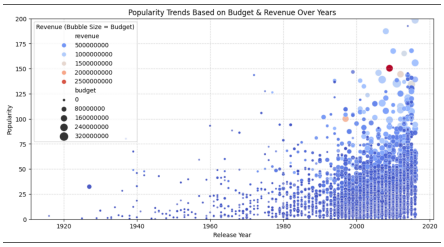
(b) Line Plot of Average Popularity over Time for Selected Genres



(c) Line Plot of Average Popularity over Time on No. of spoken languages



(d) Top 10 Years with Highest Average Popularity

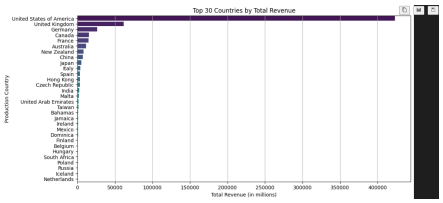


(e) Popularity Trends based on Budget and Revenue over Years

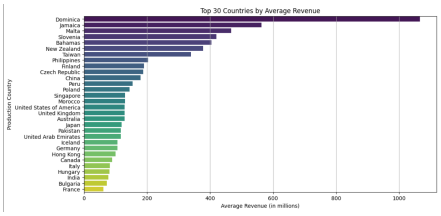
3.9.1 Observations:

- Significant popularity spikes around 1940, 1960, and 2015-2018.
- Baseline popularity has gradually increased from 1920 to 2020.
- 2010s show the most dramatic upward trend in average popularity.
- Animation, Horror, and Science Fiction have genre-specific trends.
- Movies with more spoken languages (4-6 languages) generally achieved higher popularity, especially after 1980.
- 1937, 2014, and 2015 stand out as particularly strong years for movie popularity.
- Higher budgets and multilingual productions correlate with popularity.
- Movie popularity has dramatically increased in the digital era.
- There were notable golden ages around 1940 and 2015.

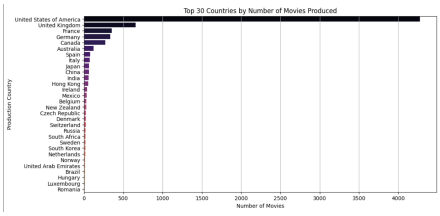
3.10 Q10: How do production countries affect global revenue?



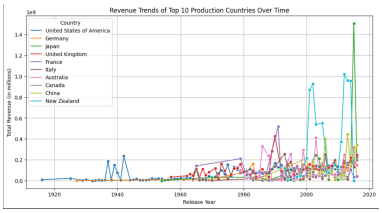
(a) Top 30 countries by Total Revenue



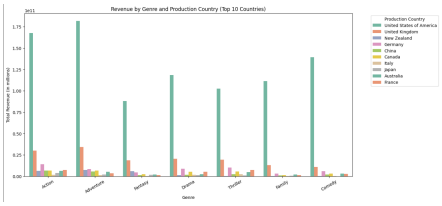
(b) Total 30 countries by Average Revenue



(c) Top 30 countries by Number od Movies Produced



(d) Revenue Trends of Top 10 Production Countries Over Time

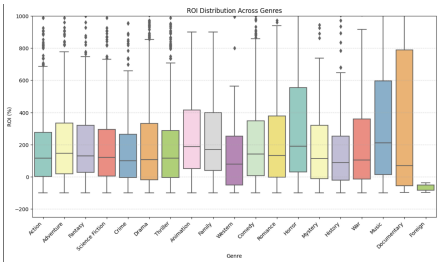


(e) Revenue by Genre and Production Country (Top 10)

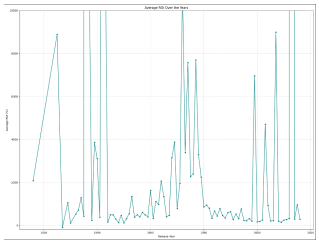
3.10.1 Observations:

- The US dominates global movie revenue.
- Smaller countries have higher average revenue per film like Dominica and Jamaica.
- The US produces significantly more movies followed by UK.
- All countries show revenue growth since the 1990s-2000s.
- Japan shows an exceptional revenue spike around 2015-2020.
- The US has consistently led in revenue throughout film history.
- Adventure films generate the highest revenue for the US.
- Film industry infrastructure correlates with revenue.

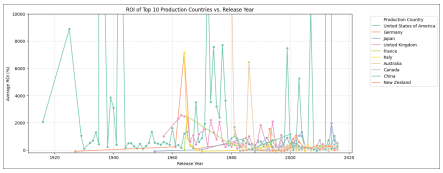
3.11 Q11: What's the ROI (Return on Investment) distribution across genres?



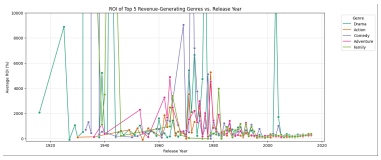
(a) Box plot of ROI distribution across Genres



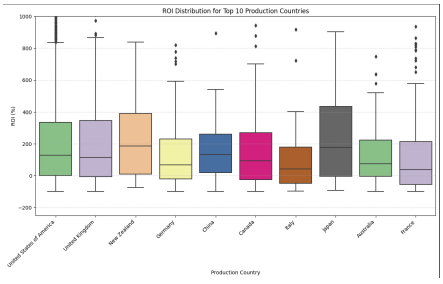
(b) Line Plot of Average ROI over Time



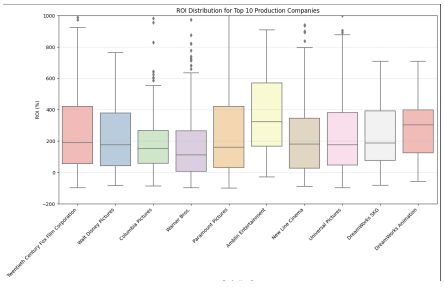
(c) Line Plot of ROI of top 10 Production Countries over Year



(d) Line Plot of ROI of top 5 Revenue Generating Genres vs Year



(e) Box Plot of ROI of top 10 Production Countries

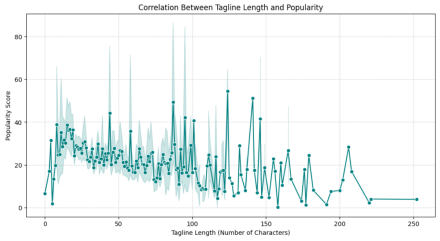


(f) Box Plot of ROI of top 10 Production Companies

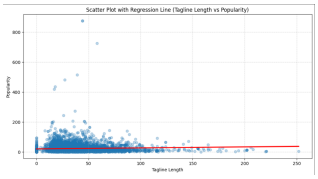
3.11.1 Observations:

- Documentary and Music have the highest median ROI followed by Horror.
- Foreign films have the lowest and most constistent ROI distribution.
- Horror has high median returns.
- The US dominates high ROI productions historically.
- Japan has the highest median ROI among countries.
- Amblin Entertainment has higher median ROI among companies.
- Warner Bros. shows significant outlier successes despite a moderate median ROI.
- ROI volatility has decreased since the 1980s.
- Trend toward lower, more stable ROI productions.

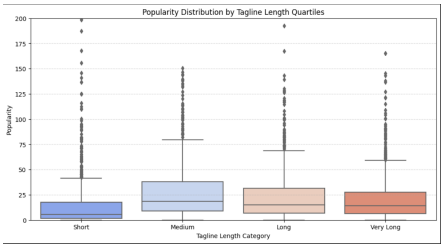
3.12 Q12: Is there a correlation between movie tagline length and popularity?



(a) Correlation between Tagline and Popularity



(b) Scatter Plot with regression line (Popularity vs Tagline)



(c) Popularity Distribution by Tagline length Quartiles

3.12.1 Observations:

- Downward trend in popularity as tagline length increases.
- Highest popularity in the 10-60 character range.
- Medium-length taglines show the highest median popularity.
- Short taglines have the lowest median popularity.
- Negative correlation but not strong.
- Optimal tagline length is in the medium range.