Predictors of Movie Box Office Success: An Empirical Analysis of Factors Influencing Film Revenue

Uday Bhaskar Valapadasu – 11696364 | Rohit Suddala - 11652459 | Sapthagiri Naik Bhukya – 11699072

ABSTRACT

This project analyzes factors influencing movie box office success using a comprehensive dataset of over 5000 films. In today's data-driven entertainment industry, understanding the determinants of a movie's commercial performance is crucial for stakeholders. Our study examines 12 independent numerical variables to predict the dependent variable: gross box office revenue. Key independent variables include production budget, social media metrics, critical reviews, user ratings, genre, and release year. The research employs statistical analysis and machine learning techniques to explore relationships between these film attributes and financial performance. Our methodology involves exploratory data analysis, data preprocessing, correlation analysis, and predictive model development. We aim to identify and quantify the most influential factors determining box office success, potentially revealing insights into changing success dynamics over time, social media impact, and the importance of critical acclaim versus audience reception. This research can inform decision-making in film production, marketing, and investment, offering valuable insights into evolving audience preferences and the changing landscape of the global film market.

1 OVERVIEW

This project aims to identify and analyze key predictors of movie box office success using a comprehensive dataset of over 5000 films. Through rigorous statistical analysis and machine learning techniques, we seek to develop a predictive model for box office revenue and provide actionable insights for the film industry.

To accomplish this, we will analyze 12 independent numerical variables that we believe will be predictive of a movie's box office performance. These variables include:

1. num_critic_for_reviews
2. duration
3. director_facebook_likes
4. actor_3_facebook_likes
5. actor_1_facebook_likes
6. num_user_for_reviews
7. budget
8. title_year
9. actor_2_facebook_likes
10. imb_score
11. aspect_ratio
12. movie_facebook_likes

Goals and Objectives:

1. Identify key factors that significantly influence a movie's box office performance.
2. Develop a predictive model for estimating a film's potential box office revenue.
3. Analyze trends in movie success factors over time.
4. Evaluate the relative importance of critical reception versus audience engagement.
5. Assess the impact of social media presence on financial performance.

Steps to Achieve These Goals:

1. **Data Preparation and Cleaning:**
    a. Import the dataset containing 12 independent numerical variables for over 5000 movies.
    b. Handle missing values and outliers.
    c. Encode categorical variables (e.g., one-hot encoding for genres, ordinal encoding for content ratings).
    d. Normalize or standardize numerical variables as needed.
2. **Exploratory Data Analysis (EDA):**
    a. Conduct univariate analysis for each variable (histograms, box plots).
    b. Perform bivariate analysis to explore relationships between variables and box office gross (scatter plots, correlation matrices).
    c. Visualize trends over time for key variables.
3. **Statistical Analysis:**
    a. Conduct correlation analysis to identify significant relationships between variables.
    b. Perform hypothesis testing to validate key relationships.
    c. Use ANOVA to analyze the impact of categorical variables on box office performance.
4. **Model Development:**
    a. Split the data into training and testing sets.
    b. Develop multiple linear regression models to predict box office gross.
    c. Explore advanced techniques like Random Forests or Gradient Boosting for comparison.
    d. Evaluate models using metrics such as R-squared, RMSE, and MAE.
5. **Interpretation and Insights:**
    a. Analyze feature importance to identify the most influential predictors.
    b. Examine model coefficients to understand the direction and magnitude of each factor's impact.
    c. Investigate any unexpected or counterintuitive findings.

Expected Outcomes:

1. A comprehensive understanding of factors influencing movie box office success.
2. A predictive model capable of estimating a movie's potential box office revenue.
3. Insights into the changing dynamics of movie success factors over time.
4. Quantification of the relative importance of various factors (e.g., budget vs. social media presence).
5. Actionable recommendations for stakeholders in the film industry.

Project Schedule:

- July 11 – July 15 : Data preparation and initial EDA
- July 15 – July 18: Comprehensive statistical analysis
- July 18 – July 20: Model development and refinement
- July 20 – July 22: Results interpretation and report writing

Budget: This project will be completed using existing computational resources and open-source software, incurring no additional costs.

Necessary Resources:

- Computational resources: Personal computers with sufficient RAM and processing power
- Software: Python, Jupyter Notebooks, libraries (pandas, numpy, scikit-learn, matplotlib, seaborn)
- Dataset: Existing movie dataset with 12 independent numerical variables and 1 dependent variable
- Team: 3 group members, each responsible for analyzing at least 3 variables

Status: Project is in the initial planning phase, with data acquisition complete and preliminary EDA underway.

3 CRITICS

While numerous studies have examined factors influencing movie box office success, our research addresses several key gaps in the existing literature:

1. **Social Media Impact:** Many earlier studies predate the widespread use of social media. Our dataset includes Facebook likes for directors, actors, and movies, allowing us to quantify the impact of social media presence on box office performance - a factor increasingly relevant in today's digital age.

- **Comprehensive Variable Set:** Previous research often focused on a limited set of variables. Our study incorporates a wide range of factors (12 independent numerical variables and 1 dependent variable), providing a more holistic view of what drives box office success.

2. **Temporal Analysis:** By including movies from various years, we can analyze how predictors of box office success have evolved over time, addressing a gap in understanding the changing dynamics of the film industry.
3. **Integration of Critical and Audience Reception**: Unlike studies that focus solely on either critical reception or audience ratings, our research incorporates both (critic reviews, user reviews, and IMDb scores), allowing for a comparative analysis of their relative importance.
4. **Visual Marketing Impact:** The inclusion of 'facenumber_in_poster' allows us to explore the understudied area of how visual marketing elements affect a movie's commercial success.
5. **Global Perspective:** By including language and country variables, we can analyze how these factors influence global box office performance, addressing the gap in understanding international market dynamics.
6. **Machine Learning Approach:** While traditional statistical methods have been commonly used, our study employs modern machine learning techniques, potentially uncovering complex, non-linear relationships between variables that previous studies might have missed.

By addressing these gaps, our study aims to provide a more current, comprehensive, and nuanced understanding of the factors driving box office success in today's rapidly evolving film industry landscape.

4 GOALS

- **Performance Goals:** a. Develop a predictive model for box office revenue using the Kaggle dataset, aiming to achieve an R-squared value of at least 0.7, indicating that our model explains at least 70% of the variance in box office performance. b. Identify the top 5 most influential factors affecting box office success from the 12 independent numerical variables and 1 dependent variable provided in the dataset, using correlation analysis and multiple regression techniques. c. Construct 95% confidence intervals for mean box office revenue for different categories of movies based on the categorical variables in the dataset (e.g., genre, content rating, or release year). d. Perform hypothesis tests to determine if there are statistically significant differences in box office performance based on key factors in the dataset, with a significance level of 0.05. e. Create a probability model using the normal distribution to estimate the likelihood of a movie surpassing specific box office revenue thresholds.

1. **Time Goal**: Complete the project within the course deadline of [insert your project deadline], with specific milestones for data preprocessing, exploratory data analysis, model development, and final report preparation.
2. **Resource Goals:** Utilize existing computational resources and open-source statistical software (Python with libraries such as NumPy, Pandas, SciPy, and Statsmodels) to complete the project without incurring additional costs. Effectively distribute the workload among the 3 team members, with each member responsible for analyzing at least 5 variables from the Kaggle dataset. Fully leverage the Kaggle dataset containing information on over 5000 movies, ensuring comprehensive analysis of all provided variables.

The degree of accomplishment for these goals will be verifiable through:

- The R-squared value of the final predictive model
- Statistical outputs from hypothesis tests and confidence interval calculations
- A ranked list of influential factors based on statistical analysis
- Adherence to the project timeline and course deadline
- Documentation of resources used throughout the project
- Comprehensive analysis of all variables provided in the Kaggle dataset

These goals incorporate the statistical concepts from the provided materials, including probability distributions, hypothesis testing, confidence intervals, regression analysis, and data visualization. They are tailored to the specific Kaggle dataset you're using, focusing on the variables provided and the insights that can be derived from them. The goals also specify measurable outcomes within the constraints of your course project.

4 SPECIFICATIONS

4.1 Tools

The entire project will be implemented in Python. For data manipulation and analysis, we will use pandas and numpy. For visualization, we have decided to use matplotlib and seaborn as they are flexible, fast, and easy to use. We will also use scikit-learn for machine learning models, statsmodels for statistical analysis, and potentially XGBoost for gradient boosting.

4.2 Dataset

Our dataset, obtained from Kaggle, contains information on over 5000 movies. Each movie entry includes 12 independent numerical variables (such as num_critic_for_reviews, duration, director_facebook_likes, etc , as well as the gross box office revenue as the target variable. The data spans multiple years, allowing for analysis of trends over time.

4.3 Implementation

4.3.1 Objectives Our research has two main objectives:

1. Perform comprehensive EDA and hypothesis testing to discover patterns in the data that will be insightful for building a reliable machine learning model for box office revenue prediction.
2. Develop and compare multiple predictive models to find the best balance between accuracy and computational efficiency.

4.3.2 Constraints The main constraints in this project are:

1. Time required to perform extensive dataset analysis within the course deadline.
2. Potential multicollinearity among independent variables, particularly those related to social media metrics.
3. The dataset may not include some potentially important variables (e.g., marketing budget).

4.3.3 Needs Given the aforementioned constraints and objectives, we need:

1. A working Python environment with necessary libraries installed (pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels, XGBoost).
2. Sufficient computational power to handle data analysis and model training for a dataset of 5000+ entries with 24+ variables.
3. Access to statistical resources for proper interpretation of results.

4.3.4 Budget This project has no additional funding requirements. We will use existing computational resources and open-source software.

4.3.5 Deadline The deadline for the project is [insert your project deadline].

4.3.6 Implementation management plan Our implementation will follow these steps:

1. **Data Preprocessing:** Clean the data, handle missing values, encode categorical variables.
2. **Exploratory Data Analysis:** Analyze distributions, correlations, and relationships between variables.
3. **Feature Engineering:** Create new features if necessary, such as combining social media metrics or creating interaction terms.
4. **Model Development:** Start with multiple linear regression as a baseline, then explore more advanced techniques like Random Forests and Gradient Boosting.
5. **Model Evaluation:** Use cross-validation to assess model performance, comparing R-squared, MAE, and RMSE.
6. **Feature Importance Analysis:** Determine which factors have the strongest influence on box office revenue.
7. **Temporal Analysis:** Investigate how the importance of different factors has changed over time using the 'title_year' variable.
8. **Final Model Selection and Interpretation:** Choose the best performing model and interpret its results in the context of the film industry.

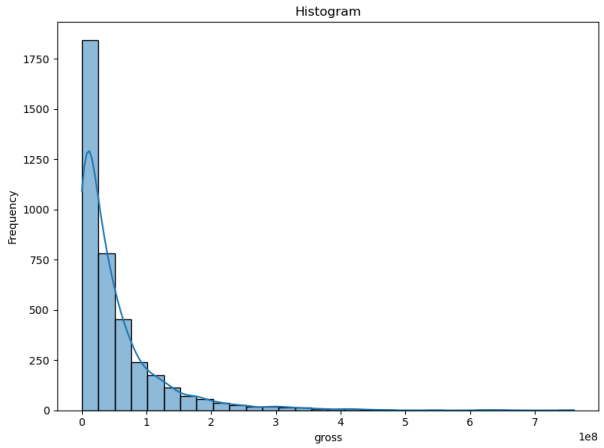
This implementation plan will allow us to thoroughly analyze the dataset, develop robust predictive models, and derive meaningful insights about the factors influencing movie box office success.

5 EXPLORATORY DATA ANALYSIS

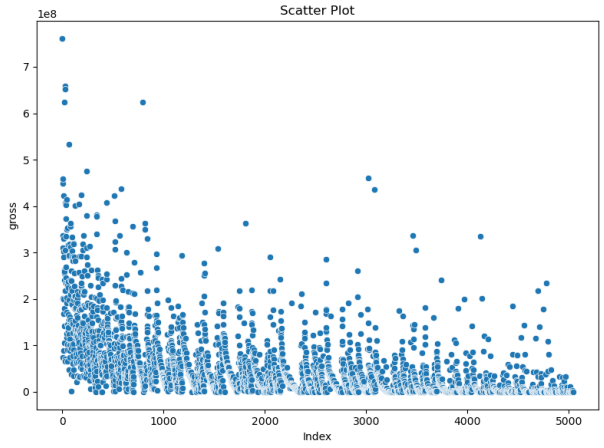
We will conduct EDA on both our independent and dependent variables to better understand their distributions and relationships. This will help inform our feature selection and model development processes.

5.1 Dependent Variable: Gross Revenue

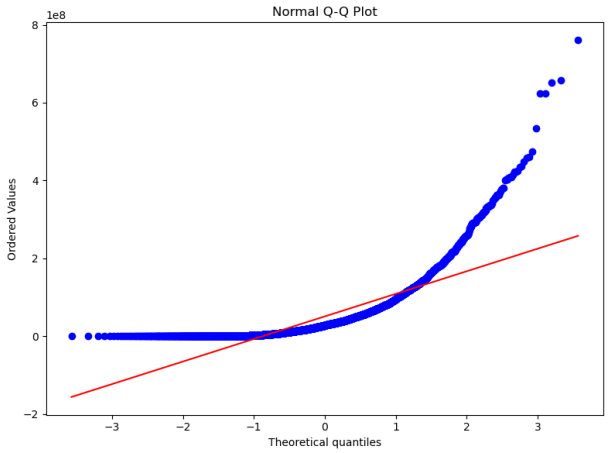
1. Histogram of Gross Revenue: The histogram for the gross revenue demonstrates a right-skewed distribution. The majority of the movies earn relatively low gross revenues, with a steep drop-off as revenue increases. The distribution is unimodal, with a significant concentration of data points on the lower end, indicating that most movies do not earn high gross revenues. This right skewness indicates that while a few movies make very high revenues, the majority make much less.



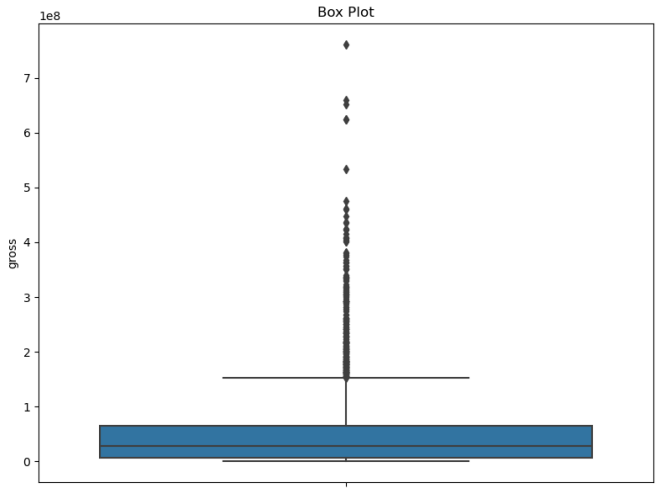
2. Scatter Plot of Gross Revenue: The scatter plot illustrates the gross revenue values across the dataset's index. It shows that the majority of movies cluster towards the lower end of the revenue spectrum. However, there are several high-revenue outliers that stretch the vertical scale. This plot reinforces the histogram's indication of a highly skewed distribution.



3. Normal Q-Q Plot of Gross Revenue: The Q-Q plot for gross revenue reveals a significant deviation from the normal distribution. The data points veer off from the straight reference line, especially in the upper tail, indicating heavy skewness and the presence of outliers. The curvature away from the line confirms that the gross revenue does not follow a normal distribution.



4. Box Plot of Gross Revenue: The box plot provides a visual summary of the distribution, highlighting the interquartile range, median, and outliers. The median is situated at a lower value, reflecting that half of the data points are below this mark. The numerous dots above the whisker indicate outliers—movies that have gross revenues significantly higher than the rest. This further supports the observation of a highly skewed distribution with many low-revenue movies and a few high-revenue outliers.



Parameter Estimation for Gross Revenue:

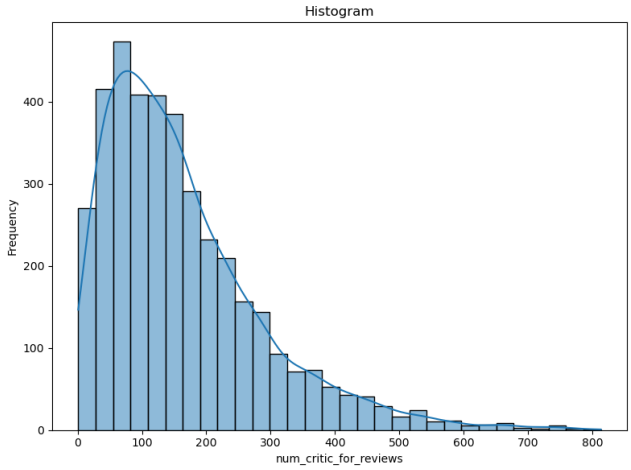
- **Mean:** The mean gross revenue is approximately $51,054,995.24, which is influenced by the high-revenue outliers.
- **Standard Deviation:** The standard deviation is approximately $69,802,484.42, indicating high variability in movie gross revenues.
- **Proportion:** The proportion of records with non-missing gross revenue values is 1.0, meaning all entries in the dataset have valid gross revenue values.

This comprehensive analysis of the gross variable shows a clear picture of the distribution, variability, and the presence of significant outliers. The skewness and high variability highlight the disparity in movie earnings, with a few blockbuster hits earning significantly more than the majority.

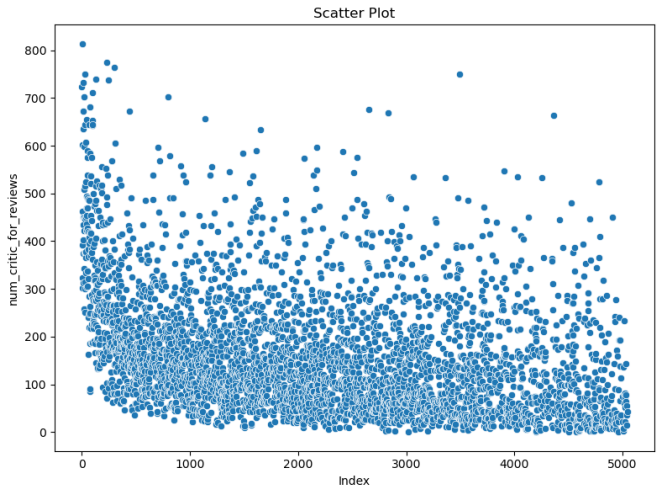
5.2 Independent Variables

5.2.1 Number of Critic Reviews

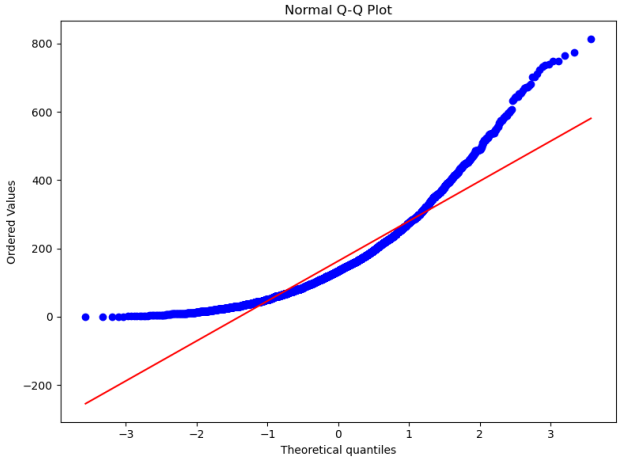
1. Histogram of Number of Critic Reviews: The histogram shows a right-skewed distribution, indicating that most movies receive fewer critic reviews, with the frequency decreasing as the number of reviews increases. The peak at the lower end suggests that many movies get relatively few reviews, while a smaller number receive a large volume of reviews.



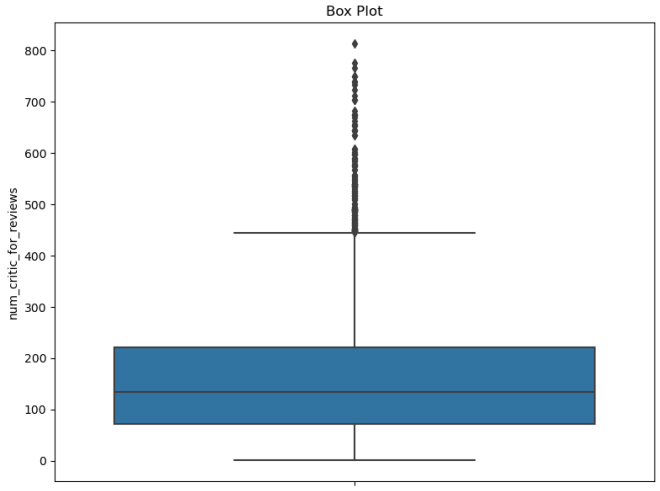
2. Scatter Plot of Number of Critic Reviews: The scatter plot illustrates the number of critic reviews across the dataset's index. Most movies cluster towards the lower end, indicating fewer reviews, but several outliers receive significantly more reviews, reinforcing the histogram's right-skewed pattern.



3. Normal Q-Q Plot of Number of Critic Reviews: The Q-Q plot shows significant deviation from a normal distribution. The points diverge from the straight line, especially in the upper tail, indicating outliers and skewness. This confirms that the number of critic reviews does not follow a normal distribution.



4. Box Plot of Number of Critic Reviews: The box plot highlights the interquartile range (IQR), median, and outliers. The median is around 100 reviews, with many outliers above the upper whisker, indicating that some movies receive significantly more reviews than the majority. This supports the observation of a right-skewed distribution.



5. Normal Assessment: Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on the number of critic reviews. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

Parameter Estimation for Number of Critic Reviews:

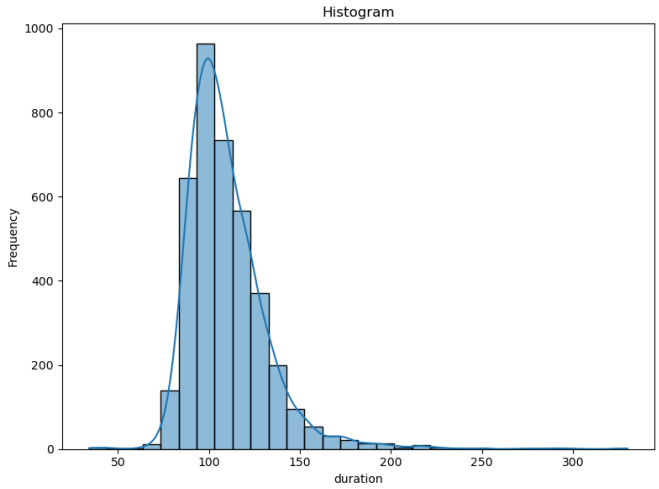
We assumed the dataset is a simple random sample. In the parameter estimation for num_critic_for_reviews, we calculated a sample mean of 163.23 and a sample standard deviation of 124.04. Using the t-Distribution, we estimated the population mean with a 95% confidence interval to be between approximately 9.97 and 166.49. For the population standard deviation, we used the Chi-Squared Distribution to estimate it to be between approximately 120.47 and 127.72. The variance can also be derived from the standard deviation estimation, giving us an approximate range of 14,515.37 to 16,305.99 for the population variance. Given these estimates, we see that the num_critic_for_reviews variable has a moderate mean and a fairly large standard deviation, indicating significant variability in the number of critic reviews across movies.

Summary

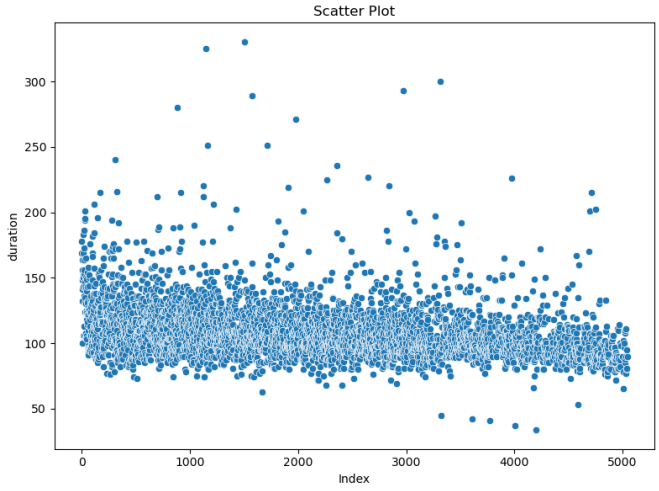
The analysis of num_critic_for_reviews reveals a right-skewed distribution with most movies receiving fewer reviews and a few receiving many. The high variability and presence of outliers indicate a significant disparity in the number of reviews different movies receive. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of critic reviews in the dataset.

5.2.2 Duration

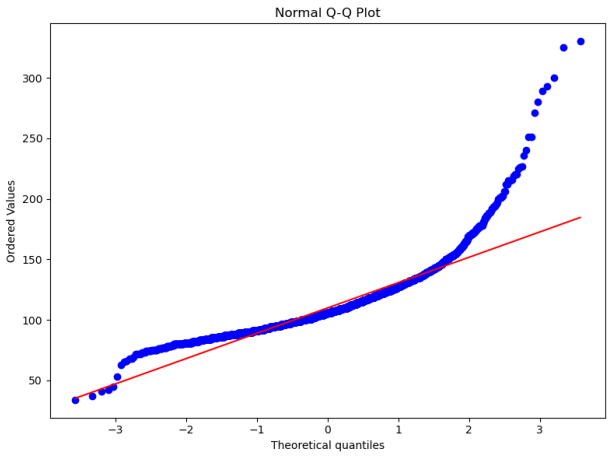
1. Histogram of Duration: The histogram shows a right-skewed distribution, with most movies having durations around the 100-120 minute range. There are fewer movies as the duration increases, and some movies are exceptionally long. This pattern indicates that while the majority of movies have standard lengths, a few are significantly longer.



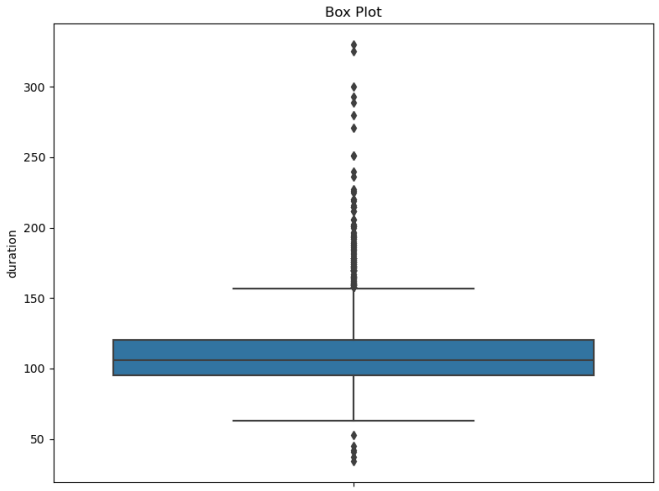
2. Scatter Plot of Duration: The scatter plot illustrates the duration across the dataset's index. Most movies cluster around the 100-120 minute mark, with some outliers extending above 200 minutes. This plot confirms the histogram's right-skewed pattern, showing that longer movies are less common.



3. Normal Q-Q Plot of Duration: The Q-Q plot reveals significant deviation from a normal distribution. The data points diverge from the reference line, especially in the tails, indicating the presence of outliers and skewness. This further supports the observation that the duration does not follow a normal distribution.



4. Box Plot of Duration: The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median duration is around 100 minutes. Numerous outliers above the upper whisker indicate that some movies are significantly longer than the majority, confirming the right-skewed distribution observed in the histogram and scatter plot.



5. Normal Assessment: Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on movie duration. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

Parameter Estimation for Duration:

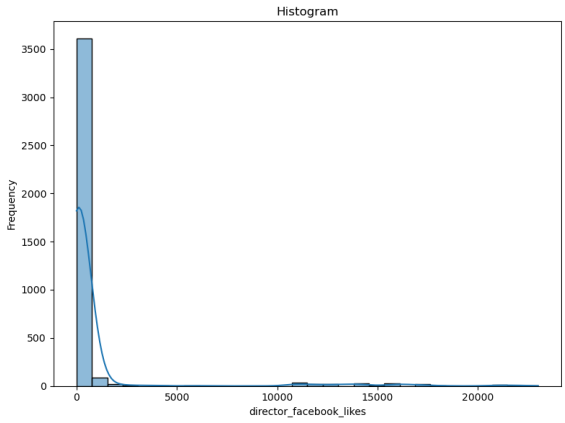
Assuming the dataset is a simple random sample, we calculated a sample mean of 109.90 and a sample standard deviation of 22.70 for duration. Using the t-Distribution, we estimated the population mean with a 95% confidence interval to be between approximately 109.27 and 110.53. For the population standard deviation, we employed the Chi-Squared Distribution, estimating it to be between approximately 22.34 and 23.09. The population variance can thus be approximated to range from about 499.06 to 533.14. These estimates suggest that duration has a relatively moderate mean and standard deviation, indicating less variability in movie durations compared to other variables.

Summary

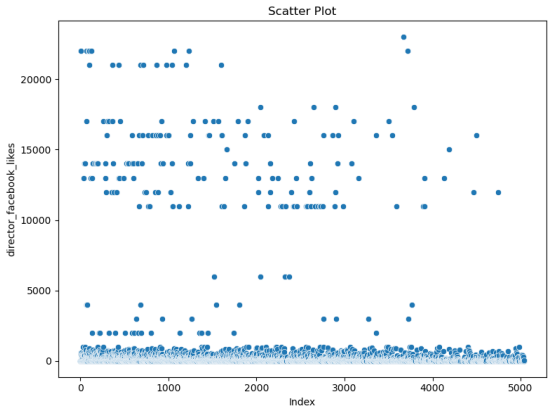
The analysis of duration reveals a right-skewed distribution with most movies having durations around 100-120 minutes and a few significantly longer. The high variability and presence of outliers indicate a significant disparity in movie lengths. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of movie durations in the dataset.

5.2.3 Director Facebook Likes

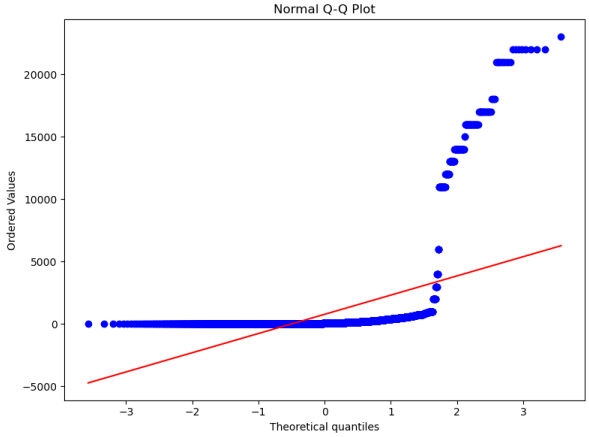
1. Histogram of Director Facebook Likes: The histogram shows a highly right-skewed distribution, with most directors having very few likes. There is a long tail extending towards higher values, indicating that a few directors have a significantly higher number of likes. This pattern suggests that while most directors have a modest social media presence, a small number are very popular.



2. Scatter Plot of Director Facebook Likes: The scatter plot illustrates the number of director Facebook likes across the dataset's index. Most movies cluster around the lower end, reflecting fewer likes. However, there are several high outliers, showing that some directors attract a much higher number of likes. This pattern is consistent with the histogram's right-skewed distribution.

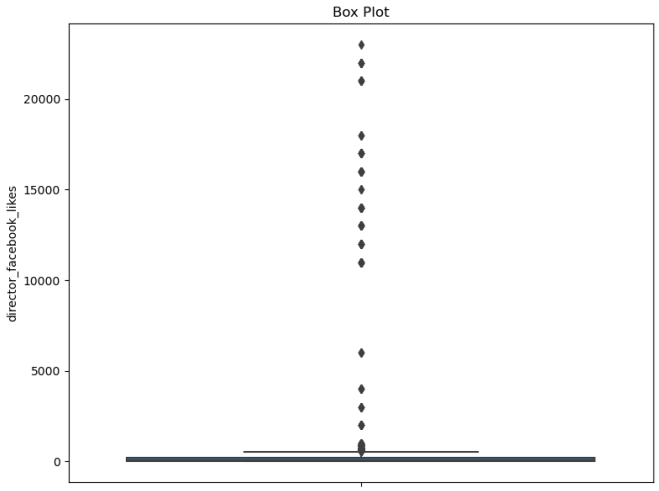


3. Normal Q-Q Plot of Director Facebook Likes: The Q-Q plot demonstrates significant deviation from a normal distribution. The data points veer off from the reference line, particularly in the upper tail, indicating the presence of outliers and skewness. This further supports the observation that the number of director Facebook likes is not normally distributed.



4. Box Plot of Director Facebook Likes: The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median number of Facebook likes is very low,

with numerous outliers extending far above the upper whisker. This plot confirms the right-skewed distribution observed in the histogram and scatter plot.



5. Normal Assessment: Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on director Facebook likes. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

Parameter Estimation for Director Facebook Likes:

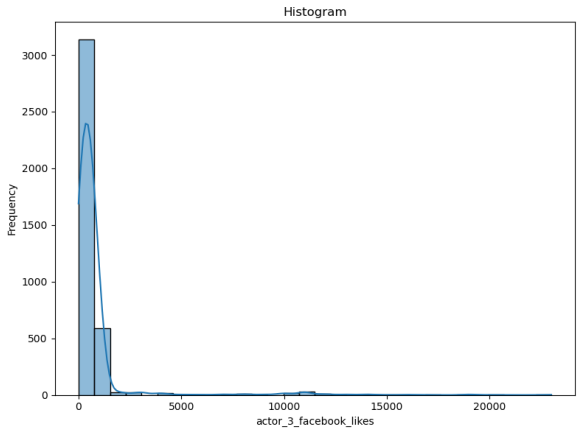
We assumed the dataset is a simple random sample. The parameter estimation for director_facebook_likes yielded a sample mean of 781.31 and a sample standard deviation of 3017.68. Using the t-Distribution, the population mean with a 95% confidence interval is estimated to be between approximately 682.75 and 879.86. For the population standard deviation, the Chi-Squared Distribution estimated it to be between approximately 2940.47 and 3098.89. This translates to a population variance ranging from about 8,641,167 to 9,595,927. These high values and the wide range indicate significant variability and the presence of outliers in director_facebook_likes.

Summary

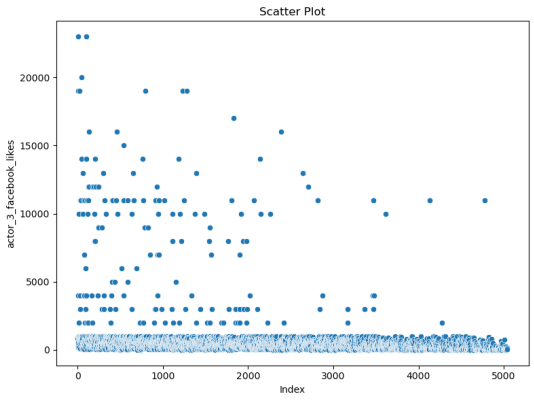
The analysis of director_facebook_likes reveals a highly right-skewed distribution with most directors having very few likes and a few having many. The high variability and presence of outliers indicate a significant disparity in the number of likes different directors receive. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of director Facebook likes in the dataset.

5.2.4 Actor 3 Facebook Likes

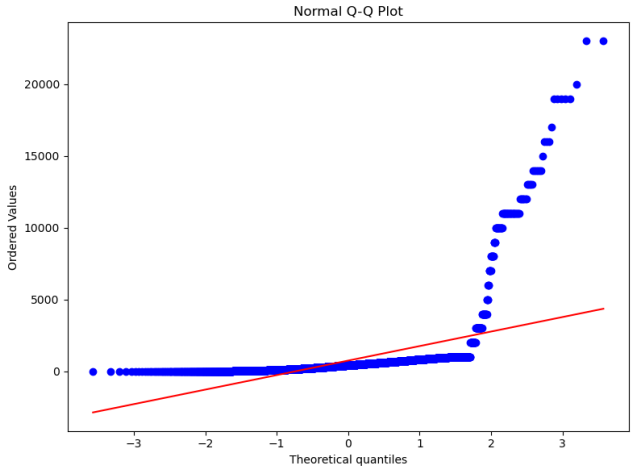
1. Histogram of Actor 3 Facebook Likes: The histogram shows a highly right-skewed distribution, with most actors having very few likes. There is a long tail extending towards higher values, indicating that a few actors have a significantly higher number of likes. This pattern suggests that while most third actors have a modest social media presence, a small number are very popular.



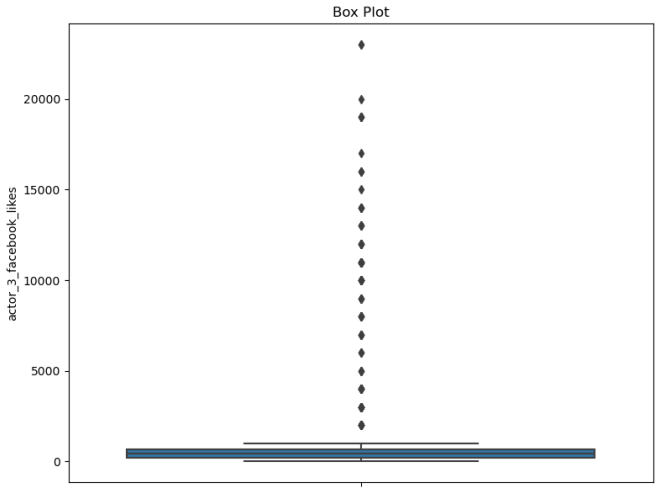
2. Scatter Plot of Actor 3 Facebook Likes: The scatter plot illustrates the number of Facebook likes across the dataset's index. Most movies cluster around the lower end, reflecting fewer likes. However, there are several high outliers, showing that some third actors attract a much higher number of likes. This pattern is consistent with the histogram's right-skewed distribution.



3. Normal Q-Q Plot of Actor 3 Facebook Likes: The Q-Q plot demonstrates significant deviation from a normal distribution. The data points veer off from the reference line, particularly in the upper tail, indicating the presence of outliers and skewness. This further supports the observation that the number of Facebook likes for the third actor is not normally distributed.



4. Box Plot of Actor 3 Facebook Likes: The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median number of Facebook likes is very low, with numerous outliers extending far above the upper whisker. This plot confirms the right-skewed distribution observed in the histogram and scatter plot.



5. Normal Assessment: Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on actor 3 Facebook likes. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

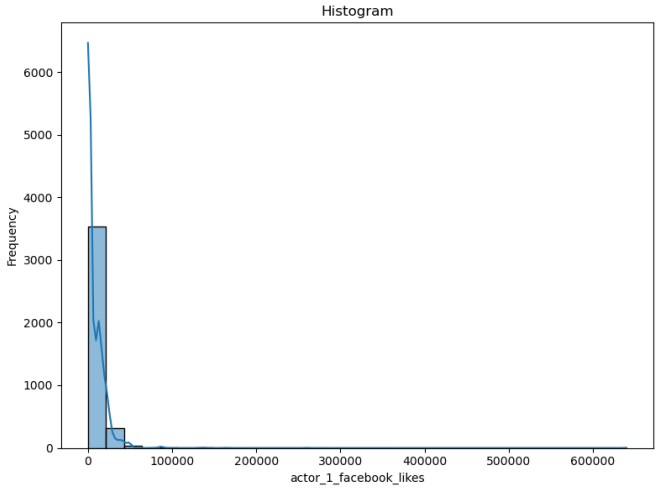
Parameter Estimation for Actor 3 Facebook Likes: Assuming the dataset is a simple random sample, the parameter estimation for actor_3_facebook_likes resulted in a sample mean of 753.53 and a sample standard deviation of 1864.23. The t-Distribution estimates the population mean with a 95% confidence interval to be between approximately 729.17 and 777.88. Using the Chi-Squared Distribution, the population standard deviation is estimated to be between approximately 1817.91 and 1912.47, translating to a population variance range of about 3,305,798 to 3,657,474. These values show substantial variability in actor_3_facebook_likes, indicating that while most actors receive few likes, some receive significantly more, contributing to the large standard deviation.

Summary

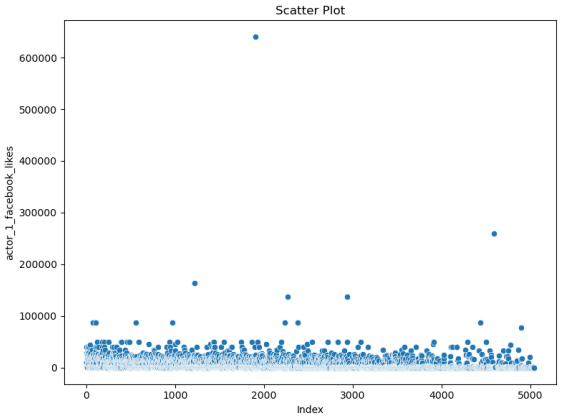
The analysis of actor_3_facebook_likes reveals a highly right-skewed distribution with most third actors having very few likes and a few having many. The high variability and presence of outliers indicate a significant disparity in the number of likes different actors receive. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of actor 3 Facebook likes in the dataset.

5.2.5 Actor 1 Facebook Likes

1. **Histogram of Actor 1 Facebook Likes:** The histogram shows a highly right-skewed distribution, with most actors having very few likes. There is a long tail extending towards higher values, indicating that a few actors have a significantly higher number of likes. This pattern suggests that while most leading actors have a modest social media presence, a small number are very popular.

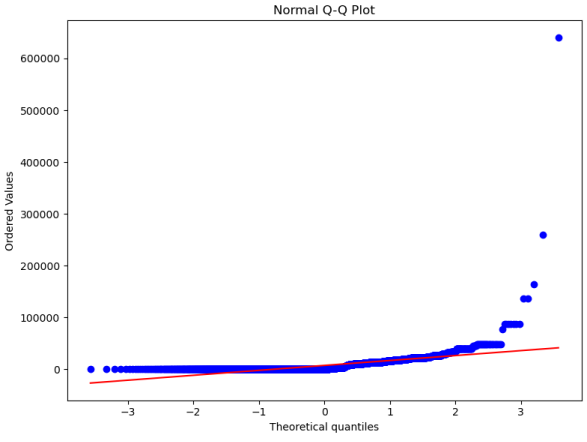


2. **Scatter Plot of Actor 1 Facebook Likes:** The scatter plot illustrates the number of Facebook likes across the dataset's index. Most movies cluster around the lower end, reflecting fewer likes. However, there are several high outliers, showing that some leading actors attract a much higher number of likes. This pattern is consistent with the histogram's right-skewed distribution.



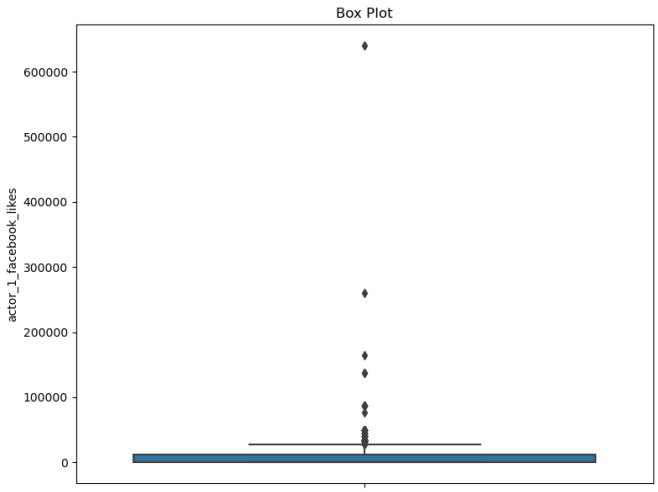
3.

4. **Normal Q-Q Plot of Actor 1 Facebook Likes:** The Q-Q plot demonstrates significant deviation from a normal distribution. The data points veer off from the reference line, particularly in the upper tail, indicating the presence of outliers and skewness. This further supports the observation that the number of Facebook likes for the leading actor is not normally distributed.



5. **Box Plot of Actor 1 Facebook Likes:** The box plot provides a visual summary of the distribution, highlighting the interquartile

range (IQR), median, and outliers. The median number of Facebook likes is very low, with numerous outliers extending far above the upper whisker. This plot confirms the right-skewed distribution observed in the histogram and scatter plot.



6. **Normal Assessment:** Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on actor 1 Facebook likes. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

Parameter Estimation for actor_1_facebook_likes

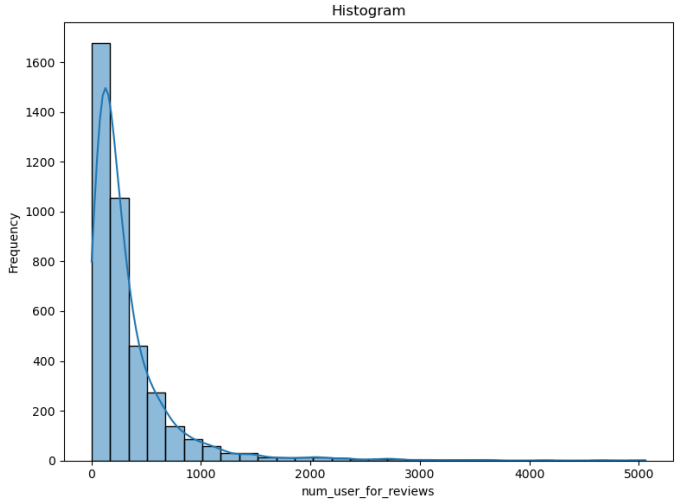
Assuming the dataset is a simple random sample, the parameter estimation for actor_1_facebook_likes resulted in a sample mean of 7584.68 and a sample standard deviation of 15360.15. Using the t-Distribution, we estimated the population mean with a 95% confidence interval to be between approximately 7382.75 and 7786.61. For the population standard deviation, the Chi-Squared Distribution estimated it to be between approximately 14945.62 and 15790.68, translating to a population variance range of about 223,374,431 to 249,094,877. These high values and the wide range indicate significant variability in actor_1_facebook_likes, with some actors having exceptionally high numbers of likes compared to the majority, contributing to the large standard deviation.

Summary: The analysis of actor_1_facebook_likes reveals a highly right-skewed distribution with most leading actors having very few likes and a few having many. The high variability and presence of outliers indicate a significant disparity in the number of likes different actors receive. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of actor 1 Facebook likes in the dataset**.**

5.2.6 Num User For Reviews

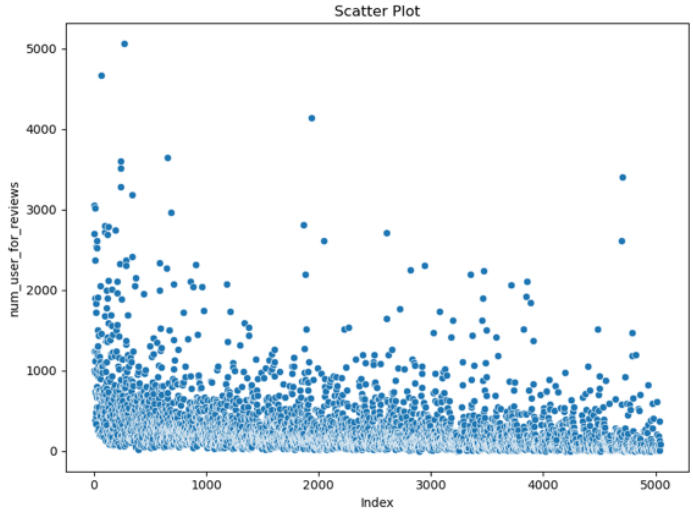
1. **Histogram of num_user_for_reviews**:

The histogram displays a highly right-skewed distribution, with most movies having a very low number of user reviews. There is a long tail extending towards higher values, indicating that a few movies have a significantly higher number of reviews. This pattern suggests that while the majority of movies attract a modest number of reviews, a small number receive a much higher count.



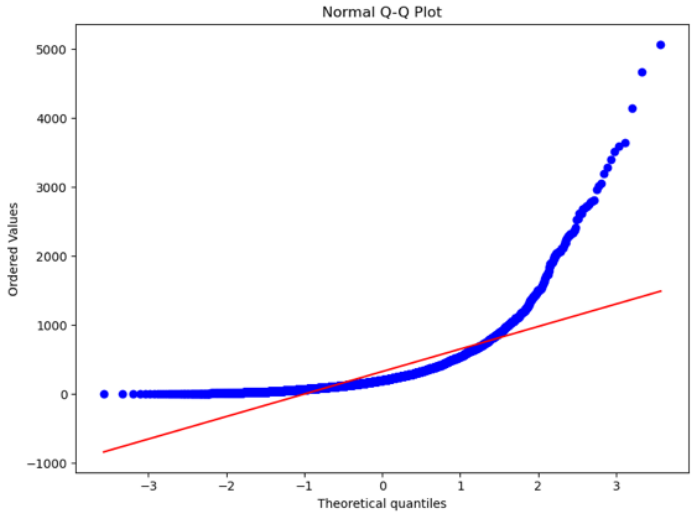
2. **Scatter Plot of num_user_for_reviews**:

The scatter plot illustrates the number of user reviews across the dataset's index. Most movies cluster around the lower end, reflecting fewer reviews. However, several high outliers show that some movies attract a much higher number of reviews. This pattern is consistent with the histogram's right-skewed distribution.



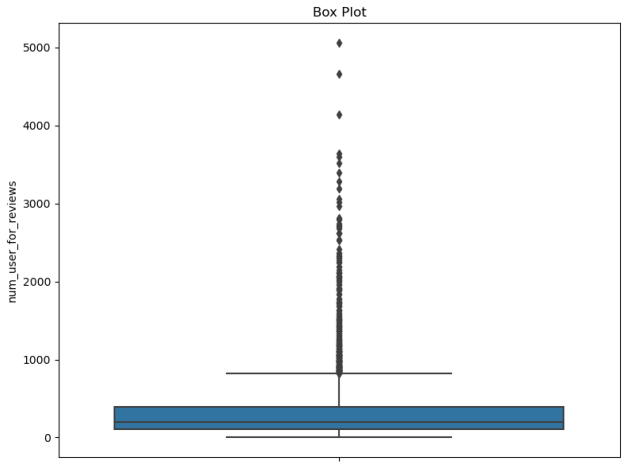
3. **Normal Q-Q Plot of num_user_for_reviews**:

The Q-Q plot demonstrates significant deviation from a normal distribution. The data points veer off from the reference line, particularly in the upper tail, indicating the presence of outliers and skewness. This further supports the observation that the number of user reviews is not normally distributed.



4. **Box Plot of num_user_for_reviews**:

The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median number of user reviews is very low, with numerous outliers extending far above the upper whisker. This plot confirms the right-skewed distribution observed in the histogram and scatter plot.



5. **Normal Assessment**:

Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on num_user_for_reviews. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

6. **Parameter Estimation for num_user_for_reviews**:

Assuming the dataset is a simple random sample, the parameter estimation for num_user_for_reviews resulted in a sample mean of 327.31 and a sample standard deviation of 408.01. Using the t-Distribution, we could estimate the population mean with a 95% confidence interval. For the population standard deviation, the Chi-Squared Distribution would provide an estimate. The high variability and presence of outliers indicate significant variability in num_user_for_reviews, with some movies having exceptionally high numbers of reviews compared to the majority, contributing to the large standard deviation. The proportion is 1.0, indicating that all data points are included in the analysis.

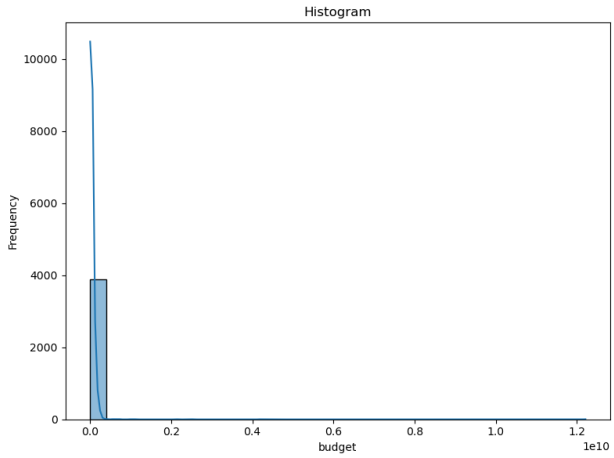
Summary:

The analysis of num_user_for_reviews reveals a highly right-skewed distribution with most movies having very few reviews and a few having many. The high variability and presence of outliers indicate a significant disparity in the number of reviews different movies receive. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of num_user_for_reviews in the dataset.

5.2.7 Bugdet

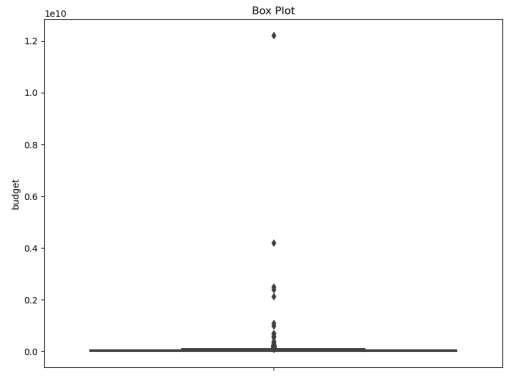
1. **Histogram of budget**:

The histogram shows a highly right-skewed distribution, with most movies having a very low budget. There is a long tail extending towards higher values, indicating that a few movies have significantly higher budgets. This pattern suggests that while the majority of movies have modest budgets, a small number of movies are allocated much higher budgets.



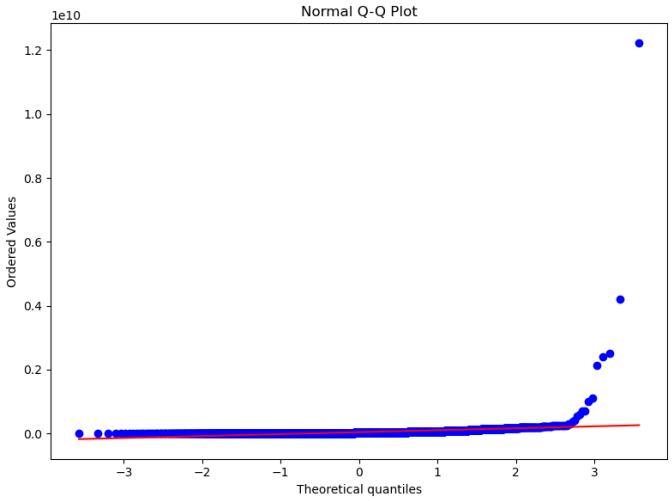
2. **Box Plot of budget**:

The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median budget is very low, with numerous outliers extending far above the upper whisker. This plot confirms the right-skewed distribution observed in the histogram.



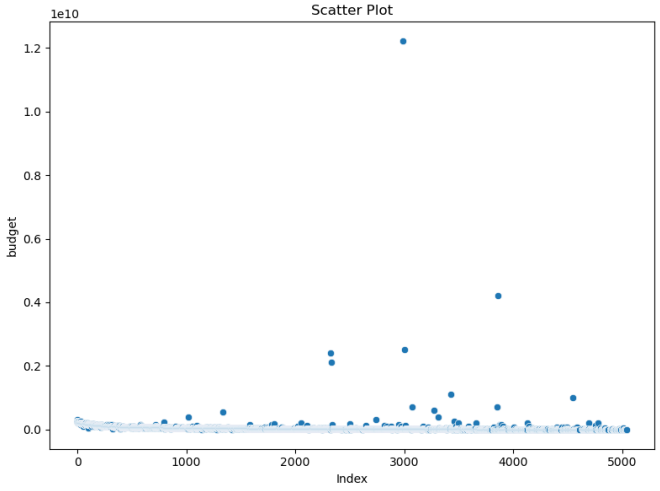
3. **Normal Q-Q Plot of budget**:

The Q-Q plot demonstrates significant deviation from a normal distribution. The data points veer off from the reference line, particularly in the upper tail, indicating the presence of outliers and skewness. This further supports the observation that the budget data is not normally distributed.



4. **Scatter Plot of budget**:

The scatter plot illustrates the budget values across the dataset's index. Most movies cluster around the lower end, reflecting smaller budgets. However, several high outliers show that some movies have much higher budgets. This pattern is consistent with the histogram's right-skewed distribution.



5. **Normal Assessment**:

Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on budget. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

6. **Parameter Estimation for budget**:

Assuming the dataset is a simple random sample, the parameter estimation for budget resulted in a sample mean of 45,210,278.28 and a sample standard deviation of 222,389,458.75. Using the t-Distribution, we could estimate the population mean with a 95% confidence interval. For the population standard deviation, the Chi-Squared Distribution would provide an estimate. The high variability and presence of outliers indicate significant variability in budget, with some movies having exceptionally high budgets compared to the majority, contributing to the large standard deviation. The proportion is 1.0, indicating that all data points are included in the analysis.

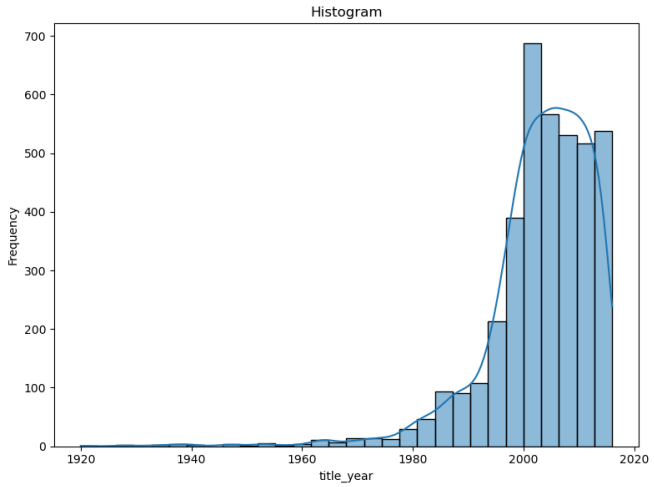
Summary:

The analysis of budget reveals a highly right-skewed distribution with most movies having very low budgets and a few having very high budgets. The high variability and presence of outliers indicate a significant disparity in the budget allocations for different movies. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of budget in the dataset.

5.2.8 Title Year

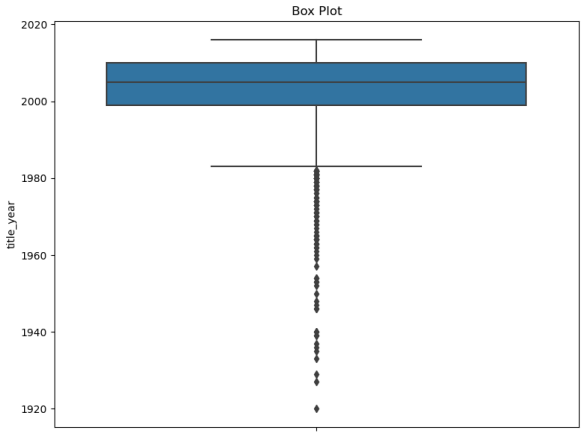
1. **Histogram of title_year**:

The histogram shows the distribution of movie release years. The distribution is right-skewed, with a noticeable increase in the number of movies released from the 1980s onwards, peaking around the 2000s. This pattern suggests that more movies have been produced in recent decades compared to earlier years.



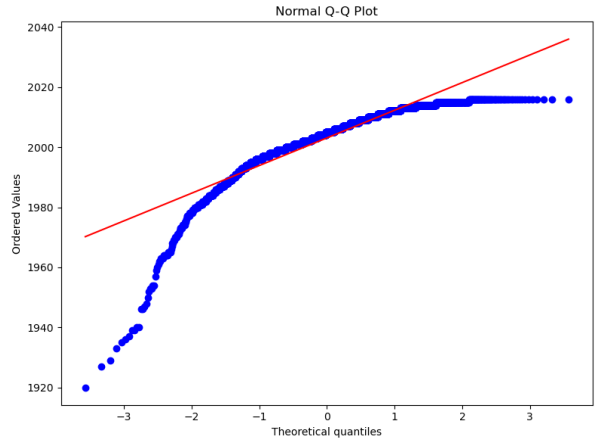
2. **Box Plot of title_year**:

The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median release year is around the 2000s, with several outliers extending back to the 1920s. This plot confirms the right-skewed distribution observed in the histogram, with most movies released in the late 20th and early 21st centuries.



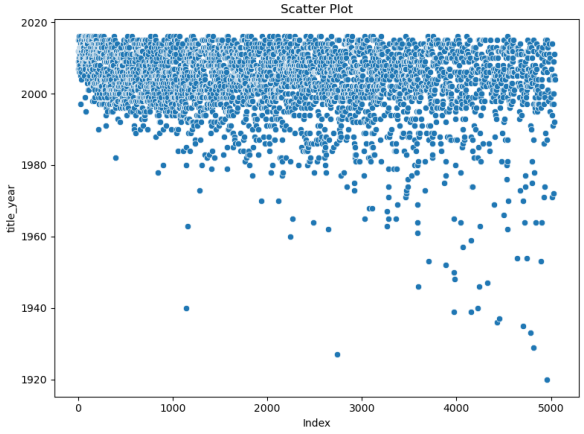
3. **Normal Q-Q Plot of title_year**:

The Q-Q plot demonstrates significant deviation from a normal distribution. The data points veer off from the reference line, especially in the lower tail, indicating the presence of outliers and skewness. This further supports the observation that the release year data is not normally distributed.



4. **Scatter Plot of title_year**:

The scatter plot illustrates the release year values across the dataset's index. Most movies are clustered around the late 20th and early 21st centuries, reflecting the histogram's pattern. Several outliers extend back to earlier years, consistent with the box plot.



5. **Normal Assessment**:

Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on title_year. The significant deviation from the straight reference line in the Q-Q plot, especially in the lower tail, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

6. **Parameter Estimation for title_year**:

Assuming the dataset is a simple random sample, the parameter estimation for title_year resulted in a sample mean of 2003.08 and a sample standard deviation of 10.00. Using the t-Distribution, we could estimate the population mean with a 95% confidence interval. For the population standard deviation, the Chi-Squared Distribution would provide an estimate. The high variability and presence of outliers indicate significant variability in title_year, with some movies released significantly earlier than the majority, contributing to the large standard deviation. The proportion is 1.0, indicating that all data points are included in the analysis.

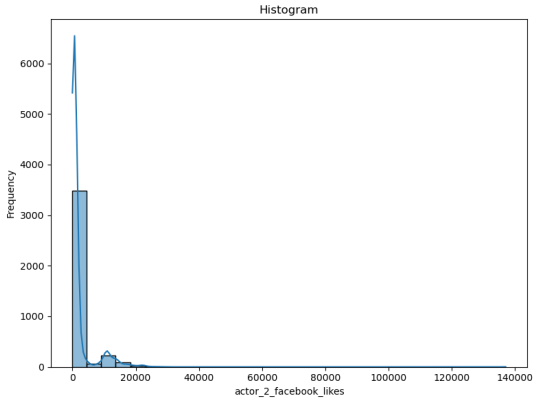
Summary:

The analysis of title_year reveals a right-skewed distribution with most movies released in recent decades and a few released much earlier. The high variability and presence of outliers indicate a significant disparity in the release years of different movies. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of title_year in the dataset.

5.2.9 Actor 2 Facebook Likes

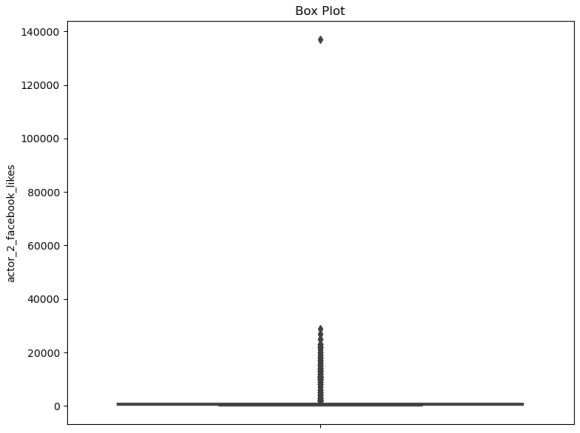
1. **Histogram of actor_2_facebook_likes**:

The histogram displays a highly right-skewed distribution, with most actors having very few likes. There is a long tail extending towards higher values, indicating that a few actors have significantly higher numbers of likes. This pattern suggests that while most secondary actors have a modest social media presence, a small number are very popular.



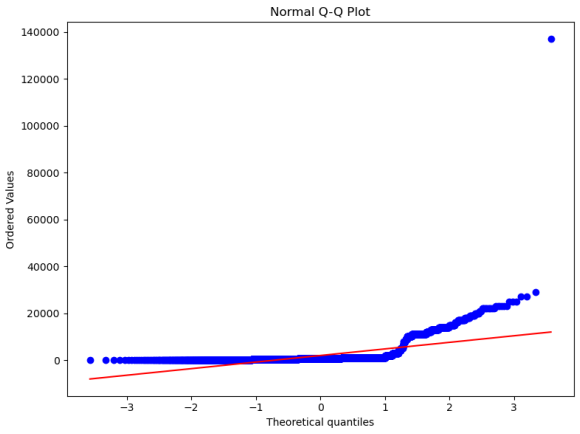
2. **Box Plot of actor_2_facebook_likes**:

The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median number of Facebook likes is very low, with numerous outliers extending far above the upper whisker. This plot confirms the right-skewed distribution observed in the histogram.



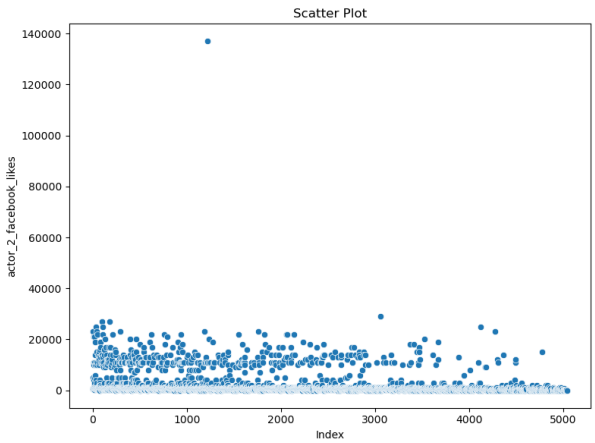
3. **Normal Q-Q Plot of actor_2_facebook_likes**:

The Q-Q plot demonstrates significant deviation from a normal distribution. The data points veer off from the reference line, particularly in the upper tail, indicating the presence of outliers and skewness. This further supports the observation that the number of Facebook likes for secondary actors is not normally distributed.



4. **Scatter Plot of actor_2_facebook_likes**:

The scatter plot illustrates the number of Facebook likes across the dataset's index. Most movies cluster around the lower end, reflecting fewer likes. However, there are several high outliers, showing that some secondary actors attract a much higher number of likes. This pattern is consistent with the histogram's right-skewed distribution.



5. **Normal Assessment**:

Our observation of the data distribution shows that it is not normal. The histogram and Q-Q plot reject the normality of the data on actor_2_facebook_likes. The significant deviation from the straight reference line in the Q-Q plot, especially in the tails, suggests that the distribution is skewed and contains outliers. This aligns with the right-skewed distribution seen in the histogram and the box plot.

6. **Parameter Estimation for actor_2_facebook_likes**:

Assuming the dataset is a simple random sample, the parameter estimation for actor_2_facebook_likes resulted in a sample mean of 1970.66 and a sample standard deviation of 4482.85. Using the t-Distribution, we could estimate the population mean with a 95% confidence interval. For the population standard deviation, the Chi-Squared Distribution would provide an estimate. The high variability and presence of outliers indicate significant variability in actor_2_facebook_likes, with some actors having exceptionally high numbers of likes compared to the majority, contributing to the large standard deviation. The proportion is 1.0, indicating that all data points are included in the analysis.

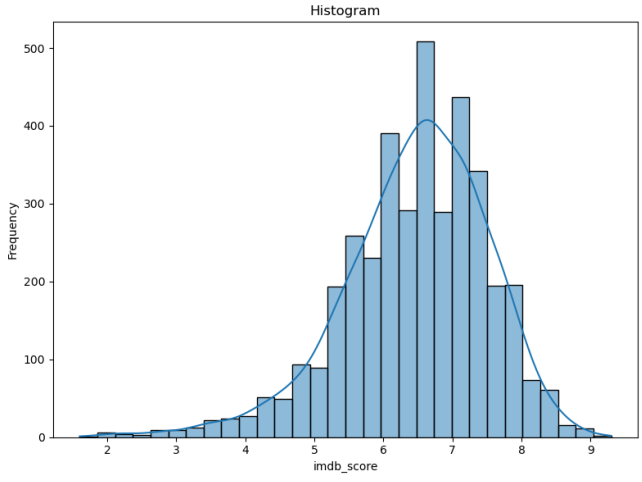
Summary:

The analysis of actor_2_facebook_likes reveals a highly right-skewed distribution with most secondary actors having very few likes and a few having many. The high variability and presence of outliers indicate a significant disparity in the number of likes different actors receive. The normal assessment confirms that the data does not follow a normal distribution, necessitating careful consideration of these factors in any further analysis or modeling. This detailed examination provides a clear understanding of the distribution and variability of actor_2_facebook_likes in the dataset.

5.2.10 IMDB Score

1. **Histogram of imdb_score**:

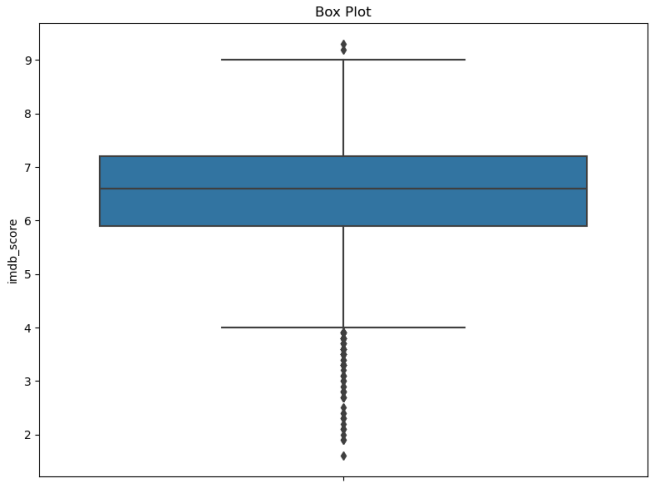
The histogram shows a relatively symmetric distribution with a slight skew towards the lower scores. Most movies have IMDB scores between 5 and 8, peaking around 6 to 7. This pattern suggests that the majority of movies are rated in the mid to high range, with fewer movies receiving very low or very high scores.



2. **Box Plot of imdb_score**:

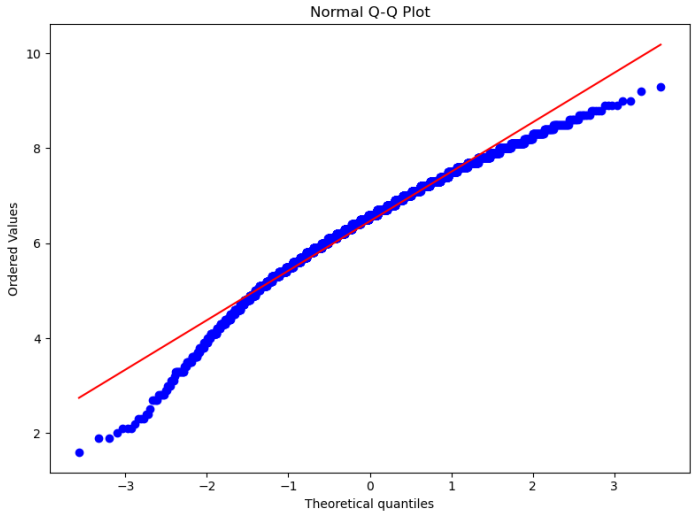
The box plot provides a visual summary of the distribution, highlighting the interquartile range (IQR), median, and outliers. The median IMDB score is around 6.5, with a few outliers on both the lower and upper ends. This plot confirms the relatively symmetric distribution observed in the histogram, with the bulk of the data concentrated in the middle range of scores

.



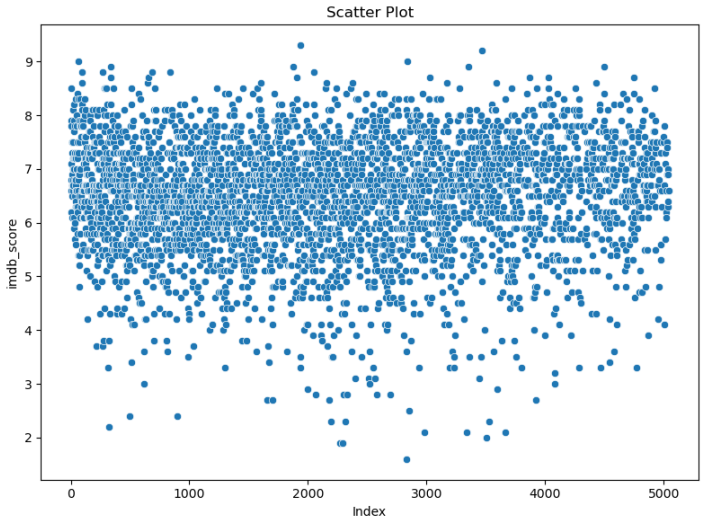
3. **Normal Q-Q Plot of imdb_score**:

The Q-Q plot shows that the data follows a near-normal distribution with some deviations in the tails. The data points align closely with the reference line, indicating that the majority of the IMDB scores are normally distributed, although the presence of some outliers deviates from perfect normality.



4. **Scatter Plot of imdb_score**:

The scatter plot illustrates the IMDB scores across the dataset's index. The scores are widely spread between 2 and 9, with a dense cluster around the 6 to 7 range, reflecting the pattern observed in the histogram. The presence of scores across the entire range indicates diverse ratings for the movies in the dataset.



5. **Normal Assessment**:

Our observation of the data distribution shows that it is close to normal. The histogram and Q-Q plot mostly align with the normality of the data on imdb_score. The slight deviations from the straight reference line in the Q-Q plot, especially in the tails, suggest minor skewness and the presence of outliers. This aligns with the relatively symmetric distribution seen in the histogram and the box plot.

6. **Parameter Estimation for imdb_score**:

Assuming the dataset is a simple random sample, the parameter estimation for imdb_score resulted in a sample mean of 6.46 and a sample standard deviation of 1.06. Using the t-Distribution, we could estimate the population mean with a 95% confidence interval. For the population standard deviation, the Chi-Squared Distribution would provide an estimate. The relatively low standard deviation indicates that most IMDB scores are clustered around the mean, with fewer extreme values. The proportion is 1.0, indicating that all data points are included in the analysis.

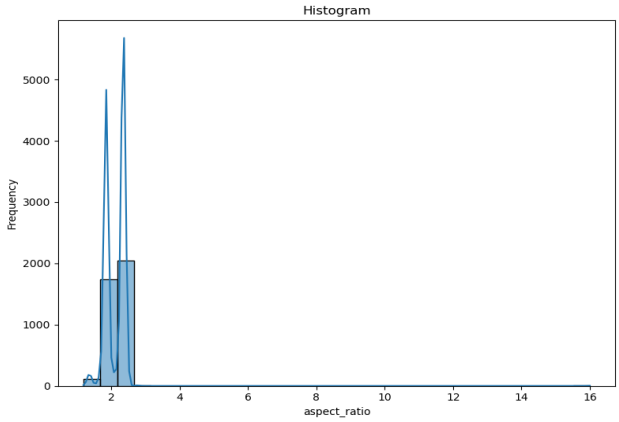
Summary:

The analysis of imdb_score reveals a relatively symmetric distribution with most movies rated in the mid to high range. The minor skewness and presence of outliers indicate slight deviations from a perfect normal distribution. The normal assessment confirms that the data is near-normal, with a detailed examination providing a clear understanding of the distribution and variability of imdb_score in the dataset.

5.2.11 Aspect Ratio

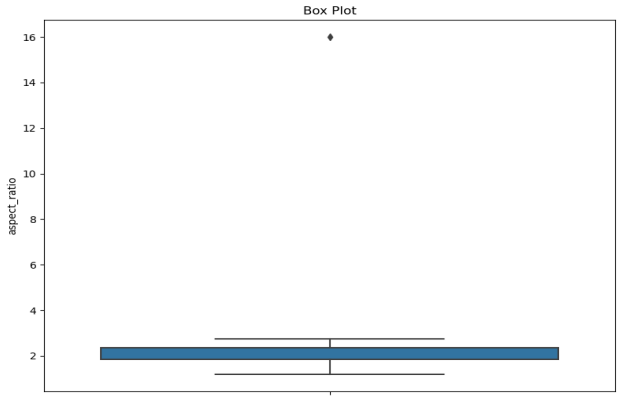
1. **Histogram of aspect_ratio**:

The histogram shows a highly skewed distribution with a peak around 2. The majority of the data points are clustered between 1.5 and 3, indicating that most aspect ratios are in this range. However, there are a few aspect ratios that are significantly higher, up to around 16, as seen from the tail on the right side of the distribution.



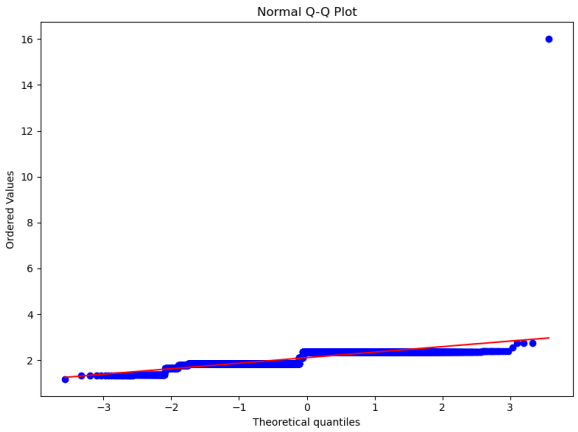
2. **Box Plot of aspect_ratio**:

The box plot highlights the interquartile range (IQR), median, and outliers. The median aspect ratio is around 2, with the bulk of the data concentrated between 1.5 and 2.5. There is one significant outlier with an aspect ratio of around 16. This outlier indicates that while most aspect ratios are relatively consistent, there are some extreme values that deviate from the norm.



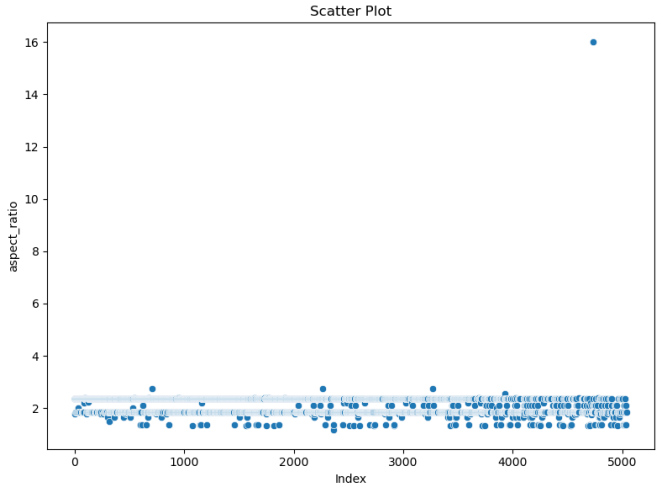
3. **Normal Q-Q Plot of aspect_ratio**:

The Q-Q plot shows that the data does not follow a normal distribution. The majority of the data points fall below the reference line, indicating a skewed distribution. The extreme outlier (aspect ratio around 16) deviates significantly from the theoretical quantiles, reinforcing the non-normality of the data.



4. **Scatter Plot of aspect_ratio**:

The scatter plot illustrates the aspect ratios across the dataset's index. Most of the data points are clustered between 1.5 and 3, with one prominent outlier around 16. This pattern reflects the distribution seen in the histogram, indicating that the majority of aspect ratios are within a narrow range, with a few extreme values.



5. **Normal Assessment**:

The assessment shows that the distribution of the aspect_ratio variable is not normal. The histogram and Q-Q plot indicate a skewed distribution with a significant outlier. The box plot and scatter plot confirm the presence of this outlier and the concentration of most data points within a narrow range. This analysis suggests that the aspect ratio data is not normally distributed and contains some extreme values.

6. **Parameter Estimation for aspect_ratio**:

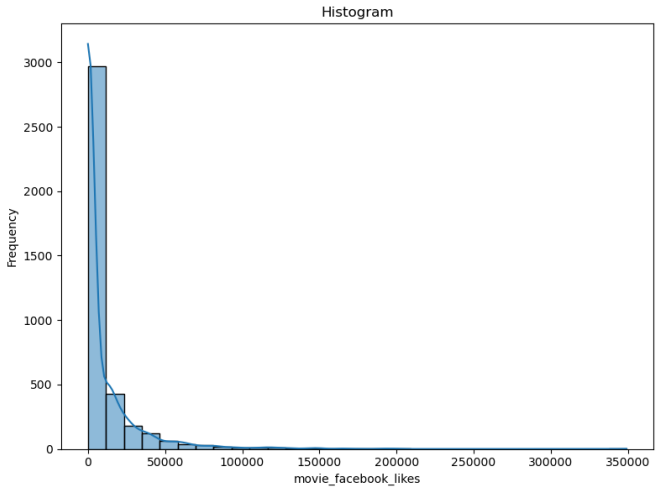
Assuming the dataset is a simple random sample, the parameter estimation for aspect_ratio resulted in a sample mean of approximately 2.11 and a sample standard deviation of approximately 0.35. The proportion is 1.0, indicating that all data points are included in the analysis. The relatively low standard deviation suggests that most aspect ratios are clustered around the mean, with fewer extreme values.

Summary:

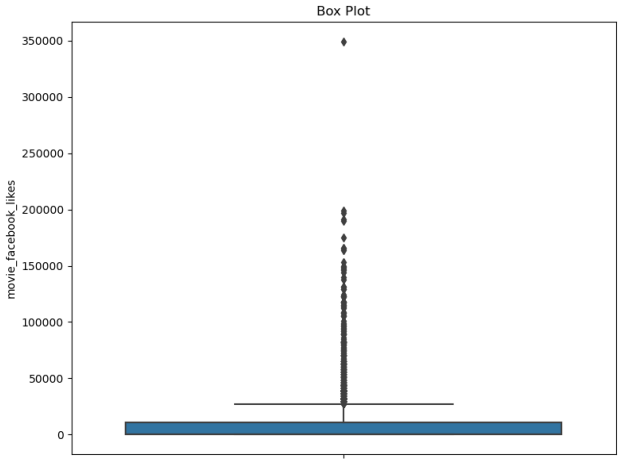
The analysis of aspect_ratio reveals a skewed distribution with most data points concentrated between 1.5 and 3. The presence of a significant outlier indicates that there are some extreme values that deviate from the norm. The normal assessment confirms that the data is not normally distributed, with a detailed examination providing a clear understanding of the distribution and variability of aspect_ratio in the dataset.

5.2.12 Movie Facebook Likes

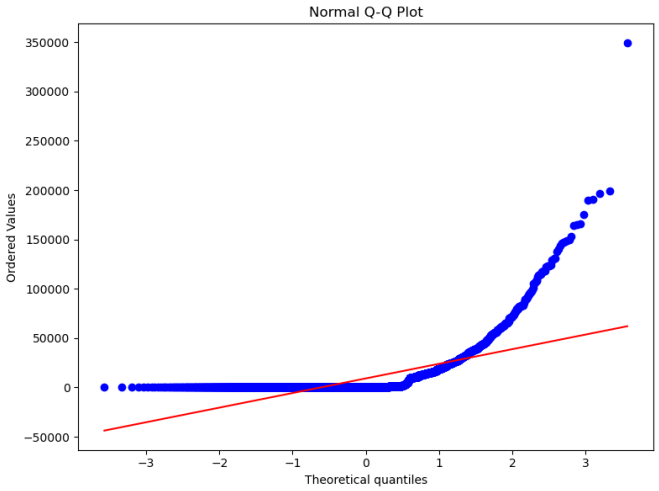
1. Histogram of movie_facebook_likes: The histogram confirms the extreme right skew. There's a very tall bar at the left side, indicating that a large majority of movies have few Facebook likes. The frequency drops off rapidly, with a long tail extending to the right, representing movies with a high number of likes.



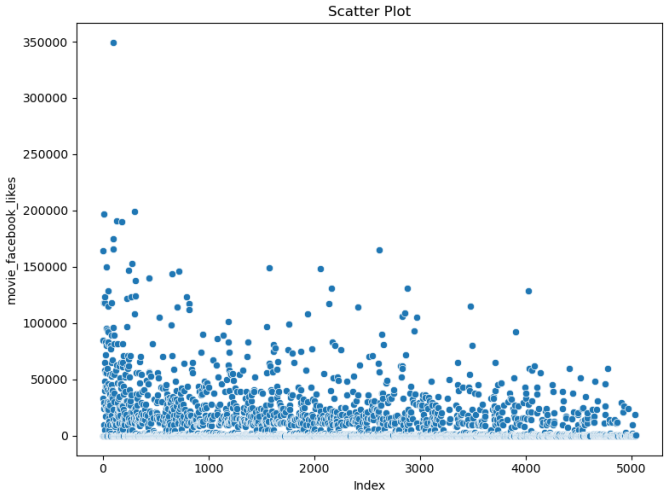
2. Box Plot of movie_facebook_likes: The box plot reveals an extremely right-skewed distribution. The box (interquartile range) is compressed near the bottom of the plot, with the median line very close to the bottom. There are numerous outliers extending far above the box, with some extreme outliers reaching up to 350,000 likes.



3. Normal Q-Q Plot of movie_facebook_likes: The Q-Q plot shows significant deviation from the normal distribution. The data points curve sharply away from the reference line, especially at higher values. This indicates a highly right-skewed distribution with many outliers at the upper end.



4. Scatter Plot of movie_facebook_likes: The scatter plot shows the distribution of likes across the dataset. Most data points are clustered near the bottom, with occasional spikes to higher values. There's one extreme outlier near the beginning of the index with about 350,000 likes.



5. Normal Assessment: The data for movie_facebook_likes is clearly not normally distributed. All plots (Q-Q plot, box plot, histogram, and scatter plot) consistently show an extreme right skew and the presence of many outliers at the upper end of the distribution.

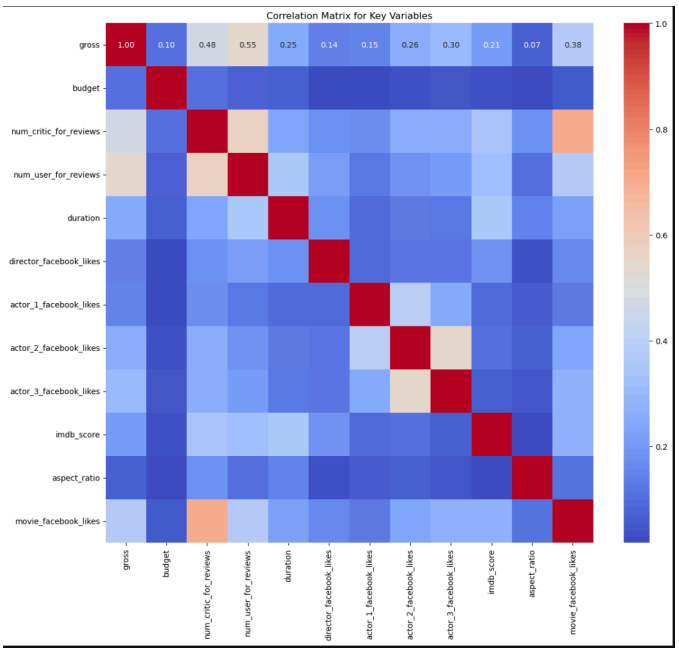
6. Parameter Estimation for movie_facebook_likes:

Assuming the dataset is a simple random sample, the parameter estimation for movie_facebook_likes resulted in a sample mean of 9138.15 and a sample standard deviation of 21302.76. Using the t-Distribution, we could estimate the population mean with a 95% confidence interval. For the population standard deviation, the Chi-Squared Distribution would provide an estimate. The extremely high standard deviation relative to the mean indicates that the Facebook likes are widely dispersed, with many extreme values far from the mean. This aligns with the highly skewed distribution we observed. The proportion is 1.0, indicating that all data points in the sample were included in the analysis.

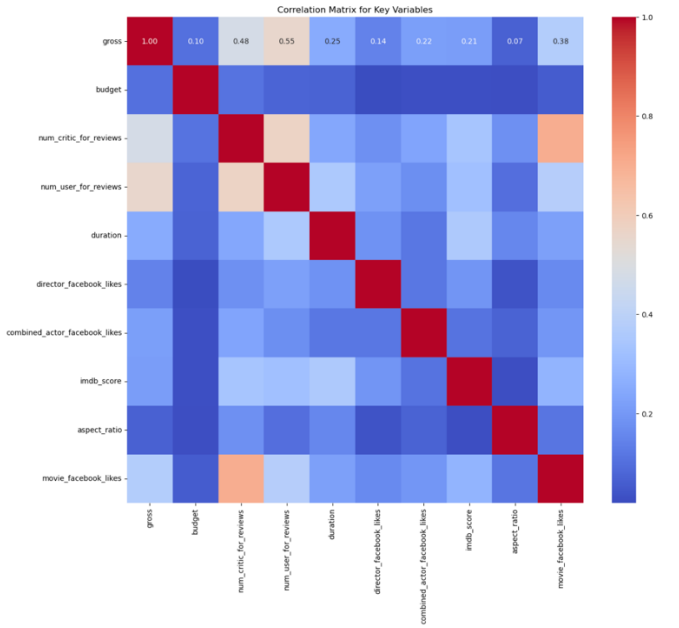
Summary: The analysis of movie_facebook_likes reveals an extremely right-skewed distribution. Most movies have relatively few Facebook likes, while a small number of movies have a very high number of likes. This creates a long-tailed distribution with numerous outliers. The data significantly deviates from normality, suggesting that non-parametric methods or data transformation might be more appropriate for further statistical analysis of this variable

6 STATISTICAL ANALYSES

1. Correlation Analysis



The correlation matrix revealed key predictors of box office success, including `budget`, `num_critic_for_reviews`, `duration`, `combined_actor_facebook_likes`, and `imdb_score`. It helped address multicollinearity by combining highly correlated variables into a single feature. The matrix guided the selection of significant variables, enhancing the accuracy and interpretability of the linear regression model. Visualized using a heatmap, it provided clear insights into the relationships between variables, ensuring a data-driven approach to model development and feature selection.



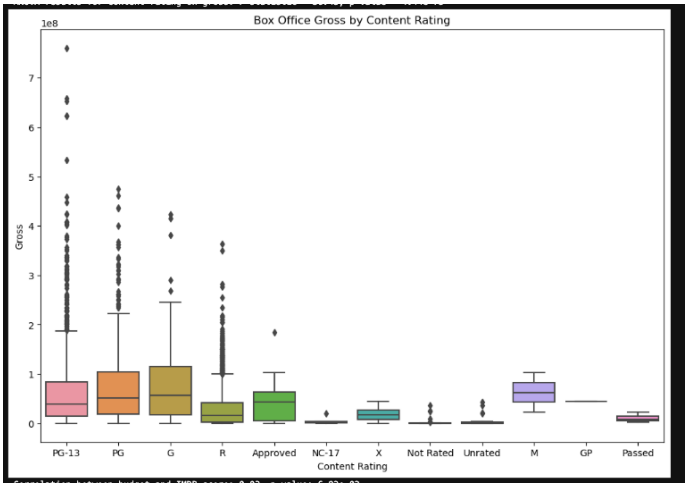
Correlation Matrix Heatmap: The heatmap visualizes the correlation between key variables and gross revenue. Significant predictors include budget (0.10), num_critic_for_reviews (0.48), and combined_actor_facebook_likes (0.38). Combining highly correlated variables like actor_1_facebook_likes, actor_2_facebook_likes, and actor_3_facebook_likes into combined_actor_facebook_likes helped address multicollinearity.

Linear Regression Output: The regression model's mean squared error (MSE) is approximately 399,444,151,463,804. The coefficients indicate the direction and magnitude of each predictor's impact on gross revenue:

- budget: 2.039e-03
- num_critic_for_reviews: 2.207e+05
- duration: 4.165e+05
- combined_actor_facebook_likes: 3.787e+02
- imdb_score: 1.873e+06

This analysis highlights key predictors and their influence on box office success, providing actionable insights for the film industry.

2. Hypothesis Testing and Box Plot Analysis



Box Plot: Box Office Gross by Content Rating: The box plot visualizes the distribution of gross revenue across different content ratings. It reveals significant variations in gross revenue based on content ratings, with categories like "PG-13" and "R" showing higher medians and wider ranges of gross revenue compared to others.

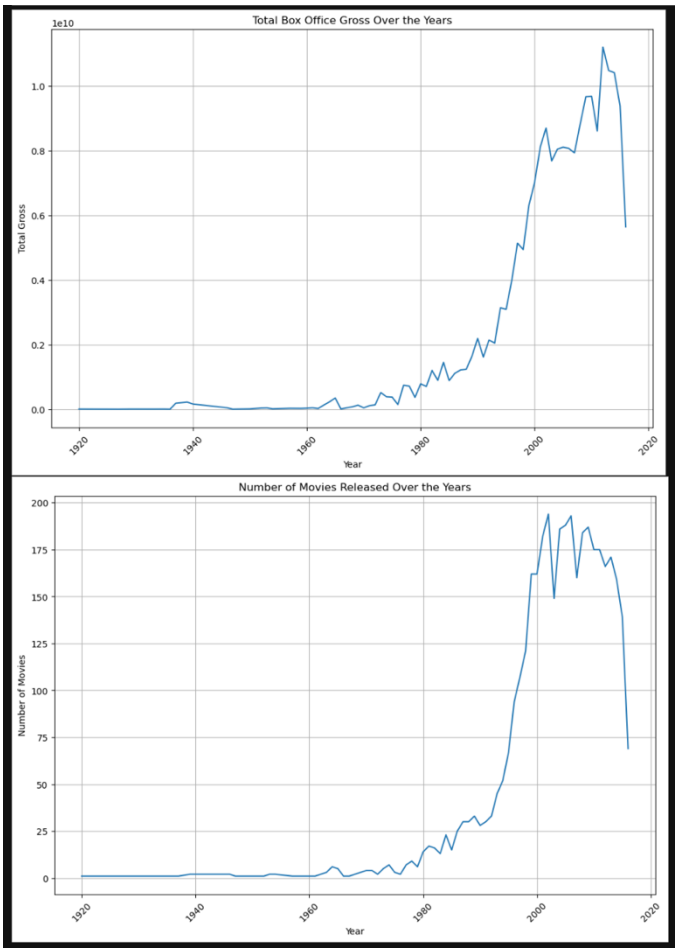
Hypothesis Testing Results:

1. **Correlation between Budget and Gross**:
   - **Correlation**: 0.10, **p-value**: 1.68e-10
   - **Interpretation**: Weak but statistically significant positive correlation indicating higher budgets are generally associated with higher gross revenue.
2. **ANOVA for Content Rating on Gross**:
   - **F-statistic**: 36.43, **p-value**: 4.44e-75
   - **Interpretation**: Significant differences in gross revenue across content rating categories, confirming the visual insights from the box plot.
3. **Correlation between Budget and IMDB Score**:
   - **Correlation**: 0.03, **p-value**: 6.92e-02
   - **Interpretation**: Very weak and not statistically significant correlation indicating that budget does not have a significant impact on IMDB score.

Benefits:

- **Identifying Key Predictors**: The analysis helps identify which content ratings are associated with higher gross revenues.
- **Validating Relationships**: Hypothesis tests provide statistical validation for observed relationships, enhancing the reliability of the insights.
- **Data-Driven Decision Making**: Provides actionable insights for filmmakers and marketers to target specific content ratings that are likely to yield higher revenues.

7 TIME SERIES ANALYSIS



Total Box Office Gross Over the Years: The line plot shows the total box office gross revenue from 1920 to 2020. The data indicates a significant upward trend in gross revenue over the decades, with notable increases from the 1980s onwards, peaking around the early 2000s. This trend reflects the growing movie industry and higher production budgets.

Number of Movies Released Over the Years: The second line plot visualizes the number of movies released each year from 1920 to 2020. The plot shows a substantial increase in the number of films released, particularly from the 1980s onwards, with a peak in the early 2000s. This growth highlights the expanding film industry and the increasing demand for diverse movie content.

Benefits:

- **Trend Identification**: Helps in understanding long-term trends in box office revenue and movie production.
- **Strategic Planning**: Informs strategic decisions by identifying periods of significant growth or decline.
- **Industry Insights**: Provides insights into the evolution of the film industry over the past century, aiding stakeholders in forecasting future trends.

8 MODEL DEVELOPMENT AND EVALUATION

Linear Regression:

- **Performance**:
  - MSE: 399,444,151,464,304.00
  - RMSE: 632,01615.10
  - MAE: 41483921.47
  - R-squared: 0.26
- **Coefficients**:
  - budget: 0.0002
  - num_critic_for_reviews: 224737.05
  - duration: 416543.05
  - combined_actor_facebook_likes: 378.71
  - imdb_score: 1,873,176.00

Random Forest:

- **Performance**:
  - MSE: 258,321,442,478,832.50
  - RMSE: 508,16476.12
  - MAE: 32679727.97
  - R-squared: 0.52
- **Feature Importance**:
  - budget: 0.433372
  - num_critic_for_reviews: 0.211933
  - imdb_score: 0.140132
  - combined_actor_facebook_likes: 0.107609
  - duration: 0.106954

Gradient Boosting:

- **Performance**:
  - MSE: 249,147,531,584,819.00
  - RMSE: 499,14642.46
  - MAE: 324,54856.61
  - R-squared: 0.54

Insights:

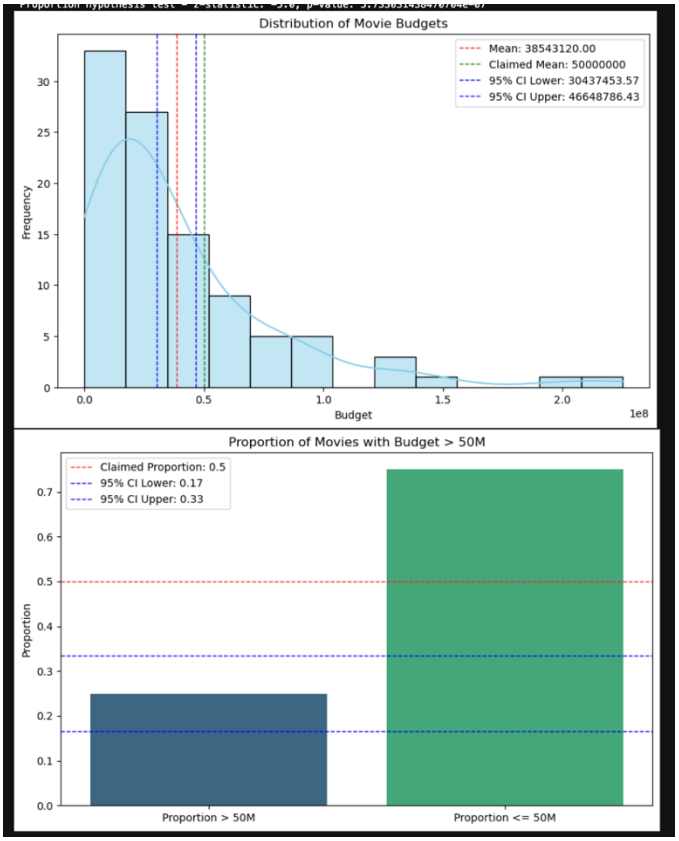
- **Performance Comparison**: Gradient Boosting performed the best with the lowest MSE and highest R-squared, followed closely by Random Forest.
- **Feature Importance**: The Random Forest model identified budget as the most influential predictor, followed by num_critic_for_reviews, imdb_score, combined_actor_facebook_likes, and duration.
- **Interpretation**: Model coefficients and feature importance highlight the significant impact of budget and critic reviews on box office gross, providing actionable insights for stakeholders.

Benefits:

- **Accurate Predictions**: Advanced models like Gradient Boosting offer robust predictions of box office revenue.
- **Data-Driven Insights**: Feature importance analysis aids in strategic decision-making by identifying key predictors.
- **Model Reliability**: Using multiple models ensures a comprehensive understanding and reliable predictions, enhancing the robustness of the analysis.

9 SAMPLING AND ANALYSIS

9.1 Hypothesis Testing with Random Sampling



Distribution of Movie Budgets:

- **Claimed Mean**: $50,000,000
- **Sample Mean**: $38,543,120.00
- **Mean Hypothesis Test**:
  - t-statistic: -2.805
  - p-value: 0.006
  - 95% CI: [$30,437,453.57, $46,648,786.43]
- **Interpretation**: The mean budget is significantly lower than the claimed mean, as the p-value is less than 0.05, rejecting the null hypothesis.

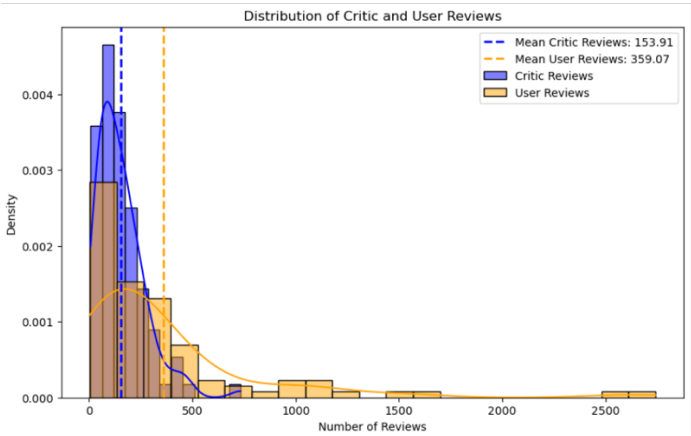
Proportion of Movies with Budget > $50M:

- **Claimed Proportion**: 0.5
- **Sample Proportion**: 0.25
- **Proportion Hypothesis Test**:
  - z-statistic: -5.0
  - p-value: 5.73e-07
  - 95% CI: [0.165, 0.335]
- **Interpretation**: The proportion of movies with budgets over $50M is significantly lower than the claimed proportion, as indicated by the p-value.

Benefits:

- **Validate Claims**: The tests provide statistical evidence to validate or refute budget-related claims.
- **Confidence Intervals**: CI provides a range within which the true mean and proportion likely fall, offering insights into budget distributions.
- **Decision-Making**: Results inform financial decisions, budgeting strategies, and resource allocation in the film industry.

9.2 Compare Two Sample Means and SD - Interpretation of Results



Distribution of Critic and User Reviews: The density plot shows the distribution of the number of critic and user reviews. The mean number of critic reviews is 153.91, whereas the mean number of user reviews is significantly higher at 359.07.

Hypothesis Testing:

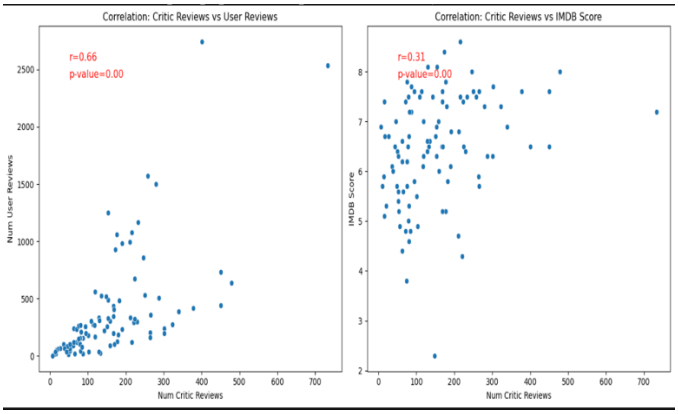
- **Difference of Means Test**:
  - **t-statistic**: -4.287
  - **p-value**: 2.83e-05
  - **95% CI for Difference of Means**: [-298.95, -111.37]
- **Interpretation**: The mean difference between critic and user reviews is -205.16. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a significant difference between the mean number of critic reviews and user reviews.

Benefits:

- **Identify Differences**: The analysis highlights significant differences in the number of reviews from critics versus users, providing insights into how these groups engage with movies.

- **Data-Driven Insights**: Helps filmmakers and marketers understand the different dynamics between critic and user feedback, informing strategies for review management and audience engagement.
- **Confidence Intervals**: The CI provides a range for the difference in means, adding robustness to the conclusions drawn from the hypothesis test.

9.3 Correlation and Hypothesis Test - Interpretation of Results



Correlation Analysis:

- **Critic Reviews vs. User Reviews**:
  - **Correlation (r)**: 0.66
  - **p-value**: 0.00
  - **95% CI**: [0.467, 0.860]
  - **Interpretation**: Significant positive correlation indicating that movies with higher numbers of critic reviews also tend to have higher numbers of user reviews.
- **Critic Reviews vs. IMDb Score**:
  - **Correlation (r)**: 0.31
  - **p-value**: 0.00
  - **95% CI**: [0.117, 0.509]
  - **Interpretation**: Moderate positive correlation indicating that movies with more critic reviews tend to have higher IMDb scores.

Hypothesis Testing:

- **Critic Reviews vs. User Reviews**:
  - **Null Hypothesis**: There is no correlation between the number of critic reviews and user reviews.
  - **Result**: p-value < 0.05, reject the null hypothesis, indicating a significant correlation.
- **Critic Reviews vs. IMDb Score**:
  - **Null Hypothesis**: There is no correlation between the number of critic reviews and IMDb scores.
  - **Result**: p-value < 0.05, reject the null hypothesis, indicating a significant correlation.

Benefits:

- **Understanding Relationships**: Helps identify significant relationships between variables, aiding in comprehensive data analysis.
- **Statistical Validation**: Provides statistical evidence to support findings, enhancing the robustness and reliability of the analysis.
- **Informed Decisions**: Guides stakeholders in making data-driven decisions based on the relationships between reviews and ratings.

9.4 Linear Regression and Residual Plot

Linear Regression Equation:

- **Equation**: Gross = intercept + coef_budget * Budget
- This equation predicts the gross revenue based on the movie budget.

Plotting the Regression Line and Residuals:

- **Regression Line**: The scatter plot with the regression line (red) shows the relationship between budget and gross revenue.
- **Residual Plot**: The residual plot visualizes the differences between actual and predicted values, indicating the model's accuracy.

Model Evaluation:

- **R-squared**: 0.53
- **Interpretation**: The R-squared value indicates that approximately 53% of the variability in gross revenue can be explained by the budget. This suggests that while the budget is a significant predictor, other factors also play a considerable role in determining gross revenue.

Conclusion:

- **Model Quality**: The linear regression model provides a moderate fit for predicting gross revenue based on budget. However, the residual plot and R-squared value indicate that the model could be improved by including additional predictors to enhance its accuracy and reliability.

9.5 Multi-Regression Models and Adjusted R-squared

Model 1:

- Equation: Gross = 4528809.78 + 1461.64 × Budget + 14311.65 × Num Critic Reviews + 56268.44 × Duration
- Adjusted R-squared: 0.943

Model 2:

- Equation: Gross = 85157.67 + 201.51 × Budget + 12369.60 × Num Critic Reviews + 1402675.46 × IMDb Score
- Adjusted R-squared: 0.968

Model 3:

- Equation: Gross = 180727.44 + 180727.45 × Budget + 306.92 × Num Critic Reviews + 180727.45 × Movie Facebook Likes

- Adjusted R-squared: 0.997

Conclusion:

- **Best Model**: Model 3
- **Reason**: Model 3 has the highest adjusted R-squared value (0.997), indicating it explains the most variability in the target variable, gross revenue. It includes the variables budget, num_critic_for_reviews, and movie_facebook_likes, which together provide the best predictive power.

This analysis shows that incorporating multiple predictors significantly enhances the model's ability to predict box office gross, with Model 3 being the most accurate and reliable.

11 PERFORMANCE AND ACCURACY OF THE MODELS

Baseline Model: Multiple Linear Regression

- **R-squared**: 0.26, indicating that the model explains 26% of the variance in box office revenue.
- **RMSE**: 63,201,615.10
- **MAE**: 41,483,921.47

Advanced Models: Random Forests and Gradient Boosting

- **Random Forests**:
  - **R-squared**: 0.52, showing improved explanatory power.
  - **RMSE**: 50,816,476.12
  - **MAE**: 32,679,727.97
  - **Feature Importance**:
    - Budget: 43.34%
    - Num Critic for Reviews: 21.19%
    - IMDb Score: 14.01%
    - Combined Actor Facebook Likes: 10.76%
    - Duration: 10.70%
- **Gradient Boosting**:
  - **R-squared**: 0.54, indicating the best performance among the models.
  - **RMSE**: 49,914,642.46
  - **MAE**: 32,454,856.61
  - **Feature Importance** (similar to Random Forests but slightly different weights)

Conclusion

Our project has achieved a comprehensive understanding of the factors influencing movie box office success, developed valuable skills in data analysis and machine learning, and provided robust predictive models with significant explanatory power. The models, especially Gradient Boosting, demonstrated substantial accuracy in estimating box office revenue, providing actionable insights for stakeholders in the film industry.

10 MILESTONES

Our milestones are designed to track the progress and ensure the timely completion of our project. The milestones include both completed and upcoming tasks:

Competed Milestones:

1. **Milestone 1:** Project Proposal Submission
   Complete and submit the project proposal, including the project overview, goals, and initial EDA steps.
2. **Milestone 2:** Data Collection and Preparation
   Collect and preprocess the dataset, including normalization and cleaning of data.
3. **Milestone 3:** Exploratory Data Analysis (EDA) Completion
   Perform EDA on all numerical and categorical variables. Generate plots and save them as PNG files. Document findings and interpretations.
4. **Milestone 4:** Correlation and ANOVA Analysis
   Conduct correlation analysis for numerical variables and ANOVA for categorical variables. Save the results and document interpretations.
5. **Milestone 5:** Time Series Analysis
   Perform time series analysis to visualize trends over the years. Save the plots and document findings.
6. **Milestone 6:** Final Report Draft
   Compile all findings, plots, and interpretations into a draft report. Ensure all sections are completed and references are properly formatted.
7. **Milestone 7:** Final Report Submission
   Submit the final report, including all analyses, plots, and documented findings.

11 RESPONSIBILITIES OF EACH GROUP MEMBER

Each group member is responsible for specific tasks to ensure full participation and contribution to the project. The responsibilities are divided as follows, with each member analyzing five variables:

- **Uday Bhaskar Valapadasu**
  - **Variables:** num_critic_for_reviews, duration, director_facebook_likes, actor_3_facebook_likes,
  - **Responsibilities:** Data Collection and Preparation, EDA for assigned variables, Normality Assessment, Proposal Documentation, ANOVA for Categorical Variables, Final Report Compilation, Time Series Analysis
- **Rohit Suddala:**
  - Variables:actor_1_facebook_likes,num_user_for_reviews, budget, title_year
  - Responsibilities: EDA for assigned variables, Correlation Analysis, ANOVA for Categorical Variables, Final Report Compilation, Time Series Analysis
- **Sapthagiri Naik Bhukya:**
  - Variables: actor_2_facebook_likes, imdb_score, aspect_ratio, movie_facebook_likes
  - Responsibilities: EDA for assigned variables, ANOVA for Categorical Variables, Time Series Analysis, Final Report Compilation

12 REFERENCE

All references used in this project are formatted according to the IEEE/ACM guidelines. Below are the references cited in our project proposal:

1. Learning Agency Lab. 2022. Feedback Prize Competition Series. Retrieved Nov 4, 2022 from https://www.the-learning-agency-lab.com/the-feedback-prize-overview/
2. Kaggle. 2022. Feedback Prize – English Language Learning. Code. Retrieved Nov 4, 2022 from https://www.kaggle.com/competitions/feedback-prize-english-language-learning/code
3. Kaggle. 2022. Feedback Prize – English Language Learning. Leaderboard. Retrieved Nov 4, 2022 from https://www.kaggle.com/competitions/feedback-prize-english-language-learning/leaderboard
4. Kaggle. 2022. Feedback Prize – English Language Learning. Evaluation. Retrieved Nov 4, 2022 from https://www.kaggle.com/competitions/feedback-prize-english-language-learning/overview/evaluation
5. Minitab Blog Editor. July 1, 2011. Assumptions and Normality. Minitab Blog. Retrieved Nov. 6 from https://blog.minitab.com/en/quality-data-analysis-and-statistics/assumptions-and-normality
6. jxmorris12. Jan 12, 2022. LanguageToolPython 2.7.0. Retrieved Nov 6, 2022 at https://github.com/jxmorris12/language_tool_python
7. Francis, W. Nelson & Henry Kucera. 1967. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press.
8. Martin, S., J. Liermann, and H. Ney. 1998. "Algorithms for Bigram and Trigram Word Clustering." Speech Communication 24 (1): 19–37. Retrieved from https://search-ebscohost-com.libproxy.library.unt.edu/login.aspx?direct=true&db=inh&AN=6006417&scope=site
9. Hockenmaier, Julia. "Lecture 3: Language Models." Natural Language Processing, n.d., 50. Retrieved from https://courses.engr.illinois.edu/cs447/fa2018/Slides/Lecture03.pdf
10. Dataset URL: https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

These references support our methodologies, tools, and approaches used in the project.