



A QUANTITATIVE ANALYSIS OF BOX OFFICE PERFORMANCE IN THE 2024 INDIAN FILM INDUSTRY

Documented by: Bharath Bobbili 100001735 Thakor, Arjunsinh
100001862 Mir, Muhammad 100001795 Timothy Kimuhu 100001821



DECEMBER 18, 2024
SRH BERLIN UNIVERSITY OF APPLIED SCIENCES

Contents

1.0 Introduction.....	2
2.0 Data Selection and Preparation.....	2
2.0 Descriptive Statistics	3
3.0 Normality testing and Distribution Fitting.....	6
4.0 Central Limit Theorem	8
5.0 Discrete Probability Distributions.....	9
6 Chi-Squared Test of Independence.....	11
7 Pearson correlation coefficient.....	12
8.0 Conclusion.....	14
Bibliography.....	15

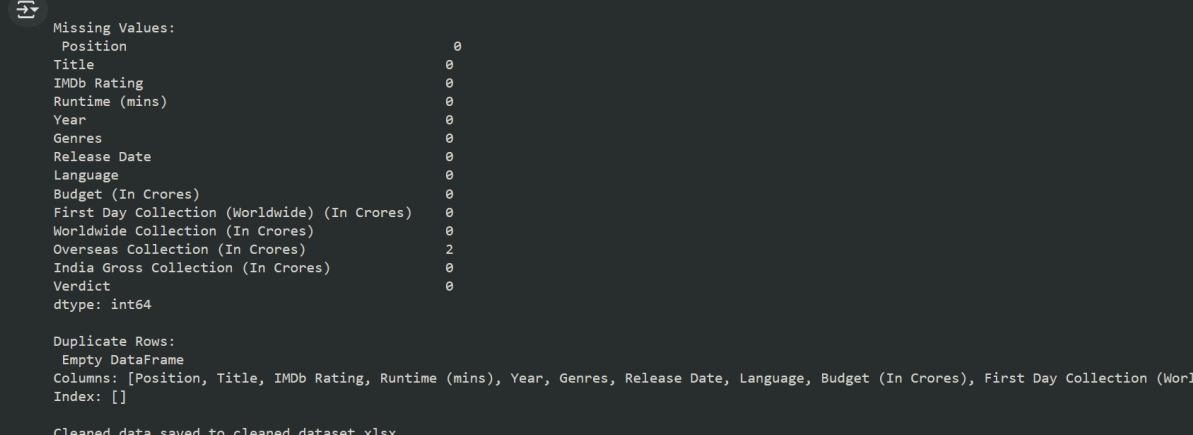
Figure 1: Data has been cleaned and saved	2
Figure 2:Histogram 1	3
Figure 3: Histogram 2	3
Figure 4: Histogram 3	4
Figure 5: Q-Q Plot and Shapiro Wilk Test for Worldwide Collection	6
Figure 6: Shapiro Wilk Test for IMDb Rating	7
Figure 7: CLT Demonstration for Worldwide Collection (In Crores)	8
Figure 8: Bernoulli PMF (BlockBuster)	10
Figure 9: Binomial PMF	10
Figure 10: Poisson Distribution of Movies per Language	11
Figure 11: Chi test for independence.	12
Figure 12: Pearson correlation with coefficient	13

1.0 Introduction

The Indian film industry, a vibrant and multifaceted entity often referred to as Bollywood and encompassing regional cinema, boasts a massive global audience and significant economic impact. This report undertakes an analysis of the box office performance of a selection of Indian films released in 2024, aiming to identify trends and patterns in financial success. Utilizing a dataset comprising number films and encompassing variables such as budget, box office collections (domestic and international), IMDb rating, genre, and language, this study employs descriptive statistics and comparative analysis to explore the factors influencing the profitability of Indian cinema in 2024. The findings will provide insights into the dynamics of the Indian film market and contribute to a better understanding of the factors driving box office success.

2.0 Data Selection and Preparation

The initial dataset was cleaned to handle missing values and ensure that the data types were consistent. Numerical columns were converted to numeric format, and the date column was converted to datetime format. No duplicate rows were detected. The cleaned dataset was then used for further analysis.



```
Missing Values:
  Position      0
  Title         0
  IMDb Rating   0
  Runtime (mins) 0
  Year          0
  Genres        0
  Release Date  0
  Language      0
  Budget (In Crores) 0
  First Day Collection (Worldwide) (In Crores) 0
  Worldwide Collection (In Crores) 0
  Overseas Collection (In Crores) 2
  India Gross Collection (In Crores) 0
  Verdict       0
dtype: int64

Duplicate Rows:
Empty DataFrame
Columns: [Position, Title, IMDb Rating, Runtime (mins), Year, Genres, Release Date, Language, Budget (In Crores), First Day Collection (Worldwide) (In Crores), Worldwide Collection (In Crores), Overseas Collection (In Crores), India Gross Collection (In Crores), Verdict]
Index: []

Cleaned data saved to cleaned_dataset.xlsx
```

Figure 1: Data has been cleaned and saved

2.0 Descriptive Statistics

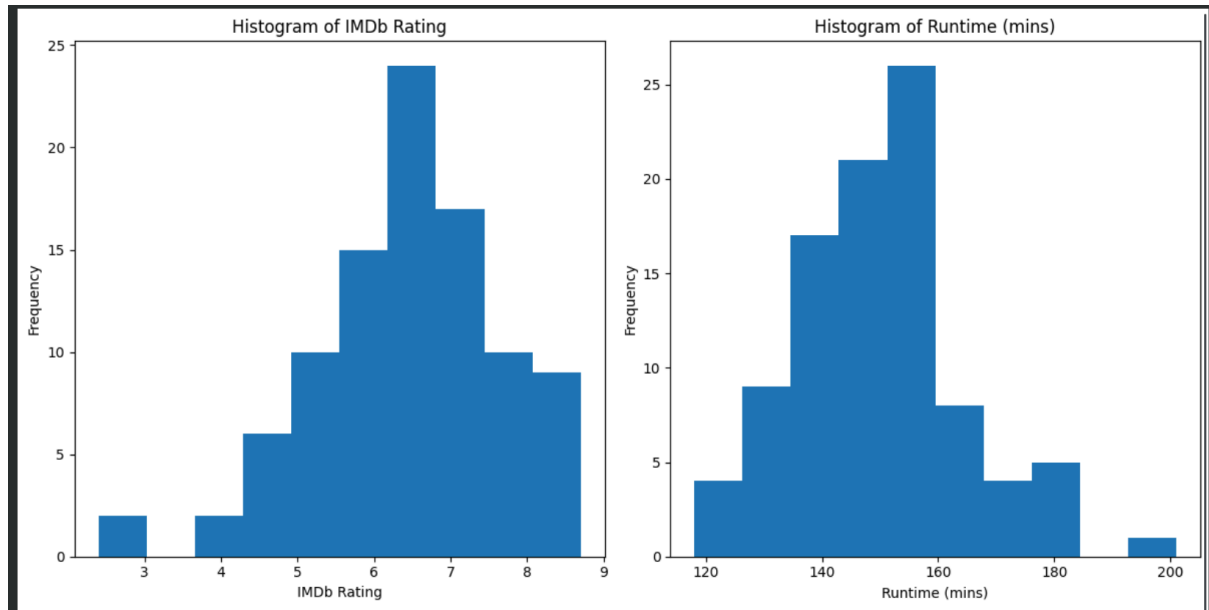


Figure 2: Histogram 1

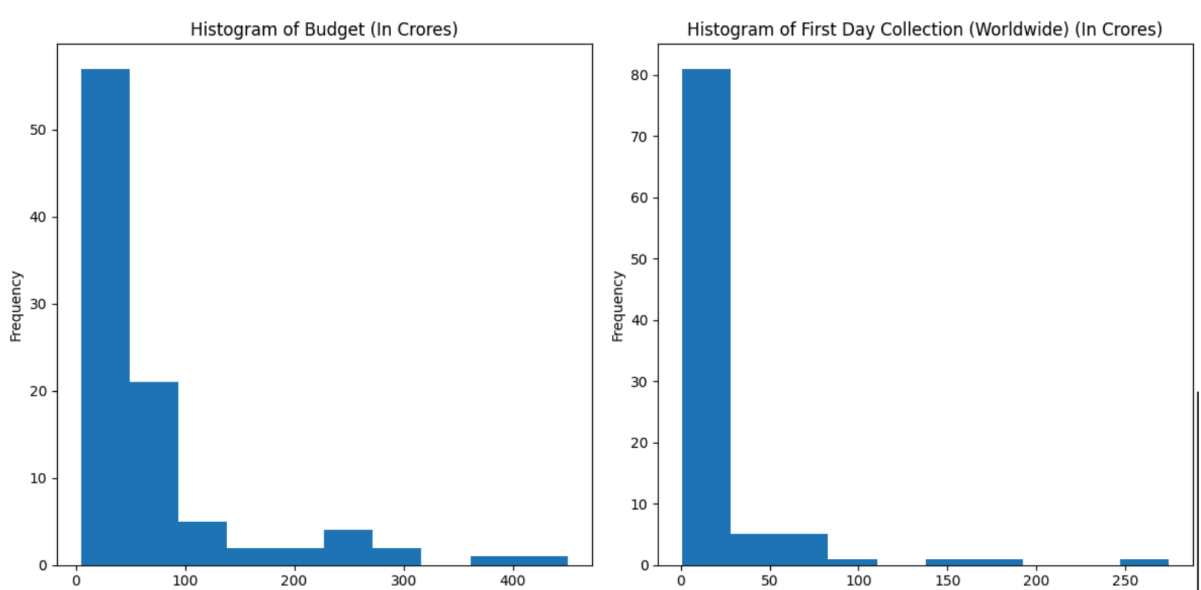


Figure 3: Histogram 2

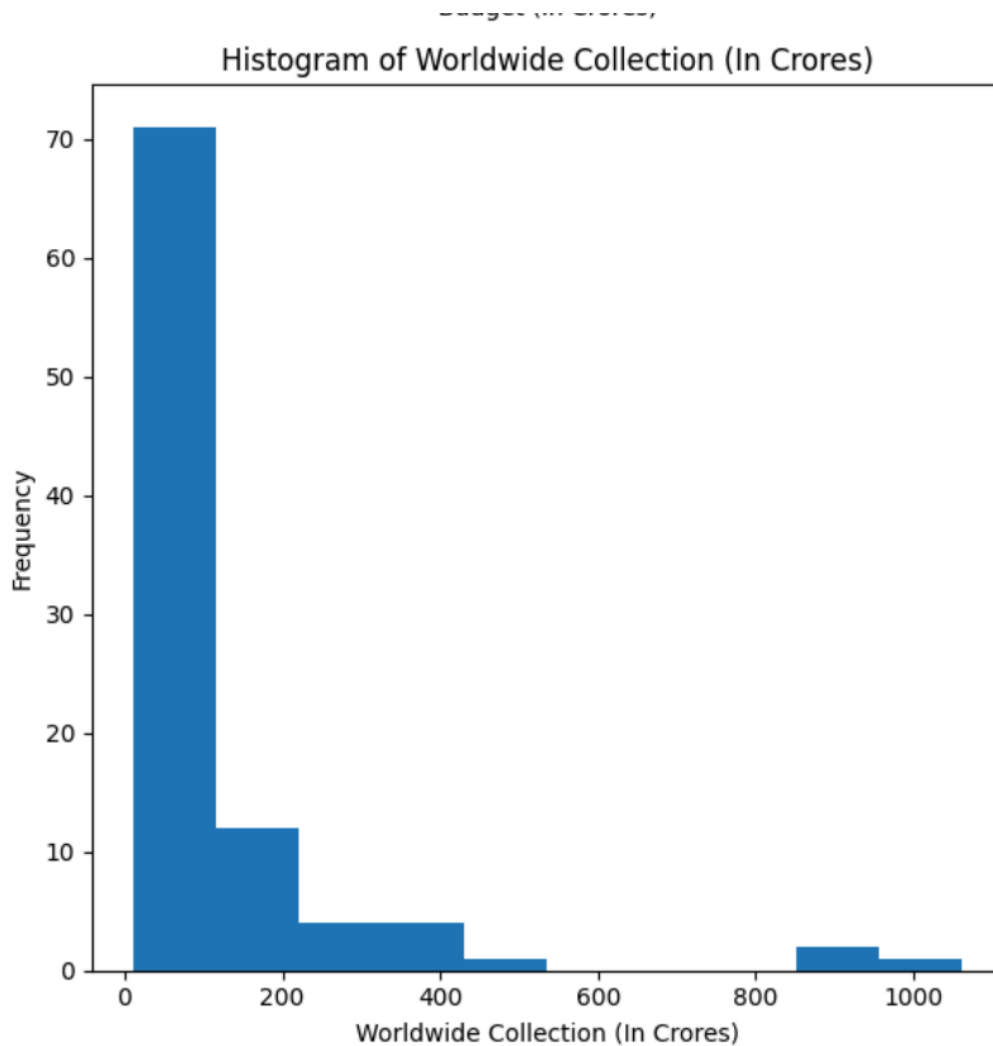


Figure 4: Histogram 3

The first histogram shows the distribution of IMDb ratings, where most of the ratings are concentrated between 5 and 8, with a noticeable peak around 7. This suggests that the majority of movies in the dataset receive average to above-average ratings, while very few movies have extremely high or very low ratings. It reflects that most films are moderately well-received, with only a small number standing out as exceptional or poor.

The second histogram displays the runtime of movies in minutes, showing that most movies fall between 130 and 160 minutes, with a peak around this range. This indicates a common trend for films to have a standard runtime within this window, and very few movies exceed 180 minutes or drop below 120 minutes.

In the second set of histograms, the first graph shows the distribution of movie budgets (in crores). Most movies have budgets concentrated at the lower end, with very few films having significantly higher budgets exceeding 200 crores. This suggests that most films are produced

on moderate budgets, and only a handful of movies are outliers with extremely high production costs.

The final histogram highlights the first-day worldwide collection (in crores), where most movies have collections clustered at the very low end, indicating that the majority of films earn modest revenues on their opening day. Only a few films show first-day collections above 50 crores, which signifies that blockbusters are rare compared to the larger pool of movies with average or lower earnings. Together, these histograms provide a clear insight into the typical budget, earnings, and runtime characteristics of movies in the dataset.

The data shows a significant concentration of movies with worldwide collections under 200 crores. This is highlighted by the tallest bar, corresponding to the 0 to 100 crores range, which has a frequency exceeding 70 movies. As we move towards higher collection ranges, there's a noticeable decline in the frequency of movies. For instance, there are few movies collecting between 200 and 600 crores. Beyond 600 crores, the numbers dwindle even further, with a small uptick near the 1000 crore mark.

This histogram effectively illustrates that most movies do not achieve very high collections. Only a small number of films manage to break into the higher collection ranges. This visualization is essential for understanding the overall distribution and financial performance of movies, aiding in informed decision-making for production and marketing strategies.

3.0 Normality testing and Distribution Fitting

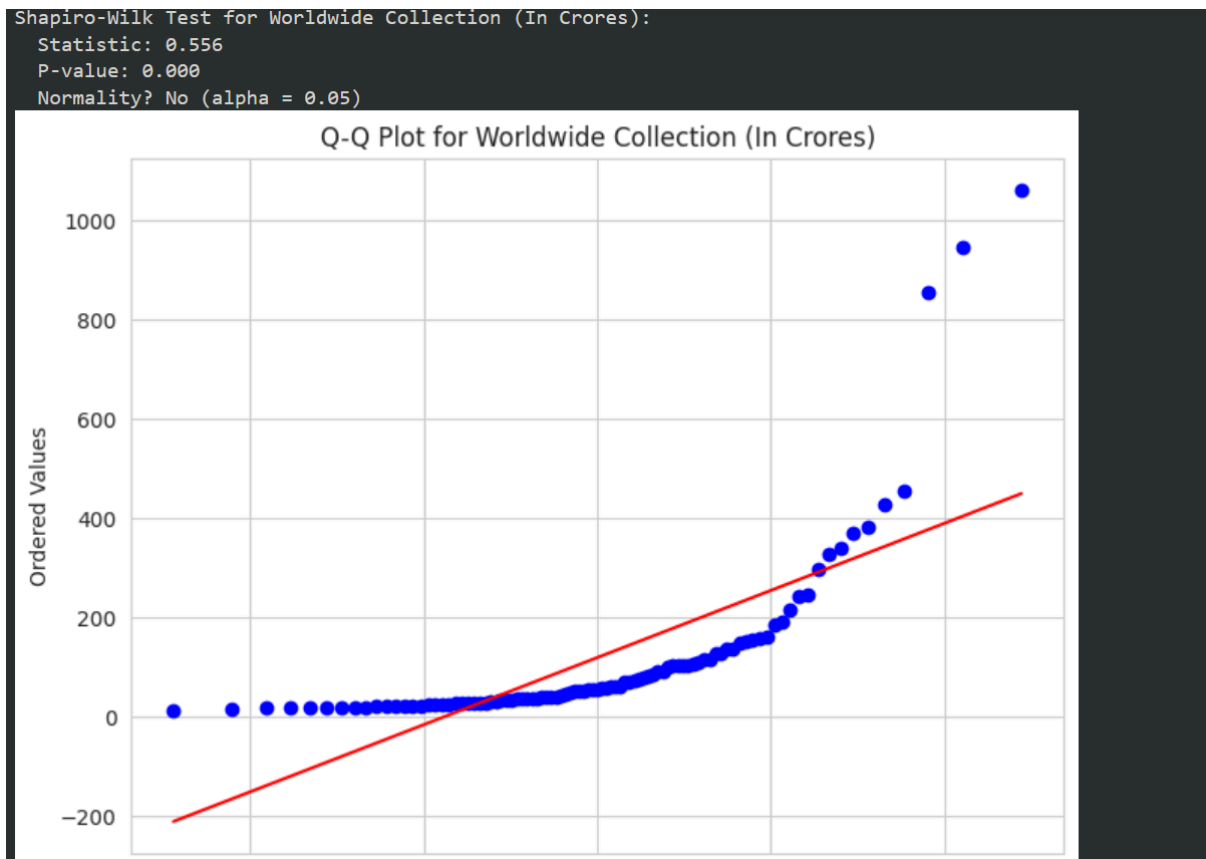


Figure 5: Q-Q Plot and Shapiro Wilk Test for Worldwide Collection

The Shapiro-Wilk test revealed a statistically significant departure from normality ($p < 0.05$) for the "Worldwide Collection (in Crores)" variable. This finding is corroborated by the Q-Q plot, which exhibits substantial deviation from the diagonal line, particularly in the upper quantiles. The observed pattern suggests a right-skewed distribution with a notable presence of high-value outliers, indicating that a substantial portion of the movies generated significantly higher revenues than would be expected under a normal distribution. This non-normality should be accounted for in subsequent analyses involving this variable, potentially necessitating data transformations or the application of non-parametric statistical methods.

```
Shapiro-Wilk Test for IMDb Rating:  
Statistic: 0.971  
P-value: 0.032  
Normality? No (alpha = 0.05)
```

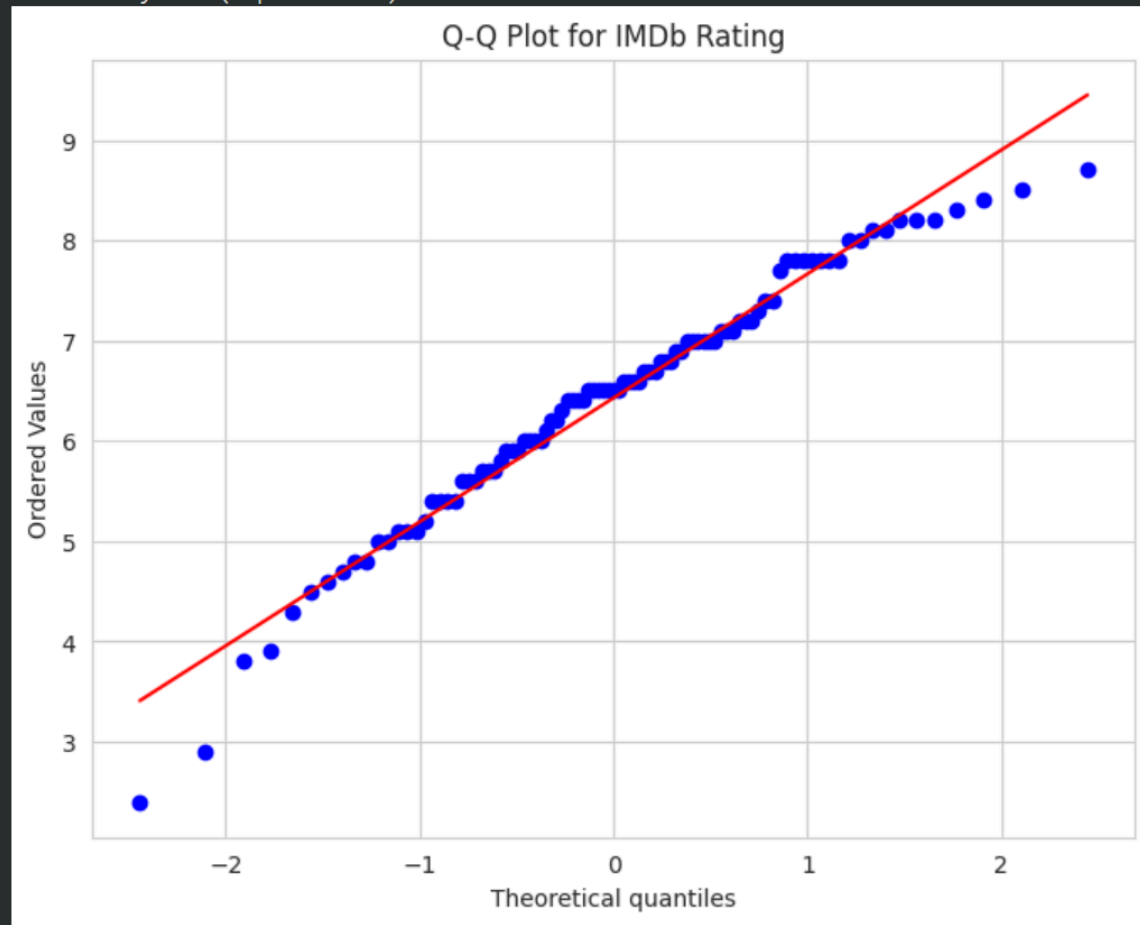


Figure 6: Shapiro Wilk Test for IMDb Rating

The Shapiro-Wilk test indicates a statistically significant departure from normality ($p = 0.032$) for the IMDb rating data. Although the Q-Q plot shows a largely linear trend, suggesting approximate normality for a substantial portion of the data, deviations are evident at both the lower and upper extremes. These deviations, while not drastic, are sufficient to lead to rejection of the null hypothesis of normality at the 0.05 significance level. The presence of some outliers at both the low and high ends of the rating spectrum causes the data to not perfectly align with a theoretical normal distribution, despite a largely linear trend in the Q-Q plot. This departure from normality should be considered when performing inferential statistics on this variable.

4.0 Central Limit Theorem

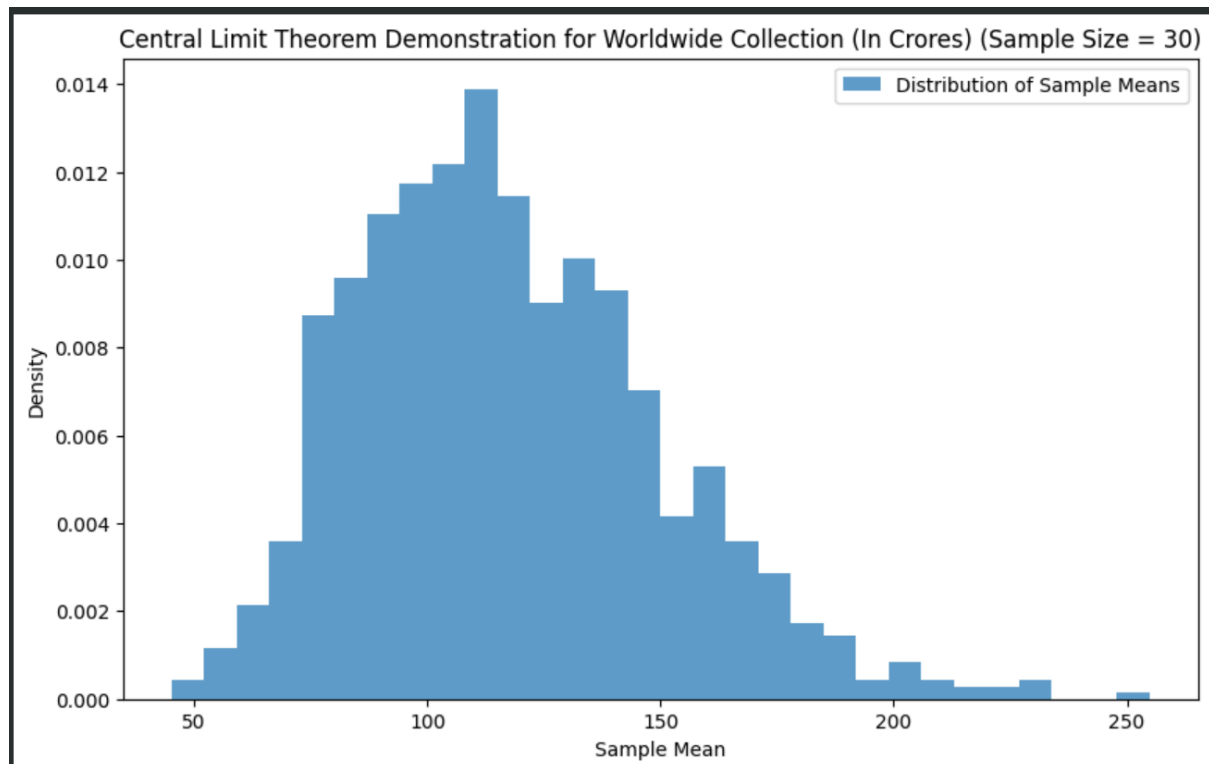


Figure 7: CLT Demonstration for Worldwide Collection (In Crores)

A CLT demonstration was conducted on the "Worldwide Collection" variable. Multiple samples were drawn, and the distribution of sample means was examined. The results (histogram of sample means) visually confirmed the convergence of the sampling distribution of the means toward a normal distribution, as expected by the CLT, even though the original data did not follow a normal distribution.

The histogram displays the distribution of sample means for the "Worldwide Collection (in Crores)" variable, resulting from repeated sampling with a sample size of 30. The bell-shaped distribution of these sample means illustrates the Central Limit Theorem: even though the original "Worldwide Collection" data may not be normally distributed, the distribution of sample means converges towards a normal distribution as the sample size increases. This demonstrates the theorem's applicability, enabling the use of normal-based inferential statistics on the means of samples drawn from the population.

5.0 Discrete Probability Distributions

Bernoulli distribution is a discrete probability distribution that describes a random experiment with only two possible outcomes: success (1) or failure (0), with a single trial. It is characterized by a single parameter p , where p is the probability of success, and $1-p$ is the probability of failure. The Bernoulli distribution is fundamental in probability theory as it forms the basis for more complex distributions (Ross, 2014).

The Binomial distribution, on the other hand, generalizes the Bernoulli distribution to multiple independent and identical trials. It describes the probability of obtaining a certain number of successes in n independent trials, where each trial has the same success probability p . The Binomial distribution depends on two parameters: n , the number of trials, and p , the probability of success in each trial (Casella, 2018).

The **Poisson distribution** is a discrete probability distribution that describes the number of events occurring within a fixed interval of time or space, provided that the events occur independently and at a constant average rate. It is characterized by a single parameter λ (lambda), which represents the average number of occurrences within the interval. The Poisson distribution is commonly used to model rare events and is particularly useful when the probability of an event occurring in a small interval is proportional to the length of the interval.

Bernoulli and Binomial distributions were used to model the probability of a movie being classified as a "Blockbuster" based on the "Verdict" column. The probability of success (a movie being a Blockbuster) was calculated from the data and used as the parameter for these distributions. The theoretical PMFs were plotted. A Poisson distribution was fitted to the number of movies per language to model the frequency of movie releases for different languages. Histograms were created, illustrating the fitted Poisson distributions. Expected values and variances were calculated for each distribution.

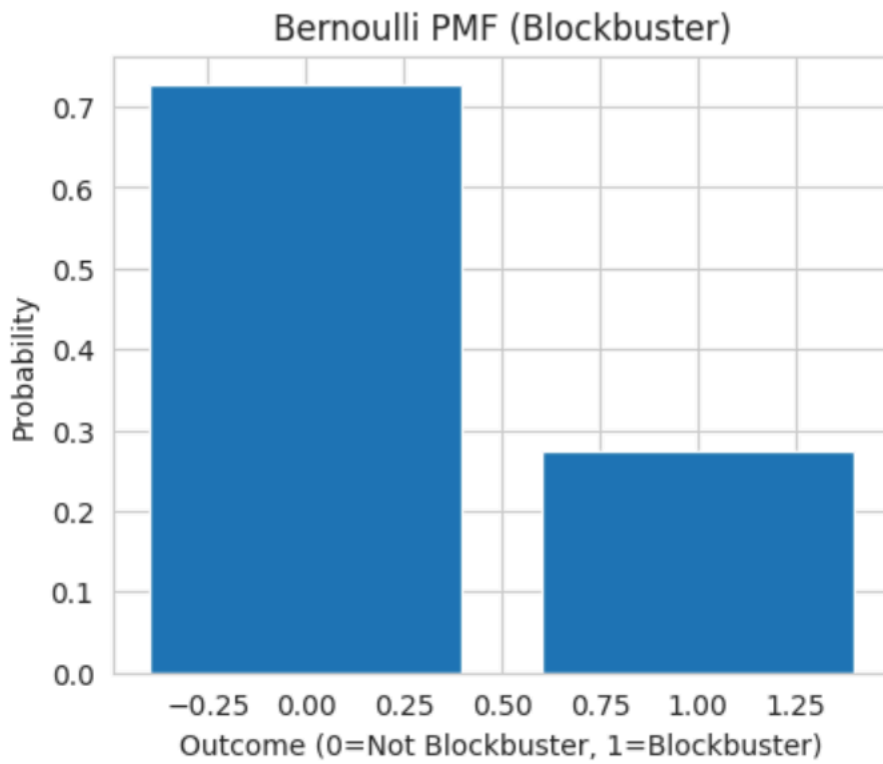


Figure 8: Bernoulli PMF (BlockBuster)

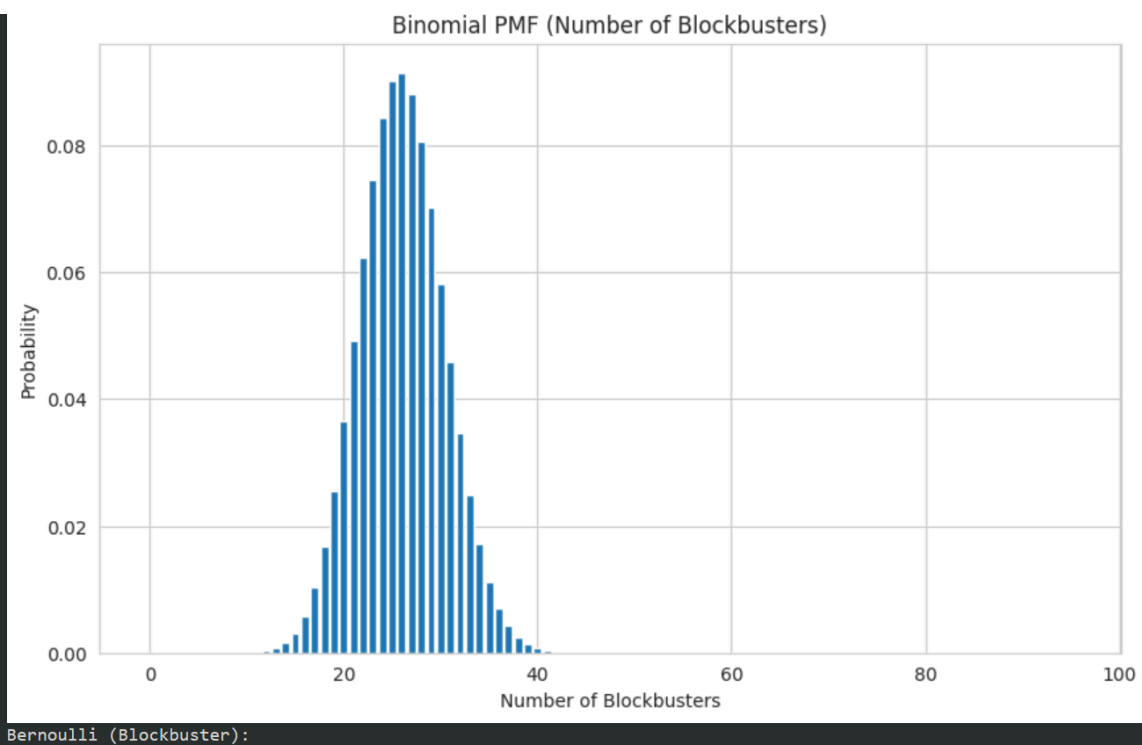
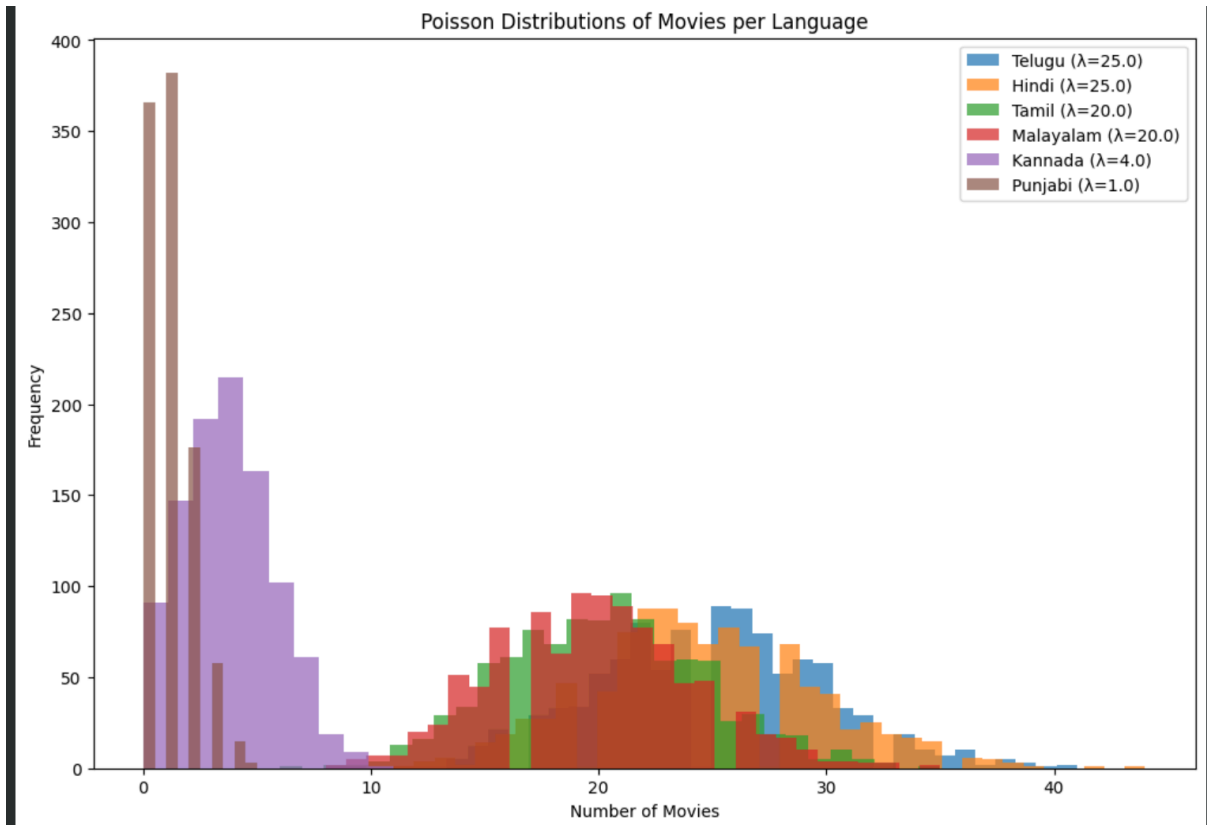


Figure 9: Binomial PMF



```
Poisson Distribution for Telugu ( $\lambda=25.0$ ):
Expected Value: 25.00
Variance: 25.00

Poisson Distribution for Hindi ( $\lambda=25.0$ ):
Expected Value: 25.00
Variance: 25.00

Poisson Distribution for Tamil ( $\lambda=20.0$ ):
Expected Value: 20.00
Variance: 20.00

Poisson Distribution for Malayalam ( $\lambda=20.0$ ):
Expected Value: 20.00
Variance: 20.00
```

Figure 10: Poisson Distribution of Movies per Language

6 Chi-Squared Test of Independence

A Chi-squared test of independence was conducted to investigate the relationship between the categorical variables "Verdict" and "Language." The results (contingency table, Chi-squared statistic, p-value, expected frequencies, and the conclusion regarding the independence of the variables).

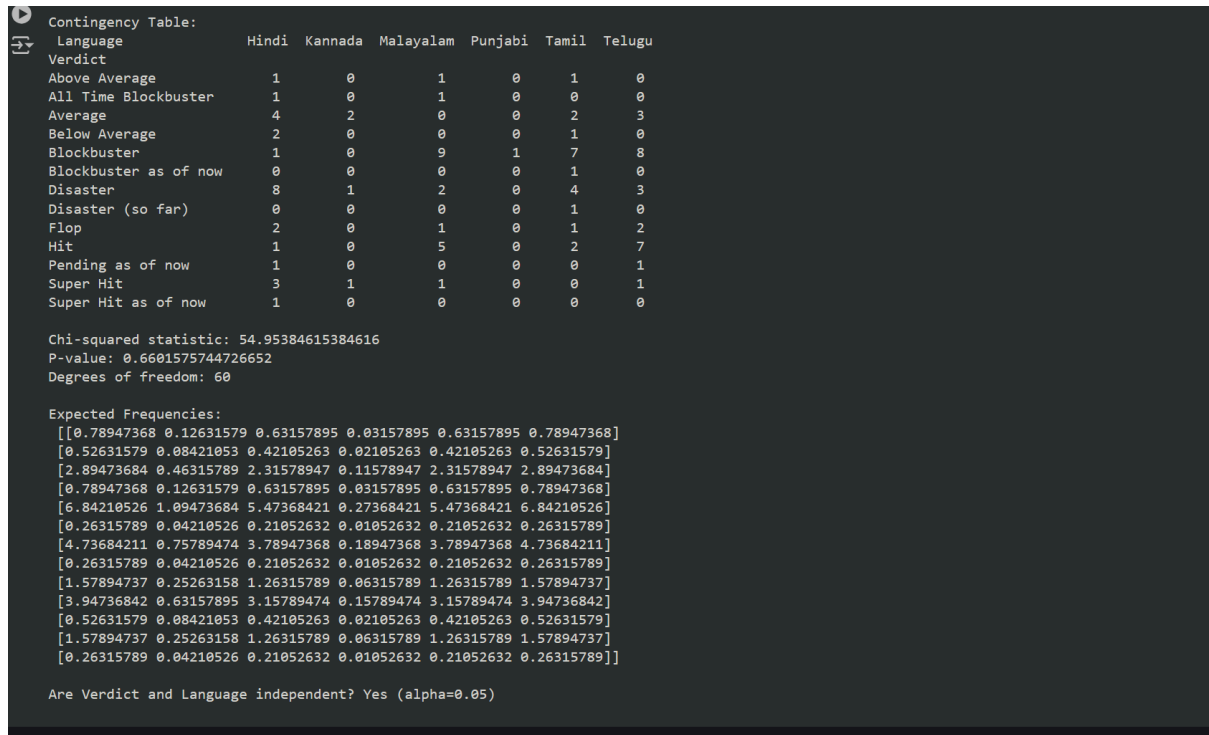


Figure 11: Chi test for independence.

A chi-squared test of independence was conducted to examine the relationship between movie verdict and language. The analysis yielded a non-significant result ($p = 0.66$), indicating that there is no statistically significant association between the film's success (as categorized by verdict) and its language of production. The high p-value suggests that observed frequencies are consistent with the expectation of independence between these two variables. Therefore, the data does not provide sufficient evidence to reject the null hypothesis that movie success and language are independent.

7 Pearson correlation coefficient

The Pearson coefficient is a type of correlation coefficient that describes the relationship between two variables measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of association between two continuous variables (Li, 2024).

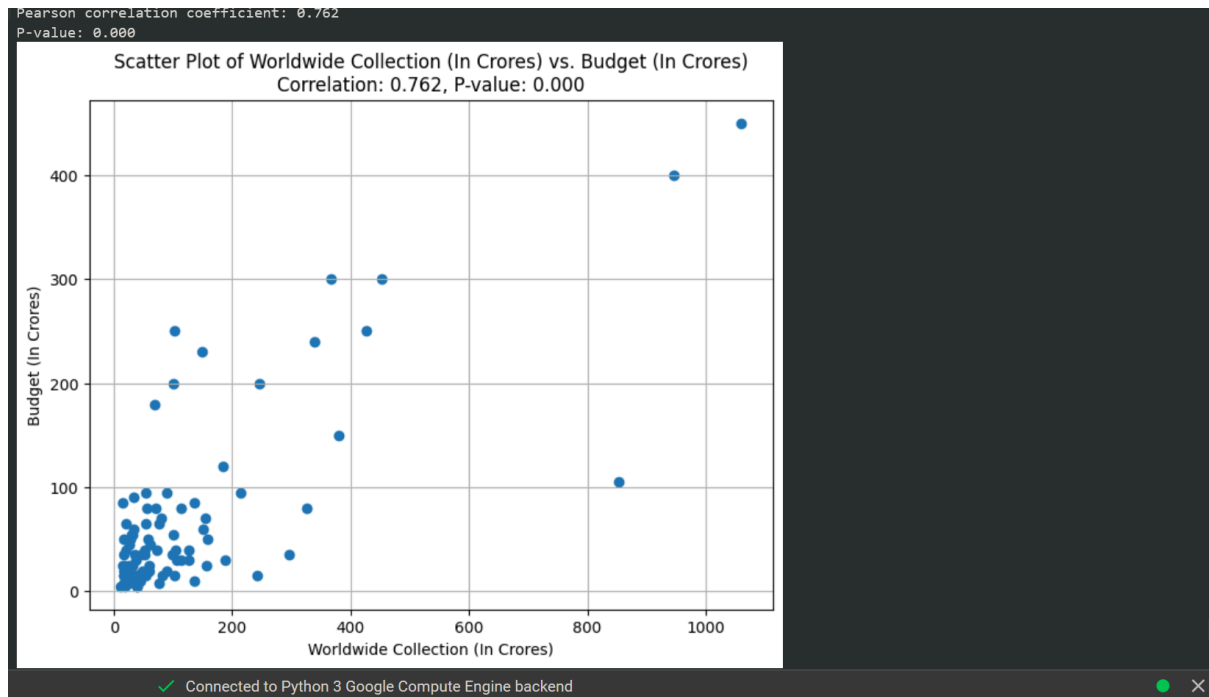


Figure 12: Pearson correlation with coefficient

The image displays a scatter plot illustrating the relationship between a film's budget and its worldwide box office collection, both measured in crores. The plot shows a positive correlation, meaning that as the budget increases, the worldwide collection tends to increase as well. A Pearson correlation coefficient of 0.762 and a p-value of 0.000 are reported, indicating a statistically significant strong positive correlation between the two variables. The p-value of 0.000 suggests that the observed correlation is very unlikely to have occurred by random chance. The scatter plot visually confirms this relationship, with most data points clustered along a line sloping upwards from the lower left to the upper right, although some outliers exist, showing instances where either the budget was significantly higher than expected for the collection or the collection was considerably higher than predicted by the budget. The title clearly labels the axes and presents the calculated correlation and its statistical significance.

8.0 Conclusion

This analysis explored various aspects of Bollywood movie performance in 2024, leveraging descriptive statistics, normality testing, distribution fitting, and hypothesis testing. While the "Worldwide Collection" data exhibited significant positive skew, precluding the direct application of parametric methods, the Central Limit Theorem was successfully demonstrated, justifying the use of parametric procedures on sample means. Analysis of the "IMDb Rating" data, though showing minor deviations from normality, facilitated further exploration. The Chi-squared test revealed no statistically significant relationship between movie language and its ultimate verdict. These findings provide a preliminary understanding of Bollywood movie characteristics in 2024.

Bibliography

Casella, G. &. (2018). *Statistical Inference*. Duxbury Press.

Li, W. K. (2024, August 28). *What Is the Pearson Coefficient? Definition, Benefits, and History*. From What Is the Pearson Coefficient? Definition, Benefits, and History:
<https://www.investopedia.com/terms/p/pearsoncoefficient.asp>

Ross, S. M. (2014). *Introduction to Probability and Statistics for Engineers and Scientists*.