Project Report: **Probability and Statistics in IMDb Movie Analysis**

1. Introduction:

The aim of this project is to apply principles of probability and statistics to analyze IMDb movies' success. We explore the relationships between movie budget, IMDb ratings, profit, and other factors like genre, language, and country of origin. By employing statistical techniques, we gain insights into what contributes to a movie's success and how different variables are interconnected.

2. Data Collection and Preprocessing:

We obtained the IMDb movies dataset containing information on movie budget, IMDb rating, profit, genre, language, and country of origin. The data was preprocessed to handle missing values and ensure data integrity for accurate analysis.

3. Defining Success Metric:

In defining success, we considered two main factors:

IMDb Rating: Movies with high IMDb ratings (typically above a certain threshold) are considered successful, indicating positive audience reception and critical acclaim.

Box Office Earnings / Profit (Revenue - Budget): Movies that achieved a specific level of box office earnings were also considered successful, indicating commercial success.

4. Joint Probability Distribution (JPD):

To understand the joint relationship between Budget and IMDb Rating; and between Budget and Profit, we modeled the data using a bivariate normal distribution. This allowed us to visualize the joint probability distribution as a 3D wireframe or contour plot. The plot provided insights into the likelihood of certain budget-IMDb rating and budget-profit combinations and highlighted regions with higher probabilities.

5. Maximum Likelihood Estimate (MLE):

We calculated the Maximum Likelihood Estimate (MLE) of the correlation coefficient between Budget and IMDb Rating; and Budget and Profit for specific genres, such as 'Action' movies. The MLE helped us quantify the strength and direction of the linear relationship between the two variables for different movie genres.

6. Confidence Interval:
Confidence intervals were employed to estimate the average IMDb rating and average Profit of movies with a budget of $10 million. The 95% confidence interval provided a range of plausible values for the average IMDb rating and average Profit, accounting for the variability in the data.

7. Likelihood and Log-Likelihood:
The likelihood function was used to model the joint probability distribution of Budget and IMDb Rating; and Budget and Profit as a bivariate normal distribution. To simplify computations and avoid numerical issues, we also computed the log-likelihood, which is the logarithm of the likelihood function.

8. Practical Insights:
Through our analysis, we gained several practical insights:
The correlation between budget and IMDb rating; and budget and profit varied across different movie genres, with some genres showing a stronger association than others.
Confidence intervals provided valuable information about the average IMDb rating and average Profit of movies with a budget of $10 million..

9. Conclusion:
In conclusion, this project demonstrated the application of probability and statistics in analyzing IMDb movie data. By examining joint probability distributions, likelihoods, and confidence intervals, we gained valuable insights into movie success factors. The findings can guide decision-making in the film industry and contribute to a better understanding of audience preferences.

10. Future Enhancements:
In the future, the analysis could be expanded to include more factors, such as language, country, actors, directors, release dates, and marketing strategies, to build more comprehensive models of movie success. Additionally, machine learning techniques could be incorporated to predict movie success based on various features and attributes.

11. Acknowledgments:
We would like to acknowledge IMDb and Kaggle for providing the dataset used in this analysis. Also a special thanks to our mentor Varun Aggarwala, his guidance made this project possible.

This report summarizes the analysis performed using probability and statistics on the IMDb movies dataset. It highlights the methodologies used and the insights gained, making it a valuable resource for understanding the impact of various factors on movie success.

Sincerely,

Satyam Chaurasiya
Soham Sanyal
Santanu Sen
Prithviraj