Project Outline – Rhea Shah, Ishika Sippy

The dataset we are going to use is from the website Kaggle and it is called “The 100 lowest ranked movies dataset on IMDb.” https://www.kaggle.com/datasets/lakshayjain611/imdb-100-lowest-ranked-movies-dataset

We chose this dataset as both of us are people who want to pursue a career in film marketing and always came across data sets of the top 100 movies and say what makes a movie successful but the lowest 100 seemed extremely interesting and intriguing. The question we would be answering is “What makes a movie undervalued/unsuccessful amongst other factors except The Rating?”, “What are the factors that are common or similar for all these 100 films.?”, “What to not do, if you are a director/producer/actor?” The statistical analysis that we are planning to use are:

1. Scatterplot between the runtime and IMDb rating variables to show potential patterns or trend in the data, additionally we can use regression analysis to create a model that predicts the rating based on the runtime variable and correlation analysis to determine the correlation coefficient between the two variables.
2. For the second question: Looking at the coefficients which variables (stars, genre, review_count) have the strongest relationship on the IMDb rating? The link between the IMDb rating and the predictor variables can be modeled using multiple regression analysis. We can obtain the coefficients for each predictor variable from the regression

analysis and show the magnitude and direction of the link between the predictor and target variable.

When deciding on the duration and content of a film, producers and filmmakers can utilize this information to their advantage. Movie fans who want to choose the right films to see can also benefit from this information.