# Data Analysis Project Proposal

Group 8 - Qishi Liu, Ly Nguyen, Van Phan

We are interested to find out if the top movies' correlation between gross income and user votes vary significantly across different movie genres. Our research aims to explore the relationship between audience engagement and financial success in the film industry. We were motivated by our group's interest in movies and a desire to understand this aspect more deeply. Specifically, our research question is: Does the top movies' correlation between gross income and user votes vary significantly among different movie genres? This question is interesting to us because it opens the door to understand the cultural and social factors that influence movie-going behavior. For instance, do action movies resonate more strongly with audiences, reflecting a societal preference for high-energy, visually stimulating content? Or do romantic movies, often rooted in emotional depth, elicit a different kind of engagement that equally translates to box office success? One potential ethical concern can be the result of our model, which should be conservatively used. In other words, if we tend to use the model to make authoritative statements about the quality of films or predicting future successes could be misleading. We aim for a general understanding, acknowledging what makes a movie successful or critically acclaimed.

The data is sourced from Kaggle, created by Mayank Ray. Our data "IMDb Top 1000 Movies Dataset" is extracted from the "IMDb" website, a comprehensive online database renowned for its extensive information related to movies. Not only does IMDb collect verified data from filmmakers and studios, but also it has information submitted by the general public. The dataset provides data of 12 variables of 1000 observations corresponding to 1000 movies. The variables of the dataset include genres of films, IMDb ratings, Metascores (a single score that represents the critical consensus of movies), the number of user votes, their gross earnings, and etc. The method employed for selecting these movies is purposive sampling. The top 1000 movies that are chosen based on user ratings. By focusing on the top-rated movies, each representing an observation unit, this dataset offers a lens through which to examine cinematic success as a function of both critical reception and audience engagement. Key variables of interest such as IMDb ratings, user votes, and gross earnings, provide a multifaceted view of what drives a movie's profitability.

The main variables we are choosing to answer the question are "votes", "genre" and "gross in $". We will begin with a simple linear regression to estimate the relationship between the number of votes a movie receives and its gross income. Then, we compare the results among different genres of the movies. Our null hypothesis states that there is not a significant linear relationship between the number of votes and the gross income of top 1000 movies. The alternative hypothesis is that there is a significant linear relationship between the number of votes and the gross income of top 1000 movies. To test these hypotheses, we will employ linear regression analysis, initially examining each genre separately. Then, we compare these models across genres, focusing on the strength of the relationship in different categories. For the linear models, we will assure that each user votes and gross revenue are independent of each other. The data variables (vote users and gross earning) are approximately normally distributed and the simple linear regression test is appropriate for both variables, which is checked by using the linear regression model visualization. For the summary statistic, there is a wide range in the number of votes received and the gross revenue as well among the top movies. Besides, in the dataset, there are outliers that might skew the data checked by boxplots created for each variable, but there are no large outliers that would cause research bias in the model after checking for homoscedasticity. Generally, there is a strong relationship between the variables indicated by correlation coefficient of them and they follow a linear pattern.

Reference: Ray, M. (2023, October 16). *IMDb top 1000 movies dataset*. Kaggle. https://www.kaggle.com/datasets/mayankray/imdb-top-1000-movies-dataset/