Group members: Isabel O'Connor, Katie Shaughnessy, Emmanuella Cann, Carson Colyer, Krishu Wadhwa
Pre-Analysis Plan
DS 3001
Professor Johnson
November 4th, 2024

# Predicting Box Office Success: Enhancing Accuracy with Keyword Analysis

## Project Goal

For our project, we plan to create a predictive model for box office revenue (in thousands of dollars) using variables like studio, genre, etc. After evaluating the accuracy of this method, we will add keywords to the model and see if they increase predictive strength. The project will be executed using supervised learning in the form of a regression model. The observation used will be a single movie record.

To summarize, we will pursue answers for the following questions:

- To what extent do keywords enhance the predictive accuracy (in terms of measured RMSE) of a box office success model built on traditional variables such as genre, studio, and budget?
    - Do genre-specific keywords add unique predictive value to a model for box office success, or are there universal keywords that enhance predictions across multiple genres?
- How does the relative importance of keywords in predicting box office success vary based on the studio producing the film?

## Strategy:

Our strategy for this project is to develop a predictive model for box office revenue (in thousands of dollars) using variables such as studio, genre, and other relevant factors. We will then introduce keywords to assess whether they enhance the model's predictive strength. We will analyze whether there are specific keywords associated with films which correlate to higher rates of popularity, user rating, and box office success. For instance, we predict that movies associated with keywords such as "Marvel", "fantasy", "comedy", or "adventure" might enjoy greater box office success. Using supervised learning with a

regression model, where each observation represents a single movie record, we aim to capture the influence of these variables on box office outcomes.

A central question driving our analysis is the extent to which keywords improve the predictive accuracy of box office success models that traditionally rely on genre, studio, and budget. Specifically, we will measure improvements in model performance using metrics like R-Squared and RMSE (Root Mean Squared Error) to quantify the impact of keywords. Our investigation will also explore whether genre-specific keywords add unique predictive value compared to universal keywords that may enhance predictions across multiple genres. Moreover, our analysis will consider potential shifts in keyword relevance over time. Specifically, assessing how changes in audience preferences and industry trends impact the effectiveness of certain keywords. Finally, we aim to understand how the importance of keywords in predicting box office success varies based on the producing studio, potentially offering insights into tailored marketing and production strategies.

**Feature Engineering:**

After handling missing data through imputation or deletion, we will split the dataset into training and testing sets. Prior to this, categorical variables like title, genres, keywords, production companies and production countries will be one-hot encoded to provide insight into their relationship with the predicted variables. Principal Component Analysis (PCA) will also be used for numeric variables that are highly correlated. Furthermore, we will account for interaction effects, creating a single variable for highly correlated categorical variables. We will also perform any necessary transformations to our data to refine our model. To address potential overfitting from high-dimensional encoding, we will apply regularization techniques like LASSO. Through these analyses and our planned feature engineering, we aim to build a robust linear regression model that can effectively predict box office success based on the strategic use of keywords alongside other relevant variables. Through this project, we hope to provide insights into movie trends and to inform effective marketing and production strategies based on data-driven analysis.

**Measuring Results:**

To evaluate the model’s accuracy, we will use metrics like R-Squared and RMSE. The R-Squared value, or coefficient of determination, will indicate the proportion of variance in box office revenue explained by our predictor variables, showing how well the model captures trends in the data. RMSE, on the other hand, will measure the average magnitude of prediction errors, giving us insight into how close our model's predictions are to actual values.

Together, the R-Squared and RMSE metrics will be used to test the success of our model. Since models can be overfitted or underfitted, comparing the training and testing data RMSE can provide insight into how well our model will perform on novel data, as well as insights on how well our model is performing with current data. A high RMSE in the training and testing data will indicate our data is underfitted. On the contrary, a low RMSE in the training data and a low RMSE in the testing data will indicate the data is overfitted. A RMSE in which both the training and testing data RMSE are similar will indicate a well-trained model. Additionally, the R-Squared will give insight into how much of the variance in our predicted value is explained by our predictors. A low R-Squared where there are multiple variables in our model will signal us to a simpler, more refined model with more relevant variables.

**Potential Design Weaknesses:**

Given the large number of variables in the movie dataset, it could prove difficult to design the optimal regression for predicting box office revenue. Some variables, such as budget and production company, should be interacted within the regression to ensure maximal predictive outcomes. To inform the variables, interaction terms, and power terms included in the regression model, a short meta-analysis will be conducted to better understand the dynamics of movie revenue outcomes. If the model design still yields suboptimal predictive accuracy, further research and model engineering will be pursued.

Another potential challenge is our reliance on linear regression, which may not effectively capture the nonlinear and complex relationships affecting movie success. While factors like budget, director, keywords, runtime, and revenue play a role, there are also many other influences, such as cultural trends, social issues, marketing strategies, franchise popularity, and availability on streaming platforms or in theaters. To tackle this challenge, we will be using feature engineering to develop new variables that better reflect these complexities. For instance, we can create combined variables that look at the relationship between marketing spending and release timing to see how they impact box office results. Additionally, we can apply transformations to key variables, such as using logarithmic scales for budget and revenue, to help smooth out any extreme values. This way, we can improve our model's ability to understand the factors that contribute to a movie's success.

Additionally, certain categorical variables in the dataset have large numbers of potential values, which could affect the difficulty of dummy variable transformation. For instance, the “director” variable has a broad range of values, which could result in “messy” one-hot

encoding. To avoid this, we could engineer the director dummy variable such that directors associated with a large revenue impact (ex: Spielberg, Nolan, or Tarantino) are given their own value within the dummy variable, whereas all other directors are lumped into the “0” value. This method could also be applied to other categorical variables, where appropriate.