Movie Earnings Background Page

Throughout our work on the project, we manipulated our data, performed calculations using the data, and created machine learning algorithms based on the data. We started by manipulating the data in order to create new variables, better organize our dataframe, and categorize our data by changing some of the variable types. We then calculated values like r, R^2, adjusted R^2, RMSE, and averages using the work from our data manipulation in order to help us begin to find relationships in the data. After playing around with different combinations of one or many variables and what their R^2 and RMSE values come out to be, we decided to move more into machine learning algorithms like linear regression, decision trees, and random forest classifications. At the start, we tried using various combos of variables for our linear regression models in order to try and improve the results. Later, we were able to run wider tests encapsulating all of our variables to see if we could find better relationships or better predict which factors affect how much a movie makes.

Now, as we didn't really explain in specifics what each model really does or what exactly our value types are, we will now walk through the values, each type of model, and how they function. The value r is a correlation coefficient that determines the closeness of the association of points in a scatter plot. R^2 is simply the proportion of the variation in the dependent variable that is predicted by the independent variable. Adjusted R^2 is simply an R^2 value that slightly decreases depending on how many explanatory variables there are. RMSE (Root Mean Squared Error) is a measure of how far predictions are from the true values. A linear regression is a type of modeling method that assumes a linear relationship between the independent and dependent variables, and aims to find the best-fitting line that describes the relationship by finding the line that most reduces the SSR (Sum of Squared Residuals). A decision tree is a non-parametric supervised learning algorithm that functions similar to a flowchart of decisions that put together create a model. Finally, random forest classification is an estimation technique that fits many decision trees on various samples of the dataset and uses averages to improve the predictive accuracy and control over-fitting.