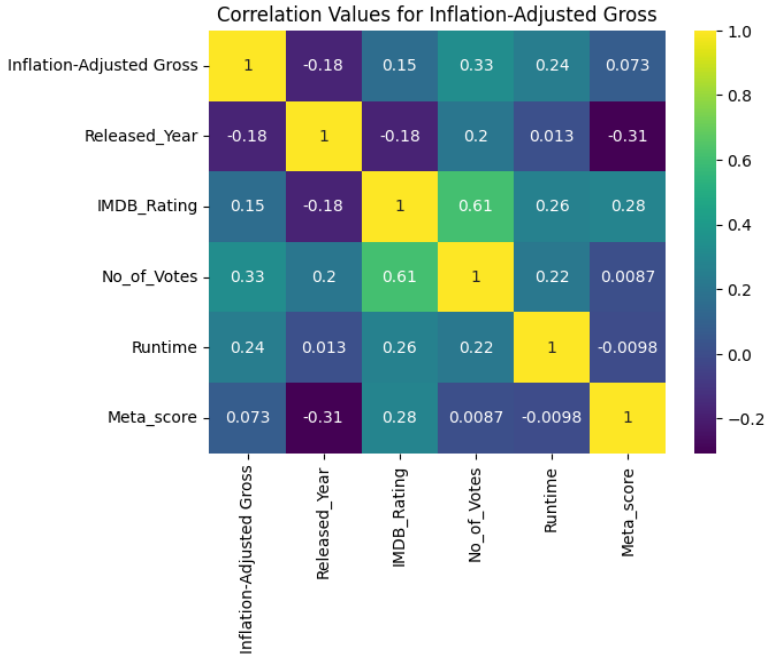
Movie Earnings Results Page

Our exploratory data analysis yielded several notable results regarding what correlates with inflation-adjusted movie earnings. First, we measured the correlation between each of our quantitative explanatory variables and inflation-adjusted gross. We found that the number of votes on a film's IMDB rating had the highest correlation.

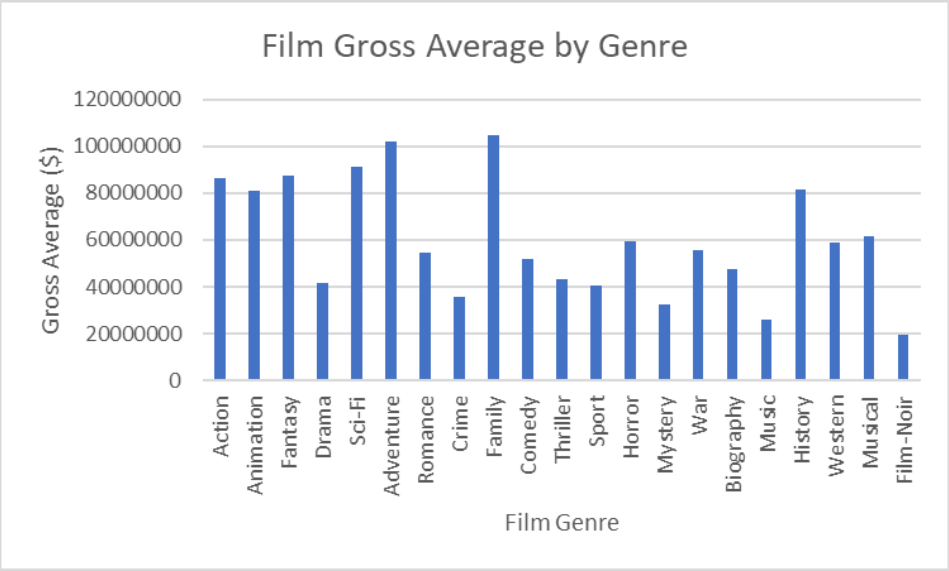


We also calculated the R^2 value for each of these variables, as well for the release year of the movie and the unadjusted gross. The only variable with an R^2 value greater than .10 was the number of votes.

	Release Year vs Gross	Release Year vs Adjusted Gross	IMDB Rating vs Adjusted Gross	Number of Votes vs Adjusted Gross	Runtime vs Adjusted Gross	Runtime vs Adjusted Gross
R^2 value	0.055	0.031	0.023	0.106	0.057	0.0052

In our exploratory data analysis stage, we looked to measure the relationship between our categorical variables and earnings. The average inflation-adjusted gross of a film made by a famous director was $69,344,856.78, while the average inflation-adjusted gross of a film made by a non-famous director was $52,474,195.54. Conducting a two-sample t-test yielded a test statistic of 1.999 with 748 degrees of freedom, corresponding to a p-value of .023. This meant

that the difference was significant at the 5% level. We also examined whether certain genres consistently earn more money than others. Graphing the average earnings by genre showed major differences between them.

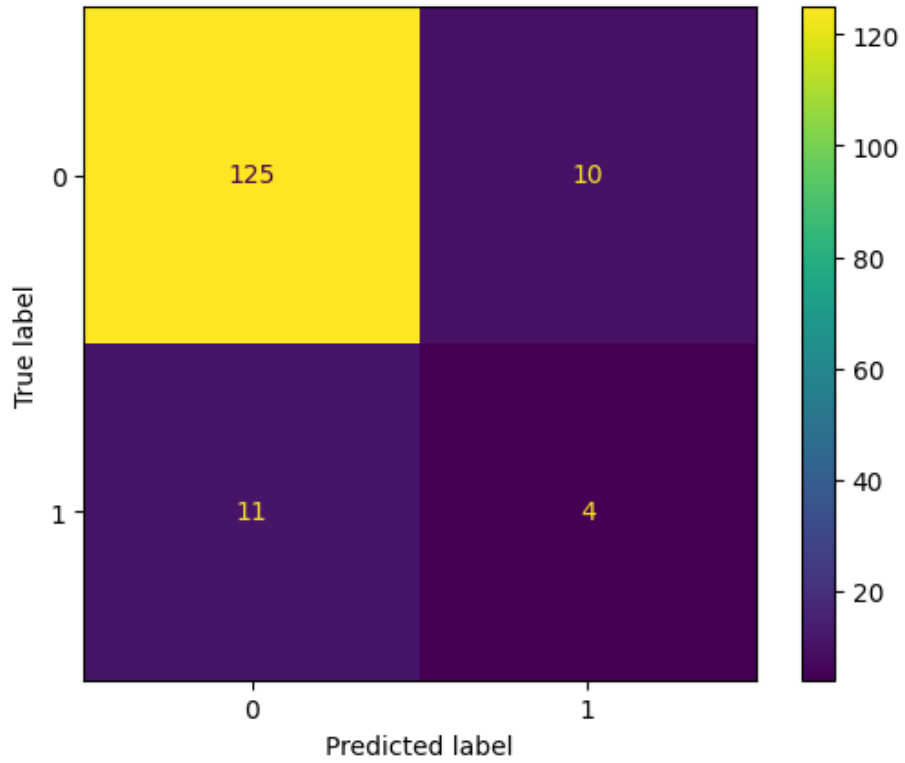


When creating a regression model to predict inflation-adjusted earnings, none of our models were particularly accurate. We tried a number of different combinations of variables to try to build an effective model, but very few even had an R^2 value greater than 0.1. The best performing linear model used the number of votes, IMDB rating, a dummy variable to indicate whether a movie was directed by a famous director, and a dummy variable to indicate whether a movie was in the Action or Adventure genres. For this model, the R^2 was 0.179 and the adjusted R^2 was 0.157. The root mean square error was quite high at 61,700,609.54. These unimpressive results suggested that linear regression was not the best choice for predicting film earnings.

Taking into account that most of the movies in our dataset earned relatively little money, while a minority earned a great deal. We decided to see if we could develop a model to predict what movies would be "blockbusters" and earn in the 90th percentile or higher. Using the variable IMDB rating, and indicator variables for whether the movie was directed by a famous director, whether it was an Action or Adventure movie, and whether it was a Family movie, we developed both a Decision Tree and a Random Forest model to classify movies as blockbusters or not. The Decision Tree model had an F_1 score of 0.6.

	Precision	Recall	F_1 score	Support
Not blockbuster	0.92	0.93	0.92	135
Blockbuster	0.29	0.27	0.28	15

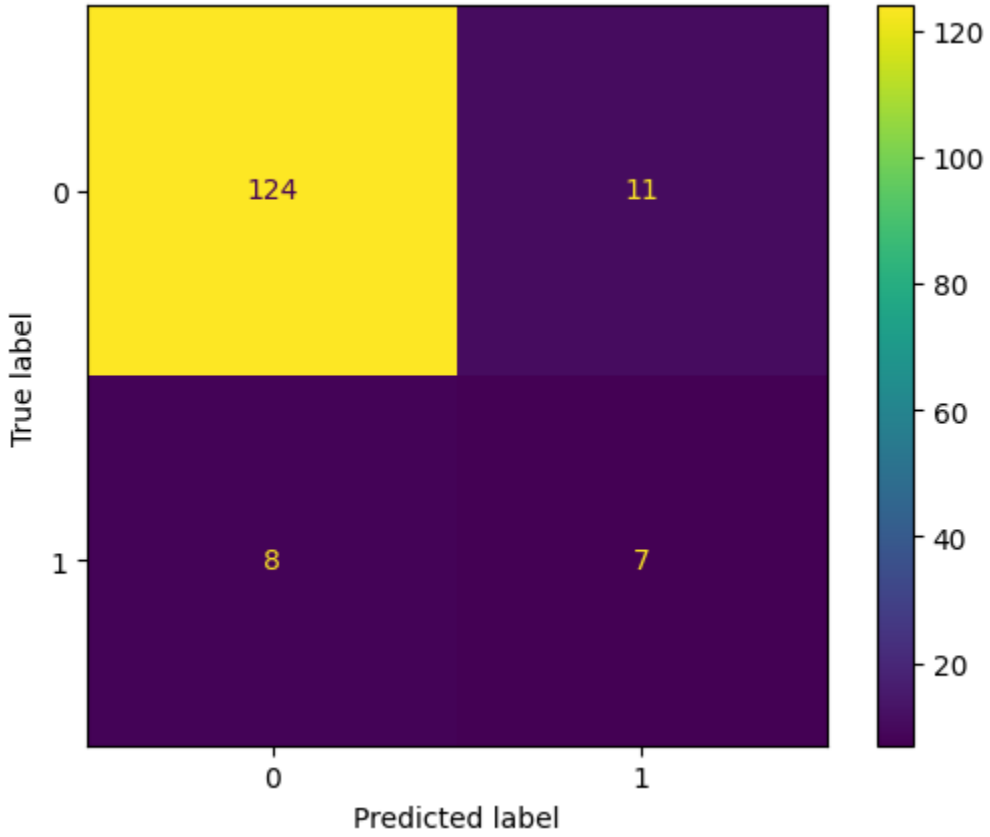
Accuracy			0.86	150
Macro Average	0.60	0.60	0.60	150
Weighted Average	0.86	0.86	0.86	150



This shows that our model was very accurate at filtering out movies that were not blockbusters. It still struggled, however, to identify those that were. The Random Forest model we developed was about as accurate in correctly classifying non-blockbusters, while it was slightly better at predicting which movies were blockbusters.

	Precision	Recall	F_1 score	Support
Not blockbuster	0.94	0.92	0.93	135
Blockbuster	0.39	0.47	0.42	15
Accuracy			0.87	150
Macro Average	0.66	0.69	0.68	150

Weighted Average	0.88	0.87	0.88	150



This result was encouraging. Using our Random Forest model, we have a way to identify which movies are more likely to become big earners that is fairly accurate.

The relative success of the classification models in comparison to the regression models indicates that the top earning movies have distinctive, common characteristics that set them apart from less commercially successful films. These characteristics are related to the film's IMDB rating, whether the movie was directed by a famous director, whether it was an Action or Adventure movie, and whether it was a Family movie. We looked at the feature importances of the variables and saw that the most important variable was IMDB rating, followed by the action indicator variable, the family indicator variable, and finally the famous director indicator variable. We concluded that these variables were particularly important in determining how much money a movie will earn.