Nick Borchich, Kurt Brown, and Matt Latzke
4/20/23
Professor Beagley
DATA-151

Introduction

For our DATA-151 project, we have been researching what factors are most important when determining how much money a movie will earn. We believe that this topic is important because movies are a major part of modern culture. By looking into what leads to a movie earning more money, we can get a better understanding of what modern culture values, what stories appeal to people, and what people want out of entertainment. It is also possible that the results of this research may lead us to finding the most desirable movie to produce next, hoping it to be a massive moneymaker (assuming all of the factors could reasonably be assembled). We are not the only people who think this research is important though. Researchers like Joseph, Sydney, and Gayirah all suggest that this topic is important for reasons like making sure that movie producers and companies stay in business so that they can keep making movies, so that the movies they continue to make remain interesting and desirable to all audiences, and in order that we still have a variety of movies being released. Our research question is "What makes a movie earn more money?"

Data and Methods

We found and used a dataset containing the top 1000 ranked movies by IMDB rating on Kaggle. This dataset contains information about movies spanning from 1925 to 2019. It also includes a number of characteristics about each of these movies such as the release year, runtime, genres, director, and domestic earnings. We believed that this dataset gave a good overview of the film industry and provided sufficient information to determine what factors are most important in predicting how much money a movie will earn. To analyze and model this data, we used Python in the Google Colaboratory IDE.

To conduct our experiment, we began by thoroughly cleaning the data. We removed unnecessary columns, such as the Poster Link and Overview columns. We also removed data points that did not have recorded domestic earnings (our response variable) or Metascore ratings. This left us with 750 data points. We also calculated the inflation-adjusted domestic earnings for each remaining movie in our dataset indexed to 1983 dollars. We did other minor tasks (such as formatting) to make the data suitable for exploratory data analysis and modeling. In our exploratory data analysis we examined the correlations of our individual explanatory variables with domestic earnings. We also calculated the mean earnings for movies with a famous director versus movies without a famous director, and the mean earnings for each genre. Following this exploratory data analysis, we developed a series of multiple linear regression models using various combinations of explanatory variables and examined the coefficient of determination and root mean square error of each model.

Our measures of success were the coefficients of determination for each of the explanatory variables on the response variable, domestic earnings, as well as the root mean square error and coefficient of determination for calculated models. We also observed the direction of coefficients with regards to such predictive models to determine the direction of these explanatory variables effects.

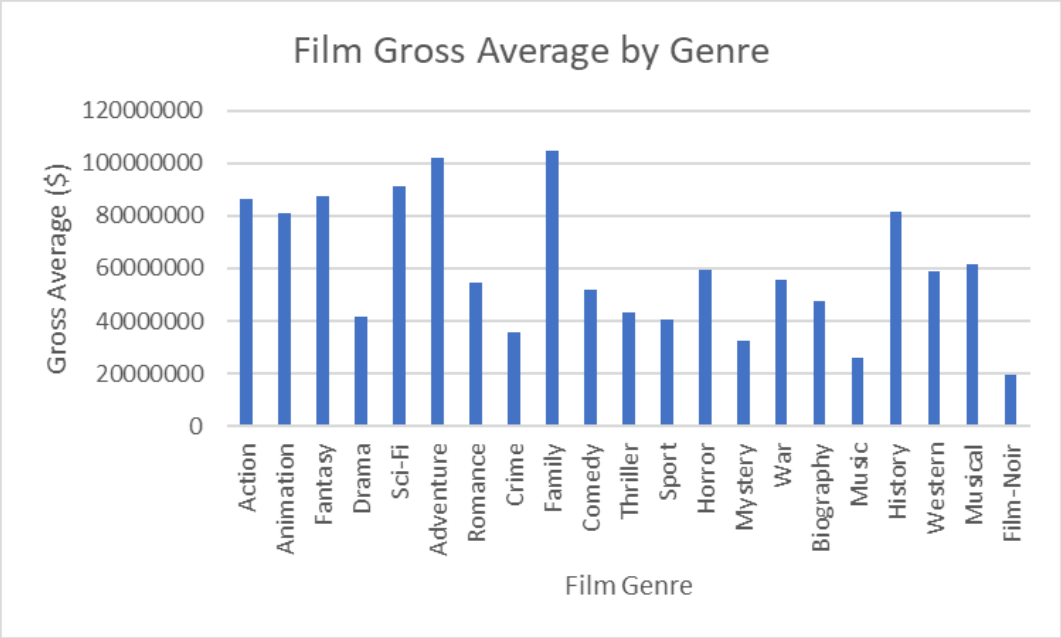
Results

When performing exploratory data analysis, we calculated the correlation coefficient and R^2 value for each of our quantitative predictor variables when used to predict the inflation-adjusted gross of a movie. Since we did not conduct a multiple linear regression, we did not control for the effect of all other variables. We were simply trying to see which variables seemed to correlate the most with earnings at a quick glance.

	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
Correlation Coefficient	0.236	-0.179	0.153	0.327	0.240	0.072
R^2 Value	0.055	0.031	0.023	0.106	0.057	0.0052

To our surprise, many of the predictor variables did not seem to have a large impact on domestic earnings. The only variable with an R^2 value greater than .10 was the number of votes on a film's IMDB rating.

After conducting exploratory analysis on our quantitative predictor variables, we examined whether certain characteristics for our categorical predictor variables seemed to have an effect on earnings. First, we tried to determine whether movies made by famous directors consistently made more money than movies not made by famous directors. With a definition of famous director as having five or more films in our dataset, we found that 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. We also analyzed the effect of genre upon gross earnings in an attempt to see if there was much variability based on genre. Our data, thankfully, did show some significant differences in gross earning average between the genres.



After this exploratory data analysis, we began work on building a model to predict movie earnings. Our results were a bit disappointing. Early versions of our models had a negative r-squared value, implying that it was less effective than guessing the mean value of revenue every time. It also implied that the fit was wrong and that a different fit could actually explain some of the variance. Later revisions to our model provided us with r-squared values just under 0.2, meaning that under 20% of the variance is explained by these models. This value was still lower than would be preferable. With an r-squared value so low, there was little room for further PCA analysis.