Final Project

Christina Yu, Damian Kim

Read in the data

Introduction and data

With the rise in the age of information, the processing speeds of data are accelerated to impossible levels. One of the densest forms of media content out there is the transcription of words into pixels. A book that previously took 2 weeks to finish now fits in the span of 2 hours in the insanely digestible form of: movies. The movie industry is worth 95 billion dollars in the US alone, and indeed there is a great demand for quality movies. Though it is not easily apparent to tell what makes a movie great, there are some basic data that might help determine some of the factors of a great movie. We decided to try and tackle this heavy yet essential question. What factors might help make a movie successful? Throughout this project we can consider this question through the lens of a potential movie investor.

This dataset was scraped from IMDb (Internet Movie Database). There are 6820 movies in the dataset (220 movies per year, 1986-2016). Each movie has the following attributes:

– `budget`: the budget of a movie. Some movies don't have this, so it appears as 0

– `company`: the production company

– `director`: the director

– `genre`: main genre of the movie.

– `gross`: revenue of the movie

– `name`: name of the movie

– `rating`: rating of the movie (R, PG, etc.)

– `released`: release date (YYYY-MM-DD)

– `runtime`: duration of the movie

– `score`: IMDb user rating

– `votes`: number of user votes

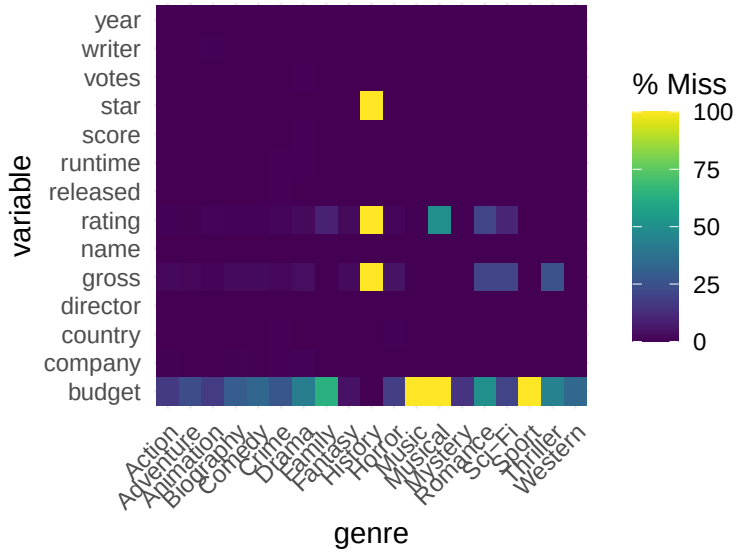
– `star`: main actor/actress

– `writer`: writer of the movie

– `year`: year of release

We will explore the factors that make a movie successful through examining the effects of our variaables of interest: `gross`, `budget`, `genre`, `rating`,`score` `votes`, and `runtime` for individual movie.

First we will explore missingness in our dataset.



The dataset includes movies that haven't been published yet (which causes a missing gross value). In our project we are considering gross revenue of a movie as the defining indicator of success. Considering that, we removed all observations that haven't been published yet or are missing gross values, because we are interested only in movies with a quantitative measure of success.

In fact, we also decided to rid our dataset of missing values for the other predictor variables as well. About 28% of the observations were missing a value for budget, whereas the next most missing variable was gross, with 2%. From elementary missingness analysis there did not seem to be very strong, numerically significant relationships between the missingness of each variable with other variables of interest. We decided to do a complete case analysis.

The Predictor Variables

We will use the `budget`, `genre`, `rating`,`score` `votes`, `runtime`, and `country` variables as predictors. Among them, `budget`, `score` `votes`, and `runtime` are numerical variables, while

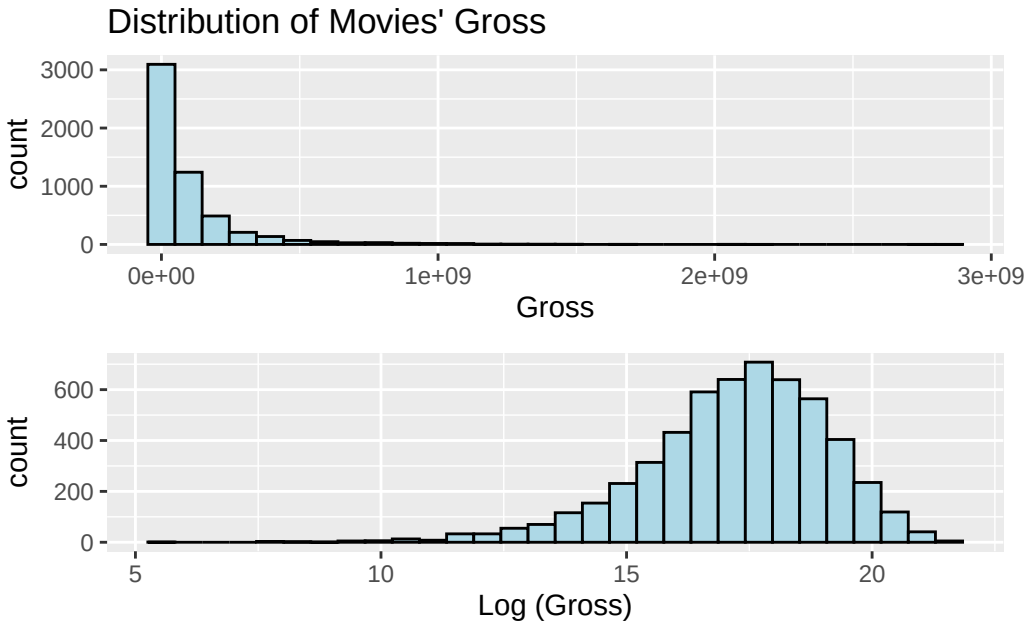
`genre`, `rating`, and `country` are categorical variables.

The Response Variable

As described earlier, the `gross` numerical variable is our response variable. 1. Summary of the `gross` variable:

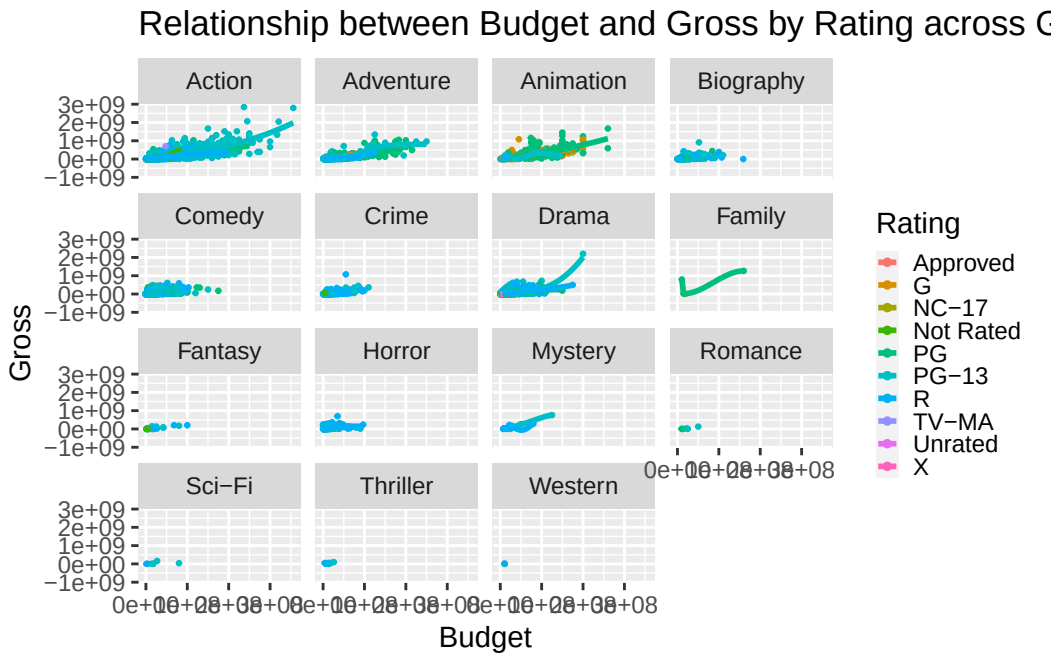
mean_gross	median_gross	sd_gross	min_gross	max_gross
103192280	36850101	187278279	309	2847246203

2. Log-transformation and Distribution of the `gross` variable:



Since the response variable is significantly right skewed, we apply a log-transformation to it and will use log(gross) as our new response variable in the future analysis. Now, our response variable is unimodal, following a roughly normal distribution, with a mean at 17.2102, and there exist some outliers on the left end.

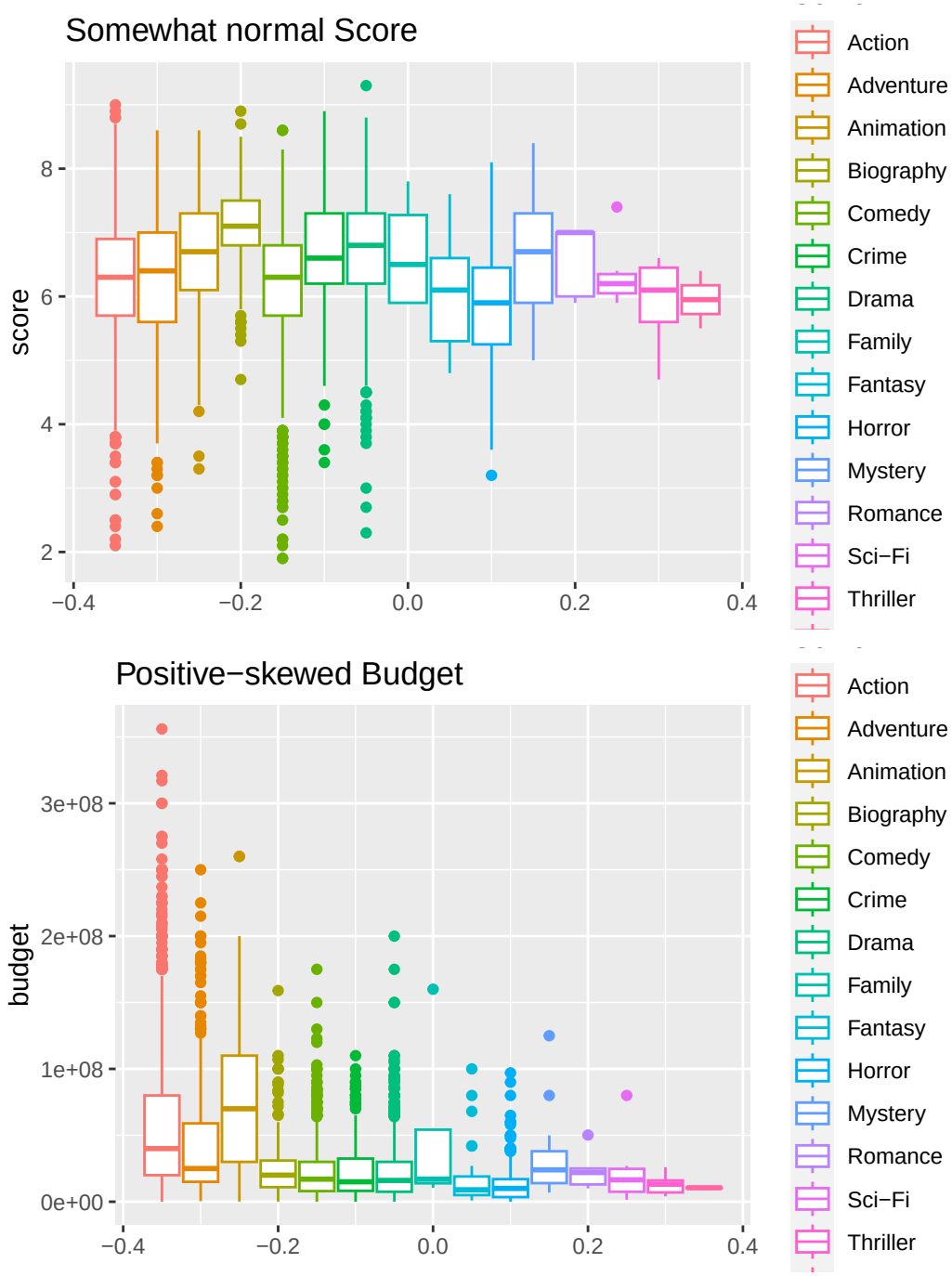
3. Relationship between Gross and Budget based on different Genres & Ratings:



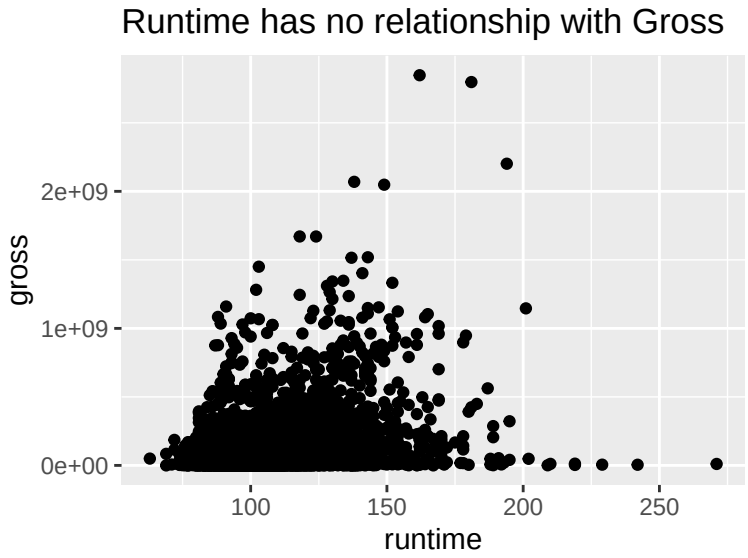
We observe that the relationship between budget and gross is vastly different across genres: action, adventure, animation movies have a steep slope and generally high budget ranges, with outliers which have exceedingly high budget and relatively high gross values. Noticeably, there are also a lot more movies in these three categories, with the most in action. On the other hand, genres such as horror, mystery and romance have a visibly flatter slope, which corresponds to the industry knowledge that certain genres are more conducive to low-budget film making than others, even when provided with additional funding. Thus, we're interested in further exploring the relationship between budget, genre, and success.

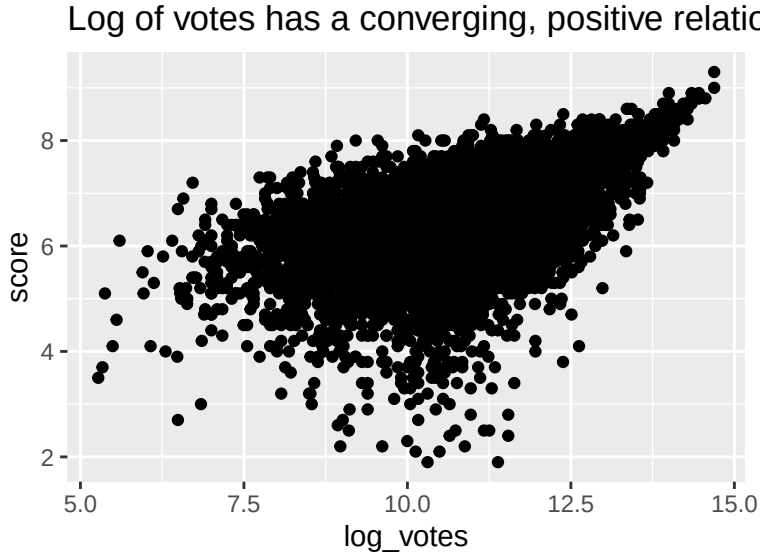
EDA: Visualizatioins and Summary Statistics

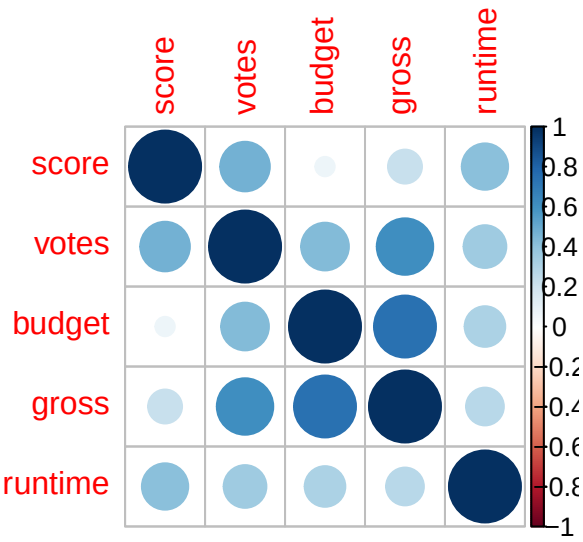
The dataset contains many interesting variables that we want to explore fully. We will first do some elementary exploratory data analysis on our datasets to show potential insights that we can explore further.



As can be seen, the score variable is relatively unskewed, whereas the votes, budget, and runtime variables are heavily positively-skewed.


Runtime has no relationship with Gross
gross
2e+09
1e+09
0e+00
100
150
200
250
runtime


Log of votes has a converging, positive relatic
score
8
6
4
2
5.0
7.5
10.0
12.5
15.0
log_votes



As can be seen, we have some correlated variables between our predictor variables, but none are particularly strong (0.8+) so we don't have to remove any of these predictor variables on the basis of correlation for our later models. Though there seem to be no particularly strong correlations between predictor variables, the strongest correlation is between votes and score. We may consider adding this into our model because it also fits with our domain knowledge that a movie that is rated by more people might generally have a higher score because, in our experience, people are more motivated to rate a movie they enjoy than one they dislike.

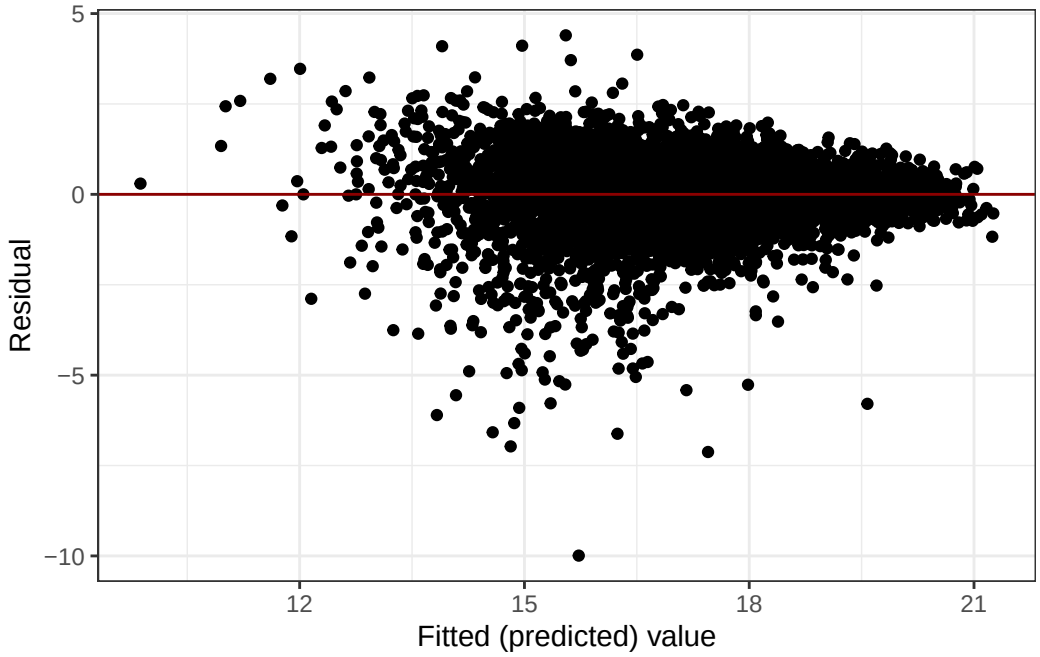
Methodology

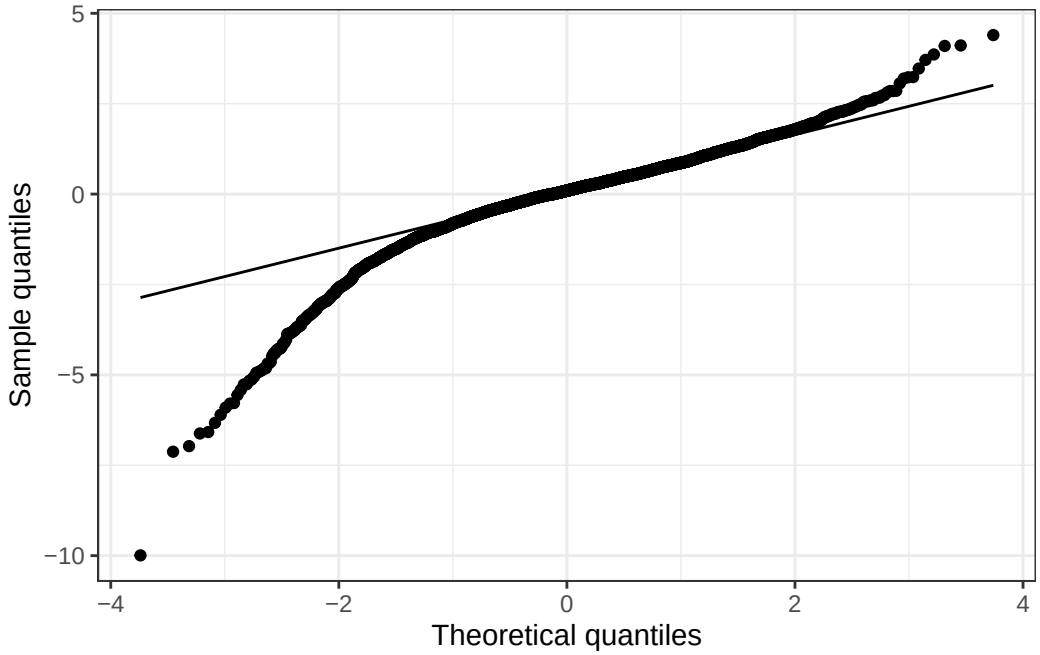
As a potential movie investor, we're obviously curious about the prediction of movies' success, and the factors that lead to success. Therefore, we want to explore the following topics in detail in order for us to be better investors: 1) Prediction of movies' gross value 2) the factors that affect the success of the movie. Some of the methods by which we are going to tackle this is through linear regression, residual plots, linear mixed models, and repeated k-fold validation.

Linear Regression of all variables

Gross revenue is a continuous response variable, so we will first examine a linear regression model for our data. We've gained the equation of XX.

Residual plots/Assumptions





We may assume that each movie is independent of each other. We can see that the linear assumption is violated since the data is not symmetrically distributed observations around the

horizontal axis. The constant variance and normality assumptions are not satisfied from the above graph. Even after log transforming our gross response variable to fix the normality, the constant variance condition was still not met. The validity of our model results are thus going to be not perfectly reliable.

Linear Mixed Model & Random Effect

Because variables such as genre, rating, and country likely have some variance in their effect on the model across their categories, we decided to create random effects for these variables. Our fixed effects are ones which have constant effects on the model which doesn't change for different "categories", namely budget, score, votes, and runtime.

While adjusting for the random intercept based on country variable and holding all other variables constant, we noticed that the estimated variance is 0.1623 which means that all the country-specific intercepts are distributed around the model's overall intercept within estimated variance of 0.1623. Given that our gross unit is log dollars, and our estimated variance is quite small, we may expect that the random effects of the country on our model is relatively similar. In terms of our interest in knowing whether a movie is successful or not, production of country may not be a big effect on the movie's success.

Additionally, we noticed that genre has an even smaller estimated variance (0.08254), and rating has a marginally higher variance. The variable effects of the different genres, countries, and ratings don't seem to be hugely related to a movie's success.

```
[1] 1.352356
```

```
[1] 1.357526
```

Our linear model has a smaller RMSE than our mixed effects model: $1.3523 < 1.3575$. The difference is very minor, however, we conclude that our linear model is a better fit for our data and has better performance.

Thus our best model is the linear regression based model: $\widehat{Gross} = \beta_0 + \beta_1 * log(budget) + \beta_2 * score + \beta_3 * log(votes) + \beta_4 * log(runtime)$

interaction terms

Since we conclude that the linear regression is better, we continue to apply linear regression and we're curious about whether the relationship between gross and budget depends on score/votes, while adjusting for all other variables.

For interaction terms: The coefficient for score is 3.320e-01 and the coefficient for votes is 5.073e-06. Since the coefficient for votes is too low, we would only consider score here.

Interested Hypothesis Test 1:

Question 1: Is there evidence to suggest that log(Budget) has an effect on the success of the movies? Null hypothesis: $p_1 = 0$ There isn't sufficient evidence to suggest that budget is associated with movies' gross, while controlling for all of the variables. Alternative hypothesis: $p_1 \neq 0$ There is sufficient evidence to suggest that budget is associated with movies' gross, while controlling for all of the variables.

We use significance level of 0.05. Since the t-statistics is 23.227 and the p-value is $<$ 2e-16 which is much smaller than our significance level, so we reject the null hypothesis since there's sufficient evidence, and thus there's sufficient evidence to suggest to log(budget) does have an effect on the success of the movies (log(gross)).

In terms of qualitatively addressing this issue: in our model, we know that every 1 million increase in our budget, the gross value is expected to increase by e^{cc} dollars. ## Interested Hypothesis Test 2: Aim 2: Is there evidence to suggest that log(Score) has an effect on the success of the movies? Null hypothesis: $p_2 = 0$ There isn't sufficient evidence to suggest that Score is associated with movies' gross, while controlling for all of the variables. Alternative hypothesis: $p_2 \neq 0$ There is sufficient evidence to suggest that Score is associated with movies' gross, while controlling for all of the variables.

We use significance level of 0.05. Since the t-statistics is 23.227 and the p-value is 3.06e-09 which is much smaller than our significance level, so we reject the null hypothesis since there's sufficient evidence, and thus there's sufficient evidence to suggest to genre does have an effect on the success of the movies.

Summary based on Hypothesis:

Based on our previous two hypothesis, we noticed that XXXXX.

Predicting moves' gross

Based on all the models we've done above, we've analyzed the relationship between different variables that we're interested in. In order to make our findings applicable for future use, we'd like to make a prediction on movies' gross based on our current variables.

After testing a linear model with and without log-transformed predictor variables, with 10-fold validation with five repeats, the average RMSE without log transformed predictor variables was 1.3736, whereas the average RMSE with transformed predictor variables waas 1.0831. The second one has better predictive power, and thus we determine that our final, best model is our linear model with log-transformed predictor variables.

Results

Explain the reasoning for the type of model you're fitting, predictor variables considered for the model including any interactions. Additionally, show how you arrived at the final model by describing the model selection process, interactions considered, variable transformations (if needed), assessment of conditions and diagnostics, and any other relevant considerations that were part of the model fitting process.

Discussion & Limitations

Summary + statistical arguments to support my conclusions + future limitations/future ideads

Variable selection, specifically the lasso for categorical variables with 2+ levels, Rescaling predictor variables, linearity assumptions,

Sources

https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a https://stackoverflow.com/questions/13646654/root-mean-square-error-in-r-mixed-effect-model https://psyarxiv.com/wc45u/