Predicting TMDb Top Film Ratings

STAT 410
Amy Cao & Melody He

Problem Statement

With media and entertainment being such a visible industry, we were interested in exploring the factors that influence voters' ratings on popular movie databases such as IMDb and TMDb. Our primary research question was: what features and characteristics cause people to "like" or "dislike" a movie, and how does it influence their rating? The main objective of our project was to investigate the factors that affect movie ratings and to determine the extent of their influence.

Introduction

The Movie Database (TMDb) consists of nearly 800,000 movies and 2.8 million contributors. Its rating system scores movies on a scale from 1 to 100, with 100 being the highest possible score. These ratings are calculated using a Bayesian average of site users' ratings, which depends on factors such as the number of votes that a movie has (Bell). Like IMDb, its ratings are commonly used by people around the world to help determine what they should watch next.

To begin our analysis, we conducted exploratory data analysis by looking at visualizations of our response variable, movie rating, as well as that of several regressors of interest (such as vote count, genre, release time, etc.). Next, we conducted a series of simple linear regression (SLR) models to analyze the individual effect of quantitative regressors on rating.

Based on the significant results from SLR and insights from our exploratory data, we decided to fit three multiple linear regression (MLR) models: 1) a model with only quantitative regressors, 2) a "full" model with quantitative and genre regressors, and 3) a model with regressors relating to prominent cast and crew. To test the effects of our categorical variables (genre, major cast/crew members), we incorporated dummy regression into our model by creating binary variables for each group. To test for significance of regression and difference between group means, we performed an ANOVA test to investigate whether each coefficient is greater than the significant F-score. Our process in model selection to reduce potential multicollinearity and residual analysis to evaluate goodness of fit is detailed in later sections.

Data Overview

Our dataset is titled "TMDB 5000 Movie Dataset" and is sourced from Kaggle. For each of the top 5000 movies on TMDb from 2015 and prior, we have information such as **rating**, **runtime, budget and revenue, genre, cast and crew, plot summary**, etc. Although TMDb ratings on the site are currently scaled from 1 to 100, our dataset has scaled ratings proportionally from 1 to 10. Some of the movies were out of our range of interest or had faulty information (i.e. budget or revenue = 0 when this was clearly not the case). To be included in the model and exploratory data, we required each movie to have:

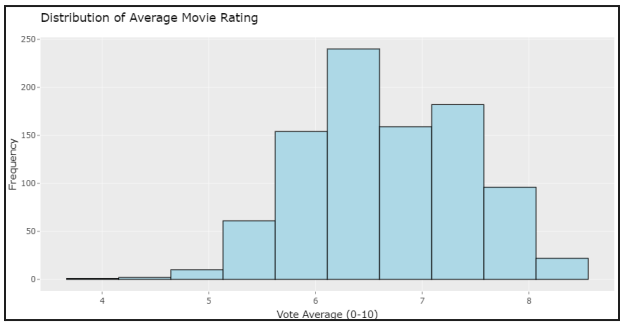
- been released during or after 1990
- more than 900 votes

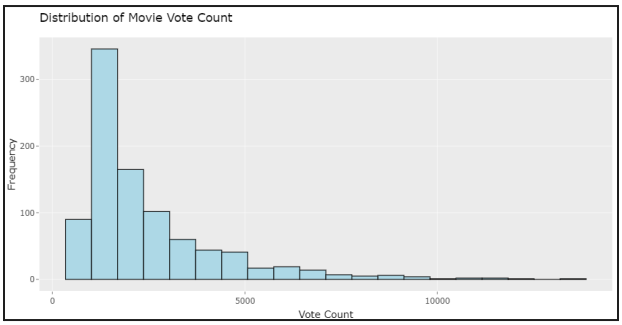
- a positive budget and revenue

Logically, these gave us the higher-quality movies and the movies that were truly "top or popular films" within the original dataset. We had 927 observations after filtering, which we felt was a satisfactory number of response observations to regress over.

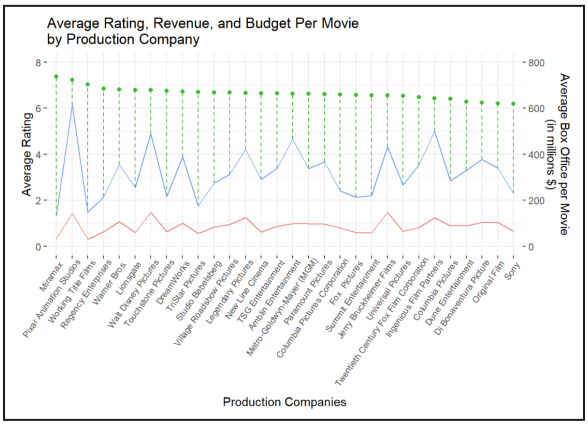
Exploratory Data

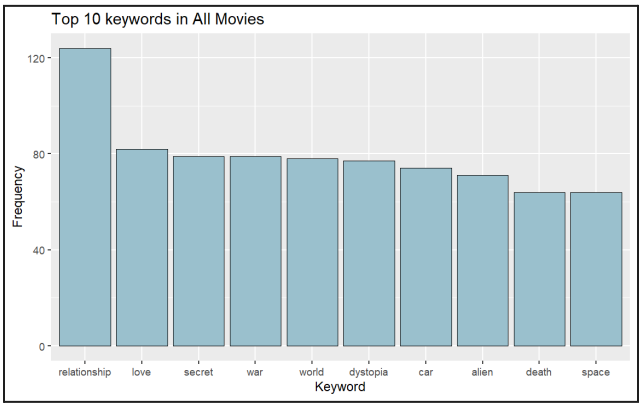
In summarizing our data, we found that these films have a mean rating of 6.7 from an average of 2,459 user votes, and gross a median of $201 million in revenue. From experience, we expected the mean rating to be around 7.5 or higher. The lower average may be due to TMDb's large international audience and smaller overall audience (which may result in critical ratings having greater weight), as it is a newer platform than rating systems such as IMDb or Rotten Tomatoes.





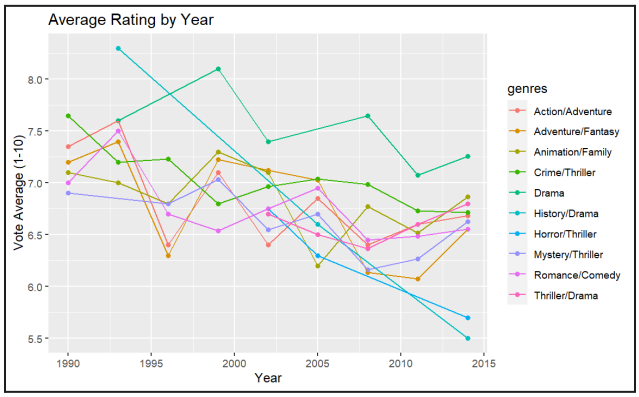
Broadly, we are interested in what contributes to movie popularity. We created visualizations based on some factors potentially related to rating. Below, we see a plot for the average rating (green), revenue (blue), and budget (red) per movie for top production companies. The revenue and budget fluctuate greatly across the companies with the best average ratings. This points towards the later result that regression coefficients for these variables may be close to 0 or otherwise difficult to predict. From this fluctuation, we can also see that there is a good balance of independent as well as major studios represented within the top-rated companies.

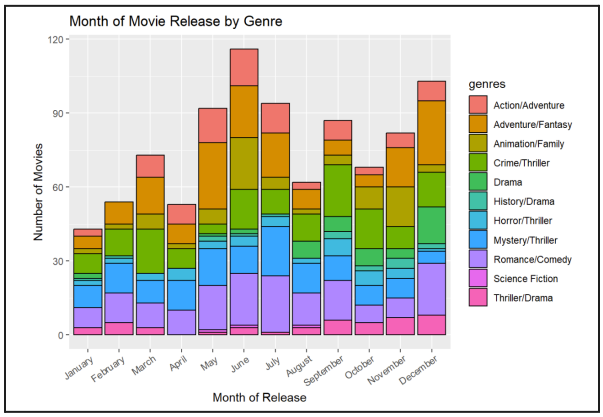




We were also interested in the popularity of certain plots. After evaluating the top ten keywords from our movie summaries, we noticed popular plot points included "love," "alien," and "space." These words hinted that genres play a big role in popularity, so we grouped each movie by broad categories–similar genres such as horror and thriller or romance and comedy were put together.

Our analysis on genre yielded no easily discernible trends across the years, although movies in the drama category seem to remain consistently more favorable across the years. We were also curious about the timing of movie releases, and found that June and December, the holiday and break months, had the highest number of releases.





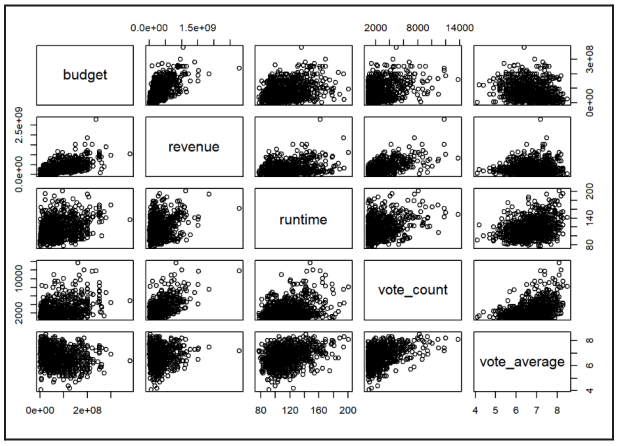
Our exploration into our data ultimately fueled many of the interesting models we decided to fit.

Analysis & Results

Simple Linear Regression

We identified variables of potential interest for simple linear regression using the pairwise plot.

The bottom row of our pairwise plot displays the correlation between our regressors of interest and vote average. Based on the plots, we saw that there was no strong correlation between our regressors of interest and average vote rating. We decided to test the most promising (budget, runtime, and number of votes). Additionally, we observed a strong correlation between budget and revenue, indicating potential multicollinearity issues if both were included as regressors in our final multiple regression model.



Next, we fitted simple linear regression models to explore the relationship between vote average and three regressors: **budget, runtime,** and **vote count**. The regression coefficients, standard error, t-values, and p-values for our models are displayed below.

Budget

beta hat extremely close to 0

R^2: 0.04
Adj. R^2: 0.04

	Est.	S.E.	t val.	p
(Intercept)	6.86	0.04	182.79	0.00
budget	-0.00	0.00	-6.30	0.00

Standard errors: OLS

Runtime

beta hat ≅ 0.01

R^2: 0.11
Adj. R^2: 0.11

	Est.	S.E.	t val.	p
(Intercept)	5.32	0.13	41.38	0.00
runtime	0.01	0.00	10.67	0.00

Standard errors: OLS

Number of Votes

beta hat extremely close to 0

R^2: 0.17
Adj. R^2: 0.16

	Est.	S.E.	t val.	p
(Intercept)	6.28	0.04	170.95	0.00
vote_count	0.00	0.00	13.53	0.00

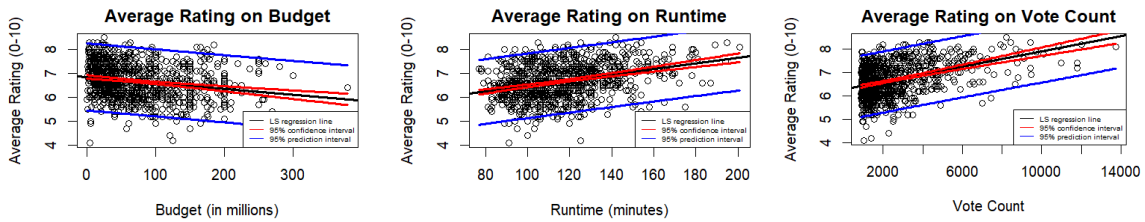
Standard errors: OLS

We observed that the $\beta 1$ estimates appear to be very close to zero for all three models, yet all are statistically significant due to their small standard errors. The R^2 were 0.04, 0.11, and 0.17 respectively, indicating that a low proportion of the variance in vote average can be explained by budget, runtime, and number of votes individually. In other words, no one quantitative variable can reliably predict a movie's overall rating. This makes sense, since the quality of movies can vary widely despite having similar quantitative characteristics.

We then calculated the confidence interval for our regression coefficients, mean response, and predicted interval using $\bar{x}$ as our x_0 estimate. The results below confirmed that all three regressors are statistically significant, as 0 does not lie within any of the confidence intervals.

Confidence Intervals for Regression Coefficients, Mean Response, and Predicted Response

Regressor	B0_CI	B1_CI	Mean_Response_CI	Predicted_Response_CI
Budget	[6.785, 6.932]	[-0.0032, -0.0017]	[6.63, 6.72]	[5.27, 8.08]
Runtime	[5.07, 5.575]	[0.0095, 0.0138]	[6.63, 6.72]	[5.32, 8.03]
Vote Count	[6.203, 6.347]	[0.0001, 0.0002]	[6.63, 6.72]	[5.37, 7.98]



We also plotted vote average against all of our potential regressors including the regression line, 95% confidence interval lines, and 95% prediction interval. We were surprised to discover that budget appears to be negatively correlated with average rating, as we thought that the more money and resources a company puts into the movie, the higher the production quality, thus resulting in a higher vote rating. However, in general, we cannot predict rating from any individual regressor.

MLR: Quantitative Regressors

Model Selection

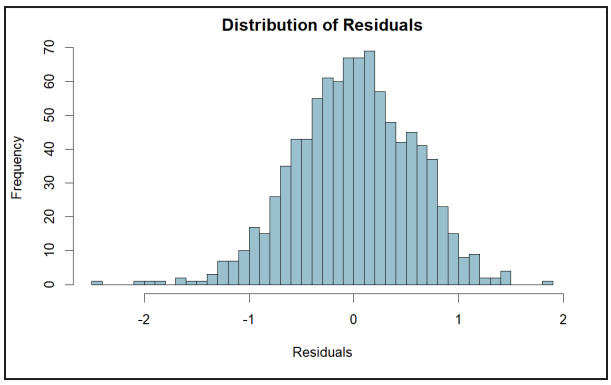
To select our model, we used forward selection and each of the criteria–since they all agreed–to choose all available regressors we considered for our MLR model. Because the coefficient for vote count had the lowest p-value on our SLR models, we started our selection with that as our first regressor. From then, we added budget and runtime, and saw relatively large changes in our adjusted R^2, AIC, corrected AIC, and BIC criteria. We also included revenue and an interaction term between budget and revenue, which increased adjusted R^2 marginally. Our final model is written in full on the following page.

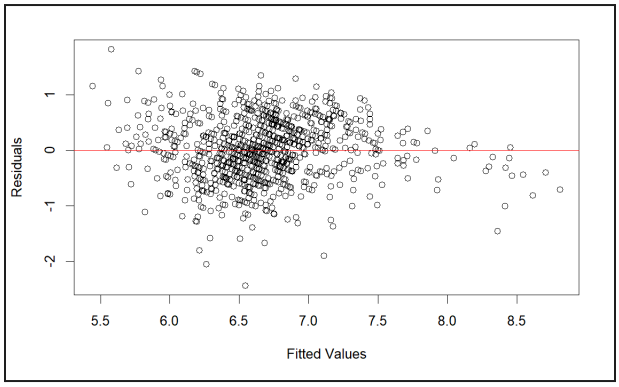
	Subset size	Predictors	R^2_adj	AIC	AICc	BIC
A	1	vote_count	0.164	-751.01	-750.98	-741.34
B	2	vote_count, budget	0.327	-950.3	-950.26	-935.81
C	3	vote_count, budget, runtime	0.403	-1061.01	-1060.95	-1041.68
D	4	vote_count, budget, runtime, budget * revenue	0.405	-1071.89	-1071.77	-1042.9
E	5	vote_count, budget, runtime, budget * revenue, revenue	0.411	-1071.89	-1071.77	-1042.9

Final Model:
vote_average = β0 + β1 vote_count + β2 budget + β3 runtime + β4 (budget * revenue) + β5 revenue + ε

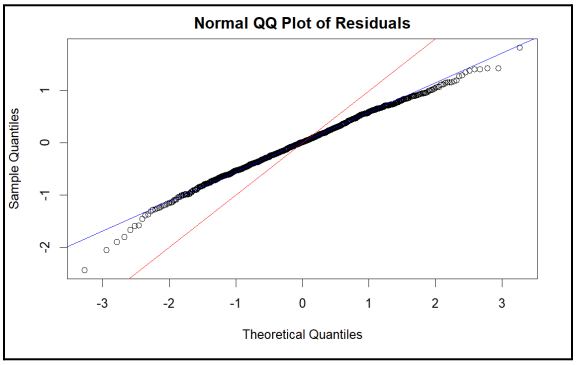
Residual Analysis

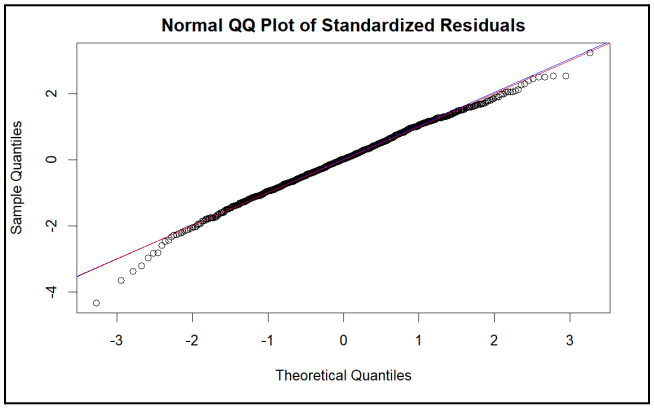
To evaluate the goodness of fit of our model, we performed a variety of residual analysis to test for our error assumptions and identify potential issues with heteroscedasticity, nonlinearity, and outliers. First, we plotted the distribution of residuals to check for normality and plotted fitted values against residuals to assess for constant variance.





The above plot on the left shows that the model residuals are mostly normally distributed for the most part, while the plot to the right indicates that the residuals are uncorrelated and have relatively constant variance except for a few outliers, which eliminates the issue of heteroscedasticity. We also created QQ plots for our residuals and standardized residuals to check the standard normal assumption for residuals, displayed below.





Based on the above two QQ plots, we can conclude that the distribution of our standardized residuals closely resembles that of a standard normal distribution. Therefore, we can move onto conducting tests and calculating confidence intervals.

Results & Testing

The table below on the left displays the regression coefficient estimates, standard error, t-values, and p-values of our multiple regression model. All five regression coefficients, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\beta}_5$ are statistically significant with p-values well below the significance level of 0.05. We proceeded to

conduct an ANOVA test to test for significance of regression, the results of which are displayed below on the right.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.051	0.029	-1.779	0.076
vote_count	0.516	0.034	15.251	0.000
budget	-0.483	0.034	-14.034	0.000
runtime	0.295	0.027	10.838	0.000
revenue	-0.111	0.046	-2.442	0.015
budget:revenue	0.076	0.020	3.772	0.000

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vote_count	1	153.034	153.034	259.995	0.000
budget	1	150.874	150.874	256.325	0.000
runtime	1	71.219	71.219	120.997	0.000
revenue	1	0.394	0.394	0.670	0.413
budget:revenue	1	8.376	8.376	14.230	0.000
Residuals	921	542.103	0.589	NA	NA

It is important to keep in mind that we only have access to the revenue figure in hindsight. In other words, it would be difficult to predict the rating of a movie that is still in theaters, since its revenue figure would be continuously changing. However, this is still a model with many highly significant regressors that we would be able to apply in most situations for prediction.

Full MLR

Next, we aimed to incorporate the variable genre into the existing quantitative regression model to determine the potential impact of movie genre on movie rating. We generated 11 binary variables, each representing a different genre category, resulting in 16 total possible regressors. Utilizing the same approach to our initial multiple regression model that focused solely on quantitative variables, we used forward selection and various metrics including adjusted R^2, AIC, AICc, and BIC to determine our final model. Our final model contained 10 regressors.

	Subset size	Predictors	R^2_adj	AIC	AICc	BIC
A	1	vote_count	0.164	-751.01	-750.98	-741.34
B	2	vote_count, budget	0.327	-950.3	-950.26	-935.81
C	3	vote_count, budget, runtime	0.403	-1061.01	-1060.95	-1041.68
D	4	vote_count, budget, runtime, ani_fam	0.446	-1128.4	-1128.31	-1104.24
E	5	vote_count, budget, runtime, ani_fam, drama	0.456	-1145.62	-1145.5	-1116.63
F	6	vote_count, budget, runtime, ani_fam, drama, crim_thr	0.466	-1152.59	-1152.43	-1118.76
G	7	vote_count, budget, runtime, ani_fam, drama, crim_thr, horr_thrill	0.474	-1155.22	-1155.02	-1116.56
H	8	vote_count, budget, runtime, ani_fam, drama, crim_thr, horr_thrill, revenue	0.479	-1156.71	-1156.47	-1113.22
I	9	vote_count, budget, runtime, ani_fam, drama, crim_thr, horr_thrill, revenue, myst_thrill	0.481	-1158.3	-1158.01	-1109.98
J	10	vote_count, budget, runtime, ani_fam, drama, crim_thr, horr_thrill, revenue, myst_thrill, budget*revenue	0.483	-1184.45	-1184.11	-1131.3

Final Model:

$$\textbf{vote\_average} = \beta 0 + \beta 1\ \textbf{vote\_count} + \beta 2\ \textbf{budget} + \beta 3\ \textbf{runtime} + \beta 4\ \textbf{ani\_fam} + \beta 5\ \textbf{drama} + \beta 6\ \textbf{crim\_thr} + \beta 7\ \textbf{horr\_thrill} + \beta 8\ \textbf{revenue} + \beta 9\ \textbf{myst\_thrill} + \beta 10\ \textbf{budget*revenue} + \varepsilon$$

The regression coefficients and the corresponding t-tests for each individual coefficient are displayed in the table to the right. Notably, all of the regression coefficients are statistically significant except for crime/thriller movies and action/adventure movies. Additionally, we observed negative correlations between budget, revenue, horror/thriller movies, mystery/thriller movies, and action/adventure movies and vote average. It's worth

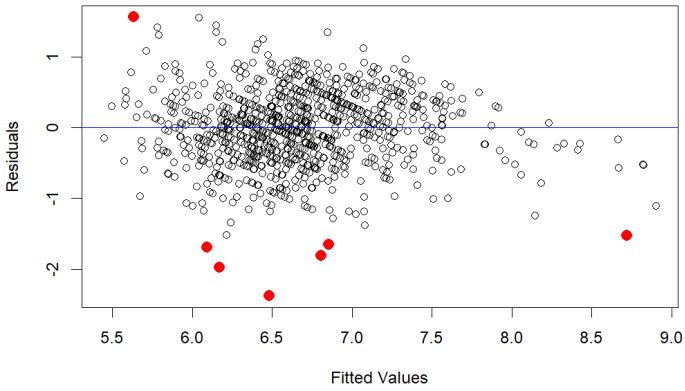
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.300	0.113	47.052	0.000
vote_count	0.000	0.000	16.516	0.000
budget	-0.007	0.000	-14.931	0.000
runtime	0.012	0.001	12.621	0.000
ani_fam	0.642	0.070	9.149	0.000
drama	0.329	0.086	3.826	0.000
crim_thr	0.072	0.052	1.401	0.161
horr_thrill	-0.251	0.086	-2.917	0.004
revenue	-0.001	0.000	-5.337	0.000
myst_thrill	-0.094	0.052	-1.811	0.071
budget:revenue	0.000	0.000	5.315	0.000

mentioning that the $\beta 0$ coefficient represents the baseline expected value for all unaccounted genres in our model: action/adventure, adventure/fantasy, history/drama, romantic comedy, science fiction, thriller/drama. Our model exhibited an R^2 coefficient of 0.4871 and an adjusted R^2 coefficient of 0.4815.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vote_count	1	81.281	81.281	295.137	0.000
budget	1	80.134	80.134	290.971	0.000
runtime	1	37.827	37.827	137.352	0.000
ani_fam	1	21.102	21.102	76.623	0.000
drama	1	5.570	5.570	20.224	0.000
crim_thr	1	2.560	2.560	9.296	0.002
horr_thrill	1	1.312	1.312	4.763	0.029
revenue	1	0.986	0.986	3.579	0.059
myst_thrill	1	1.008	1.008	3.661	0.056
budget:revenue	1	7.780	7.780	28.251	0.000
Residuals	916	252.268	0.275	NA	NA

The table to the left displays the results of our ANOVA test for the difference between group means. We noted that there is no statistically significant evidence of variation between groups for mystery/thriller movies, as the p-value for this regressor exceeds the significance level. However, all other genres, including animation/family, drama, crime/thriller, horror/thriller, and action/adventure, show statistically significant evidence of variation between groups.

Our next objective was to eliminate any potential outliers to potentially improve the accuracy of our model. To accomplish this, we utilized the studentized residuals to detect outliers, eliminating any data points with a studentized residual greater than 3. We identified and removed a total of 7 outliers, reducing our data set to 920 observations. Refitting the model resulted in an R^2 coefficient of 0.5132 and an adjusted R^2 coefficient of 0.5078, demonstrating an improvement in model accuracy . The figure to the right displays the residuals plotted against the fitted values, with the outliers highlighted in red.



Furthermore, we wanted to standardize our regressors and response variable, as many of our regressors such as revenue, budget, and vote count were not on the same scale as our response variable. We utilized unit normal scaling, and the results of our new model with standardized regressors and response variable is displayed below. Our standardized model had an R^2 coefficient of 0.5132 and an adjusted R^2 coefficient of 0.5078. The table to the right displays the regression coefficients for the standardized model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.088	0.027	-3.299	0.001
vote_count	0.522	0.031	16.936	0.000
budget	-0.537	0.033	-16.204	0.000
runtime	0.357	0.027	13.020	0.000
ani_fam	0.246	0.026	9.404	0.000
drama	0.094	0.025	3.835	0.000
crim_thr	0.037	0.025	1.475	0.140
horr_thrill	-0.081	0.024	-3.352	0.001
revenue	-0.167	0.042	-3.975	0.000
myst_thrill	-0.040	0.024	-1.634	0.103
budget:revenue	0.130	0.020	6.575	0.000

MLR: Cast & Crew

We wanted to test if the involvement of certain actors or directors influenced a movie's ratings. Originally, we included the 21 categorical variables (assigned 1 to involvement of that person and 0 for no involvement) seen in the tables below in our full MLR. However, we ran into the issue of high multicollinearity, since the regression coefficients assigned to highly-rated directors such as Christopher Nolan and Quentin Tarantino were negative and insignificant. The involvement of prominent people may correlate in varying ways with regressors such as revenue and budget, popular genres, and vote counts. More than surprising, this was suspicious and led us to conduct a separate MLR on a subset of our overall movie data. The tables of our estimates as well as t-tests and ANOVA tests are included below on the next page.

We selected these 21 people through a combination of examining recent Oscar-winning actors and directors, as well as people we believed were well-regarded or well-known in the film industry (see Appendix). Since our subset of Oscar winners was all White, many of our hand selected figures were intentionally people of color. After running an initial regression, we cut down on some of these hand selected figures due to large p-values and for the sake of having a digestible model.

We found the following individuals to have statistically significant positive influence on their movies' ratings: Christopher Nolan (director), Quentin Tarantino (director), Wes Anderson (director), Tom Hanks, Leonardo DiCaprio, Robert Downey Jr., and Jennifer Lawrence. Adam Sandler was the only significant negative influence. We were happy to see that some of the variables we expected to be significant turned out to be so. Most of the directors we included were significant, but most of the actors were not. This result indicates that renowned directors likely have more influence on their ratings, which makes sense since they have more control over the final product. However, it also shows that the popularity of any one actor or actress doesn't typically have a large influence on the rating.

We also noticed that in general, even if not statistically significant, actors or directors of color had negative regression coefficients. This is likely due to the bias implicit in the film industry with the types of topics, style of marketing, etc. that these movies undertake. Although an interesting result, since the coefficients were not significant, we cannot draw conclusions in this area or perform further analysis with the information in our dataset.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.000	0.051	0.000	1.000
tom_hanks	0.272	0.062	4.383	0.000
spielberg	0.030	0.059	0.509	0.611
anne_hathaway	0.085	0.060	1.416	0.158
julia_roberts	-0.048	0.055	-0.873	0.384
meryl_streep	-0.024	0.057	-0.416	0.678
robin_williams	0.071	0.059	1.199	0.232
sandra_bullock	0.067	0.057	1.177	0.240
rob_downey_jr	0.119	0.058	2.032	0.043
tom_cruise	0.029	0.061	0.473	0.637
nolan	0.223	0.057	3.890	0.000
tarantino	0.292	0.057	5.121	0.000
leo_dicaprio	0.228	0.058	3.927	0.000
jlaw	0.097	0.058	1.675	0.095
gal_gadot	-0.036	0.056	-0.643	0.521
michaelbjordan	-0.068	0.054	-1.264	0.207
chris_rock	-0.067	0.061	-1.094	0.275
adam_sandler	-0.103	0.064	-1.624	0.106
kevin_hart	-0.084	0.055	-1.520	0.130
therock	-0.065	0.060	-1.084	0.280
will_smith	-0.061	0.062	-0.989	0.324
anderson	0.226	0.057	3.991	0.000

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tom_hanks	1	13.414	13.414	19.726	0.000
spielberg	1	0.932	0.932	1.371	0.243
anne_hathaway	1	0.648	0.648	0.953	0.330
julia_roberts	1	1.321	1.321	1.943	0.165
meryl_streep	1	0.012	0.012	0.018	0.892
robin_williams	1	0.312	0.312	0.459	0.499
sandra_bullock	1	0.138	0.138	0.203	0.652
rob_downey_jr	1	1.288	1.288	1.894	0.170
tom_cruise	1	0.193	0.193	0.283	0.595
nolan	1	11.843	11.843	17.417	0.000
tarantino	1	25.360	25.360	37.294	0.000
leo_dicaprio	1	15.533	15.533	22.843	0.000
jlaw	1	4.238	4.238	6.233	0.013
gal_gadot	1	0.373	0.373	0.549	0.459
michaelbjordan	1	0.635	0.635	0.933	0.335
chris_rock	1	3.141	3.141	4.619	0.033
adam_sandler	1	1.421	1.421	2.089	0.150
kevin_hart	1	2.389	2.389	3.513	0.062
therock	1	1.191	1.191	1.751	0.187
will_smith	1	2.587	2.587	3.805	0.052
anderson	1	10.830	10.830	15.926	0.000
Residuals	240	163.200	0.680	NA	NA

Summary & Discussion

Through running a series of simple linear regression models, we discovered that no single regressor can accurately explain the variation in movie average vote rating, despite the regression coefficients being statistically significant. Moving onto our quantitative multiple regression model, we determined that vote count had the most significant impact on vote average, followed by runtime and budget. Interestingly, budget and revenue both were negatively correlated with vote average in both the SLR and MLR models, indicating that blockbuster movies and franchises may prioritize special effects or marketing over quality of the script or plot.

Next, we investigated the effect of genre on average movie rating, and concluded that the regressors for animation/family, drama, crime/thriller, horror/thriller, and mystery/thriller genres were statistically significant. Of these genres, horror/thriller and mystery/thriller movies were negatively correlated with average movie rating, likely due to their focus on eliciting audience reactions rather than character and plot development. On the other hand, animation/family and drama movies tend to have more complex and well-written characters and plots, resulting in their positive correlation with average movie rating. To improve the accuracy of our model, we eliminated outliers and standardized our regression coefficients, resulting in an R^2 coefficient of 0.5132 and an adjusted R^2 coefficient of 0.5078 for our final model.

Finally, we investigated the effect of prominent cast and crew members on average movie ratings. We found that directors who typically take great amounts of artistic control over their movies have a high positive influence on their movies' ratings. This is probably due to their focus on creating movies with engaging concepts and plots to add to their portfolios, rather than trying

to create the highest-grossing picture. While there were a few influential actors, most do not have a significant effect on rating. This indicates that most actors have limited impact on the ratings of their movies, even if they are very famous or have won awards.

Our analysis of what makes movies popular has reaffirmed that the entertainment industry is complex. We hope that some of the factors we have identified as influential to a movie's ratings highlight both what is important to viewers and what goes into constructing a well-rated final product.

Sources

Bell, Travis. "Notice: New method of calculating image votes." *The Movie Database Support*, The Movie DB, https://www.themoviedb.org/talk/50a41d59760ee34f2c000218. 21 Apr. 2023.

"TMDB 5000 Movie Dataset." *Kaggle*, 2016, https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata. 21 Apr. 2023.

Appendix

Some of the variable names in "MLR: Cast & Crew" may not clearly indicate the actor or director they are referring to. In order, here are the full names of the people we were interested in:

- Tom Hanks
- Steven Spielberg
- Anne Hathaway
- Julia Roberts
- Meryl Streep
- Robin Williams
- Sandra Bullock
- Robert Downey Jr.
- Tom Cruise
- Christopher Nolan
- Quentin Tarantino
- Leonardo DiCaprio
- Jennifer Lawrence
- Gal Gadot
- Michael B. Jordan
- Chris Rock
- Adam Sandler
- Kevin Hart
- Dwayne "The Rock" Johnson
- Will Smith

The dataset from which we filtered for recognizable and relevant Oscar winners can be found **here**.