# MGSC 661 – Multivariate Statistical Analysis

Midterm Project 2022

Presented by:
R You Not Entertained


**Group Members:**

Tarek Cheaito 260886446

Amr Maraqa 261101609

Michael Murphy 261060598

Robert Prattico 260867390

Raman Vakil 261110545


November 9th, 2022

# Introduction

IMDB is an established database and social-networking site dedicated to film, television, and other forms of visual media. Since its inception, it has developed a long-standing reputation as the go-to online platform for all information related to the entertainment industry where film enthusiasts can convene to share their passions, view crowd-sourced film ratings, or read up on the latest in entertainment news.

One of the core features of an IMBD page is the rating assigned to the film or television the page pertains to. IMDB ratings are scores generated from the votes of hundreds of thousands of users who attribute a rating between 1 and 10 based on their personal opinion of the film. With such a high quantity of feedback being sourced to compute them, IMDB ratings are highly trusted as reference to judge the quality of a film. As such, it is not unusual to infer the artistic merits of a film based on its rating on IMDB.

In this project, we seek to predict the success of an unreleased film by forecasting its rating on IMDB. Using the film's production and storyline characteristics as predictors, we will seek to predetermine which of the 12 most highly anticipated films premiering in November 2022 will garner the most acclaim from IMDB users.

The goal of our project is to develop a statistical model that will compute an estimation of a film's IMDB rating that is as close to the true rating as possible. The success of our model will be derived from the difference between rating estimation and its true value; the smaller the difference, the greater the success. More specifically, a successful model is one that can yield a prediction that is within 25% of the true value.

In order to build our model, we will use an expansive dataset consisting of the known IMBD ratings and characteristics of nearly 2,000 films. We will begin the model's development by first individually exploring the variables, focusing on their distributions, collinearity, and potential correlations among them. Subsequently, we will examine the predictive relationship between each characteristic and the film rating, our target variable. This is where we will decide which predictors to include in the model as well as their optimal polynomial degree. Once the relationships are established, we will construct our model by piecing them together and running multiple linear regressions to study the aggregated effect of the predictors on the target variable. Our final results will be summarized in the conclusion of this report.

## Data Description

Our given dataset includes 1930 observations of 42 variables, which will act as our potential predictors. In the first phase of the model development, we needed to get a sense of the data. We then started exploring variables individually, mainly focusing on their distributions, collinearity, and the potential correlations among them. This provided us with an initial idea of the variables that might not be significant in coming up with a prediction for our target variable. We continued the process of analyzing the relationships between the target variable and potential predictors by running simple linear regressions. Based on the results of this step, we decided whether or not to include the variable as a predictor. The next step was to find the type of the relationship between predictors and the target variable. We decided to use polynomial regressions for this task. Our reasoning behind this choice has been described in the model selection section. In alignment with our decision to use polynomial regressions, we ran non-linearity tests to find the optimal polynomial degree for each predictor.

Having completed our initial data exploration, we identified the variables to be incorporated into our model. The chosen numerical variables were *movieBudget*, *duration*, *nbNewsArticles*, and *movieMeter*. The chosen categorical variables were *productionCompany*, *director*, *cinematographer*, and the *drama* genres.

For our inherently numeric variables, including *movieBudget*, *duration*, *nbNewsArticles*, *drama*, *movieMeter*, we ran linearity tests. In the case of nonlinearity of the predictor, we tested different polynomial degrees and ran anova tests to find the optimal degree for each predictor. According to our test results, the optimal degree for *duration*, *nbNewsArticles*, and *movieMeter* are 2, 5, and 8 respectively. However, having a polynomial of $8^{th}$ degree would significantly increase the chance of overfitting. Moreover, when we made predictions using a model that had these degrees, we ended up with a predicted rating that was far beyond the 1-to-10 scope. For example, the optimal degree for *duration* was 2. However, squaring a movie length of 90 added 8100 points to our score. The model's coefficient for *duration* was not small enough to return

the predicted score to with 10, prompting our model to yield a score of nearly 10,000 once all other variables were considered. In order to make sure our predictions remained in the appropriate range of 10, we decided to maintain the relationship between these predictors and our target variable as linear. It is also worth mentioning that we ran collinearity tests between all the numeric variables before adding categorical variables to our model. According to this test, there was no collinearity among the selected numeric predictors.

Implementing categorical variables into the model posed another challenge. Fields like *director*, *productionCompany* and *cinematographer* contained hundreds of unique director or cinematographer names. We recognized that creating a dummy for each of these values would be highly impractical. Instead, we decided to devise a ranking system. Using external data, we identified the most important values for each of these categories. For *directors and cinematographers*, we used a credible film review source to identify the top 30 directors and cinematographers with the most highly acclaimed films. For *productionCompany*, we reasoned that production companies whose movie generated the most box office revenue were likely to produce higher quality movies. We followed these rankings to create a top 30 ranking of the directors, cinematographers, and production companies in our dataset. We then create 3 "top 30" dummies: if a film was directed by a director ranked in our top 30, the dummy would be attributed a value of 1. The same logic was applied to the top 30 cinematographers and production company dummies. Because these were variables with values of either 0 or 1, we recognized fitting polynomials to these dummies would be futile

Lastly, based the heteroskedasticity tests we ran on all the predictors, we concluded that all the predictors are heteroskedastic. Using the methods learned in class, we reconciled the heteroskedasticity issues at the final stage of the model's development.

## Model Selection

To build our prediction model, we started with selecting predictors. Based on the available information in the dataset, we had 42 potential predictors. However, we quickly recognize that building a model with 42 predictors was highly impractical and could very well lead to overfitting. Therefore, we intuitively removed irrelevant columns that evidently had no relation to film rating. These included *movieTitle*, *movieID*, and *imdbLink*. From there, we studied the relationships between numerical predictors and our target variable. As a starting point, we ran linear regressions to obtain an initial understanding of the relationship between the selected variable and our target variable.

At this stage, we took note of the p-values and the adjusted $R^2$ values each linear regression yielded. We then ran a non-linearity test on our each of these regressions to either confirm that the linearity of the variables with an already strong relationship (p-value < 0.05, the lower the better) or that the weakness of other relationships wasn't due to the non-linearity of the predictors.

Having better defined the relationship of our predictors with the target, we kept the linear variables intact and ran ANOVA tests for the non-linear predictors to determine their optimal polynomial degrees. Preliminarily, we compared polynomials of degrees 1 to 5. If the increase from the $4^{th}$ to the $5^{th}$ degree showed positive significance (p-value <= 0.05), we continued our analysis up to degree 10 for that specific predictor.

Once the linearity of predictors was identified and adjusted accordingly, we conducted heteroskedasticity tests to ensure there's no inconsistency in the variance of the predictors and that their p-values are not biased. To do so, we the non-constant variance test on all the predictors and corrected the heteroskedasticity if needed.

Finally, we dummified categorical variables and ran regressions to decide which of them to incorporate into our final model. Thereafter, we ran an outlier test on your

model and removed any observations we found exceeding four standard deviations from the mean.

In the following stage of our model's development, we had to decide whether to use splines or continue with polynomials. Keeping in mind that we had eliminated periodic variables (*releaseDate* and *releaseMonth*) from our list of potential predictors due to their insignificant relationship with the target variable, we concluded that using splines instead of polynomials did not add value to our model because we are not trying to follow a time-reliant trend.

After making that decision, we continued our trial and error with polynomials to find an optimal model. In selecting the optimal model, we needed to make sure that our model had significant predictive power without reaching an unreasonable level of complexity. Indeed, in certain instances, we compromised the best polynomial degree for some predictors in favor of reducing the complexity of our model, which in turn would reduce the risk of overfitting.

## Results

In the end, our completed model was able to successfully compute predictions for all 12 test movies. The model itself was entirely linear and incorporated 9 predictors into its prediction calculation. Based on the coefficients, we found that our director dummy variable had the greatest positive impact on our predicted rating. To be more precise, if a test movie was directed by director in our Top 30 Directors list, it gained 0.64 to its rating. *The Fabelmans*, directed by Steven Spielberg, was the only movie that was able to profit from this coefficient.

In fact, after inputting the characteristics of our test movie into our model, we found that *The Fabelmans* was predicted to have the highest rating of all, scoring a commendable 7.8/10. *¡Que viva México*! (7.4/10) and *Devotion* (6.18/10) placed second and third respectively. Interestingly, *Black Panther: Wakanda Forever* (4.91/10) and *Strange World* (4.81/10) – two highly anticipated movies – were among the lowest rated of all movies.

*movieMeter_IMDBpro* had a detrimental effect on the predicted rating, which would intuitively make sense: the lower a movie is ranking (ie. the greater its rank number), the poorer one would expect its rating to be. What was more interesting to discover was that *movieBudget* also had a negative impact on our predictions. That is to say the more money that a movie had available to spend on its production, the poorer the rating is ultimately received. This was a fascinating point to discover as one would expect a movie's quality to improve with a greater number of resources at its disposal. We can argue that sometimes producers and/or directors try to compensate for the weaknesses of their movies by spending more on other aspects of it. However, this overspending on some aspects cannot neutralize the effects of poor performance in other sections.

Another intersting finding about the significance of predictors is that the coefficient for top 10 production companies is positive whereas the coefficient for 11-20 companies is negative. It means having a producer which is ranked among the top 10 contributes to the increase in the movie's *imdbScore*. Initially, we expected having a producer which is ranked between 11-20 also increase the score. However, according to the coefficients, the opposite is the case. It means that aving a producer which is ranked between 11-20 not only would not increase the score, but also would lower it.

Also, we expected the genre of the movie to have a considerable impact on the movie's rating. However, the results showed that only the drama genre could make a difference. According to these results, whether a movie is *drama* or not has a significant impact on its *imdbScore*. However, if we keep everything else consistent, we cannot expect a significant difference between movies from different genres. It can be interpreted as drama is more popular among viewers. Otherwise, they will not care about the genre of the movie.

On the other hand, the result for the impact of top directors and cinematographers is in line with our initial expectations. We expected having a top director or a top cinematographer increase the *imdbScore* of the movie. Based on the regression results, this hyphthesis can be accepted since the coefficient for top predictors and top cinematographers are 0.67 and 0.23 respectively. It also corroborates the expectation that the director contribute to the score of the movie significantly more than the cinematographer.

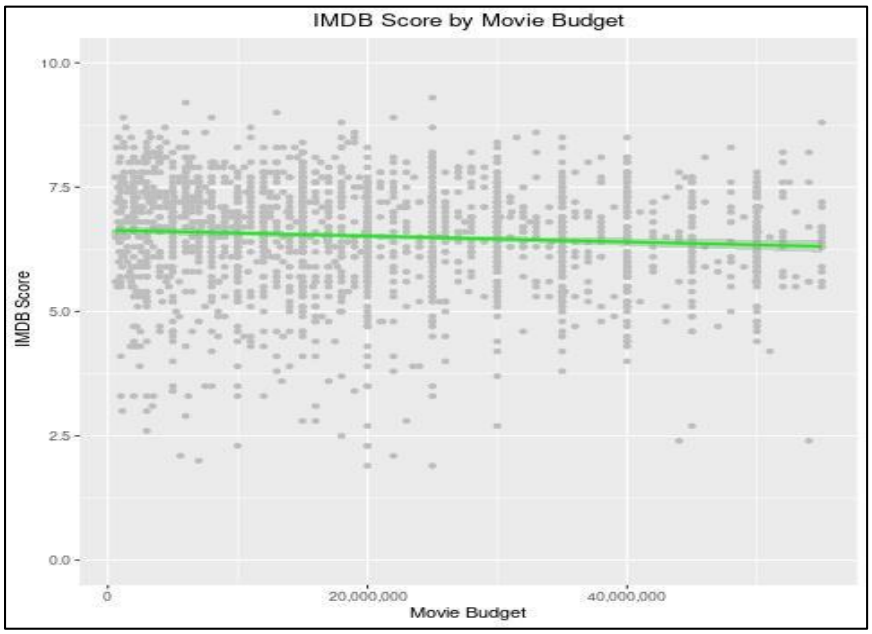
Also, as expected, the number of news articles has a positive relationship with the *imdbScore*. It is intuitive since the more articles about a movie the more attention it would receive from viewer. So, we can expect an increase in the score as a result of receiving more attention from the media. A considerable point is the low coefficient of this predictors which is resulted from the great variance among the observations. We have hundreds of movies that receive only a few (or even one) articles meanwhile some top movies receive thoudands of articles.

Admittedly, the adjusted R squared of our model is disappointingly low. With a score of just over 0.33 and an MSE of 0.78, our model clearly has room for improvement. There are many avenues by which improvement in adjusted R squared can potentially be attained. Predcitors were divided into categorical and numerical.We acknowledge that dummifying many predictors in this case is vital. In addition, the numerical predcitors, such as *duration* (in minutes) and *movieBudge*t, vary in their ranges and are not standardized. This will therefore cause misleading answers in our final model, in the case they are not standardized or scaled appropriately. For example, when the optimal degree of *duration* tuned out to be 2, and we chose it to be linear as a result of the huge increase in our final answer when squaring this variable, the team could have figured out a method to scale it to avoid these drastic increases/decreases. In the submitted version of our code, you'll find a conceptual outline of how these changes could be implemented. Nevertheless, we believe our model performs at a satisfactory level.

In conclusion, using the modelling methodologies learned in class, we were able to successfully develop a predictive model that could yield movie rating forecasts that are reasonable when considering the face value of a film's characteristics and hopefully accurate to their true value.

# Appendix

## Movie Budget




**IMDB Score by Movie Budget**

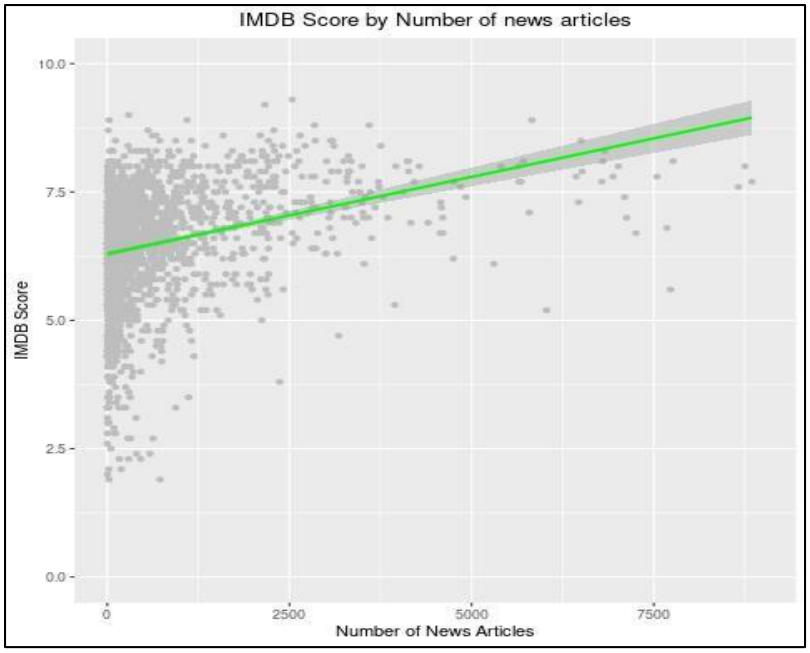
| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Movie Budget | -0.000*** |
| | (0.000) |
| TRUE | 6.636*** |
| | (0.044) |
| Observations | 1,930 |
| $R^2$ | 0.006 |
| Adjusted $R^2$ | 0.006 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

## Duration


**IMDB Score by duration**

| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Duration | 0.021*** |
| | (0.001) |
| TRUE | 4.179*** |
| | (0.120) |
| Observations | 1,930 |
| $R^2$ | 0.169 |
| Adjusted $R^2$ | 0.168 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

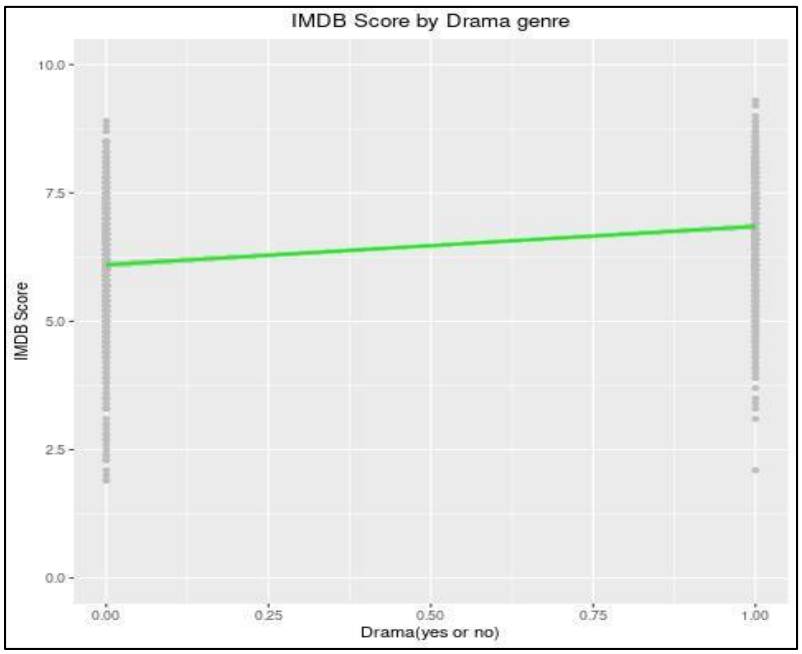
## Number of News Articles




**IMDB Score by Number of News Articles**

| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Number of news articles | 0.0001*** |
| | (0.00001) |
| TRUE | 6.409*** |
| | (0.026) |
| Observations | 1,930 |
| $R^2$ | 0.051 |
| Adjusted $R^2$ | 0.050 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

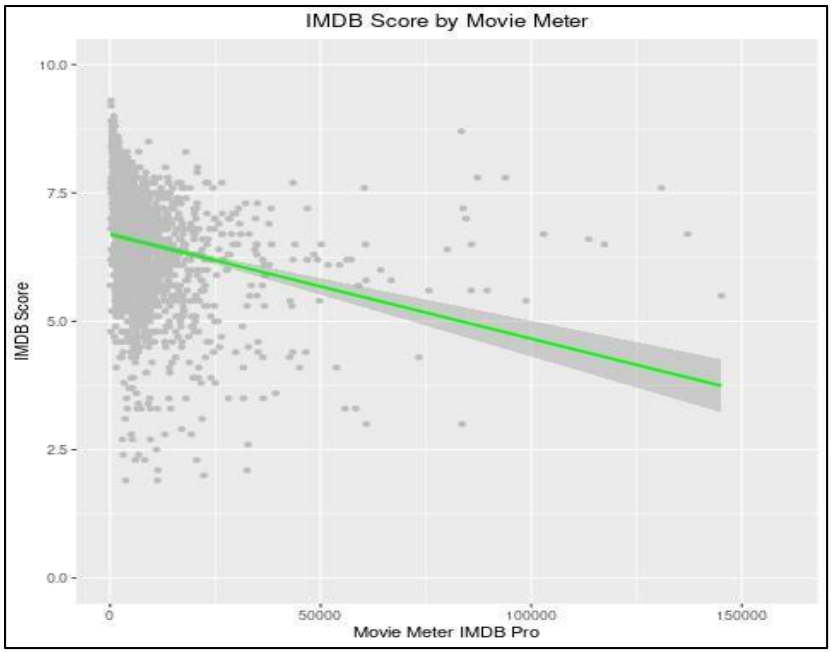
## Genre (Drama)




**IMDB Score by Drama Genre**

| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Drama | 0.748*** |
| | (0.047) |
| TRUE | 6.101*** |
| | (0.035) |
| Observations | 1,930 |
| $R^2$ | 0.114 |
| Adjusted $R^2$ | 0.114 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

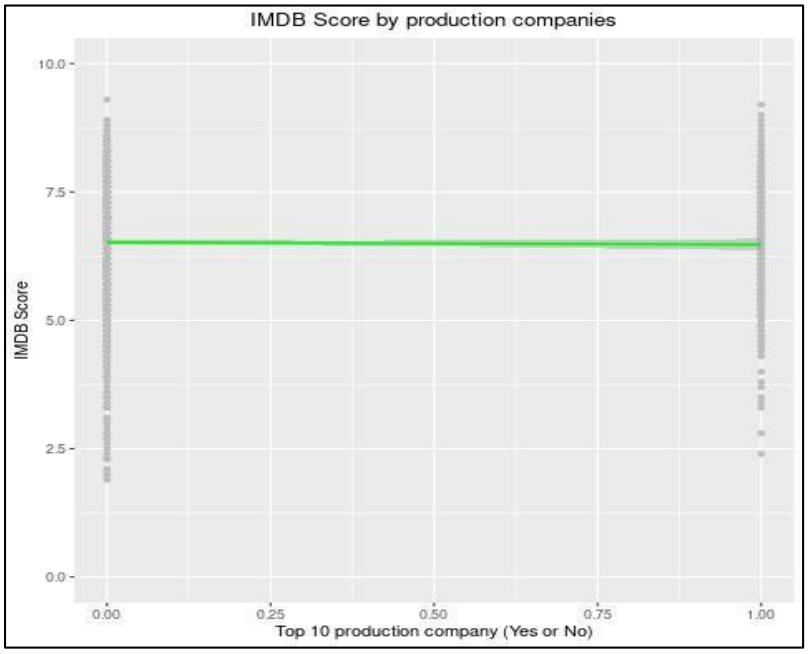
## Movie Meter Pro




**IMDB Score by Movie Meter Pro**

| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Movie Meter Pro | -0.00000*** |
| | (0.00000) |
| TRUE | 6.540*** |
| | (0.026) |
| Observations | 1,930 |
| $R^2$ | 0.008 |
| Adjusted $R^2$ | 0.008 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

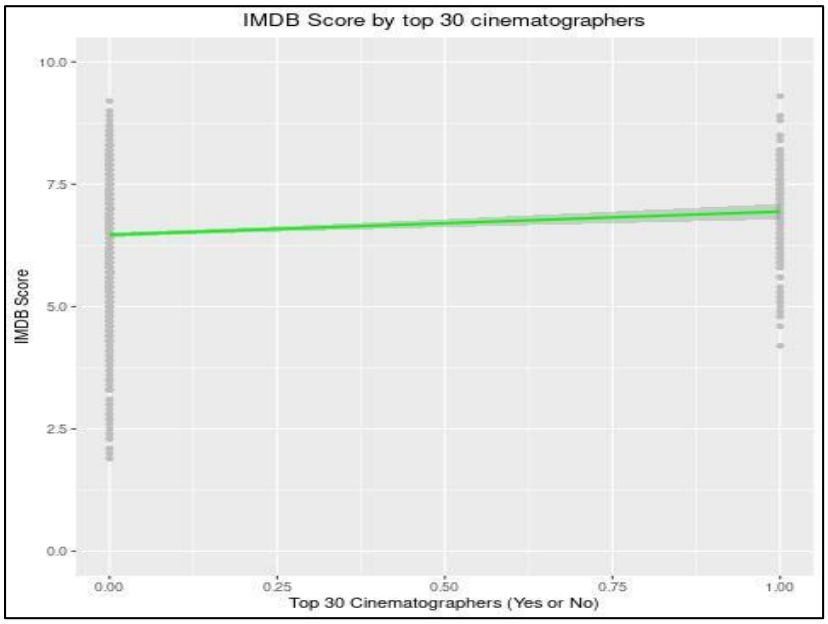
## Production Companies




**IMDB Score by Top 20 Production Companies**

| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Top 10 Production Company | -0.046 |
| | (0.063) |
| Production Companies 11 to 20 | -0.119 |
| | (0.210) |
| TRUE | 6.523*** |
| | (0.028) |
| Observations | 1,930 |
| $R^2$ | 0.0004 |
| Adjusted $R^2$ | -0.001 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

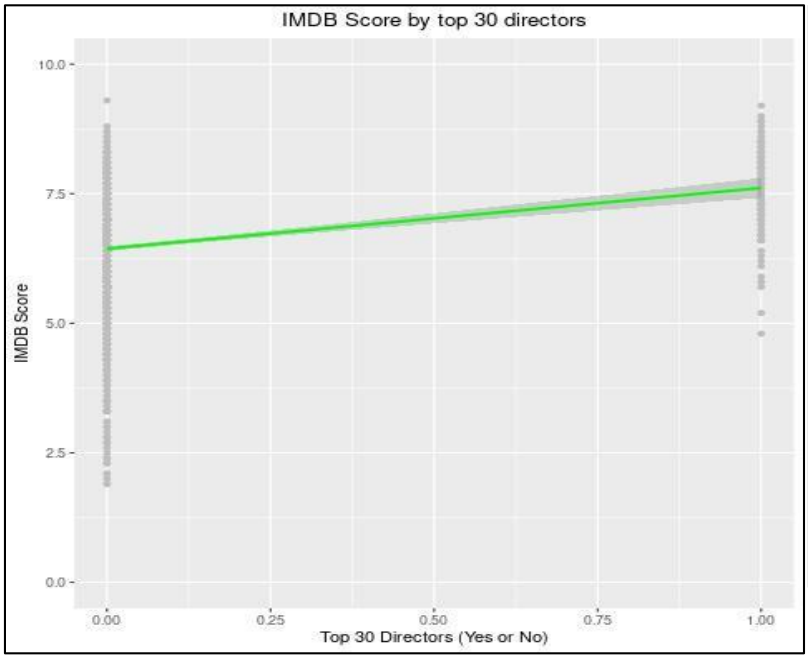
Cinematographer




**IMDB Score by Top 30 Cinematographers**

| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Top 30 Cinematographers | $0.474^{***}$ |
| | (0.086) |
| TRUE | $6.468^{***}$ |
| | (0.026) |
| Observations | 1,930 |
| $R^2$ | 0.016 |
| Adjusted $R^2$ | 0.015 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

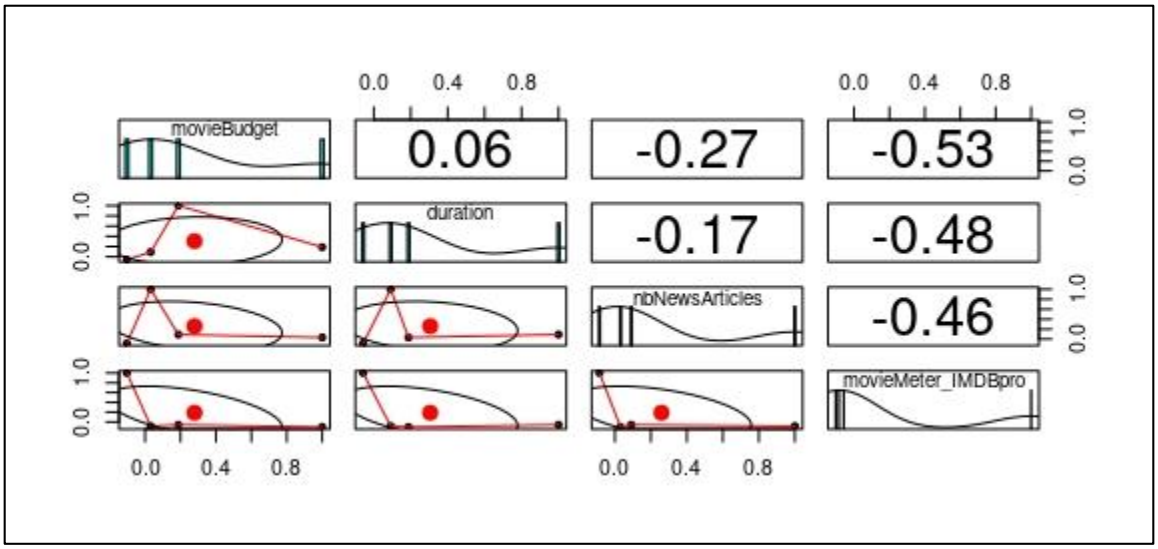
Director




**IMDB Score by Top 30 Directors**

| | *Dependent variable:* |
|---|---|
| | IMDB Score |
| Top 30 Directors | $1.172^{***}$ |
| | (0.101) |
| TRUE | $6.440^{***}$ |
| | (0.025) |
| Observations | 1,930 |
| $R^2$ | 0.066 |
| Adjusted $R^2$ | 0.065 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Correlation Matrix – quantitative variables



# Works Cited

Nash Information Services, LLC. "Top Grossing Director at the Worldwide Box Office." *The Numbers*, Nash InformationServices,LLC,https://www.the-numbers.com/box-office-star-records/worldwide/lifetime-specific-technical-role/director.

Nash Information Services, LLC. "Top Grossing Cinematographer at the Domestic Box Office" *The Numbers*, NashInformationServices,LLC,https://www.the-numbers.com/box-office-star-records/domestic/lifetime-specific-technical-role/cinematographer

Nash Information Services, LLC. "Movie Production Companies" *The Numbers*,NashInformationServices,LLC,https://www.the-numbers.com/movies/production-companies/