Date: July 8, 2020
From: Nasir Zaidi
RE: Is there a relationship between a movie's IMDB score and its director, genre, budget and content rating?

Cinema in the 21st century has extremely evolved due to new technological capabilities. With these new capabilities to produce high quality films, comes a hefty price tag. These budgets often decide how much work will be done, often the movies with a vast budget have the ability to produce the best possible film. Although there are many more factors that go into producing a high rated film other than the budget, a movie's director, content rating and genre play a huge part in a movie's IMDB score. I believe there is a distinct relationship among a movies director, genre, budget and content rating that affects its IMDB score. The findings in my study will serve as a basis for RCA records to use as they decide potential budgets for future cinematic endeavors.

I hypothesize that a movie's director, genre, budget, and content rating all have an influence on the IMDB score of a movie. This memo shows how the effect of each variable compares against another, and shows that biographical films, directed by Christopher Nolan, which are also R-rated, produced the highest IMDB scores.

To properly investigate the hypothesis, I used a comprehensive movie data library with over 5000+ movies sourced from the IMDB Movie Data. With such a large dataset I decided to narrow my data to movies only made in America from directors such as: James Cameron, Quentin Tarantino, Martin Scorsese, Steven Spielberg, and Christopher Nolan. This is due to there being nearly 2400 unique director names to process, so to focus on our analysis we used directors of similar status that are from the same era.

The dependent variable(DV) used in the analysis is the IMDB score of a movie, this score is a quantitative variable measuring the score of a movie (1 being the lowest, 10 being the highest.) The sample size(n) of scores I looked at was 66, with a mean of 7.8 and a median of 7.9. This shows that there is a high concentration of higher scoring films in this subsetted data that I am analyzing, and this can be seen in the relationship with the 4 contributing independent variables(IV) that I have selected.

Table 1: Quantitative Summary

	Freq	Min	Max	Med.	Mean	SD
Budget ($)	66	$500,000	$250,000,000	$52,000,000	$69,214,076	$60,732,083
IMDB Score	66	5.9	9.0	7.9	7.8	0.7

The IV's used in the analysis are the Budget(in millions $), Genre, Content Rating, and Director of the movies in our dataset. Our budget variable was kept as its quantitative self and held a mean of $69.2 million and median of $52 million which shows that the vast majority of the movies in our dataset are multi-million dollar productions with high potential. The genre variable is separated in categories of: Action, Adventure, Biography, Crime, Drama, Thriller. Each of these categories was coded into dichotomous dummy variable's measuring which genres were applied to what movies. Each category held a value of 1 if the data entry matched that genre, if not it held a value of 0. Since there could be multiple

genres for one movie, there is a larger sample size represented in table 2. Drama had the highest frequency of index tagging with 40 different movies under its category. The content rating variable was separated by the only 3 content rating's represented in the data: PG, PG-13, and R ratings. Each of these ratings were coded into dichotomous dummy variables. PG rated movies had the highest frequency with 37 data entries, and was used as the reference in our analysis. Also, I not only used directors to subset our data but I also used them in our analysis, the categories of this variable are each of the directors in our subset. Steven Spielberg had the highest frequency of films in our analysis with 24 data entries, and was also the reference in our analysis.

Table 2: Categorical Summary

	Freq.	Percent
Genre		
Action	19	15%
Adventure	19	15%
Biography	10	8%
Crime	15	12%
Drama	40	32%
Thriller	22	18%
Total	125	100%
Content Rating		
PG	37	41%
PG-13	25	27%
R	29	32%
Total	91	100%
Director		
Nolan	8	12%
Cameron	6	9%
Tarantino	8	12%
Scorsese	20	30%
Spielberg	24	36%
Total	66	100%

Using our dependent and independent variables from our subsetted data, I conducted a multiple ordinary least squares regression, all OLS regression assumptions are accounted for and an alpha of 0.05 was used in this analysis.

The Regression model can be shown by: IMDB Score = 8.224 -0.009(Budget) + 0.155(Genre:Adventure) + 0.161(Genre:Action) + 0.409(Genre:Crime) + 0.244(Genre:Drama) - 0.215(Genre:Thriller) + 0.429(R rated) + 0.078(PG-13 rated) - 0.675(Dir:Cameron) - 1.188(Dir:Scorsese) - 0.810(Dir:Tarantino) - 0.986(Dir:Spielberg).

After performing this regression, our entire model was proven to be statistically significant, our F-statistic(2.33) had a ($p<0.05$). Due to this we are able to reject the null hypothesis and accept the alternate hypothesis that there is in fact a relationship among a movie's budget, genre, content rating, director and the IMDB score that movie received. I am able to support my hypothesis and generalize my findings to the target population because my findings show there is a non-zero correlation for all variables beside budget.

For every movie tagged within the adventure genre category corresponds to an IMDB score increase of 0.16, holding all other variables constant.

For every movie tagged within the action genre category corresponds to an IMDB score increase of 0.16, holding all other variables constant.

For every movie tagged within the crime genre category corresponds to an IMDB score increase of 0.41, holding all other variables constant.

For every movie tagged within the drama genre category corresponds to an IMDB score increase of 0.24, holding all other variables constant.

For every movie tagged within the thriller genre category corresponds to an IMDB score decrease of 0.22, holding all other variables constant.

	Dependent variable:
	imdb_score
genres_ad	0.155
	(0.217)
genres_action	0.161
	(0.247)
genres_c	0.409*
	(0.237)
genres_d	0.244
	(0.232)
genres_t	-0.215
	(0.194)
content2_R	0.429
	(0.285)
content2_PG13	0.078
	(0.265)
budget.m	-0.009
	(0.018)
director_nameJames Cameron	-0.675*
	(0.356)
director_nameMartin Scorsese	-1.188***
	(0.341)
director_nameQuentin Tarantino	-0.810**
	(0.385)
director_nameSteven Spielberg	-0.986***
	(0.305)
Constant	8.224***
	(0.454)
Observations	64
R2	0.354
Adjusted R2	0.202
Residual Std. Error	0.628 (df = 51)
F Statistic	2.326** (df = 12; 51)
Note:	*p<0.1; **p<0.05; ***p<0.01

For every PG movie rating, a R rated movie has an increased IMDB score of .43, holding all other variables constant.

For every PG movie rating, a PG-13 rated movie has an increased IMDB score of .08, holding all other variables constant.

Budget had an almost zero related correlation coefficient, so I decided it is unimportant in the IV's relationship to the IMDB score of a movie.

For every Christopher Nolan film, James Cameron's average IMDB Score decreases by .68 compared to Nolan, holding all other variables constant.

For every Christopher Nolan film, Martin Scorsese's average IMDB Score decreases by 1.19 compared to Nolan, holding all other variables constant.

For every Christopher Nolan film, Quentin Tarantino's average IMDB Score decreases by .81 compared to Nolan, holding all other variables constant.

For every Christopher Nolan film, Steven Spielberg's average IMDB Score decreases by .99 compared to Nolan, holding all other variables constant.

I believe that the coefficients mentioned above are meaningful because they accurately show that IMDB scores are affected by the variables I tested significantly enough. The Adjusted R-Squared value is 20%, this means that 20% of the variation in IMDB scores can be explained by knowing the budget, genre, content-rating, and director of a movie. Potential weaknesses in my study could include outlier data and confounding variables. When examining my regression diagnostics I saw that there are a few outliers, however they are not significant enough to cause a change in my analysis. An example of a confounding variable that could affect IMDB scores is content-rating, as it only had data for 3 out of 4 content-ratings therefore this could limit its application to younger audiences that fall under the 'G' content rating.

Regression Diagnostic Plots