To: President of Marketing, Universal Studios
From: Jose Aguirre-Mori
Re: Do IMDB votes, budget, duration, and total Facebook likes predict movie gross-earnings?

Movies are always highly publicized during their production. There are press releases for casting, announcements for budget, and interviews with directors tend to reveal the duration of the film. After the movie has released, there are reviews and quantitative and qualitative receptions from the public viewers. Therefore, it is important to examine the effects of these various factors on the gross earnings of a movie.

I hypothesize that the duration, budget, total Facebook likes received from the cast, and the total amount of IMDB votes casted for the movie all have a positive increase for the movie's gross earnings. This memo may prove that accurate, but not without some questions regarding the validity of some aspects of the study. Figure 1 summarizes each variable prior to scaling.

Variables	Observations	Min	Max	Median	Mean	SD
Gross Earnings(Dollars)	3890	162	760505847	27996968	5106807	69806681
Number of Voted Users	3890	5	1689764	50469	102610	150726.6
Duration (Minutes)	3890	34	330	106	109.9	22.7
Total Cast Facebook Likes	3890	0	656730	3888	11266	18927
Budget (Dollars)	3890	218	12220000000	24000000	45200000	222417694
Figure 1						

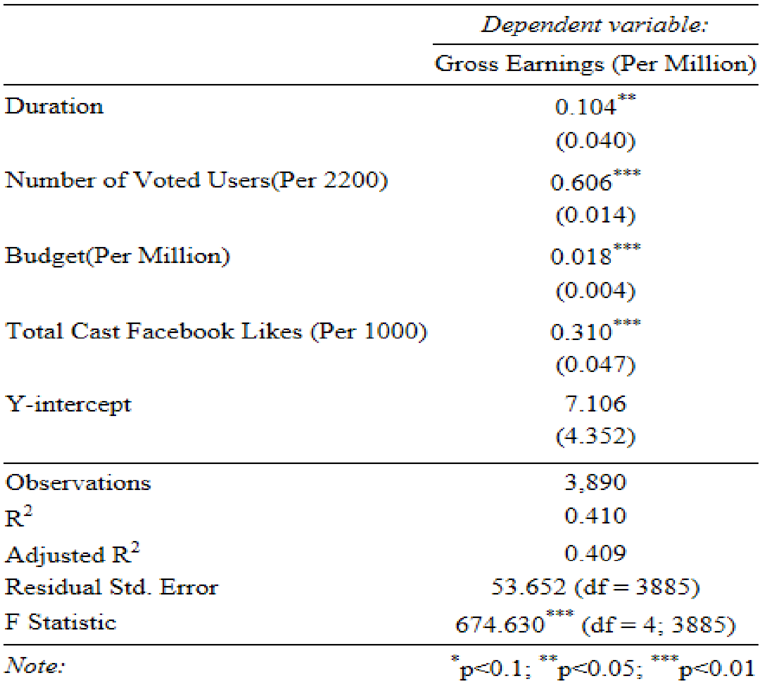
The data was modified where any entries that had missing values for any of the examined variables and gross earnings were omitted. The model was also scaled, gross earnings and budget were scaled down to millions. The total number of IMDB votes were scaled to every 2200 user votes. Total cast Facebook likes was scaled down to every 1000 likes. The target population of the model is the same as the sample size. With an alpha of 0.05, the regression model is as follows in figure 2 at the bottom of the memo:

$$\widehat{Gross\ earnnings(in\ millions)} = 7.11 + 0.61\ (Number\ of\ voted\ users(per\ 2200)) + 0.10\big(Duration(In\ Minutes)\big) + 0.02\ \big(Budget(In\ Millions)\big) + 0.31(Total\ Cast\ Facebook\ Likes(Per\ 1000))$$

The regression model means that a 2200 increase in the number of Voted users on IMDB corresponds to an average gross earnings increase of 610,000 dollars, with all other varaibles constant. $b_{number\ of\ voted\ users\ on\ IMDB} = 0.61, P < 0.05$, so the coefficent is significant. A one minute increase in duration corresponds to an average 100,000 dollar increase in gross earnings for the movie, with all other variables constant. $b_{duration} = 0.10, P < 0.05$, therefore the coefficient is significant. A one million dollar increase in budget corresponds to a 20,000 dollar average increase in gross earnings, holding all other varaibles constant. $b_{budget} = 0.02, P < 0.05$, therefore the coefficient is significant. Lastly, every 1000 Facebook likes earned by the cast in total corresponds to a predicted average 310,000 dollar increase of gross earnings for the movie. $b_{Cast\ Facebook\ Likes} = 0.31\ , P < 0.05$, therefore it is significant. The entire model is significant as p<0.05. That being said, not all of the independent varaibles may be meaningful. It would be impossible to have a movie budget or duration of 0. The coefficient of determination is 0.4093, meaning that the model accounts for 40.93% of the variance in gross earnings. With a variance inflation factor for all of the independent variables being less than 2, our coefficient of determination is safe from artificial inflation and is accurate. It can be generalized that a movie's budget, duration, number of IMDB users who voted, and the total amount of Facebook likes the cast receives does positively affect the movie's gross earnings.

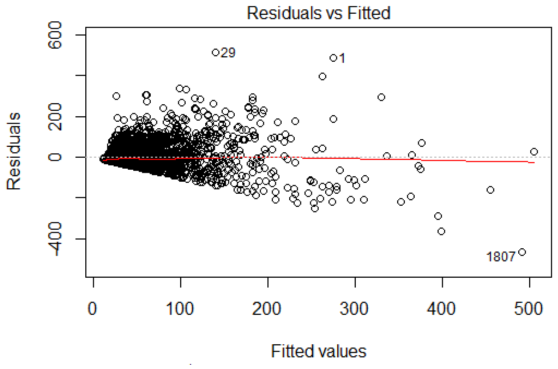
There is some skepticism about the data based on plot diagnostics. Heteroscedasticity and linearity are potential problems based on Figure 3 (see last page). The residuals seem to follow a pattern, indicative of unequal variance. Unequal variance is further evidenced by Plot Figure 3 (see last page) as the plot points are not evenly distributed along the line. Figure 5 (see last page) is also a concern as the plot points do not follow the 45-degree line. It can be stated that the model is not normal based on plot 4. Influential outliers are included in the model as some data observations are contained beyond cooks' distance as seen in Plot Figure 6. This can lead to bias estimates.

Lastly, not all movies are created equal. There are short films, feature length films, documentaries, and possibly even student films that could be contained in the database. It is common-knowledge that these classes of movies are not created equally, from casting to marketing. Considering the class of movie being produced should be considered for future analysis.



	Dependent variable:
	Gross Earnings (Per Million)
Duration	0.104**
	(0.040)
Number of Voted Users(Per 2200)	0.606***
	(0.014)
Budget(Per Million)	0.018***
	(0.004)
Total Cast Facebook Likes (Per 1000)	0.310***
	(0.047)
Y-intercept	7.106
	(4.352)
Observations	3,890
R^2	0.410
Adjusted R^2	0.409
Residual Std. Error	53.652 (df = 3885)
F Statistic	674.630*** (df = 4; 3885)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 2



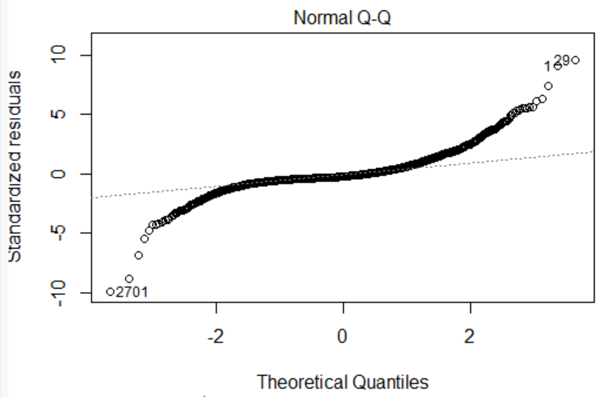
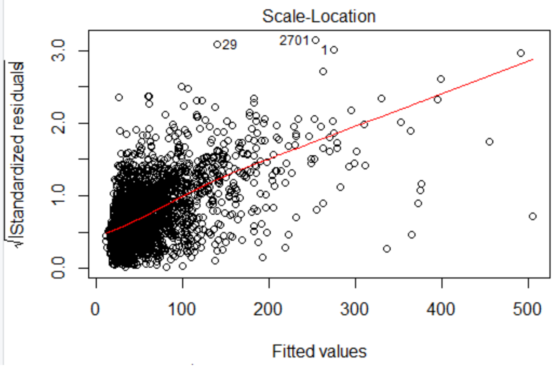


Figure 3

Figure 4



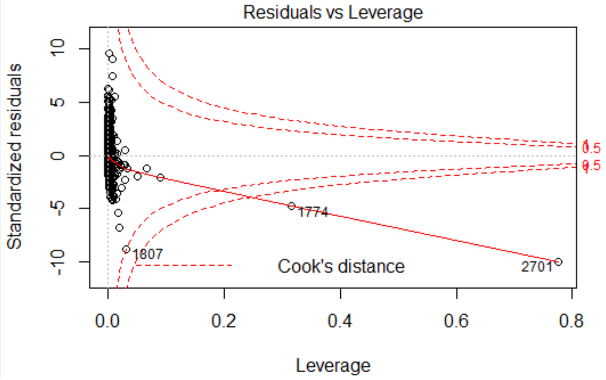
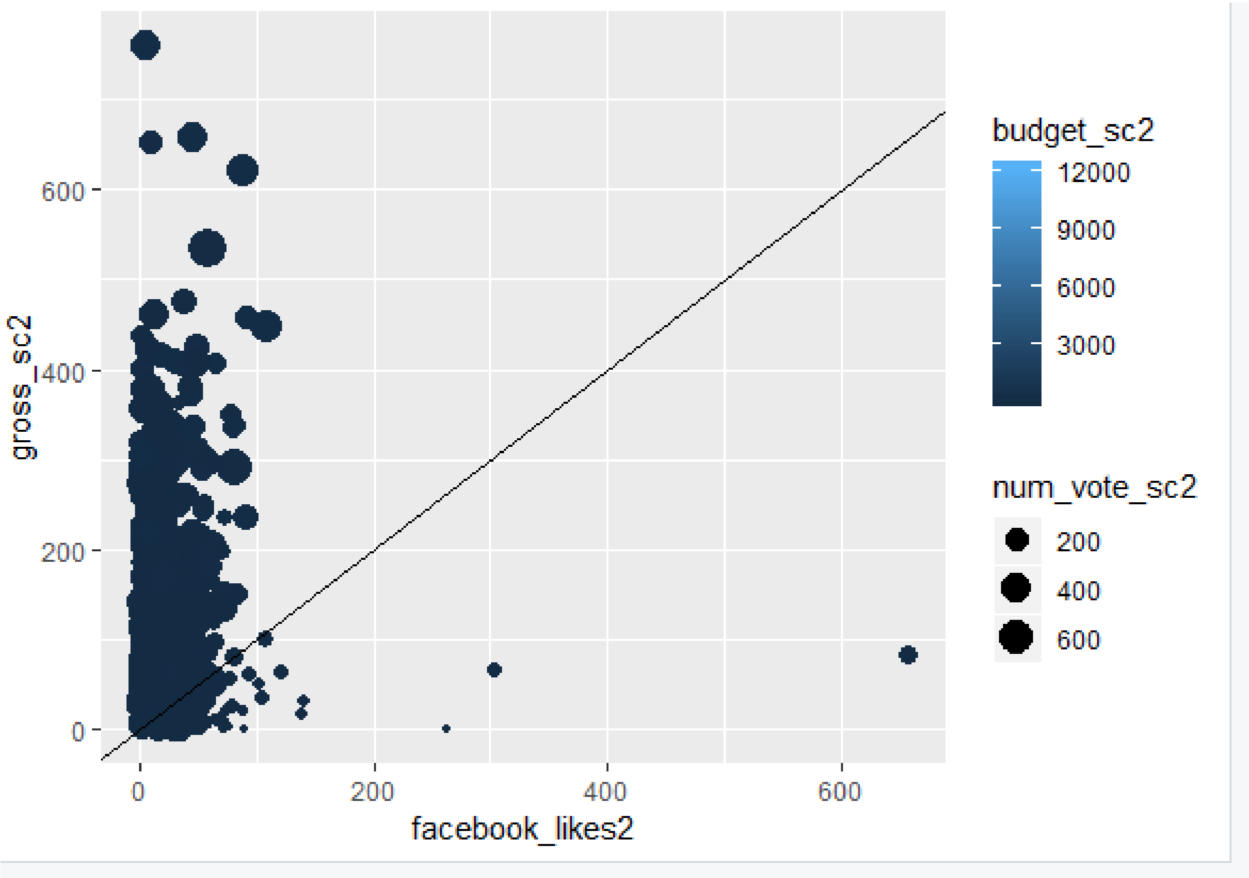


Figure 5

Figure 6


gross_sc2
facebook_likes2
budget_sc2
num_vote_sc2