Memo

To: Prof. Patricia Cortes
Consulting Manager
QM716 Questrom Consulting Firm

From: Bulgamaa Batdelger
Karollyne Zortea
QM716 Questrom Consulting Firm

Date: December 1, 2021

Re: Statistical analysis of the movie industry for QM716 Consulting Firm

This memo includes information about (1) Methodology and fit of predictive models; (2) Budget and animation genre impact on movie gross revenues; (3) Predicted gross revenue for a 2019 movie release; (4) Scene consideration to maximize revenues and IMDb ratings; (5) Correlation between movie duration and gross revenue; (6) Recommendations for hiring movie director; (7) Variables' suggestions to improve the fit of the prediction model.

(1) Methodology and fit

With data on 2090 American movies between 2000 and 2016, we created a predictive model for gross revenues on movies. We included all variables from the data and created dummy variables (i.e., numerical variable, 0 or 1, for a categorical variable) for the content rating (G-rated as baseline) and genre (Action as baseline). We started by eliminating the IMDb rating variable, because the predictive model is for a movie that is to be released. We also eliminated the variable that was not statistically significant and reproduced the model without the variable and continued doing the same thing until all variables were statistically significant (i.e., more than just a chance occurrence). When eliminating the first variable (drama), this changed the baseline variable (i.e., what we are comparing) to the average of Action or Drama movies. This process resulted in the following statistically significant variables (intercept included): duration of movie, Facebook likes for cast, budget, content rating (R), and genre (comedy, thriller-horror, animation, adventure sci-fi) and Facebook likes for director.

In order to determine how well the model fits the data, we took in consideration the adjusted R^2 (0.52), which tells us that 52% of the variation in the outcome of gross revenues can be explained by the variables included in the model, and the Standard error (SE) for the regression line (51.74), which tells us how far in general the points fall vertically from the regression line.
Based on the regression:

$$\widehat{gross\ revenue} = -47.61 + 0.5267 * (duration) + 0.00018 * (cast\ Facebook\ likes) + 0.9310 * (budget) - 7.85 * (R) + 16.39 * (comedy) + 18.98 * (thriller) + 34.28 * (animation) + 8.15 * (adventure\ sci - fi) + 0.00083 * (director\ Facebook\ likes)$$

When keeping all other variables constant, animation and thriller movies tend to do well in the box office. An animation movie generates $34.28 million more in gross revenues than an action movie and a thriller movie generate $18.98 million more in gross revenues than an action movie.

(2) Budget and animation genre

Our predictive model tells us that budget and animation movie genre influence the gross revenue of a movie. When keeping all other variables that impact gross revenues constant, the model tells us that with an increase of $1 million in budget, the predicted gross revenue on movies increases by about $0.93

million. When comparing animation and action genres, while keeping all other variables constant, an animation movie generates $34.28 million more in gross revenue than an action movie.

(3) 2019 movie release – predicted gross revenue
For an adventure sci-fi movie released in 2019, given a budget of $135 million, rated PG-13, with a sum of Facebook likes of 357 for director and 2705 for cast, and a duration of 121 minutes (about 2 hours long), the predicted gross revenue is $150.79 million with a 95% confidence interval between $47.31 million and $254.26 million**.** This means that there is a 95% probability that the predicted gross revenue for this type of movie falls in between this range.

(4) Scene consideration to maximize gross revenues and IMDb ratings
To have gross revenues maximized, the editing company should consider dropping scenes of rated-R content. Based on our model, while holding all the other variables constant, rated-R movies generates $7.85 million less in revenue than G-rated movies.
To have IMDb ratings maximized, the company should consider including rated-R scenes because every additional scene will increase the IMDb rating, on average, by 0.398 points than PG-13-rated movies.

The adjusted R^2 (0.52 for gross revenues and 0.31 for IMDb ratings) seen in our models indicate that the variation in the predicted gross revenues can be explained by 52% of the variables in the regression and variation in the IMDb ratings can be explained by 31% of the variables in the regression.

We are aware of the shortcomings while taking our approach. There are some confounding factors that can potentially influence the gross revenues and IMDb ratings of movies that were not taken into consideration while running our models.
Gross revenues confounders – The number of movie theaters and location that showed the movies in our data, opening gross revenues.
IMDb confounders – The numbers of movies that are sequels and the amount of money spent in advertising before and within the first week that the movie is released.

(5) Correlation between movie duration and gross revenue
As seen in our two model regressions, our conclusion is that the correlation between movie duration and gross revenue is **linear**. In model 1, the variable of *duration* has a tstat of 7.25, which is statistically significant. In model 2, the variable of *duration* has a tstat of 2.99, which is statistically significant and $duration^2$ has a tstat of 1.45, which is not.

(6) Recommendations for hiring movie director
Our recommendation is to hire Robert Stromberg. Based in our prediction model, the movie directed by him was expected to generate $182.16 million in revenue, when in fact, his movie outperformed the prediction by $59.25 million, resulting in a gross revenue of $241.41 million. In contrast, our model predicted that the movie directed by Cedric Nicolas-Troyan would generate $139.62 million in revenue, but in reality, his movie underperformed by $91.66 million, resulting in a gross revenue of $47.95 million.

(7) Variables' suggestions to improve the fit of the prediction model
As we can learn a lot from our prediction models, there are still some factors, which were not in our data, that puts some restrains in our predictions. In order to improve the fit of our prediction models, we believe that supplemental data collection is paramount on marketing strategy (e.g., advertising, budget), awards and nominations (e.g., if a movie was nominated for an Oscar), release date and seasonality, sequel or franchise status (e.g., collection of films in succession that share the same fictional universe) and "Buzz", which is a combination of social chatter, reviews, impressions and editorial mention.

Table 1. Dependend variable: Gross revenues of movies				Table 2. Dependend variable: IMDb rating
	(1),(2),(3),(4A),(5A),(6)	(5B)		(4B)
gross revenues				0.004125805
				(0.000366274)
year				
duration	0.526794777	1.004397718		0.016700316
	(0.072630124)	(0.335198282)		(0.001228511)
duration_squared		-0.001876424		
		(0.00128567)		
cast_total_facebook_likes	0.000187974	0.000185576		3.41108E-06
	(5.09205E-05)	(5.09332E-05)		(8.53392E-07)
budget	0.931099442	0.92722869		-0.003995389
	(0.033868058)	(0.033962571)		(0.000660595)
R	-7.855570966	-7.862402005		0.398952678
	(2.528377422)	(2.527695171)		(0.042333643)
PG				
PG-13				
comedy	16.39036619	16.92845785		-0.560217519
	(3.184320652)	(3.204733997)		(0.053530773)
drama				
thriller_horror	18.98412737	19.51756487		-0.429997329
	(4.521887961)	(4.535411137)		(0.075856082)
other				
animation	34.2846895	35.92719915		0.495455258
	(5.531011321)	(5.642871315)		(0.093243276)
drama_comedy				
adventure_Sci_fi	8.159081659	8.350618491		-0.216963866
	(3.781667285)	(3.782917434)		(0.06324221)
director_facebook_likes	0.000833096	0.000848508		3.57121E-05
	(0.000380602)	(0.000380646)		(6.36516E-06)
Intercept	-47.61606265	-76.67640543		4.328028207
	(7.748484375)	(21.3650585)		(0.130605701)
Observations	2090	2090		2090
SSE	51.73762558	51.72357616		0.864260757
Adjusted R_squared	0.51961983	0.519880691		0.307299931
Note: SEE is Standard Error of the Regression. Standard errors of coefficients in parenthesis.				