What makes a movie successful?

Team Going Merry (glee39, schen74, mwang33, slim20)

Hypothesis

We hypothesized that the budget and GDP of the country a movie was produced in would collectively be positively correlated with revenue, but not correlated with rating.

Data

We downloaded the IMDB data and used the TMDB Movie API. Every movie has a unique ID overlapping between the two databases that was created to facilitate joining, so that is what we used to join the IMDB and TMDB data together. The only data not directly from these sources was GDP data, for which we used the World Bank database to retrieve GDP for each movie based on the year it was released (year found through TMDB). We removed data points that had missing or NULL values for any of our necessary data categories; unfortunately, our sources had a lot of missing data for budget and revenue so we eventually ended up with 513 data points, about 1/10 of our original volume.

Findings

Claim #1: First, we wanted to individually inspect the relationships between our independent and dependent variables.

Results for Claim #1: We ran a single regression with (1) budget vs revenue, (2) budget vs rating, (3) GDP vs revenue, and (4) GDP vs rating. The only significant coefficient and p-value was between budget and revenue.

To clarify the units - budget, GDP, and revenue were all in terms of USD while rating was a scale of 1 to 10. For example, a coefficient of 3.4450 between budget and revenue would imply that for every $1 increase in budget, we may expect a corresponding $3.45 increase in revenue. A coefficient of 8.171e-09, nearly zero, between budget and rating, would imply that for each $1 increase in budget, we would expect basically no increase in rating.

The following table shows the coefficient for each regression:

	Revenue (USD)	**Rating (1-10)**
Budget (USD)	3.4450	8.171e-09
GDP (USD)	-1.841e-07	2.133e-14

Table 1: coefficients for single regression

The following table shows the p-value for each regression:

	Revenue (USD)	Rating (1-10)
Budget (USD)	0.000	0.422
GDP (USD)	0.756	0.135

Table 2: p-values for single regression

Claim #2: Since we ran our single regressions separating our independent variables for our first claim, we wanted to run multiple regressions instead to control for the variables. We believe that budget and GDP will collectively have a positive correlation with revenue, but will have little to no correlation, positive or negative, with rating.

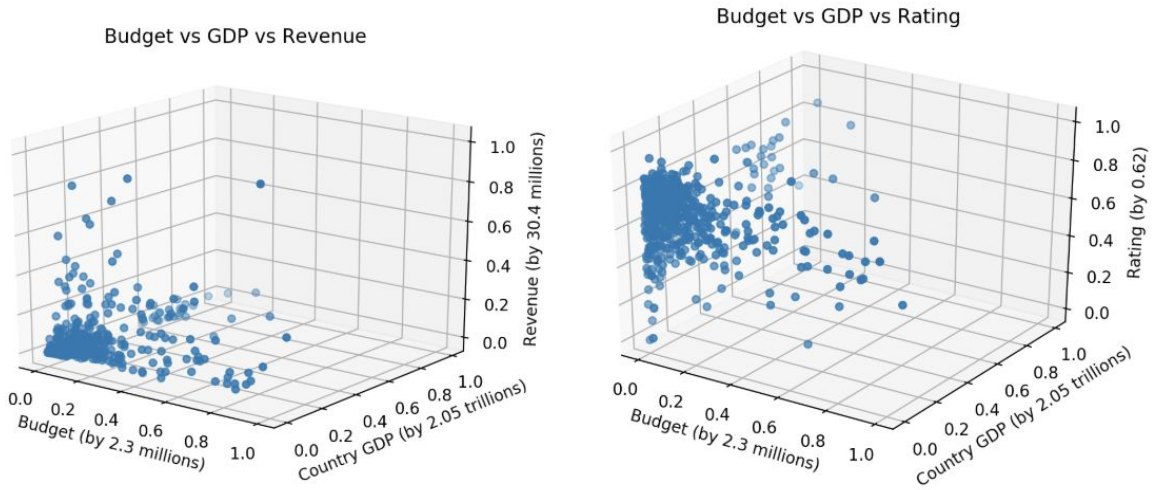
Results for Claim #2: We ran multiple regression with two cases: (1) budget and GDP as independent variables, revenue as the dependent variable, (2) budget and GDP as independent variables, rating as dependent variable. We found the budget's coefficient (3.4447) and p-value (0.000) to be significant for revenue, but not for rating (coeff = 8.211e-09, p = 0.419). We found GDP to not be significant for either dependent variable based on coeff and p values. We believe there are other variables, such as genre or actors, that would allow for additional explanation of variation if included. This would help us more confidently comment on whether the relationship we observe is truly causal or not. Since our highest r-squared, from when our dependent variable is revenue, is only 0.13023, we are not capturing a high level of variation as is.

The following table details our results in full:

	Revenue (USD)	Rating (1-10)
R-squared value	0.13023	0.00563
Budget coeff	3.4447	8.211e-09
Budget p-value	0.000	0.419
GDP coeff	-1.715e-07	2.136e-14
GDP p-value	0.756	0.134

Table 3: r-squared, coeff, p-values for multiple regression

The 3D plots for these scenarios are visualized below. For the sake of readability, we normalized all the values between 0 and 1 and provided the scale per tick mark for each axis.



Figures 1 & 2: 3D plots for multiple regression

Claim #3: Lastly, we wanted to see how these relationships could have evolved over time. To investigate this, we decided to run multiple regressions once again, but split our data points into five ranges of years and regress within each era.

Results for Claim #3: We divided the 513 data points roughly into 5 categories so each category had about 102 data points each. Here is a plot for the average value of each variable, divided into the ranges of years:

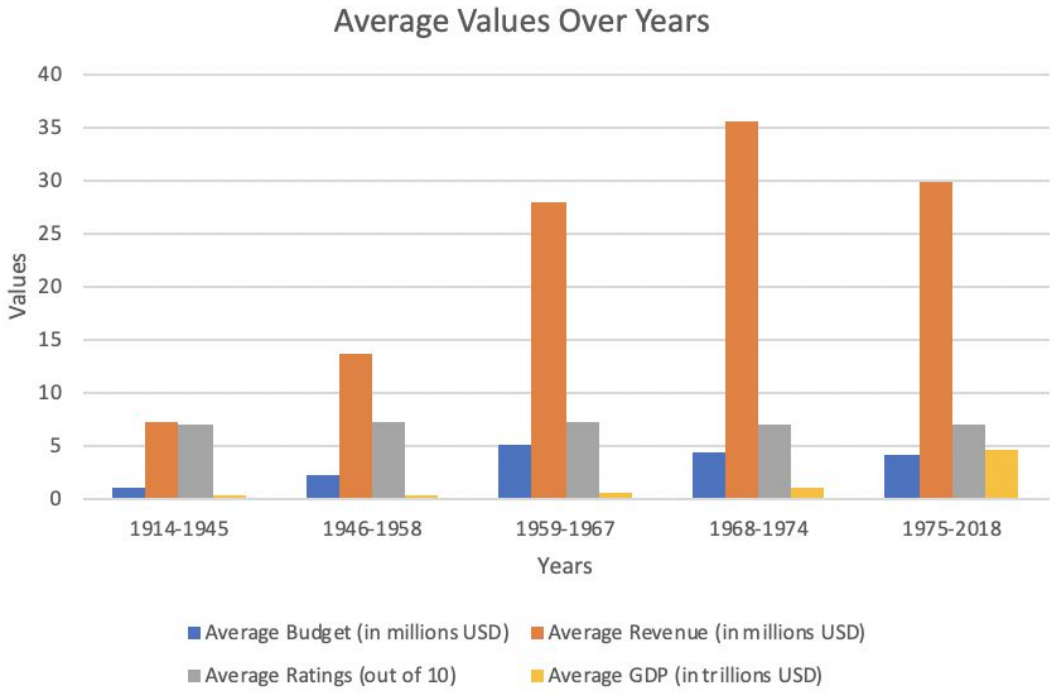


Figure 3: Averages for each time period

Regressing with revenue, budget remained significant for almost every era except 1968-1974. GDP was insignificant for all time periods. We did notice that there is a decent amount of variation between time periods.

Here is a table detailing results for our regression with **revenue** as the dependent variable:

	1914-1945	**1946-1958**	**1959-1967**	**1968-1974**	**1975-2018**
R-squared value	0.08531	0.16534	0.07692	0.02605	0.23769
Budget coeff	8.0254	6.4992	2.3514	1.6732	3.7693
Budget p-value	0.003	0.000	0.007	0.102	0.000
GDP coeff	-1.354e-05	-4.479e-06	-1.533e-05	3.024e-06	-1.202e-06
GDP p-value	0.619	0.845	0.344	0.749	0.112

Table 4: results for categorized multiple regression against revenue

Regressing with ratings, we saw that now, there actually were two time periods (1914-1945, 1946-1958) where budget had a significant p-value (0.000 and 0.061, respectively). GDP actually also had an era (1946-1958) where the p-value, 0.004, was significant.

Here is a table detailing results for our regression with **rating** as the dependent variable:

	1914-1945	**1946-1958**	**1959-1967**	**1968-1974**	**1975-2018**
R-squared value	0.12541	0.09684	0.00238	0.03443	0.05982
Budget coeff	4.233e-07	9.758e-08	-1.608e-09	3.797e-08	-2.951e-08
Budget p-value	0.000	0.061	0.926	0.058	0.132
GDP coeff	-4.009e-13	-2.461e-12	-1.463e-13	3.93e-14	3.053e-14
GDP p-value	0.732	0.004	0.652	0.831	0.106

Table 5: results for categorized multiple regression against rating

Although budget and revenue are most strongly correlated in modern times compared to the other combinations of variables, this was not always necessarily the case.