Business Statistics Data Analysis Report

Mara Riley and Megha Rao

MGSC2301 Business Statistics - Sec 12

Professor Zaarour

7 Dec 2018

Executive summary/intro:

With this project, we set out to determine if there is a significant relationship between the budget and total U.S. gross of movies. We initially hypothesized that there *is* a significant positive relationship between budget and total U.S. gross. We narrowed our focus to American made movies that were released in 2017 by gathering data on the entire population of movies that fit those two constraints and taking a random sample of 50 of those movies. In order to gathering meaning from the data we collected and attempt to prove or disprove our hypothesis, various categories of data and calculations needed to be gathered, such as descriptive statistics, including both measures of central tendency and measures of variability. Additionally, many calculations and values were gathered in order to find the equation of the regression line to enable us to predict the total gross of different budgets.

In our calculations, we found that the coefficient of determination, which shows the percentage of changes in y that are *explained* by changes in x, was very low, at 0.243. This value explains that only 24.3% of changes in total gross are explained by changes in the budget, which is low enough to lead us to doubt that budget and total U.S. gross are highly correlated. Despite this, when conducting both a t-test and an F-test, the null hypothesis, that there is no relationship between budget and total U.S. gross was rejected, providing significant evidence that there is a relationship between budget and total U.S. gross. These test proved our initial hypothesis correct, however there were many variables that we did not take into consideration including the number of days each movie was shown in theaters and the level of fame of actors and directors of each movie. Failing to consider these and many other variables may have skewed our data and resulted in less accurate results and conclusions than we could have obtained.

Body of Report:

For our statistics project, we decided to determine if there is a direct correlation between movie budgets and total U.S. gross. Our data selection process was comprised of three steps. First, we gathered data on the total population of 703 American made movies released in 2017. By controlling these two variables, country of production and year of release, we were enabled to compare movies more fairly and reduce two confounding variables. The data gathered pertaining to each movie included genre, total budget, first weekend gross, and total U.S. gross. Next, we decided on a sample size of 50 movies. We used a random number generator to select 50 movies from the population of 703 movies to make up our random sample. Finally, we needed to decide which pieces of data to compare. We ultimately decided to compare the total budget and total U.S. gross, as we expected there to be the highest positive correlation between those two variables. Once we decided on the variables to compare, we hypothesized that there is a significant relationship between budget and total U.S. gross.

To provide a sense of the movies that were randomly selected to be in our sample, we determined the percentage makeup of each genre represented in our sample. We determined that our sample was comprised of 34% action, 32% drama, 30% comedy, and 6% horror. Despite horror films not being highly represented in our sample, the other three genres each made up 30-34% of the sample, providing us with confirmation that our sample was random and also was a good representation of the population as a whole.

To begin, we calculated some descriptive statistics, including the measures of central tendency. We started off with the minimum and maximum values for both the budget and the total U.S. gross. The movie *Get Out* had the lowest budget, at $5 million, while *Sleight* had the highest budget, at $250 million. As for total U.S. gross, *The Bad Batch* had the lowest gross, at $177,680, and *Wonder Woman* had the highest gross, at $412.56 million. There was a large

range in budgets and total U.S. gross across our random sample. Based on this data, we determined that our sample's total budget has a range of $245 million, while the total U.S. gross has a range of $412.39 million. Both ranges are extremely high, however the fact that total U.S. gross has a range of $412.39 million shows how difficult it is for movie producers and executives to predict their movie's gross. Additionally, the movie that had the lowest budget was not the same movie that had the lowest total U.S. gross. The same can be said of the movies with the highest budget and total U.S. gross. For example, the *Wonder Woman*, the film that had the maximum total U.S. gross of $412.39 million, had a budget of $149 million, which is $101 million less than the maximum budget represented in our sample. On the other hand, *Sleight* had the highest budget, at $250 million. This movie ultimately only grossed a total of $3.85 million. As you can see there were many discrepancies with what we originally hypothesized: movies with the highest budget will have the highest total U.S. gross (that there is a direct correlation between budget and total U.S. gross).

There are many more calculations and measures of central tendency that can help to paint a better picture of the distribution of movie budgets and total U.S. gross. The budget had an average of $60.64 million. Additionally, the median budget (also known as 50th percentile, meaning that 50% of budgets were greater than and 50% were less than that value) was $36.5 million. Similarly, the 25th percentile (Q1) was $21.5 million and the 75th (Q3) percentile was $72.5 million, which explains that 25% of the sample had budgets lower than $21.5 million and 75% of budgets were lower than $72.5 million. Using the Q1 and Q3 percentiles, we were able to determine the interquartile range (IQR = Q3 - Q1), which shows the spread of the middle values. The IQR of the budget was determined to be $51 million, which means that the middle 50% of budgets have a range of $51 million. Finally, the mode budget was $5 million, meaning that a budget of $5 million appeared most frequently in our total sample of 50 random movies.

We also determined the measures of central tendency values for the total U.S. gross. The mean of the total U.S. gross was $81.78 million. The median, which as previously stated as the same value as the 50th percentile, was determined to be $47.39 million. This means that 50% of the sample total U.S. gross were above $47.39 million, and the other 50% were below $47.39 million. Continuing with percentiles, the 25th percentile(Q1) was $20.12 million and the 75th percentile(Q3) was $105.63 million, which shows that 25% of the sample had a total U.S. gross lower than $20.12 million and 75% of the sample had a total U.S. gross lower than $105.63 million. Using these values, we were able to determine the IQR to be $85.51 million, which means that the middle 50% of the total U.S. gross' has a range of $85.51 million.

Some additional measures of variability are useful to understand before beginning to tackle the question of whether or not a movie's budget has an impact on its total U.S. gross and get a deeper understanding of the numbers we are crunching. Variance is the average of the squared distances from each data point to the mean, and it measures how far a set of data is spread out. The budget sample variance was calculated to be $3.772E+15 million which means that the sample budgets are $3.772E+15 million away from the budget mean of $60.64 million. Similarly, the total U.S. gross sample variance was determined to be $9.16E+15 million, meaning that the sample total U.S. gross' are $9.16E+15 million away from the total U.S. gross mean of $81.78 million. Both variances are very large, and the larger the variance, the further data points are from the mean. Standard deviation is very closely related to variance (since it is the square root of the variance) and similarly shows how much variation or spread there is between data points and the mean. The standard deviation of the budget was determined to be $61.4 million, while the standard deviation of the total U.S. gross was determined to be $96.7 million (both of which were the square roots of their variances). As both standard deviations are

high, it shows that data points for both budget and total U.S. gross are spread out over a large range of values in relation to their respective means.

Once the individual values for each variable were calculated, we could move on to calculating values that would help us analyze and answer our original question. First, we calculated the covariance (S_{xy}), which is a measure of the linear association between the budget and total U.S. gross. The covariance was determined to be 2.895E+15. A positive value indicates a positive linear association, which means that as budget increases, total U.S. gross increases. As previously pointed out with the Sleight example, even though total U.S. gross did not always increase as budget increased, the value of the covariance showed us that overall the two variables followed this trend. The next value we calculated was the correlation coefficient (R_{xy}), which is a measure of the relationship between two variables. The value of a correlation coefficient can range from -1 to 1. A value of ∓ 1 means that the two variables have a perfect linear correlation or perfectly increase or decrease at the same rate, while a value of 0 means that the two variables have no linear correlation at all. The correlation coefficient we calculated was 0.492 which is right between 0 and 1. This means that there is no linear correlation or perfect linear correlation, showing that there is some correlation between budget and total U.S. gross, but it is not a perfect correlation by any means.

In order to determine the coefficient of determination, we needed to calculate some more values, namely the sum of squared total (SST), the sum of squared error (SSE), and the sum of squared regression (SSR). The SST is the sum of the squared differences between the total U.S. gross data points and the total U.S. gross mean. The SSE is the sum of the squared differences between the total U.S. gross data points and the estimate while the SSR is the sum of the squared differences between the estimate and the mean. The SSE accounts for all changes in y that are *unexplained* by changes in x. The SSE accounts for all changes in y that

are *explained* by changes in x. It is important to note that the sum of the SSR and SSE is always equal to SST, which our values did. Once these values are calculated, it is easy to see that in order to calculate the coefficient of determination, or the percentage of explained changes, the changes in y that are explained (SSR) is divided by the total changes in y (SST).

After getting the values above, we calculated the coefficient of determination (R^2). The coefficient of determination is a measure of the variability in y that is explained by variability in x. In other words, the coefficient of determination shows the percentage of changes in total U.S. gross that are explained by changes in budget. We calculated the coefficient of determination to be 0.243, meaning that our data shows us that only 24.3% of changes in total U.S. gross are explained by changes in budget.

Next, we needed to determine the equation of the regression line. In order to determine the equation, we needed to calculate more values. First, we calculated the mean of the squared error (MSE, measures the average of the squares of the errors) and the mean of the squared regression (MSR, measures the average of the squares of the regression). To do this, we simply found the average of the SSE and SSR, respectively. Then we determined the standard error(S_e) of the estimate, which shows the deviation around the regression line, or the spread of values around the regression line. This value was calculated to be $84.16 million, which shows that the data points are very spread out away from the regression line. Once we knew the standard error of the estimate, we could determine the standard error of the regression slope(S_{b1}), or the amount of spread around the slope. This was determined to be 0.196. Using the prior values calculated, we determined that the equation of the regression line was $\hat{y} = 0.767x + 35248705.9$ (where $\hat{y} = b_1x + b_0$ is the equation of the regression line).

When we placed this regression line on a scatter plot with the budget on the x-axis and total U.S. gross on the y-axis, it is clear that the data does have a slightly positive slope,

confirmed by the calculated slope in the equation of 0.767 (b_1). On the lower end of the budget, most of the data points are more centralized around the regression line, however as the budget increases, most of this pattern disappears. In summary, there is significantly more deviation from the regression line when the budget is higher than there is when the budget is lower.

Using the previously calculated values, we were also able to determine a confidence interval (a range of values so defined that there is a specified probability that the value of a parameter lies within it) and prediction interval (an estimate of an interval in which a future observation will fall, with a certain probability, given what has already been observed). For both intervals, we chose to test at 95% confidence and use \$82 million as our x_p value. This means that we were calculating the range of values that we were 95% confident that the total U.S. gross would fall into when the budget was \$82 million. The x_p was chosen as \$82 million because it fell within our range of budgets, and it was not an actual budget of any of our data points, which made it a good value. When the budget of *any movie* is \$82 million, the confidence interval determines that, with 95% confidence, the total U.S. gross should fall somewhere between \$74,207,755.17 and \$122,077,656.60. When the budget of *one specific movie* is \$82 million, the prediction interval determines, with 95% confidence, that the total U.S. gross will fall between -\$72,787,031.20 and \$269,072,443.30. The prediction interval will always be wider than the confidence interval, because the confidence interval determines the range for any instance of x, while the prediction interval determines the range for one specific instance of x. The prediction interval also has a wider range due to the fact that we are estimating an interval in which a *future* value will fall. Additionally, the prediction interval we calculated ranges from a negative value to a positive value. Because it is not possible to have a negative total gross, this interval is not an accurate prediction of the total gross when the budget is \$82 M.

To further determine whether or not there is a relationship between the budget and the total U.S. gross of movies, we conducted both a t-test and an F test, which are both hypothesis testing tools to check if there is a significant difference between the means of two groups which may be related in certain features. They are used for normal distribution graphs. We used an alpha of 0.05 for both tests. For the t-test, the $t_{\alpha/2}$ value was $\mp$ 1, and the test statistic t value was 3.91. Because t was greater than $t_{\alpha/2}$, the test statistic fell in the rejection area (also known as significance level area), and thus the null hypothesis, that there is no slope and no relationship between x and y, was rejected. This means that there is a significant relationship between x and y, or between budget and total US gross. To confirm whether or not our results from the t-test were accurate, the F test was conducted as well. F_{α} was determined to be 4.043, and the test statistic F was determined to be 15.37. Because F was significantly greater than F_{α}, it fell in the rejection area. Once again the null hypothesis was rejected. This means that there is significant evidence that there is a slope and a relationship between x and y. The t-test and F test produced the same results, giving us confidence in our conclusion that there is a significant relationship between budget and total US gross.

Conclusion

Based on the data we gathered, our initial hypothesis that a movie's budget impacts its total U.S. gross has been confirmed. Both the t-test and F-test confirmed that there is a significant relationship between the two variables. The way we proceeded with this, was first by calculating the measures of central tendency. This included the minimum, maximum, mean, median, mode, Q1 and Q3 for both the x (budget) and y (total U.S. gross) variables. We then proceeded to calculate the measures of variability which included the range, IQR, variance, standard deviation and coefficient of variation for both the budget and total U.S. gross variables. We analyzed these numbers even further by calculating S_{xy}, R_{xy}, R^2, SST, SSE, SSR, MSE, MSR, S_e, S_{b1}, and the equation of linear regression $\hat{y}$. All of these calculations, helped us finally conduct the hypothesis testing with the t-test and F-test which both confirmed that there was a direct positive correlation between the movie's budget (x) and the total U.S. gross (y).

As with many statistical analyses, there were many variables that we did not take into account that could have skewed our data and ultimate findings. Some variables that we did not take into account include the extent to which the movies had been marketed, the net worth and popularity of actors and directors of each movie, the number of days each movie was in theaters, the number of and location of theaters each movie was shown in, and whether movies were part of a series that already had a large following and fan base. Had we accounted for these variables during data selection, our sample would have had a much more narrow focus and we would have been comparing movies more fairly, which likely would have produced results that more confidently supported our initial hypothesis that movies with higher budgets have a higher total U.S. gross. Upon further analysis, we would like to discover which, of many, variables have a stronger direct correlation with total U.S. gross than the correlation between budget and total U.S. gross.

Appendix

Table 1: Measures of Central Tendency

	Minimum ($)	Maximum ($)	Mean ($)	Median ($)	Mode ($)	Q1 ($)	Q3 ($)
Budget (x)	5 M	250 M	60.64 M	36.5 M	5 M	21.5 M	72.5 M
Total U.S. Gross (y)	177,680	412.56 M	81.78 M	47.39 M	75.47 M	20.12 M	105.63 M

Table 2: Measures of Variability

	Range ($)	IQR ($)	Variance ($)	Standard Deviation ($)	Coefficient of Variation
Budget (x)	245 M	51 M	3.772E+15 M	61.41 M	101.26%
Total U.S. Gross (y)	412.38 M	85.51 M	9.160E+15 M	95.7 M	117.02%

Table 3: Linear Correlation Values

Covariance (S_{xy})	Correlation Coefficient (R_{xy})	Coefficient of Determination (R^2)
2.895E+15	0.492	0.243

Table 4: Simple Linear Regression Values

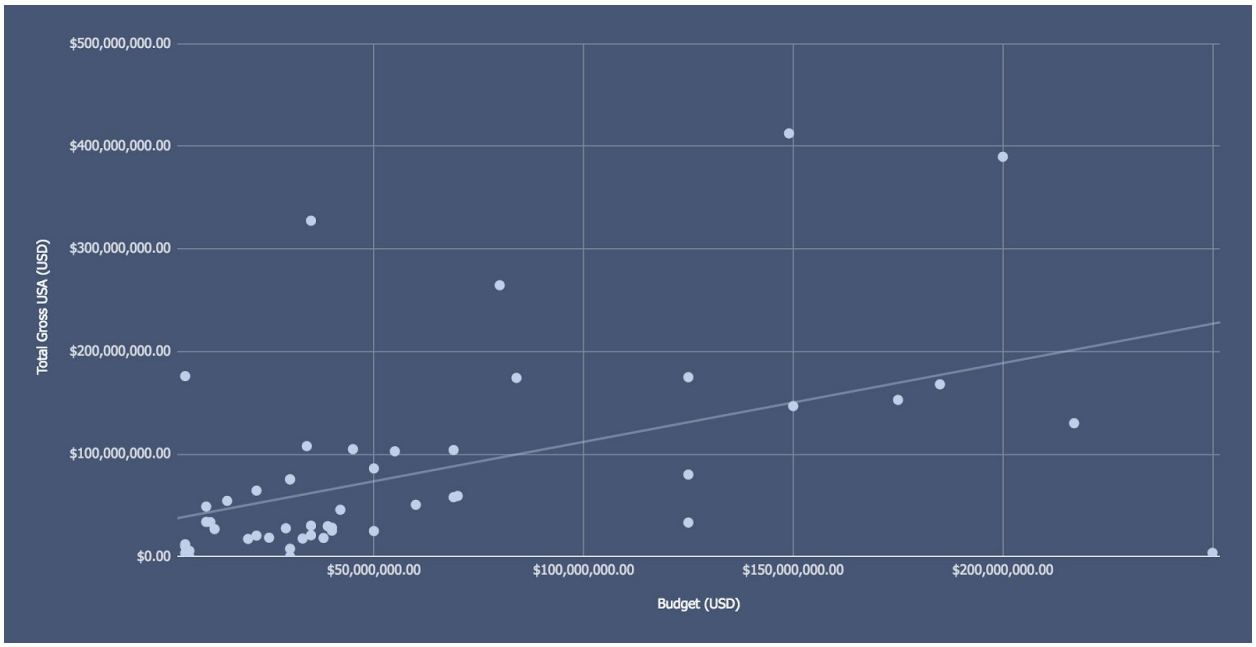
SST	SSE	SSR	MSE	MSR	S_e	S_{b1}	b_1	b_0
4.5E+17	3.4E+17	1.1E+17	7.1E+15	1.1E+17	84.16 M	0.196	0.767	35.2 M

Using these values, we were able to determine the equation of the regression line: $\hat{y} = 0.767x + 35248705.9$

Table 5: Intervals

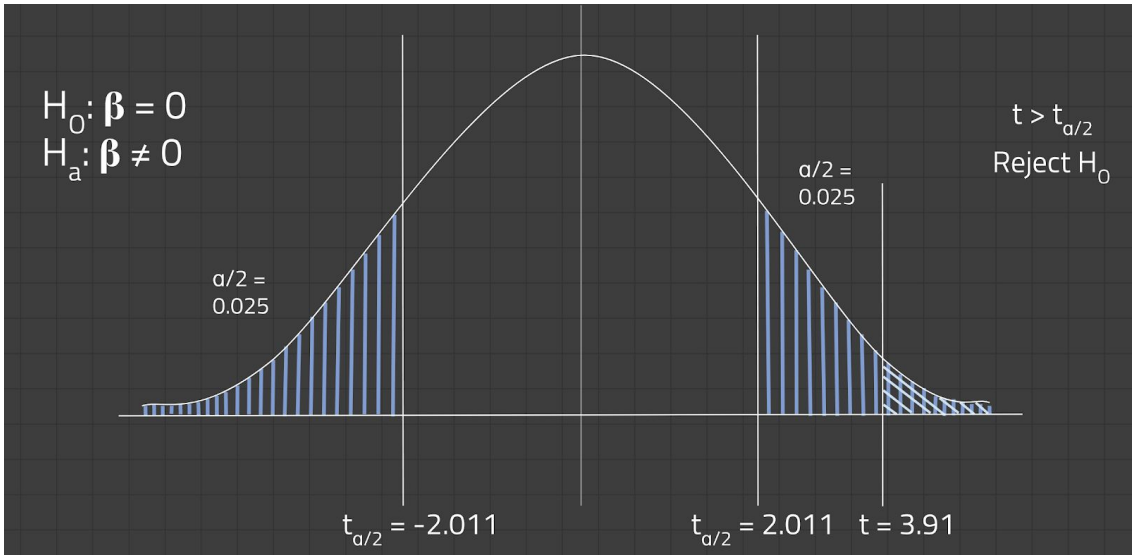
Confidence Interval 95%, X = $82 M	Prediction Interval 95%, X = $82 M
$\$74,207,755.17 \leq Y_{\$82M} \leq \$122,077,656.60$	$-\$72,787,031.20 \leq Y_{\$82M} \leq \$269,072,443.30$

Figure 1: Budget and Total Gross USA (USD)



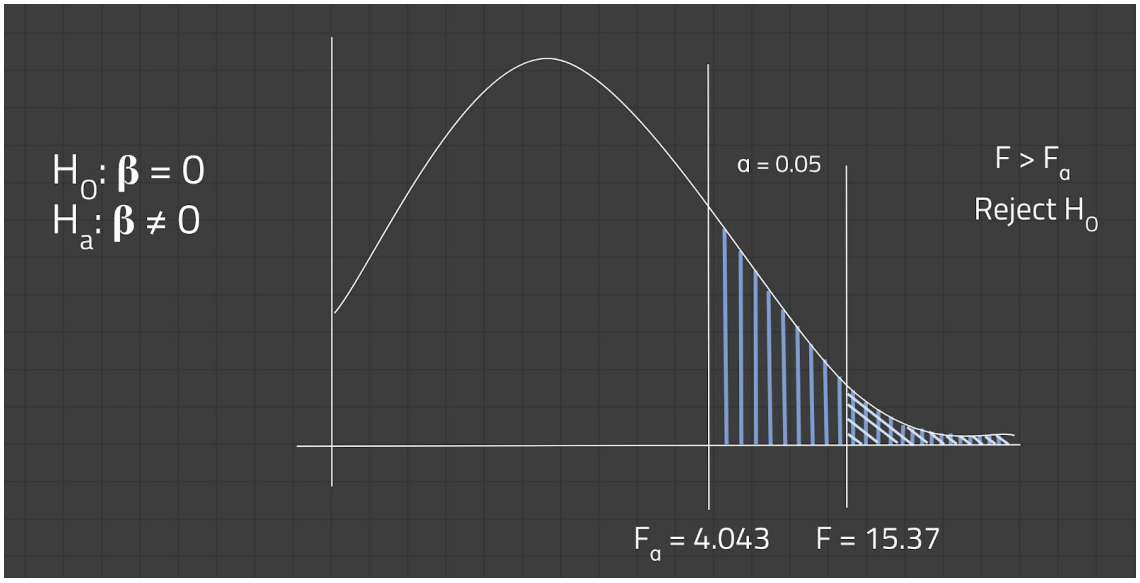
This is a scatter plot showing the budget and total U.S. gross for each movie. As you can see, higher budget does not necessarily mean that there is a higher total U.S. gross.

Figure 2: t-Test



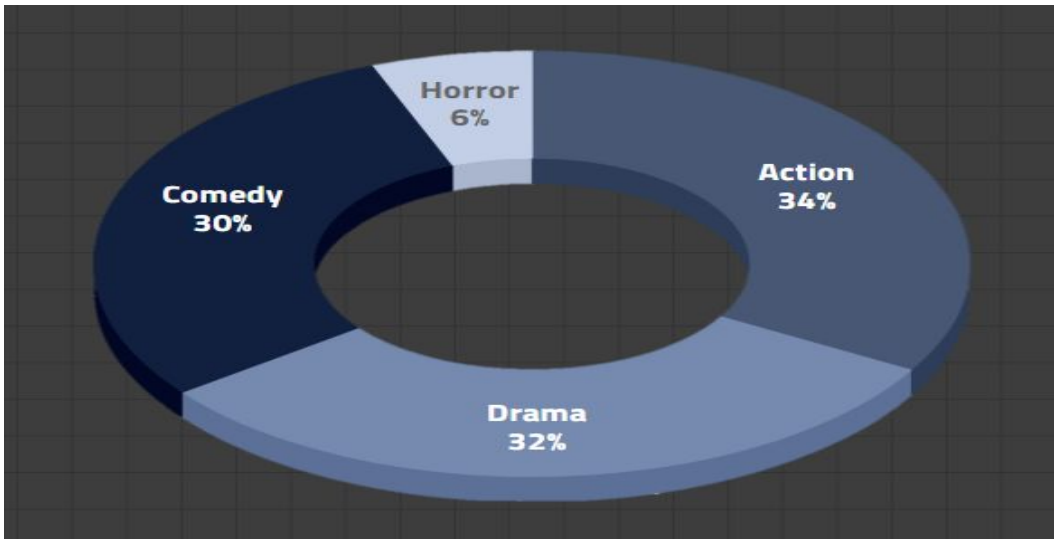
Since t lies in the rejection area, it is clear that there is a significant relationship between budget and total U.S. gross.

Figure 3: F-Test



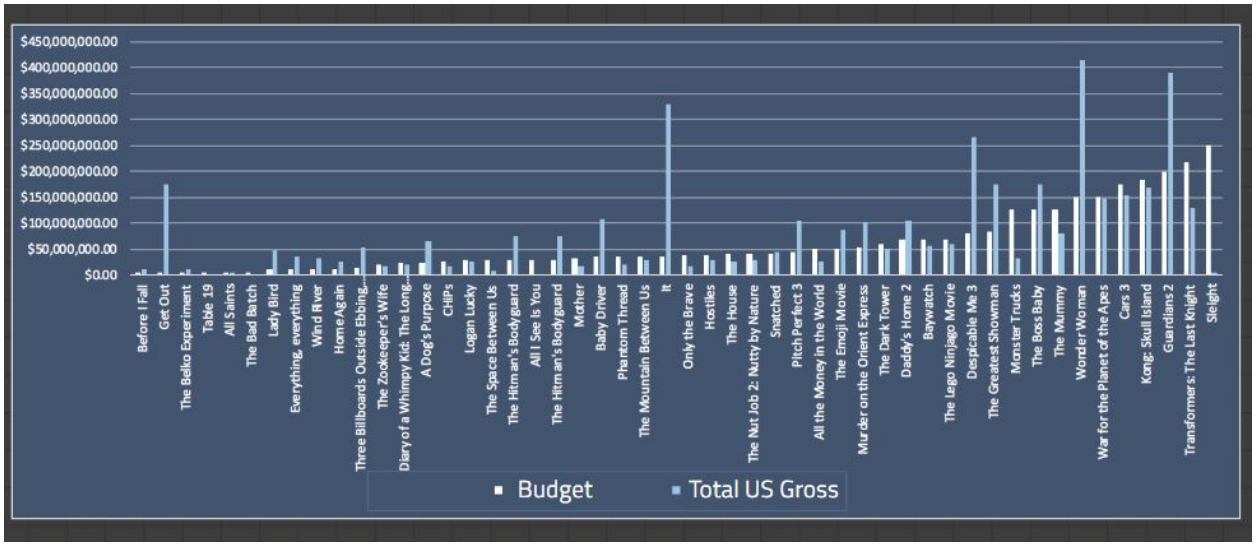
Since F lies in the rejection area, it is clear that there is a significant relationship between budget and total U.S. gross. Both the t-test and F-test produced the same results.

Figure 4: Pie Chart showing Distribution of Genre



Although there are only 6% of horror movies, the other genres take up 30-34% of the sample size, reassuring us that the sample is random.

Figure 5: Ascending Budget and Total U.S. Gross



Sleight had the highest budget, at $250 million. This movie ultimately only grossed a total of $3.85 million. As you can see there were many discrepancies with what we originally hypothesized: not all movies with high budgets had high total U.S. gross value.

References

Box Office Mojo, IMDb.com, Inc, www.boxofficemojo.com/.

Internet Movie Database, IMDB.com, Inc, www.imdb.com/search/.

"Movie Budgets." *The Numbers - Where Data and Movies Meet*, Nash Information Services, LLC, www.the-numbers.com/movie/budgets/all.

Statistics for Business & Economics, Revised, 13th Edition. Cengage Learning, 2018. [Cengage].