TEXTUAL ANALYSIS FINAL REPORT

GROUP 3 MOVIE REVIEW ANALYSIS

Jason Chemaly, Paige Macmillan, Claire McCollough, Blaire Greenblatt

---

Abstract

Review aggregation sites have become increasingly popular over the last two decades. Intentionally or otherwise, they have become indicators and predictors of consumer spending. This paper and our subsequent research will look to determine if movie review sites, such as Rotten Tomatoes, can be predictive of box office revenue. Our secondary research will uncover if movie review sentiment is predictive of movie review ratings. We will perform regression analyses to determine the correlation between movie review sentiment, review score, and gross revenue. We will be able to "hold constant" independent variables such as production budget and movie studio to ensure the most accurate correlation output.

Introduction

"There is no question that there is some correlation to box office performance — critics matter…" (Barnes, 2017). But how much? Since the popularization of review sites, there has been much speculation around not only their effect on consumer behavior, but also regarding the parameters and considerations used to determine final ratings. Critics of movie review sites, and Rotten Tomatoes in particular, speculate whether the sites' reviewers are consistent in regards to specific considerations across all movies and genres. Furthermore, could the sentiment of the review have any bearing on the final movie "score"?

To answer these questions, we amassed over 44,000 movie reviews from the Rotten Tomatoes website for 1000 movies from 2015 to 2019. We calculated each review's sentiment score using Loghran McDonald dictionaries and took the average per movie. We analyzed multiple metrics per movie and how they correlated or could predict box office revenue.

Data

Our first step was to collect all the necessary data. We scraped lists of the 200 most successful movies per year from 2015 to 2019 according to Box Office Mojo (such as https://www.boxofficemojo.com/year/2015/?grossesOption=totalGrosses). We chose these years because they reflect recent trends without muddling the statistics by including the pandemic. From this list of one thousand movies across five years, we generated possible Rotten Tomatoes URLs for the reviews. This was a surprisingly complex process since movies have a wide variety of special characters in the titles. We performed extensive cleaning on these titles before placing them in the URLs. Furthermore, some Rotten Tomatoes URLs are just the movie title, such as https://www.rottentomatoes.com/m/the_martian. Others contain the release year, such as https://www.rottentomatoes.com/m/inside_out_2015. We looped through all of the 1000 movies, first trying the plain URL and secondly trying the movie title and year URL. We conditionally marked each in a Boolean column indicating if the URL succeeded or not. Interestingly, Rotten Tomatoes allows some movies to be valid at both URLs, so in those instances, we kept the one with the simpler URL. At the end of this process, we found 855 movies with valid Rotten Tomatoes URLs and corresponding Box Office Mojo information.

Once we found the correct Rotten Tomatoes URL, we downloaded all of the available reviews and movie details on the information page (https://www.rottentomatoes.com/m/

the_martian), the All Critics reviews page (https://www.rottentomatoes.com/m/the_martian/reviews), the Top Critics page (https://www.rottentomatoes.com/m/the_martian/reviews?type=top_critics), top "fresh" (positive) reviews (https://www.rottentomatoes.com/m/the_martian/reviews?sort=fresh), top "rotten" (negative) reviews (https://www.rottentomatoes.com/m/the_martian/reviews?sort=rotten), the Verified Audience page (https://www.rottentomatoes.com/m/ the_martian/reviews?type= verified_audience), and the General Audience page (https://www.rottentomatoes.com/m/the_martian/reviews? type=user) and saved these to distinct files by movie title. We then parsed information from each of these files containing HTML code from the websites. We chose to collect as much information as we could from each file. Please see Appendix 1 for a complete list of the data fields we gathered. Most importantly, we gathered the text and scores for the variety of reviews and the score information by movie. Since we scraped from several different URLs for the same movie, we made sure to deduplicate the reviews. In total, we scraped the text and scores of 44,158 reviews.

The last external data we gathered were movie budgets from The Numbers website (https://www.the-numbers.com/movie/budgets/all/101). We derived the page numbers to put in the URL, and looped through the pages until we extracted budget information for the first 6,258 movies. Due to inconsistencies in movie title formatting by website, only 649 of the 855 movies with review information (76%) matched our budget dataset.

The last major hurdle within the data was converting the review score information into a standardized numeric format. The raw data allowed reviewers to put whatever score they wanted, which included the following, to name a few:

- A
- C- -
- ¾
- 5/12

- 85%
- 5.42042
- 2 of 5 stars

- Not Recommended

We took any scores that did not contain a forward slash or a whole number and assigned percentages to those on a case-by-case basis. For whole number scores, we assigned either out of 5 or out of 10, depending on the size of the number. For scores containing a forward slash, we performed the division to get a percentage. Once we converted the review score to a percentage and performed basic data cleaning on the remaining fields, we completed the last few regression preparation steps in Excel. We utilized pivot tables and VLookup Excel functions in order to group average sentiment scores per movie as well as average percent scores per movie as opposed to per review. This final data clean-up allowed us to analyze various regression relationships.

Methodology and Results

We started our analysis by calculating sentiment score and identifying common words in each review. In creating the corpora and TermDocumentMatrix, we converted everything to lowercase and removed number characters, punctuation, special non-ASCII characters, and English stopwords. There were 1.3 million total evaluated words across all 44 thousand reviews. We utilized the Loughran McDonald positive and negative dictionaries to calculate sentiment. We chose to use this dictionary despite the risk that it may not accurately or cohesively capture movie-industry-specific words that comprise the reviews. Per our expectations, only 3.7% of words in the reviews had a match in Loughran McDonald dictionaries.

Since this dictionary was only fractionally successful, we started exploring what it would take to create a custom movie industry word dictionary. Fully creating this dictionary would be outside the scope of this project, but we did look at common words found in poorly rated and highly rated reviews. For this analysis, we recalculated the sentiment based on “fresh” vs “rotten” reviews out of the total number of reviews. For example, “refreshingly” was in 30 fresh reviews and 2 rotten reviews, so it received a score of 0.875. On the other end, “flat” was in 19 fresh reviews and 124 rotten reviews, so it received a score of -0.73. After calculating these revised quasi-sentiment scores, we further filtered the results to only contain words that were found in at least 25 reviews.

Creating a dictionary requires more complex consideration than just a threshold, but for the scope of this project, we defined positive words as having a score greater than 0.7 and negative words a score of less than -0.7. This resulted in 104 remaining words. This is insufficient for an entire dictionary, but it is a step in the right direction. When we scale down the Loughran McDonald results as if it had 104 words as well, it would match approximately 0.14% of words in reviews. These 104 words had a 0.9% match, which shows that the movie words already perform better in analyzing sentiment than Loughran McDonald. Here is a sample of our most common words, and for full results, please see Appendix 2.

positive word	score	reviews
gloriously	1	29
parasite	1	25
transcends	1	33
heartbreaking	0.95238095	42
performed	0.93939394	33
standout	0.93548387	31
hopeful	0.93103448	29
admission	0.92592593	27
joyous	0.92592593	27
exhilarating	0.92	50

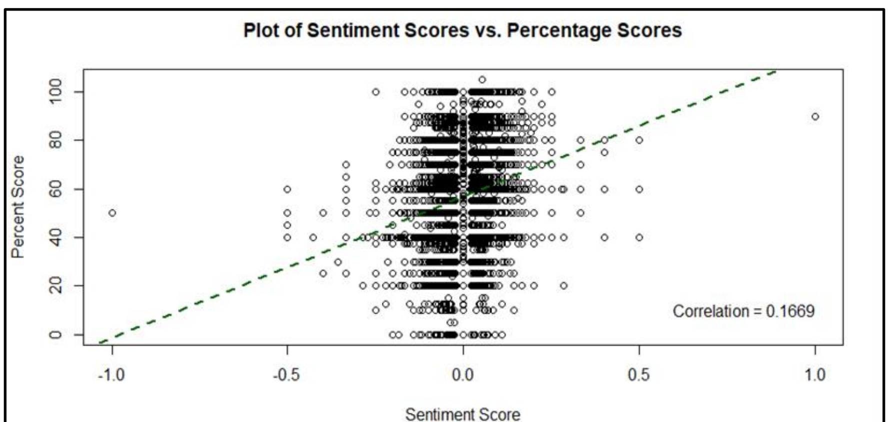
negative word	score	reviews
joyless	-1	25
wastes	-1	26
lifeless	-0.952381	42
slog	-0.9512195	41
unfunny	-0.9069767	86
misfire	-0.8947368	38
uninspired	-0.877551	98
bland	-0.872093	172
laughable	-0.8571429	28
tepid	-0.8518519	27

After investigating common words, we started on the primary focus of our analysis: regressions. We used various single and multiple regressions to gain an understanding of the relationship that exists between the selected variables. For all of the following regressions analyzed below, the default null hypothesis is **H_0: B_1= 0**; meaning that there is no correlation between the X (independent) variable and the Y (dependent) variable.

<u>Model 1: Single Regression - Y: Percent Score per Review vs. X: Sentiment Score per Review</u>

$$Percent\ Score\ =\ \boldsymbol{B0}\ +\ \boldsymbol{B1}Sentiment\ Score\ +\ \mathcal{E}$$

The first analysis we conducted was drawn from 44,157 reviews that each had a sentiment score as well as a percentage score. In this analysis the independent variable is the percent score and the dependent variable is the sentiment score for each review. Our regression output can be seen in model 1 below:




```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          56.82886    0.08969  633.61   <2e-16 ***
merged_dat$sentiment 58.18077    1.78265   32.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.29 on 37186 degrees of freedom
  (6969 observations deleted due to missingness)
Multiple R-squared:  0.02785,   Adjusted R-squared:  0.02782
F-statistic:  1065 on 1 and 37186 DF,  p-value: < 2.2e-16
```

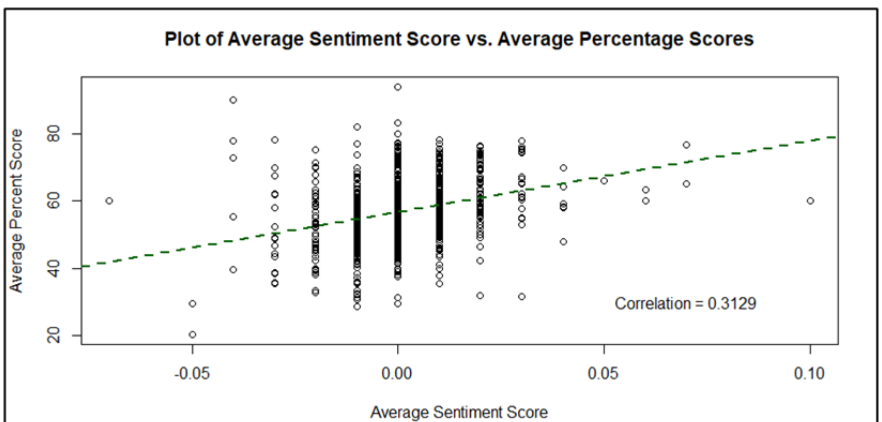
At a 95% confidence threshold, with the p-value < 5%, the data is statistically significant. We can therefore reject the null hypothesis that there is not a linear relationship between the two variables. Therefore, we can conclude that the alternative hypothesis is true, that there is a relationship between sentiment score and percent score. We found the correlation between sentiment score and percent score is 0.1669, which is a weak positive relationship.

In our analysis, the estimated slope on sentiment is 58.18077, implying a positive relationship between Sentiment Score and percent score. Thus, as sentiment score increases by 0.1, we predict that percent score will increase by 5.818 percent.

<u>Model 2: Single Regression - Y: Average Percent Score vs. X: Average Sentiment Score</u>

$$Average\ Percent\ Score\ =\ \boldsymbol{B0}\ +\ \boldsymbol{B1}Average\ Sentiment\ Score\ +\ \mathcal{E}$$

Our second analysis compared the average sentiment score to the average percent score. In order to conduct this analysis, an average sentiment score and an average percent score per each movie was combined. The regression output can be seen in model 2 below:



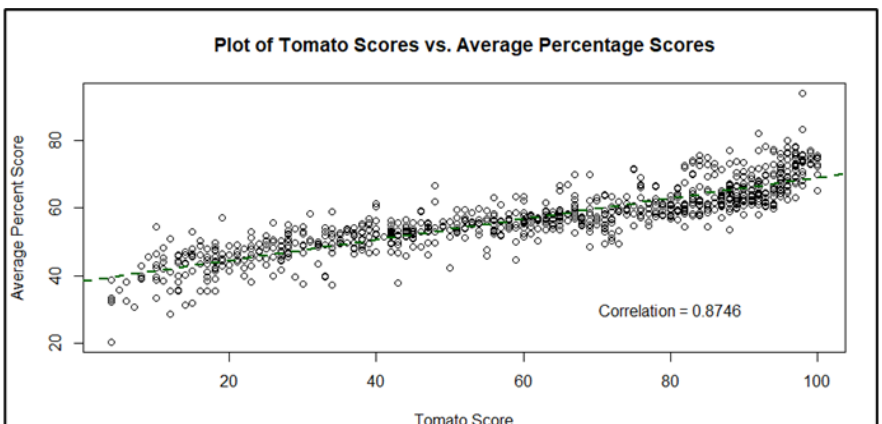

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       56.7914     0.3242 175.201   <2e-16 ***
regression_analysis$AverageSentiment 211.2989    22.2264   9.507   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.328 on 833 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.09788,	Adjusted R-squared:  0.09679 
F-statistic: 90.38 on 1 and 833 DF,  p-value: < 2.2e-16
```

At a 95% confidence threshold, with the p-value < 5%, we found statistically significant data that indicates that the null hypothesis can be rejected and the alternative hypothesis is true, stating that there is a relationship between average sentiment score and average percent score. The correlation between average sentiment score and average percent score is 0.3129, which is a weak positive relationship. Although this is still a weak and positive relationship, there is evidence that indicates the consolidation of reviews per movie with regards to average sentiment and average percentage has a stronger correlation in comparison to sentiment score and percent score per individual review.

Model 3: Single Regression - Y: Average Percent Score vs. X: Tomato Score

$$Average\ Percent\ Score\ =\ \boldsymbol{B0}\ +\ \boldsymbol{B1} Tomato\ Score\ +\ \mathcal{E}$$

Our third analysis compared the average percent score to the tomato score. In order to conduct this analysis, we compared the average percent score per movie aggregated from each review to the Rotten Tomatoes score. The regression output can be seen in model 3 below:




```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  38.44078    0.39639   96.98   <2e-16 ***
regression_analysis$TomatoScore  0.30567    0.00591   51.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.662 on 822 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.7649,    Adjusted R-squared:  0.7646
F-statistic:  2675 on 1 and 822 DF,  p-value: < 2.2e-16
```

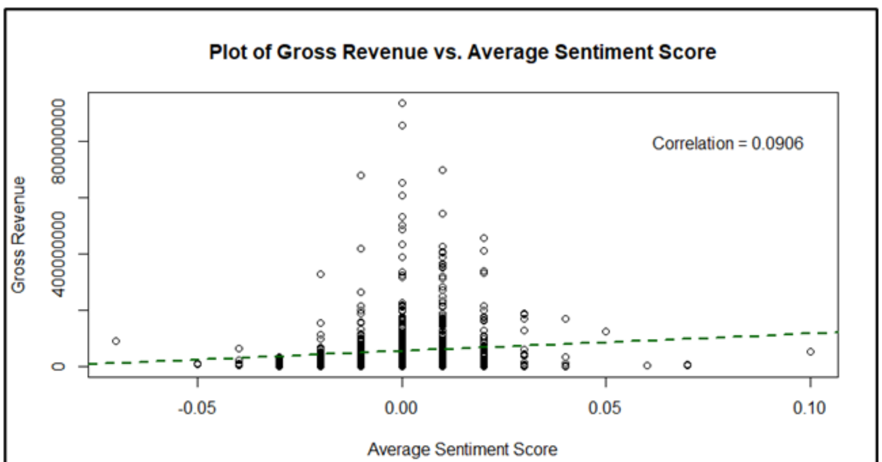
At a 95% confidence threshold, with the p-value < 5%, we found statistically significant data that indicates that the null hypothesis can be rejected and the alternative hypothesis is true, stating that there is a relationship between tomato score and average percent score. The correlation between tomato score and average percent score is 0.8746, which is a strong positive relationship. This indicates that the average score per Rotten Tomatoes reviewer is similar to the overall Tomato score given to each movie.

All regressions below will now be utilized to understand the relationship between the gross revenue (as the dependent variable) and other factors.

<u>Model 4: Single Regression - Y: Gross Revenue vs. X: Average Sentiment Score</u>

$$Gross\ Revenue\ = \boldsymbol{B0}\ +\ \boldsymbol{B1} Average\ Sentiment\ Score\ +\ \mathcal{E}$$

Our fourth analysis compared the gross revenue to the average sentiment score. In order to conduct this analysis, we compared the gross revenue per movie to the average sentiment score per movie. The regression output can be seen in model 4 below:




```
Coefficients:
                                  Estimate Std. Error t value              Pr(>|t|)
(Intercept)                       56354533    3458692  16.294 < 0.0000000000000002 ***
regression_analysis$AverageSentiment 621930776  237013699   2.624              0.00885 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99470000 on 832 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.008208,  Adjusted R-squared:  0.007016
F-statistic: 6.886 on 1 and 832 DF,  p-value: 0.008849
```

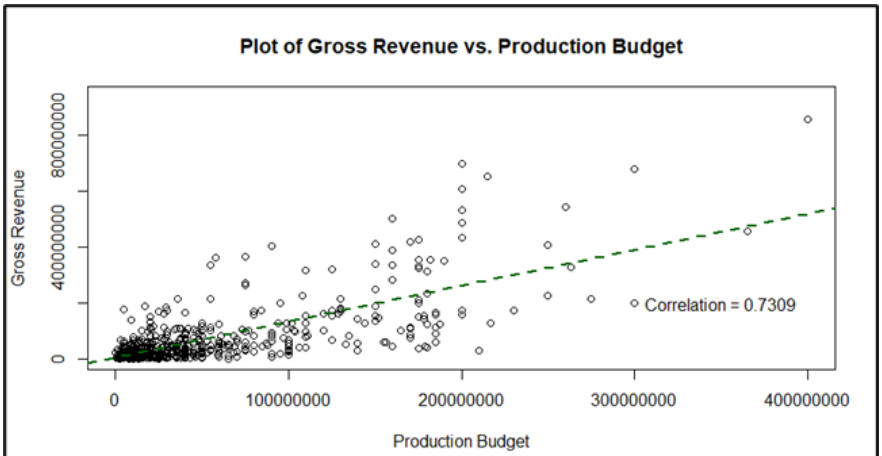
At a 95% confidence threshold, with the p-value < 5%, we found statistically significant data that indicates that the null hypothesis can be rejected and the alternative hypothesis is true, stating that there is a relationship between gross revenue and average sentiment score. The correlation between gross revenue and average sentiment score is 0.0906, which is a weak positive relationship. Although there is a statistical relationship between these variables, the correlation is virtually zero, which indicates that based on this data that there is virtually no relationship between gross revenue and average sentiment score per movie. This very low correlation may be an artifact of using a dictionary that is not specifically geared toward positive and negative words in the context of movie reviews.

Model 5: Single Regression - Y: Gross Revenue vs. X: Production Budget

$$Gross\ Revenue\ =\ \boldsymbol{B0}\ +\ \boldsymbol{B1} Production\ Budget\ +\ \varepsilon$$

Our fifth analysis compared the gross revenue to the production budget. In order to conduct this analysis, we compared the gross revenue per movie to the production budget per movie. We conducted this analysis due to a low correlation in model 4 above which indicated that gross revenue was virtually uncorrelated to the average sentiment score, so this regression gives insight into understanding the elements that drive gross revenue per movie. The regression output for model 5 can be seen below:

At a 95% confidence threshold, with the p-value < 5%, we found statistically significant data that indicates that the null hypothesis can be rejected and the alternative hypothesis is true, stating that there is a relationship between gross revenue and production budget. The correlation between gross revenue and production budget is 0.7309, which is a strong positive relationship.




```
Coefficients:
                                      Estimate    Std. Error t value             Pr(>|t|)
(Intercept)                       4670131.59720 3865807.31911   1.208                0.228
regression_analysis$ProductionBudget    1.28435       0.04925  26.077 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72100000 on 593 degrees of freedom
  (246 observations deleted due to missingness)
Multiple R-squared:  0.5342,	Adjusted R-squared:  0.5334
F-statistic:   680 on 1 and 593 DF,  p-value: < 0.00000000000000022
```

This provides statistical evidence that indicates that a higher gross revenue per movie influences the production budget on a greater scale. This suggests that the more revenue injected into the production of the movie yields a larger gross revenue.

In our analysis, the estimated slope on gross revenue is 1.284, implying a positive relationship between production budget and gross revenue. Thus, for every additional dollar spent on production, our model predicts that gross revenue will increase by $1.28 cents.

An R-squared of 0.5334 tells us that 53.34% of the variation in gross revenue can be explained by the movie's production budget.

Model 6: Multiple Regression - Y: Gross Revenue vs. X: Production Budget, holding constant the distributor (Fixed Effect)

$$Gross\ Revenue = \boldsymbol{B0} + \boldsymbol{B1 Production\ Budget} + \boldsymbol{B2} Distributor + \varepsilon$$

In Model 6, we incorporated the distributor of each movie as a categorical variable to estimate the multiple regression model. It must be noted that due to the large output from the multiple regression, an extract has been selected for the purposes of the discussion. The full extracted multiple regression can be referenced from the “Group 3 Box Office Mojo Regression Analysis” which has been attached as a supporting R Studio script. The results of the regression can be seen below:

```
Coefficients:
                                              Estimate      Std. Error t value            Pr(>|t|)
(Intercept)                              -5008072.90987  32038809.01217  -0.156             0.87585
ProductionBudget                               1.04913        0.06369  16.474 < 0.0000000000000002 ***
DistributorA24                           12962181.75698  35999693.76929   0.360             0.71894
DistributorAffirm Films                  20181410.41077  52307303.75717   0.386             0.69978
DistributorAmazon Studios                -7570916.85733  52305778.60623  -0.145             0.88497
DistributorAnnapurna Pictures            -1745939.43615  52306416.37070  -0.033             0.97338
DistributorAtlas Distribution Company     9094140.03554  78459463.98496   0.116             0.90777
DistributorAviron Pictures                3348503.57056  48047950.17000   0.070             0.94447
DistributorBH Tilt                        5543390.74611  41938700.71433   0.132             0.89489
DistributorVertical Entertainment        -1140046.33788  78457582.33866  -0.015             0.98841
DistributorWalt Disney Studios Motion Pictures  106837168.46771  35464005.96948   3.013      0.00271 **
DistributorWarner Bros.                  18075923.28699  33240555.42495   0.544             0.58681
DistributorYash Raj Films               -16566292.70951  78462079.60513  -0.211             0.83286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71620000 on 537 degrees of freedom
  (246 observations deleted due to missingness)
Multiple R-squared:  0.5837,    Adjusted R-squared:  0.5396
F-statistic: 13.21 on 57 and 537 DF,  p-value: < 0.00000000000000022
```

At a 95% confidence threshold, with the p-value < 5%, we found statistically significant data that indicates that the null hypothesis can be rejected and the alternative hypothesis is true, stating that there is a relationship between gross revenue and production budget, while controlling for the distributor. The correlation is 0.7640, which is a strong positive correlation and reflects a slightly higher correlation than in model 5 above. We calculated the correlation in this multiple regression by taking the square root of the Multiple R-squared variable. The way in which this regression can be interpreted is as follows: holding constant the distributor, as the production budget increases by $1, we predict that gross revenue will increase by $1.05. This model has allowed for a “less noisy” analysis on the true relationship between gross revenue and the production budget.

Although this fixed effect allows us to see all coefficients separately and conduct individual 2-sample t-tests, we may instead be interested in constructing a joint hypothesis test (F-test) to test whether the categorical variable of Distributor is significant to the model overall. Running the **Anova(mylm5)** command produces the following output:

```
Analysis of Variance Table

Response: GrossRevenue
                 Df              Sum Sq             Mean Sq  F value                Pr(>F)
ProductionBudget   1 3534940914349619712 3534940914349619712 689.1193 <0.0000000000000002 ***
Distributor       56  328032286507574336    5857719401920970   1.1419                0.2317
Residuals        537 2754622276058968576    5129650420966422
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The default F-test in this regression context is that all distributors have the same mean production budget. The p-value for the F-test relating to the Distributor dummy variable is > 5%, therefore there is not strong enough evidence to reject the null hypothesis that the mean of the production budget across all distributors is the same. Therefore, we conclude that the mean of the production budget per movie does not vary in a meaningful way across distributors and that the “**Distributor” variable does *not* matter** in estimating the relationship between the gross revenue and production budget.

<u>Model 7: Multiple Regression - Y: Gross Revenue vs. X: Average Sentiment Score, holding constant the distributor (Fixed Effect)</u>

$$Gross\ Revenue = \boldsymbol{B0} + \boldsymbol{B1 Average\ Sentiment\ Score} + \boldsymbol{B2} Distributor + \varepsilon$$

In Model 7, we incorporated the distributor of each movie as a categorical variable to estimate the multiple regression model. It must be noted that due to the large output from the multiple regression, an extract has been selected for the purposes of the discussion. The full extracted multiple regression can be referenced from the “Group 3 Box Office Mojo Regression Analysis” which has been attached as a supporting R Studio script. The results of the regression can be seen below:

At a 95% confidence threshold, with the p-value < 5%, we found statistically significant data that indicates that the null hypothesis can be rejected and the alternative hypothesis is true,

```
Coefficients:
                                            Estimate Std. Error t value            Pr(>|t|)
(Intercept)                                  7862525   21400514   0.367            0.713425
AverageSentiment                           261765040  215678626   1.214            0.225252
Distributor101 Studios                      -1882985   85486050  -0.022            0.982432
DistributorA24                               7050714   26853024   0.263            0.792957
DistributorAbramorama                      -10167412   85652761  -0.119            0.905541
DistributorAffirm Films                      9762250   52408507   0.186            0.852282
DistributorAmazon Studios                   -4240643   40022493  -0.106            0.915645
DistributorAnnapurna Pictures                2213158   46577924   0.048            0.962115
DistributorArea 23a                        -16544805   86035689  -0.192            0.847558
DistributorArtAffects Entertainment         -6055309   85486050  -0.071            0.943549
DistributorWalt Disney Studios Motion Pictures   258366759   24760882  10.434 < 0.0000000000000002 ***
DistributorWarner Bros.                     75794426   23202210   3.267            0.001138 **
DistributorWell Go USA Entertainment        -6895247   40062507  -0.172            0.863396
DistributorYash Raj Films                   -6518889   42835793  -0.152            0.879083
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82760000 on 750 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.3811,	Adjusted R-squared:  0.3126
F-statistic: 5.564 on 83 and 750 DF,  p-value: < 0.00000000000000022
```

stating that there is a relationship between gross revenue and average sentiment score, while controlling for the distributor. The correlation is 0.6173, which is a strong positive correlation and reflects a substantially higher correlation relative to the 0.0906 between gross revenue and average sentiment score (without controlling for the production budget) in model 4 above. We calculated the correlation in this multiple regression by taking the square root of the Multiple R-squared variable which was 0.1657.

Although this fixed effect allows us to see all coefficients separately and conduct individual 2-sample t-tests, we may instead be interested in constructing a joint hypothesis test (F-test) to test whether the categorical variable of Distributor is significant to the model overall. Running the **Anova(mylm6)** command produces the following output:

```
Analysis of Variance Table

Response: GrossRevenue
                 Df              Sum Sq            Mean Sq F value                Pr(>F)
AverageSentiment   1   68134123592365672 68134123592365672  9.9468              0.001676 **
Distributor       82 3095449226624630784 37749380812495496  5.5110 < 0.00000000000000022 ***
Residuals        750 5137412087636130816  6849882783514841
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The default F-test in this regression context is that all distributors have the same mean production budget. The p-value for the F-test relating to the Distributor dummy variable is very low and therefore we have enough evidence to reject the null hypothesis that the mean of the production budget across all distributors is the same. Therefore, there is statistically significant evidence that indicates that the mean of the production budget per movie varies in a meaningful way across distributors and that the **Distributor variable does matter** in estimating the relationship between the gross revenue and average sentiment score.

Conclusion

The objective of our research was to determine if a relationship could be drawn between movie reviews from Rotten Tomatoes critics and box office revenue. We examined movie review sentiment and review score and how these variables may or may not affect revenue. Additionally, we analyzed if review sentiment has any bearing on review score.

The regression analysis we performed produced statistically significant results across all category combinations. However, most relationships were only slightly positively correlated. The essence of our research was to determine if review sentiment correlated to review score and gross revenue. The relationship between review sentiment and review score only produced a 0.1669 correlation, which means we cannot assume sentiment is a strong predictor of score. Furthermore, we determined the relationship between sentiment score and gross revenue to be

0.0906, which again does not allow us to assume sentiment would be a predictor of revenue. However, when we ran the same regression, holding constant movie studio, the positive correlation increased to 0.6173. We went on to examine the relationship between movie production budget and revenue, holding constant movie studio, in which we found a very strong positive relationship of 0.7640.

As discussed, we had determined that the Loughran McDonald dictionaries were possibly not able to determine the sentiment of movie review specific jargon based on their lists of positive and negative words. Should we continue our research, we would create our own dictionaries which would consider movie “lingo” and may possibly produce different sentiment scores. Furthermore, continued research would allow us to look at not only top performing films, but also films that performed the worst. We would also analyze additional variables, such as movie genre, director and/or cast members to determine what effect these variables could have on review sentiment.

References

Barnes, B. (2017, September 7). *Attacked by Rotten Tomatoes*. The New York Times. https://www.nytimes.com/2017/09/07/business/media/rotten-tomatoes-box-office.html

Scorsese, M. (2017, October 10). *Martin Scorsese on Rotten Tomatoes, Box Office Obsession and Why 'Mother!' Was Misjudged*. Hollywood Reporter. https://www.hollywoodreporter.com/news/general-news/martin-scorsese-rotten-tomatoes-box-office-obsession-why-mother-was-misjudged-guest-column-1047286/

Appendix 1: Data Dictionary

Field	**Description**	**Source**	**File**
Rank	Box Office Mojo yearly rank for each movie, between 1-200	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Release	Movie title	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Gross	Total box office gross revenue	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Max.Th	Maximum number of theaters in which the movie showed	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Opening	The gross revenue from opening time period	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
X..of.Total	Percent of total revenue the opening time period generated	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Open.Th	Number of theaters in which the movie initially opened	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Open	Date the movie opened in theaters	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Close	Date the movie closed from theaters	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Distributor	The name of the distributing company	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Estimated	Boolean column indicating whether the revenue numbers were estimated by Box Office Mojo	Box Office Mojo	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Year	Year in which the movie was in the top 200 movies on Box Office Mojo's site. Parsed from URL	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
CleanName	Movie name with special characters removed and including the year, if required by Rotten Tomatoes	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv

BaseURL	Base Rotten Tomatoes URL	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Filename	Filename that will contain the base Rotten Tomatoes website HTML	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
ReviewURL	Rotten Tomatoes URL for general review page	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
ReviewFile	Corresponding filename	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
topCriticURL	Rotten Tomatoes URL for top critic review page	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
TCFile	Corresponding filename	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
verifiedURL	Rotten Tomatoes URL for verified audience review page	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
VAFile	Corresponding filename	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
FreshURL	Rotten Tomatoes URL for fresh reviews page	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
FreshFile	Corresponding filename	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
RottenURL	Rotten Tomatoes URL for rotten reviews page	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
RottenFile	Corresponding filename	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
ReviewSuccess	Boolean column indicating whether or not the HTML was successfully downloaded from the URL	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
TCSuccess	Boolean column indicating whether or not the HTML was successfully downloaded from the URL	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
VASuccess	Boolean column indicating whether or not the HTML was successfully downloaded from the URL	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
FreshSuccess	Boolean column indicating whether or not the HTML was successfully downloaded from the URL	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
RottenSuccess	Boolean column indicating whether or not the HTML was successfully downloaded from the URL	Calculated	FinalMovieIndex_DerivativeURLS_wSuccess.csv
Release	Movie title, exact copy from the FinalMovieIndex corresponding column	Calculated	All_MovieInfo.csv
Rating	Age appropriate rating from MPAA such as PG-13, R, etc	Rotten Tomatoes Base URL	All_MovieInfo.csv
AudienceScore	Rotten Tomatoes audience score	Rotten Tomatoes Base URL	All_MovieInfo.csv
TomatoState	Rotten Tomatoes status such as "fresh" or "rotten"	Rotten Tomatoes Base URL	All_MovieInfo.csv
TomatoScore	Rotten Tomatoes critic score	Rotten Tomatoes Base URL	All_MovieInfo.csv
Title	Movie title as listed on Rotten Tomatoes' website	Rotten Tomatoes Base URL	All_MovieInfo.csv
Info	Summary of year, genre, and runtime	Rotten Tomatoes Base URL	All_MovieInfo.csv
NumCriticReviews	Number of critics who reviewed the movie on Rotten Tomatoes site	Rotten Tomatoes Base URL	All_MovieInfo.csv
NumAudienceReviews	Number of audience members who reviewed the movie on Rotten Tomatoes site	Rotten Tomatoes Base URL	All_MovieInfo.csv
Synopsis	Rotten Tomatoes synopsis of the movie	Rotten Tomatoes Base URL	All_MovieInfo.csv
Genre	Movie genre(s), pipe-delimited	Rotten Tomatoes Base URL	All_MovieInfo.csv

Director	Movie director(s), pipe-delimited	Rotten Tomatoes Base URL	All_MovieInfo.csv
Release	Movie title, exact copy from the FinalMovieIndex corresponding column	Calculated	All_Reviews.csv
ReviewType	Review type by URL, such as "FreshReview" or "CriticReview"	Calculated	All_Reviews.csv
ScoreState	"fresh" or "rotten" as assigned by the reviewer. Audience reviews do not have a score	Rotten Tomatoes Reviews	All_Reviews.csv
ReviewerType	"Critic" "Top Critic" or "Audience"	Rotten Tomatoes Reviews	All_Reviews.csv
OriginalScore	Score as assigned by the reviewer, in the format of the reviewer. Audience reviews do not have scores.	Rotten Tomatoes Reviews	All_Reviews.csv
ReviewQuote	Text of the review	Rotten Tomatoes Reviews	All_Reviews.csv
ReviewerName	Name of the reviewer	Rotten Tomatoes Reviews	All_Reviews.csv
Source	Reviewer website, if applicable	Rotten Tomatoes Reviews	All_Reviews.csv
Date	Review date, cleaned within R from the original format listed on the website	Calculated	All_Reviews.csv
percent_num	The original review score converted to a standardized numeric format	Calculated	MovieInfo_ScoreNumber.csv
percent_char	percent_num divided by 100 to get a percentage	Calculated	MovieInfo_ScoreNumber.csv
Rank	TheNumbers budget rank	The Numbers	All_Budgets.csv
ReleaseDate	Movie release date, according to TheNumbers	The Numbers	All_Budgets.csv
Movie	Movie title as listed on The Numbers' website	The Numbers	All_Budgets.csv
ProductionBudget	Movie production budget	The Numbers	All_Budgets.csv
DomesticGross	Gross revenue within the United States	The Numbers	All_Budgets.csv
WorldwideGross	Gross revenue across the world	The Numbers	All_Budgets.csv
ReleaseDateClean	Cleaned release date	Calculated	All_Budgets.csv
MovieNameClean	Cleaned movie title, in an attempt to match with the previous movie titles	Calculated	All_Budgets.csv
Year	Release year, as parsed from the release date	Calculated	All_Budgets.csv
MovieNYear	Cleaned movie title with the year on the end, for matching purposes	Calculated	All_Budgets.csv
ntotal	Number of total words in the review text	Calculated	sentiment_score.csv
npos	Number of LoughranMcDonald positive words in the review text	Calculated	sentiment_score.csv
nneg	Number of LoughranMcDonald negative words in the review text	Calculated	sentiment_score.csv
pos_score	Positive words divided by total words	Calculated	sentiment_score.csv
neg_score	Negative words divided by total words	Calculated	sentiment_score.csv
sentiment	Calculated sentiment score for each review	Calculated	sentiment_score.csv

Appendix 2: Common Polarized Words - Full Results

word	fresh	rotten	total	sentiment_score	type
gloriously	29	0	29	1	positive
parasite	25	0	25	1	positive
transcends	33	0	33	1	positive
heartbreaking	41	1	42	0.952380952	positive
performed	32	1	33	0.939393939	positive
standout	30	1	31	0.935483871	positive

hopeful	28	1	29	0.931034483	positive
admission	26	1	27	0.925925926	positive
joyous	26	1	27	0.925925926	positive
exhilarating	48	2	50	0.92	positive
excels	34	2	36	0.888888889	positive
winner	34	2	36	0.888888889	positive
brilliantly	50	3	53	0.886792453	positive
gem	33	2	35	0.885714286	positive
importantly	33	2	35	0.885714286	positive
accessible	32	2	34	0.882352941	positive
wellmade	32	2	34	0.882352941	positive
refreshingly	30	2	32	0.875	positive
steals	28	2	30	0.866666667	positive
delicate	27	2	29	0.862068966	positive
panda	27	2	29	0.862068966	positive
assured	26	2	28	0.857142857	positive
gently	26	2	28	0.857142857	positive
mature	38	3	41	0.853658537	positive
richly	25	2	27	0.851851852	positive
taika	25	2	27	0.851851852	positive
deftly	37	3	40	0.85	positive
feat	37	3	40	0.85	positive
gripping	83	7	90	0.844444444	positive
silence	23	2	25	0.84	positive
poignant	80	7	87	0.83908046	positive
delightful	68	6	74	0.837837838	positive
exquisite	34	3	37	0.837837838	positive
kindness	33	3	36	0.833333333	positive
thankfully	43	4	47	0.829787234	positive
refreshing	84	8	92	0.826086957	positive
revelation	29	3	32	0.8125	positive
sensitive	38	4	42	0.80952381	positive
thoughtprovoking	36	4	40	0.8	positive
masterful	35	4	39	0.794871795	positive
kung	26	3	29	0.793103448	positive
pixars	26	3	29	0.793103448	positive
tender	43	5	48	0.791666667	positive
universal	43	5	48	0.791666667	positive
effortlessly	25	3	28	0.785714286	positive
innocence	25	3	28	0.785714286	positive
immersive	33	4	37	0.783783784	positive
intimate	57	7	64	0.78125	positive
captures	72	9	81	0.777777778	positive
enchanting	24	3	27	0.777777778	positive
terrific	111	14	125	0.776	positive
compassion	31	4	35	0.771428571	positive
deft	23	3	26	0.769230769	positive
empathetic	23	3	26	0.769230769	positive
rewarding	23	3	26	0.769230769	positive
visceral	45	6	51	0.764705882	positive
accomplished	37	5	42	0.761904762	positive
handles	22	3	25	0.76	positive

holidays	22	3	25	0.76	positive
lopez	22	3	25	0.76	positive
proving	22	3	25	0.76	positive
stewart	22	3	25	0.76	positive
sublime	22	3	25	0.76	positive
devastating	29	4	33	0.757575758	positive
glad	29	4	33	0.757575758	positive
affecting	50	7	57	0.754385965	positive
thanks	187	29	216	0.731481481	positive
timely	90	14	104	0.730769231	positive
wonderful	130	21	151	0.721854305	positive
enjoyable	264	46	310	0.703225806	positive
joyless	0	25	25	-1	negative
wastes	0	26	26	-1	negative
lifeless	1	41	42	-0.952380952	negative
slog	1	40	41	-0.951219512	negative
unfunny	4	82	86	-0.906976744	negative
misfire	2	36	38	-0.894736842	negative
uninspired	6	92	98	-0.87755102	negative
bland	11	161	172	-0.872093023	negative
laughable	2	26	28	-0.857142857	negative
tepid	2	25	27	-0.851851852	negative
poorly	6	72	78	-0.846153846	negative
unconvincing	2	24	26	-0.846153846	negative
unimaginative	2	23	25	-0.84	negative
halfbaked	3	32	35	-0.828571429	negative
garbage	3	31	34	-0.823529412	negative
disappointingly	4	38	42	-0.80952381	negative
tiresome	4	38	42	-0.80952381	negative
misguided	3	28	31	-0.806451613	negative
soulless	3	28	31	-0.806451613	negative
fake	3	27	30	-0.8	negative
problem	20	177	197	-0.796954315	negative
offensive	5	44	49	-0.795918367	negative
unfortunately	24	201	225	-0.786666667	negative
reduces	3	25	28	-0.785714286	negative
unfortunate	6	47	53	-0.773584906	negative
bore	4	30	34	-0.764705882	negative
cash	5	37	42	-0.761904762	negative
boring	27	199	226	-0.761061947	negative
lazily	3	22	25	-0.76	negative
flat	19	124	143	-0.734265734	negative
fails	40	250	290	-0.724137931	negative
dull	35	218	253	-0.723320158	negative
mediocre	16	95	111	-0.711711712	negative
disappointing	22	129	151	-0.708609272	negative