IST718: Big Data Analytics
Final Project
Team: Coders: Baskar Dakshinamoorthy, Randy Geszvain, Reporter: Reed McElfresh

Code and original datasets for this project can be found at:
https://github.com/rmcelfresh/IST718FinalProject

Does Pre-Movie Hype and Investment Predict Box Office Results?

Abstract: Can you use general information about a film along with sentiment of tweets and critic reviews to determine how a movie will perform at the box office? This review was limited to three film universes, the Marvel Cinematic Universe (MCU), the DC Extended Universe (DCEU) and the Star Wars Universe and only included films released in the last five years.

Overall, we weren't that successful in predicating box office results, partly because our selection of films was small (only 21) and this resulted in some categorical variables having a limited number of observations making the regression results not statistically significant. Based on our regression results, we predicted the next film, Spiderman Homecoming will have a worldwide gross of $611,549,260 when it is released in July.

We recommend that this data be continually added to as new franchise films are released to increase its predictive power. Additionally, if the regression results suggest a movie may underperform expectations, we recommend that the studio involved review the reasons why (i.e. month) and determine if the potential downside is worth changing the release date.

Specification:

The movie industry is an economic juggernaut, in 2018 the industry earned $96.8 billion (MPAA). With this amount of money being generated, much thought goes into what movie is produced and when it is released. Additionally, there has been an industry wide move toward producing films in franchises in order to build up a reliable audience base. In 2018, nine of the top ten best performing movies were franchise films (MPAA). However, sometimes these franchise films fail to meet analysts' box office expectations and can result in significant losses for media companies. Unfortunately, being a franchise film doesn't guarantee that a movie will do well. Two franchises' most recent films have underperformed expectations: Solo: A Star Wars Story significantly underperformed expectations, the film made $603.7 million at the box office compared to a budget of $275 million. This was significantly less than the previous Star Wars Story which grossed over $1 billion against a $200 million budget (Box Office Mojo). this failure resulted Disney reworking their entire movie release plan (Gramuglia).

The latest X-Men film, Dark Phoenix also significantly underperformed expectations. Although it is still in theaters, it has only generated box office results of $143 Million in its first two weeks with an estimated $200 million budget.

With the amount of money being invested in franchise films every year it is worthwhile to ask can movie attributes, previous box office results, early reviews and movie discussion predict upcoming franchise film's box office results? With the reviews we scraped/collected, what can we do to find the association between movies and review?

This paper will look at the results from three of the major active franchises: The DC Extended Universe (DCEU), Marvel Cinematic Universe (MUC) and Star Wars Universe. The X-Men Universe has been excluded from this analysis due to the recent purchase of the franchise by Disney and the uncertainty with the future of the property (The Wrap, 2019)

Observation

About the Data:

Data was drawn from a number of sources for this analysis. The table of films included were generated from Wikipedia's pages on the different film universes. The data captured from these tables were: Film name and release date. A list of potential 'hashtags' to scrapeTwitter was generated based on the film’s name. Best-Hashtags.com was also referenced for additional hastags.

A second data frame was created using the list of hashtags to scrape tweets relating to each movie. Tweepy was used in conjunction to create a data frame for each individual movie. These data frames were then combined using Alteryx 2019 for further processing in Python.

A third data frame was created using Beautiful Soup to scrape critical reviews from Rotten Tomatoes, a well known movie review aggregator. Again, these were combined using Alteryx 2019.

A fourth data frame was scrapped from StatCrunch.com that included the budget and box office results for movies released since 1915.

A review of the combined dataset showed that average budget per movie is fairly consistent during the period of our review (Figure 1). Generally, every year there is one movie with a budget that is an outlier.The average budget per movie is approximately $250 Million.

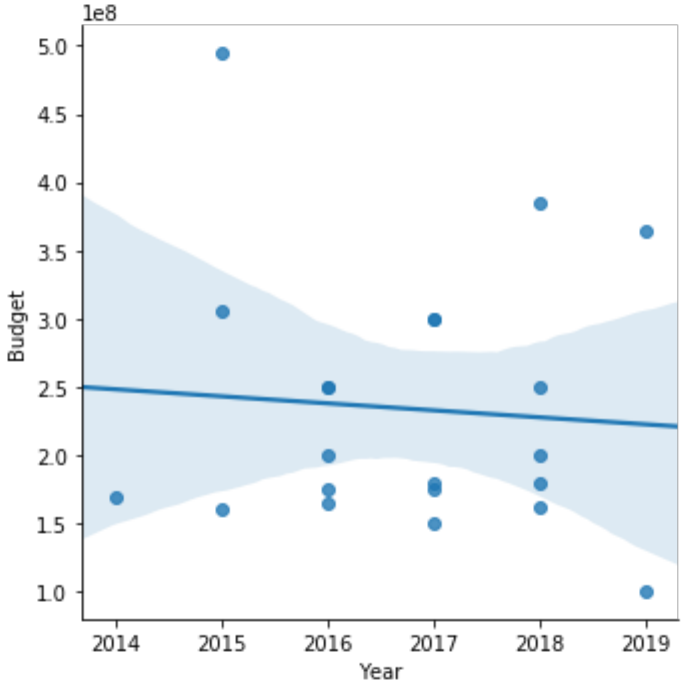


Figure 1: Comparison of Budgets over time. Generally there is one movie a year with a very high budget.

The domestic gross has been fairly consistent over time, averaging around $400 million per movie as shown in Figure 2.

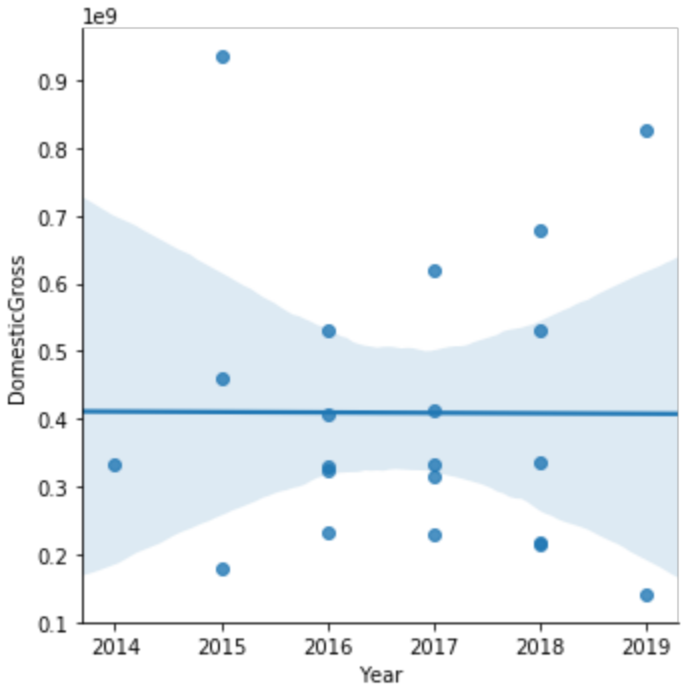


Figure 2: Comparison of domestic gross over time.

While the domestic gross has been consistent, the worldwide gross is increasing every year, This has been slightly skewed by two movies that have performed exceedingly well in the last two years, Avengers End Game and Avengers Infinity War.

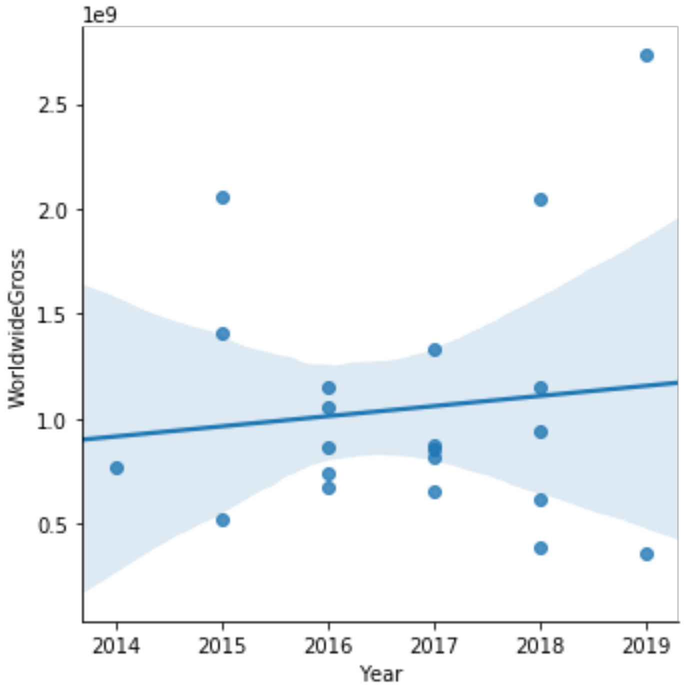
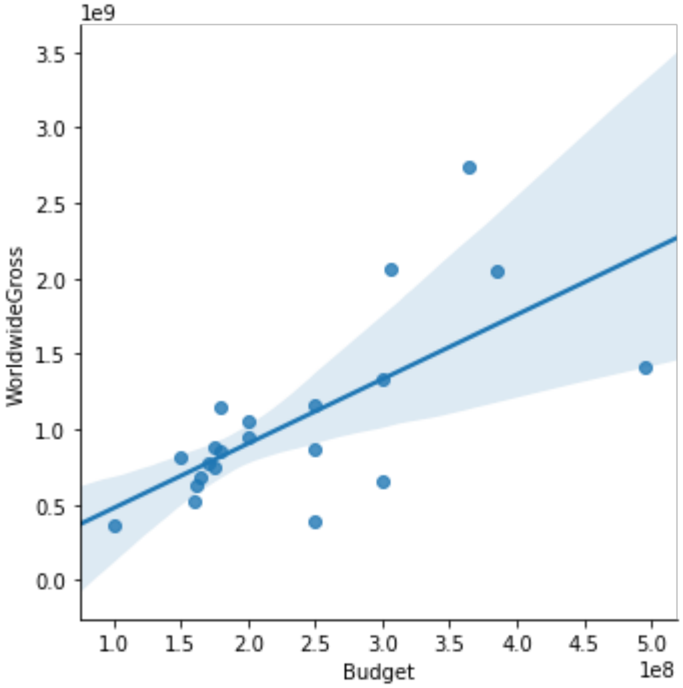


Figure 3: Worldwide Gross has been increasing over time.

Increasing a films budget does not directly tie to a higher worldwide gross. The best performing film (Infinity War) had the third highest budget. The film with the highest budget, Avengers: Age of Ultron, significantly underperformed what would be expected with a linear relationship, only grossing $1.4 billion at the box office, compared to the $2.7 million for Infinity War with a $385 Million budget as shown in Figure 4.



.

Figure 4: Comparison of Budgets to Worldwide Gross

A more positive polarity of Rotten Tomato reviews was not correlated to an increased worldwide gross as shown in Figure 5. Interestingly, a lower budget was tied to a higher Rotten Tomato polarity, suggesting that critics are not fans of ensemble movies with higher budgets.

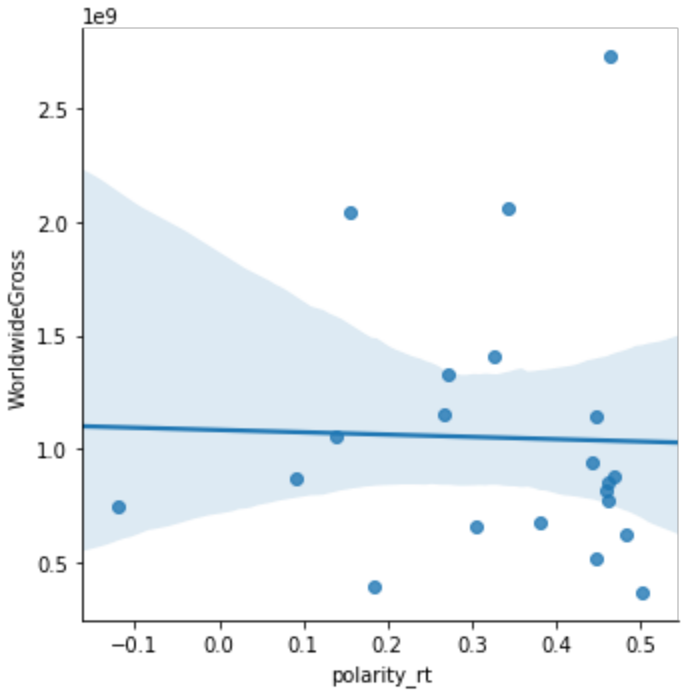


Figure 5: Comparison of Rotten Tomato Polarity to Worldwide gross.

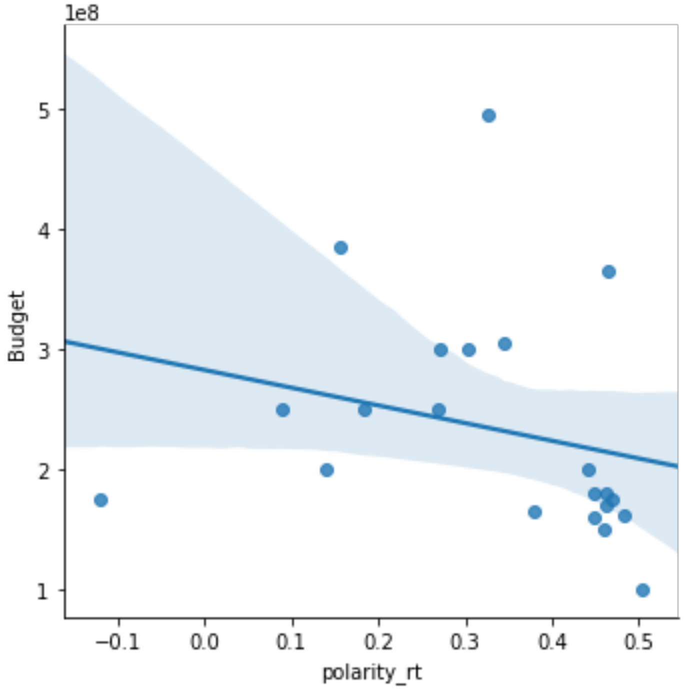


Figure 6: Comparison of Rotten Tomato Polarity to Budget.

Positive tweet polarity was correlated to a lower worldwide gross as shown in Figure 7. This suggests that dedicated fans who are likely to tweet about a movie aren't the key drivers of attendees and therefore box office results. It is hypothesized that for a movie to perform well it must connect with a wider audience. Looking at budget compared to tweet polarity, there is a slightly positive relationship, however the two movies with the highest budget was around the average tweet polarity as shown in Figure 8.

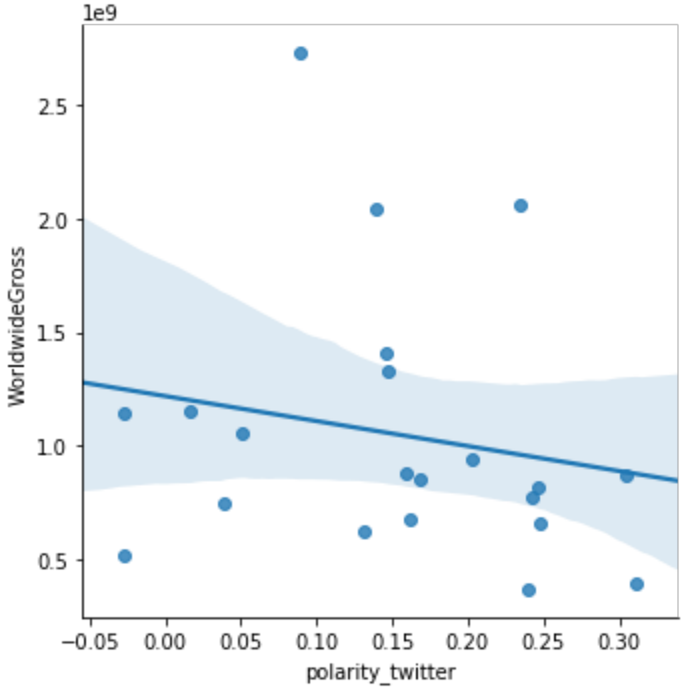


Figure 7: Comparison of tweet polarity and worldwide gross.

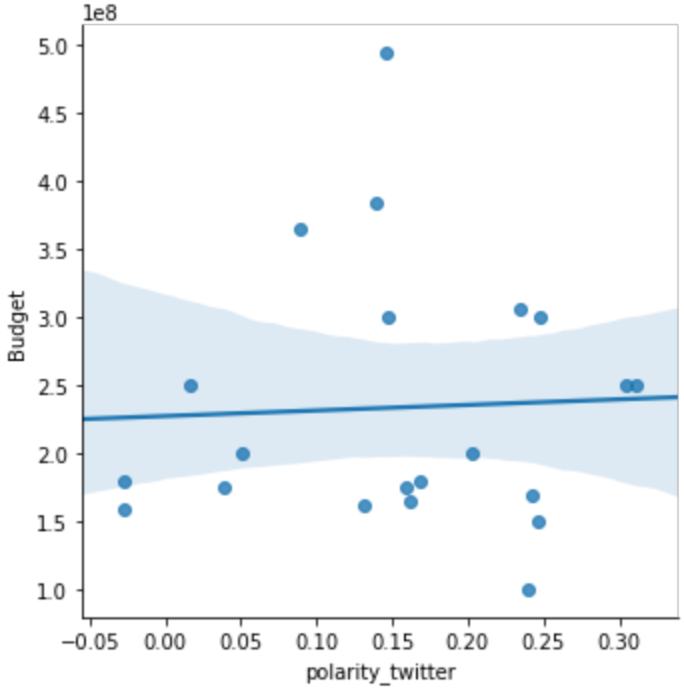


Figure 8: Comparison of tweet polarity and budget.

The average film budget was around $200 Million, with only three films budgeting more than $300 million as shown in Figure 9.

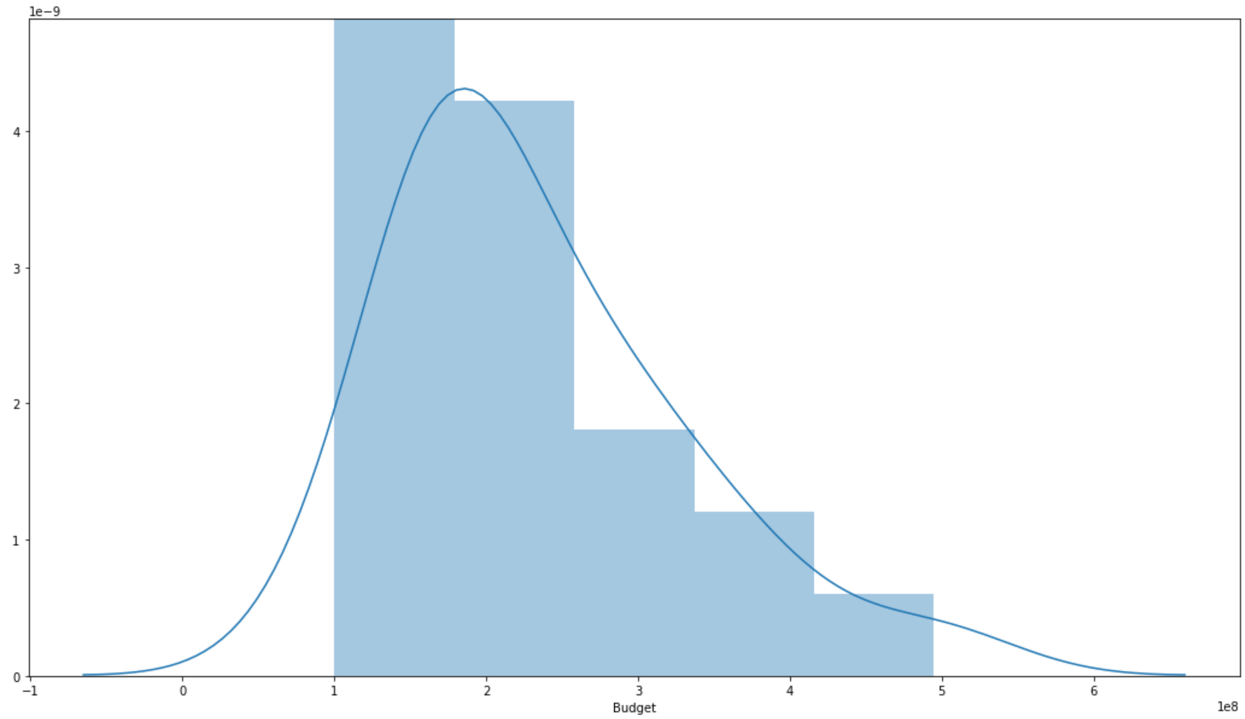


Figure 9: Distribution of Film Budgets.

Looking at budgets on a per universe basis, The MCU has had the top two highest budget movies which correlated to a higher worldwide gross, followed by Star Wars. The DCEU movies are generally cheaper to release but also have some of the lowest worldwide grosses as shown in Figure 10.

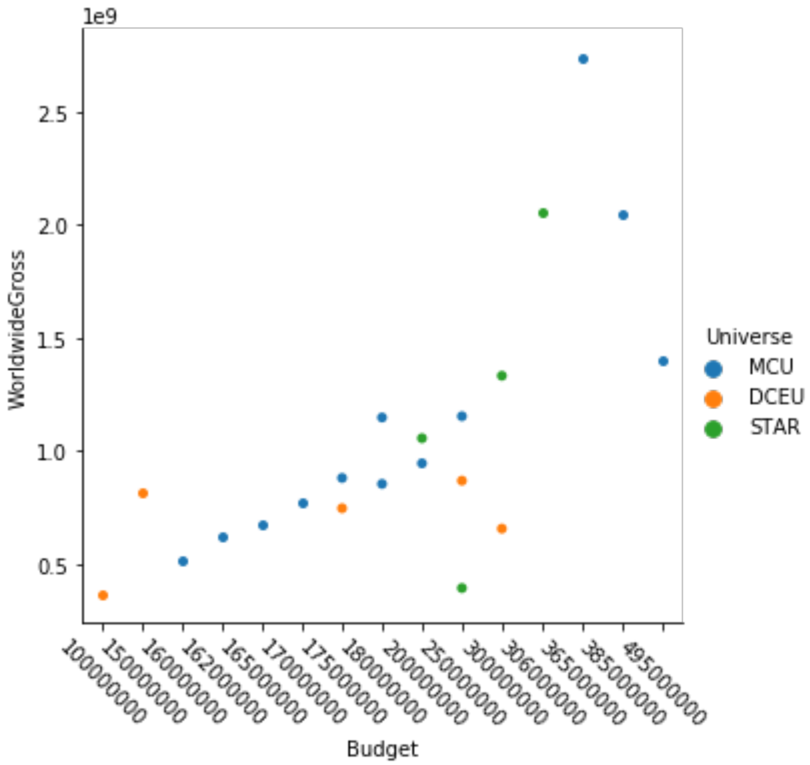


Figure 10: Universe Budgets compared to worldwide gross.

Most of the franchise films gross less than $1 Billion, however it is not uncommon for these to gross more than that, with the highest grossing films collecting over $2.5 billion as shown in Figure 11.

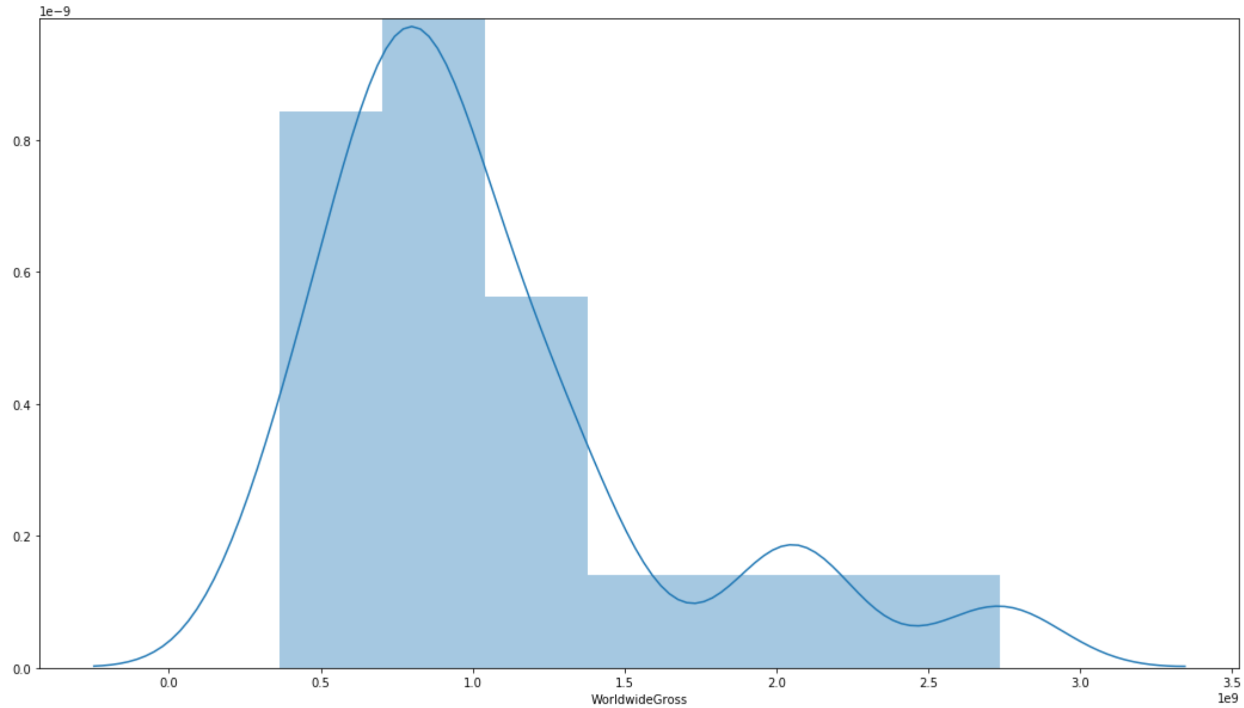


Figure 11: Distribution of Worldwide Gross.

Analysis

Data Analysis:

For each tweet and Rotten Tomato results, a sentiment score was calculated using VADER (Valence Aware Dictionary and sEntiment Reasoner). VADER is a "Lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media" (GitHub). VADER was chosen due to its high performance on social media type text and the ability to run without a training data set.

Based on this, an average sentiment of the reviews and average sentiment of the tweets were created. The overall count of sentiments was reviewed using histograms (Figures 12 and 13). The Rotten Tomato reviews were generally positive, however a large number of reviews were void of sentiment (sentiment = 0). The tweets were also generally positive, however less so than the Rotten Tomato reviews.

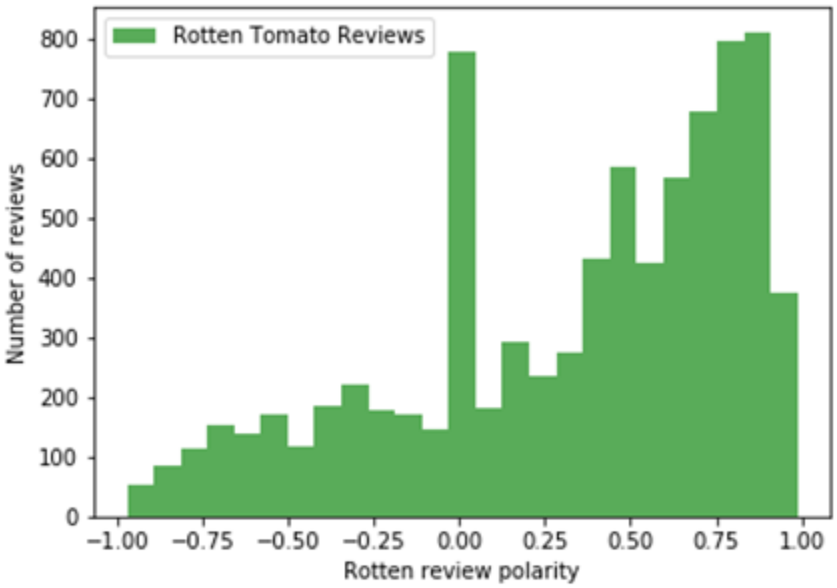


Figure 12: Rotten Tomato Sentiment Analysis. -1 is considered very negative and 1 is considered very positive. Most reviews were positive.

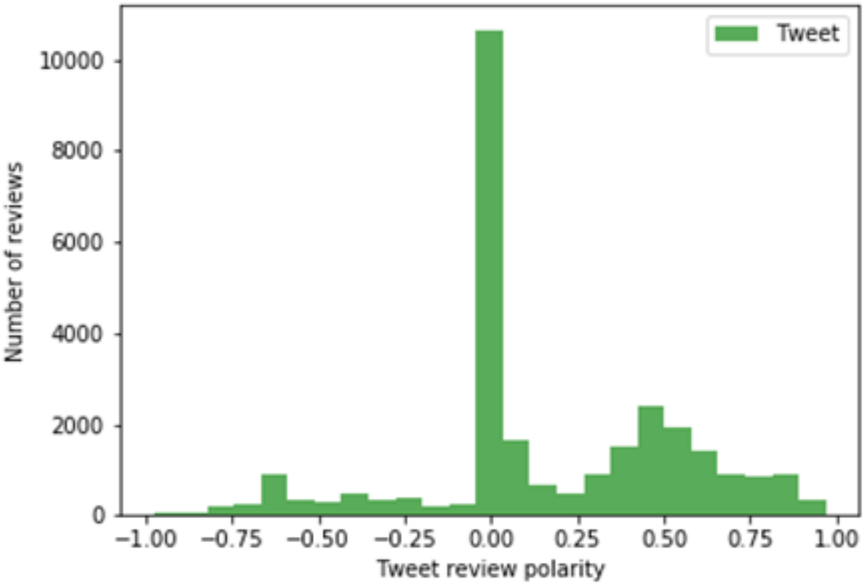


Figure 13: Tweets Sentiment Analysis. -1 is considered very negative and 1 is considered very positive.

The generally greater positivity of the Rotten Tomato Reviews is shown in the comparison in Figure 14.

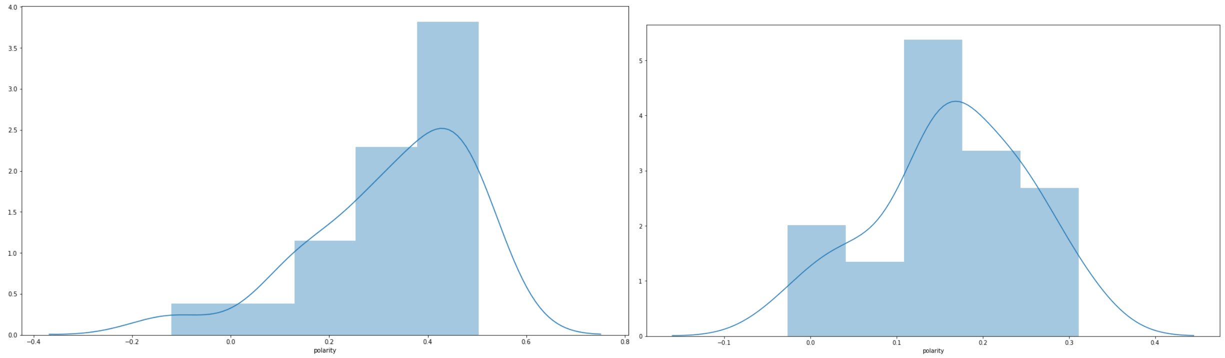


Figure 14: Rotten Tomato reviews (left) were generally more positive than the tweets (right).

All of the data frames were then combined into a master data frame to build the final model.

We conducted some additional analysis on the Rotten Tomato review. Below shows some random reviews with the highest positive sentiment polarity and the most neutral sentiment(zero) polarity.

5 random reviews with the highest positive sentiment polarity:

a. Marvel does it again in this delightful, stand-alone sequel.
b. Ranks, on every front, as one of the best from Marvel's universe!
c. Shazam! just refused to take any risks and put all its eggs in the family and comedy basket. The performances by all the leads are the best part of the film.
d. With a self-contained storyline and plenty of laughs, Ant-Man and The Wasp is the perfect follow-up to the chaos of Infinity War.
e. Works best when it's just letting this surrogate family bond over their mutual guilt and fears.

5 random reviews with the most neutral sentiment(zero) polarity:

a. We won't talk about the technobabble inelegantly cited in a bid to hold it all together - the equivalent of taking out an airplane's jackscrews and replacing them with chewing gum.
b. Chris Hemsworth walks off with the 3-hour cameo extravaganza. The talking rodent comes in second. Doesn't anybody deserve to die?
c. The forceful 'Star Wars: The Last Jedi' is a cinematic triumph.
d. Everything here is a goof.
e. It's "How I Met Your Wookiee"!

We assume that when the sentiment polarity is greater than 0, the movie is recommended. If the sentiment polarity is smaller than 0, the movie is not recommended.

Example 1:

Review context	Fans of the franchise will be pleased, but those looking in from the outside of comic-book culture may find themselves also looking at their watches.
Sentiment Polarity	0.2382
Recommend Flag	1

Example 2:

Review context	"Avengers: Age of Ultron" is a sometimes daffy, occasionally baffling, surprisingly touching and even romantic adventure with one kinetic thrill after another. It earns a place of high ranking in the Marvel Universe.
Sentiment Polarity	0.8813
Recommend Flag	1

Example 3:

Review context	The sharp, interpersonal dramedy that made the first movie such a delight is again present in flashes, but not infrequently it is drowned out by the noisy, inevitable need to Save the World.
Sentiment Polarity	-0.1655
Recommend Flag	0

Figure 15 shows, based on our assumption, how many reviews recommended the movies and how many reviews didn't recommend the movies.

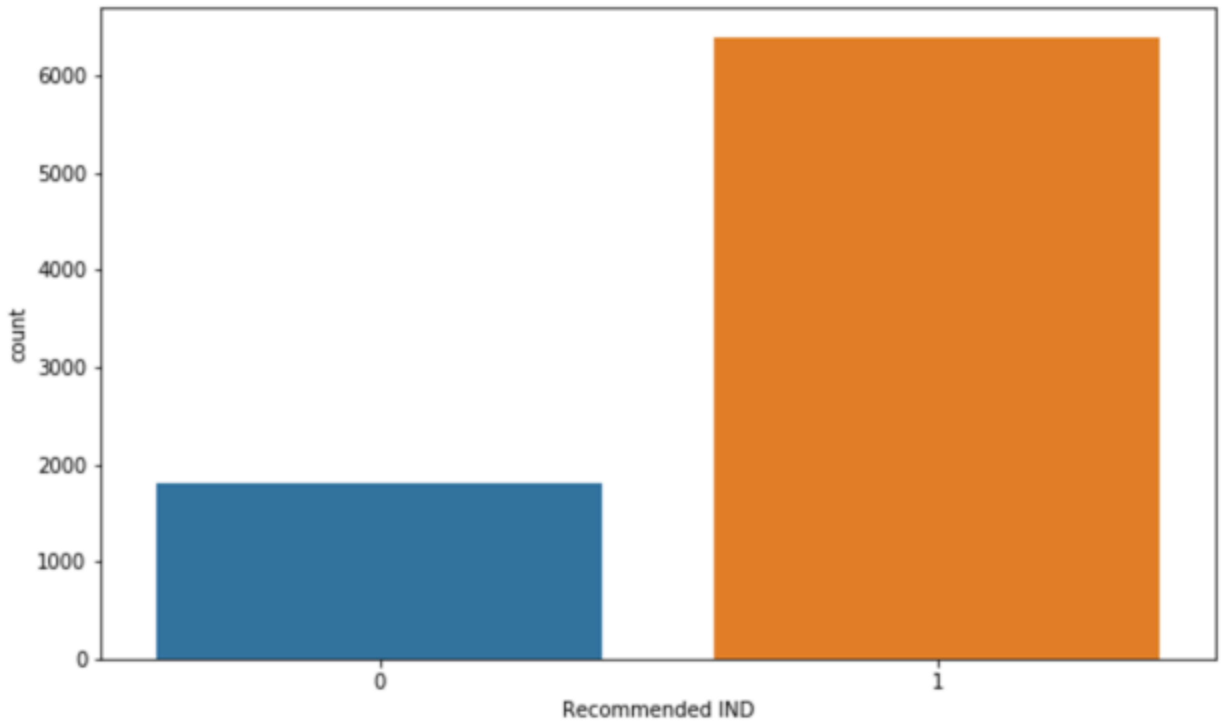


Figure 15: count of recommendations.

Figure 16 shows the distribution of the word count in reviews. The curve is very close to bell shape and it suggested a normal distribution.

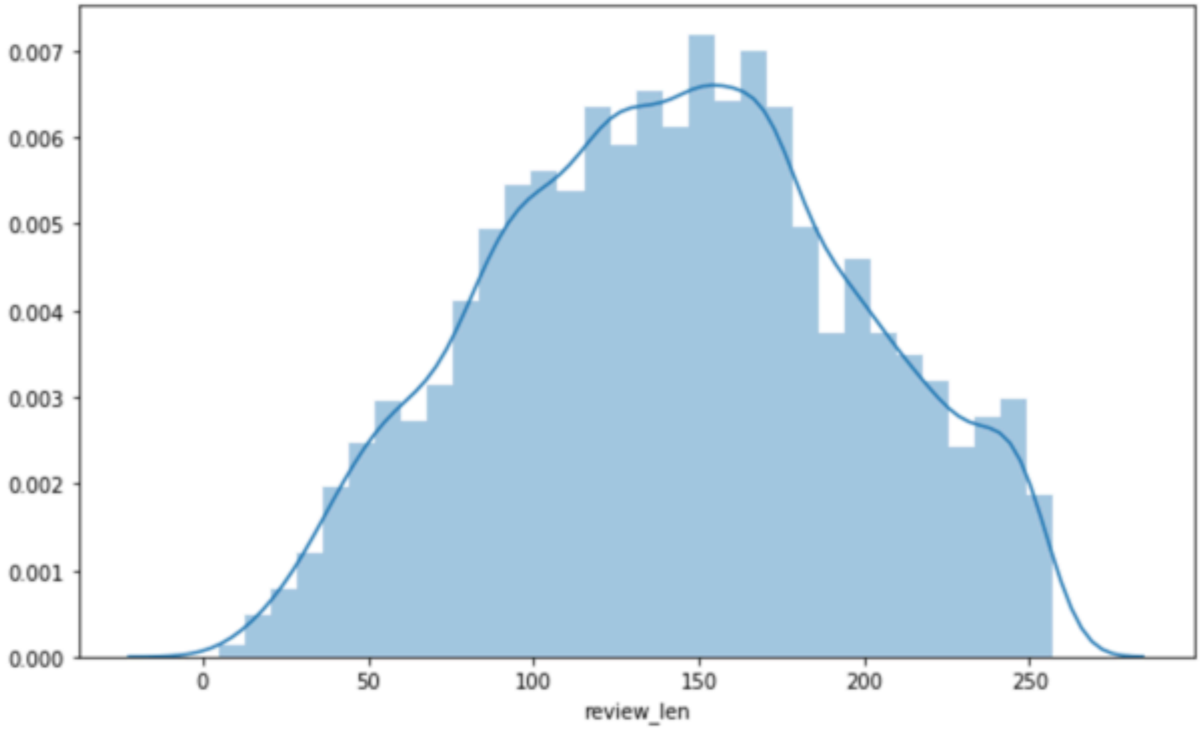


Figure 16: A distribution plot showing the word counts in each review and the polarities.

We explored the characteristic terms and their associations to each movie.

Example 1
Associated movie: AgeofUltron
Characteristic terms:

```
['ultron is',
 'whedon',
 'ultron',
 'of ultron',
 'joss',
 'director joss',
 'joss whedon',
 'age of',
 "whedon 's",
 'avengers age']
```

Example 2
Associated movie: DoctorStrange
Characteristic terms:

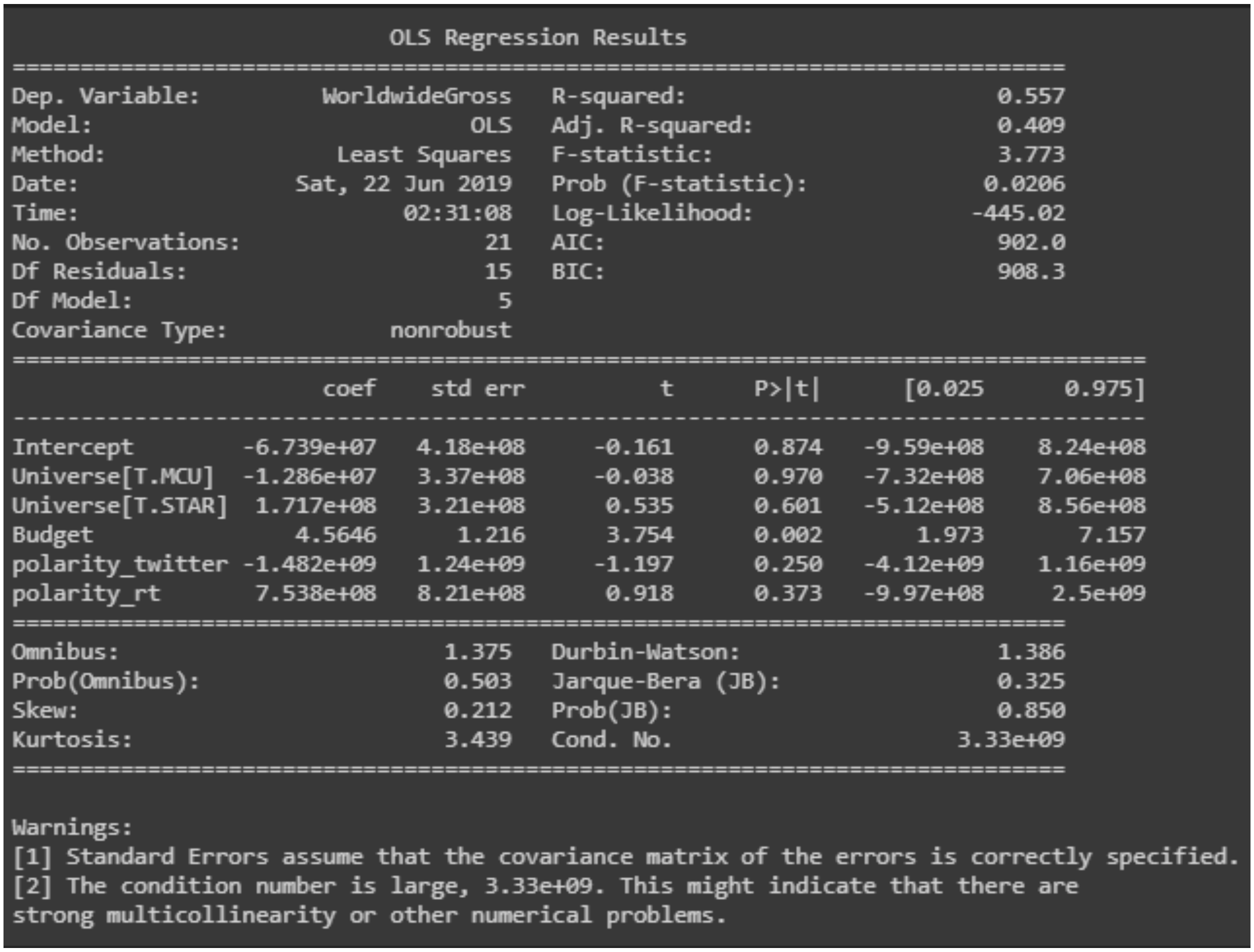
```
['strange is',
 'doctor strange',
 'cumberbatch',
 'doctor',
 'strange',
 'derrickson',
 'bending',
 'benedict',
 'mind bending',
 'benedict cumberbatch']
```

Example 3
Associated movie: SuicideSquad
Characteristic terms:

```
['ayer',
 'robbie',
 'margot',
 'margot robbie',
 'david ayer',
 'squad is',
 'harley',
 'quinn',
 'suicide squad',
 'suicide']
```

Category 1- Predicting Box Office Results

To answer the question posed, [Can we predict the results of the next francaise film (Spiderman Homecoming) based on hype?] we constructed an ordinary least squares linear regression model whose results are shown in Figure 17. The predictor variables were Universe, Budget, and the two polarities. This model was able to predict 40% of the Worldwide Gross while being significantly significant with a Probability F-Statistic of 0.02.

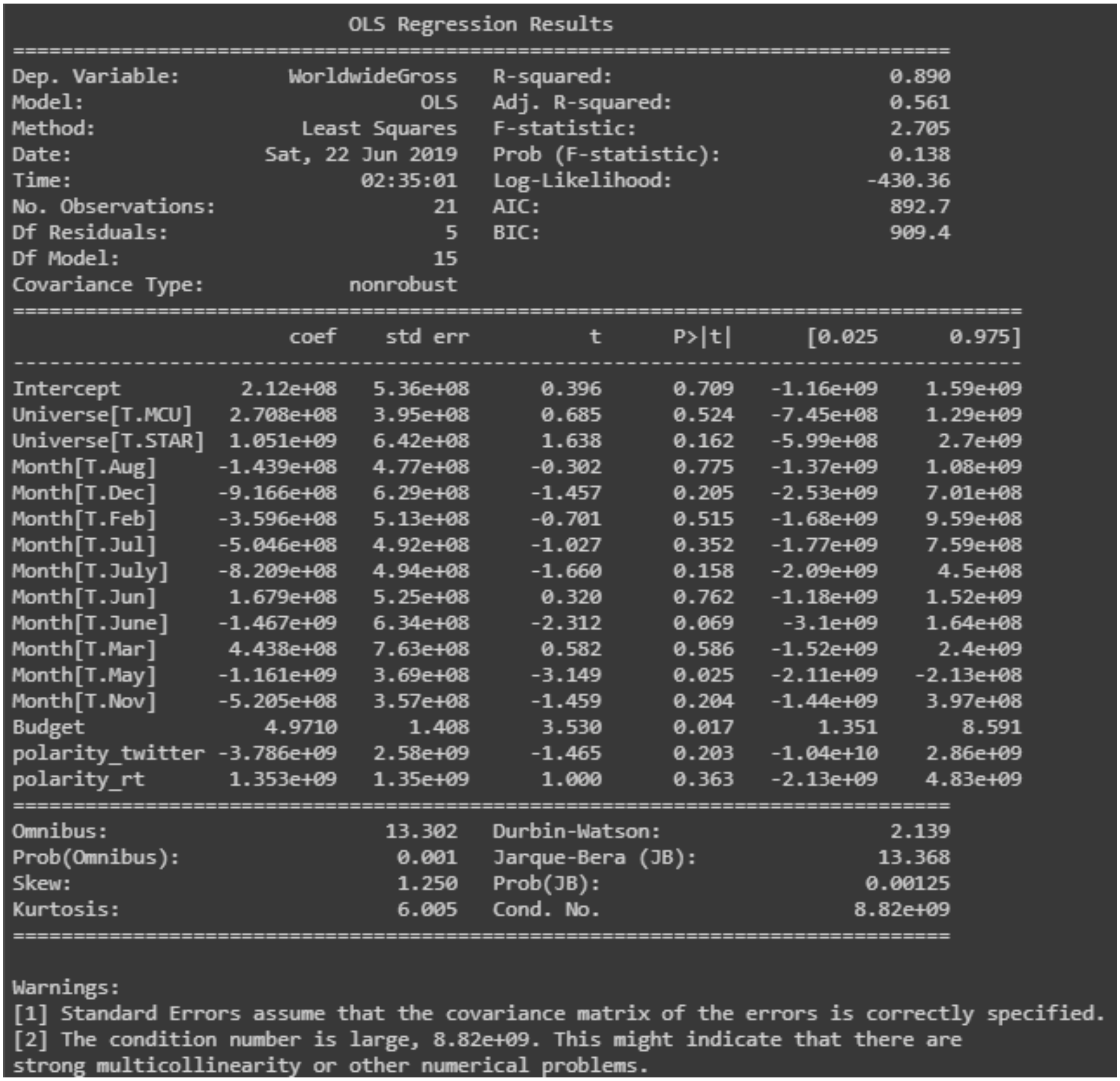


```
                            OLS Regression Results
==============================================================================
Dep. Variable:          WorldwideGross   R-squared:                       0.557
Model:                             OLS   Adj. R-squared:                  0.409
Method:                  Least Squares   F-statistic:                     3.773
Date:                 Sat, 22 Jun 2019   Prob (F-statistic):             0.0206
Time:                         02:31:08   Log-Likelihood:                -445.02
No. Observations:                   21   AIC:                             902.0
Df Residuals:                       15   BIC:                             908.3
Df Model:                            5
Covariance Type:             nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept        -6.739e+07   4.18e+08     -0.161      0.874   -9.59e+08    8.24e+08
Universe[T.MCU]  -1.286e+07   3.37e+08     -0.038      0.970   -7.32e+08    7.06e+08
Universe[T.STAR]  1.717e+08   3.21e+08      0.535      0.601   -5.12e+08    8.56e+08
Budget               4.5646      1.216      3.754      0.002       1.973       7.157
polarity_twitter -1.482e+09   1.24e+09     -1.197      0.250   -4.12e+09    1.16e+09
polarity_rt       7.538e+08   8.21e+08      0.918      0.373   -9.97e+08     2.5e+09
==============================================================================
Omnibus:                        1.375   Durbin-Watson:                   1.386
Prob(Omnibus):                  0.503   Jarque-Bera (JB):                0.325
Skew:                           0.212   Prob(JB):                        0.850
Kurtosis:                       3.439   Cond. No.                     3.33e+09
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.33e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 17: Initial OLS Regression Model.

A second regression including month was run. This increased the predictive power to 56%. Unfortunately, our probability F-statistic was no longer statistically significant using an alpha of 0.05, it instead was 0.138 as shown in Figure 18. This would be unacceptable in most circumstances, however we accepted it for the purposes of this paper due to the limited number of movies we were reviewing.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          WorldwideGross   R-squared:                       0.890
Model:                             OLS   Adj. R-squared:                  0.561
Method:                  Least Squares   F-statistic:                     2.705
Date:                 Sat, 22 Jun 2019   Prob (F-statistic):              0.138
Time:                         02:35:01   Log-Likelihood:                -430.36
No. Observations:                   21   AIC:                             892.7
Df Residuals:                        5   BIC:                             909.4
Df Model:                           15
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept           2.12e+08   5.36e+08      0.396      0.709   -1.16e+09    1.59e+09
Universe[T.MCU]    2.708e+08   3.95e+08      0.685      0.524   -7.45e+08    1.29e+09
Universe[T.STAR]   1.051e+09   6.42e+08      1.638      0.162   -5.99e+08     2.7e+09
Month[T.Aug]      -1.439e+08   4.77e+08     -0.302      0.775   -1.37e+09    1.08e+09
Month[T.Dec]      -9.166e+08   6.29e+08     -1.457      0.205   -2.53e+09    7.01e+08
Month[T.Feb]      -3.596e+08   5.13e+08     -0.701      0.515   -1.68e+09    9.59e+08
Month[T.Jul]      -5.046e+08   4.92e+08     -1.027      0.352   -1.77e+09    7.59e+08
Month[T.July]     -8.209e+08   4.94e+08     -1.660      0.158   -2.09e+09     4.5e+08
Month[T.Jun]       1.679e+08   5.25e+08      0.320      0.762   -1.18e+09    1.52e+09
Month[T.June]     -1.467e+09   6.34e+08     -2.312      0.069    -3.1e+09    1.64e+08
Month[T.Mar]       4.438e+08   7.63e+08      0.582      0.586   -1.52e+09     2.4e+09
Month[T.May]      -1.161e+09   3.69e+08     -3.149      0.025   -2.11e+09   -2.13e+08
Month[T.Nov]      -5.205e+08   3.57e+08     -1.459      0.204   -1.44e+09    3.97e+08
Budget                4.9710      1.408      3.530      0.017       1.351       8.591
polarity_twitter -3.786e+09   2.58e+09     -1.465      0.203   -1.04e+10    2.86e+09
polarity_rt        1.353e+09   1.35e+09      1.000      0.363   -2.13e+09    4.83e+09
==============================================================================
Omnibus:                       13.302   Durbin-Watson:                   2.139
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               13.368
Skew:                           1.250   Prob(JB):                      0.00125
Kurtosis:                       6.005   Cond. No.                     8.82e+09
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.82e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 18: Regression results including month. This boosted in adjusted R-squared at the cost of decreasing the probability f-statistic.

Using this model, the reported budget for Spiderman Homecoming, July as the month of release, an a current average tweet polarity of 0.159506, being a film in the MCU, we then assumed the average critic sentiment would be the average of all critic sentiments to date (there weren’t any reviews as of the date this paper was written). Based on the above, we predicted the film will gross $611,549,260 at the worldwide box office.

Category 2 - Predicting the box office result using Prophet.

We utilized the Prophet time series prediction model to run predictions for each universe. However, the result was quite disappointing. We believe the amount of data we collected and broken down into 3 universes diluted the data points for the prediction model.

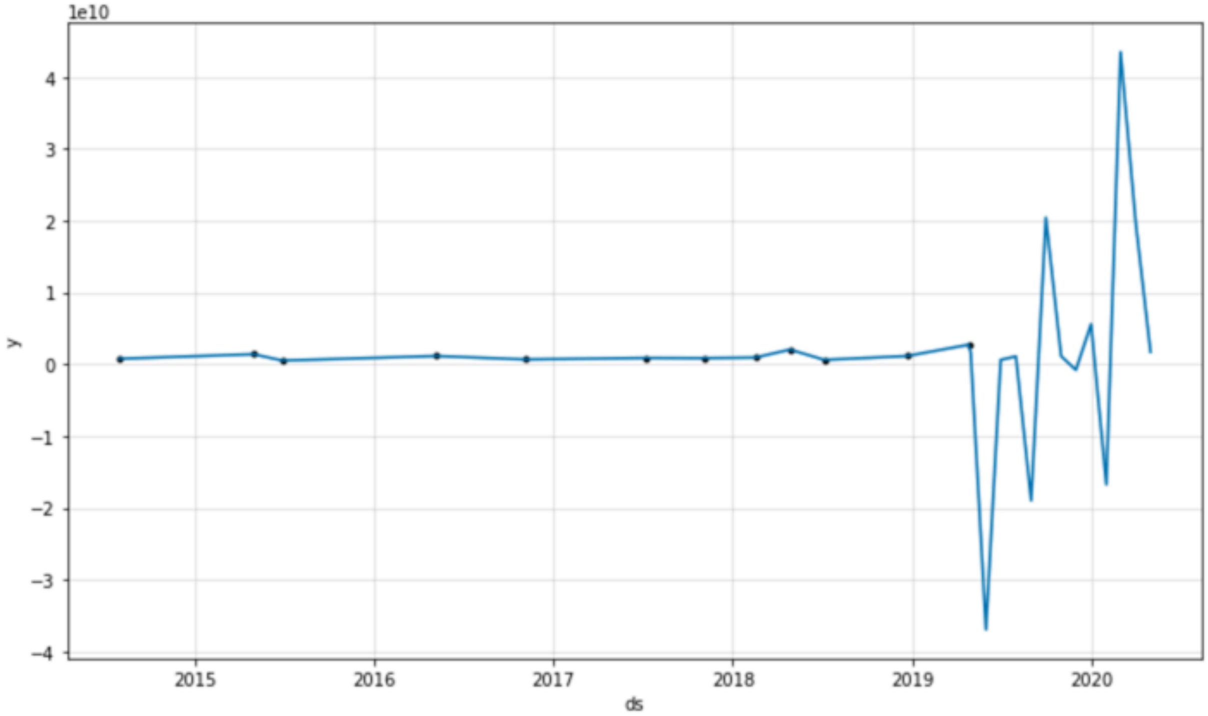


Figure 18: Time series prediction with 13-month predictions for MCU

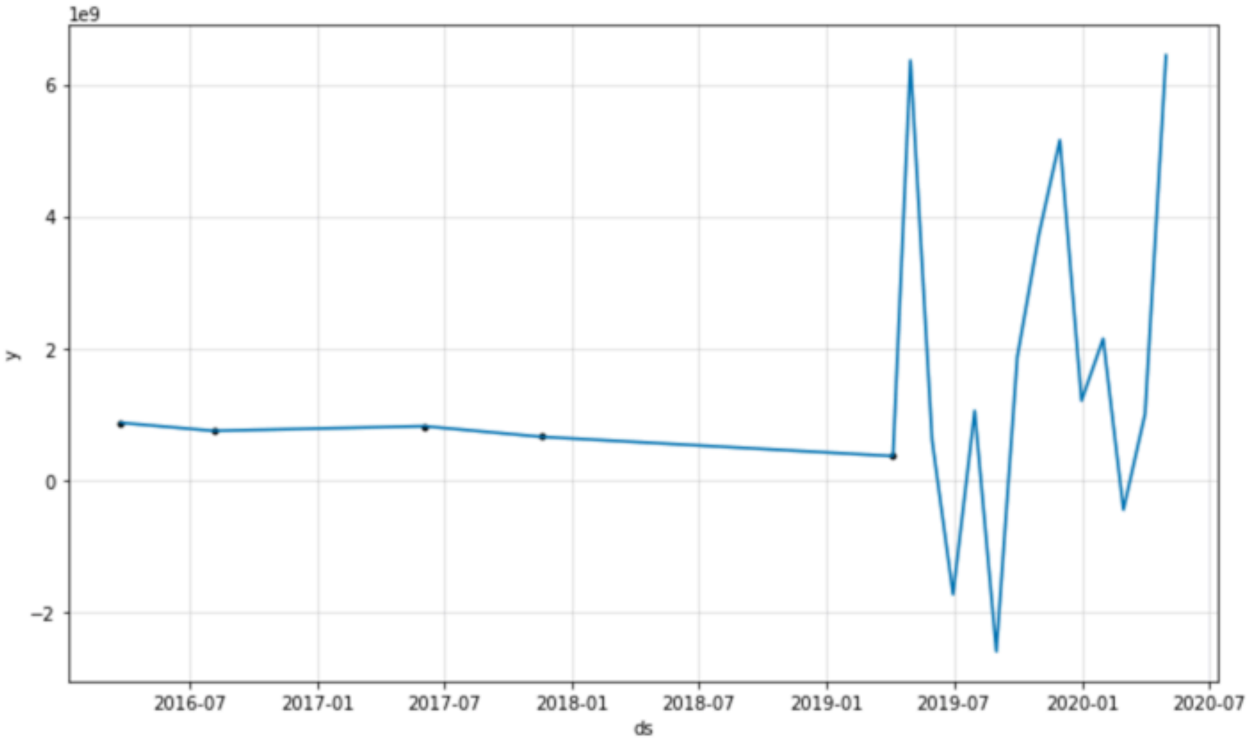


Figure 19: Time series prediction with 13-month predictions for DCEU

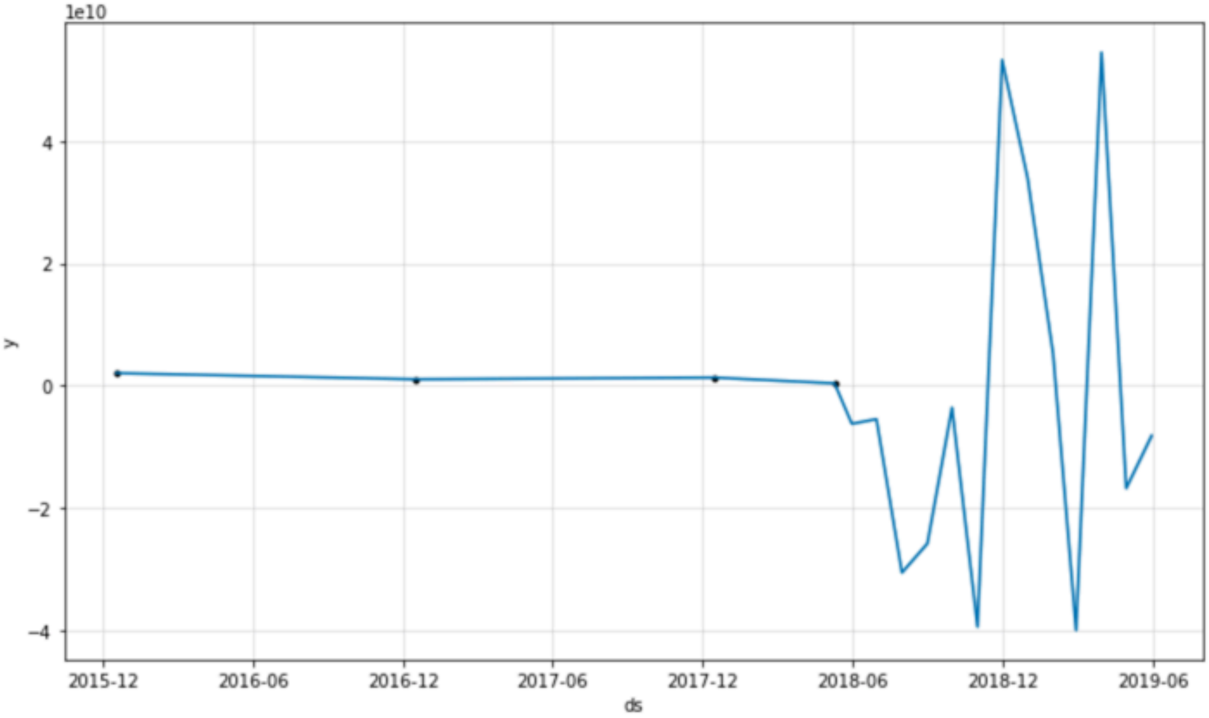


Figure 20: Time series prediction with 13-month predictions for STAR

Category 3 - Multi-Class Text Classification with Scikit-Learn

We utilized Multi-Class Text Classification concepts and with Scikit-Learn models, we build a prediction model predicting movie categories based on the movie reviews.

Figure 21 shows the text counts form reviews we scraped for each movie. We can see Shazam has the most text count and Ant-Man and the Wasp has the least.

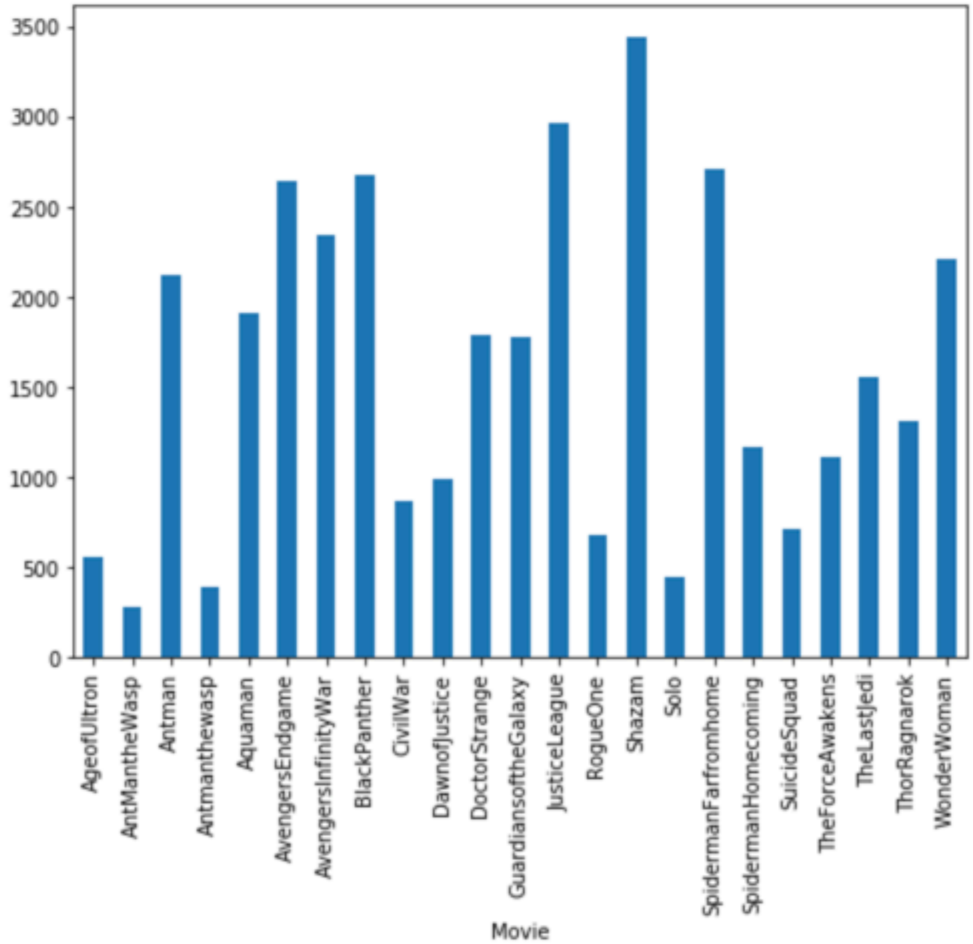
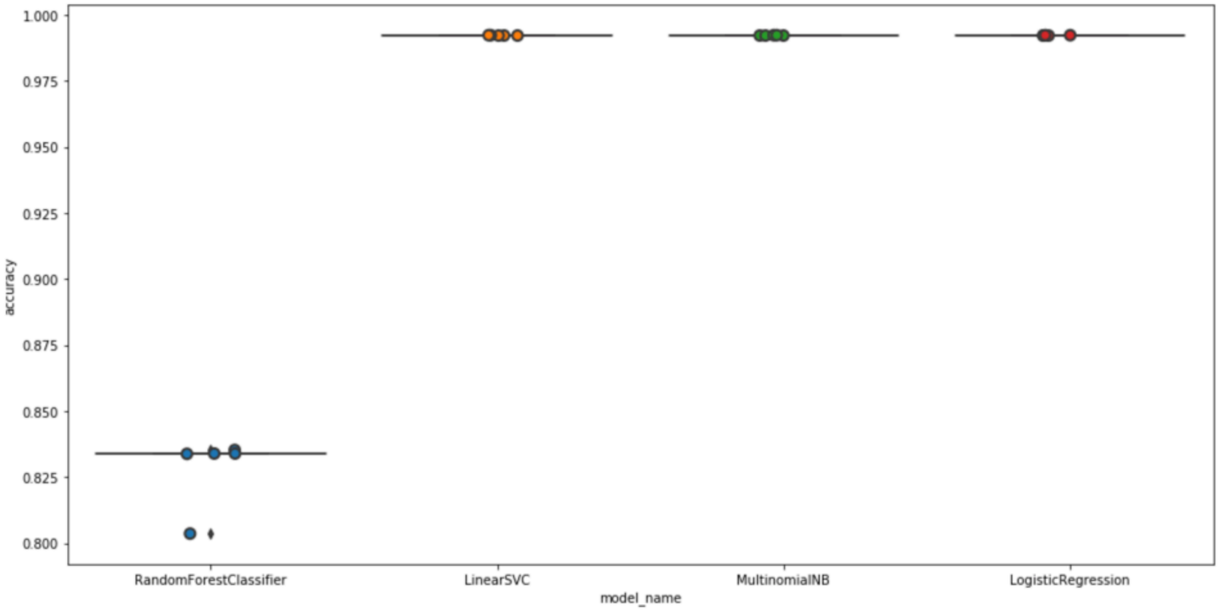
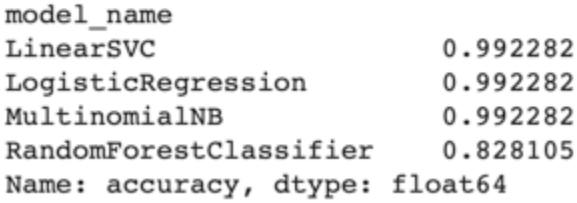


Figure 21: Text counts from reviews for each movie

Figure 22 shows the prediction accuracies from each prediction method. We got the same accuracy score (0.992282) for Linear SVC, Logistic Regression, and Multinomial NB. Random Forest Classifier has a lower accuracy score (0.828105.)





```
model_name
LinearSVC                 0.992282
LogisticRegression        0.992282
MultinomialNB             0.992282
RandomForestClassifier    0.828105
Name: accuracy, dtype: float64
```

Figure 22: The prediction accuracies from each prediction method.

When we randomly selected some review and ran through the prediction model, we organized the result as below.

Reviews	Predicted Movie	Comment
It's clear that [Alden] Ehrenreich is from having the charisma of the young Harrison Ford, but he manages.	Solo	Correct
Ayer has often been a careless writer, and so he is here..	TheLastJedi	Incorrect - it should be SuicideSquad
The fun of The Force Awakens comes from visiting an old friend and finding that little has changed; the	TheForceAwakens	Correct

excitement is that great performances from all four new leads means it's a strong foundation for whatever comes next.		

Recommendation

There were a number of compromises made in this analysis. The initial compromise being the number of movies included in the analysis. With 21 movies, there was often only one movie released in a given month. Additionally, we only looked at three of the many franchises. This was done because these franchises were among the most prolific producers, however, adding in additional franchises could have increased the predictive power of the regression while improving the statistical significance. Adding in additional variables, including information about the cast, the overall economy, number of similar films at the box office at the time of release, week to week change in results, marking if a film is a significant film in the franchise (first film in an extended period or end of a film era etc.), could have provided additional predictive power.

Accepting these compromises, the results suggest that if you want a high grossing film, you should release a Star Wars film in March and attempt to generate negative tweets but positive critic reviews. Additionally, a higher budget is predicted to increase the gross.

In conclusion, our models weren't great, however the correlation of sentiment to budget and worldwide gross provided an intriguing insight that hype has a very slight correlation to gross where positive critic reviews increase the gross but positive tweets decrease the gross. The multi-class prediction was interesting and can be utilized during the movie releases. The marketing area could utilize this sort of technique to adjust the marketing plan during the campaign. We recommend that if a proposed film is likely to underperform expectations, the studio reviews the regression results and determines what variable(s) is/are causing the decreased results and determine if it is something they can change (ex. release month) or if they are willing to accept the potential risk. Additionally, this analysis should be updated with each new release to increase the predictive power.

References:

https://www.boxofficemojo.com/

https://www.thewrap.com/will-there-be-any-more-x-men-movies-after-dark-phoenix/

MPAA 2018 Theme Report: https://www.mpaa.org/wp-content/uploads/2019/03/MPAA-THEME-Report-2018.pdf

GitHub: https://github.com/cjhutto/vaderSentiment

Gramuglia, Anthony. (2018). *CBR*., Disney's Star Wars Plans Prove It Learned From Solo's Blunder. https://www.cbr.com/disney-star-wars-learned-from-solo/

https://www.statcrunch.com