CIS 361 Project

Allie Seavers

5/1/2020

## Summary

Movies seems to come out almost every month. With this vast range of a selection, most people like to give an opinion about a certain one. Along with their own opinion, people to like to what others think. Many people are shocked to see the ratings given by different organizations. These ratings can at times be drastically different from each other. This makes me wonder which source of these ratings gives an accurate representation of how the movie did in terms of revenue. Most researchers spend time either trying to give recommendations to a user based on movies they like or predicting a movie’s potential rating. There really isn’t a way to show which organization gives a better representation for how the public feels about a movie compared to the revenue made.

This project will look through three different rating sources and compare values to see which one is a better indicator. The three are IMdB, Rotten Tomatoes Critic, and Rotten Tomatoes Audience. My theory is that IMdB will be the better indicator for the revenue reflecting by the ratings, as it does not have any strict guidelines to leave a review.

After some analysis, it shocked me to see that Rotten Tomatoes actually fit the model best with the data I had collected, proving my theory wrong.

## Literature Review

After doing some research thanks to Google Scholar, there was not a lot of research done about the particular piece I want to focus on. Most of the articles were about how to predict a movie’s rating or a movie’s revenue.

I was able to find one similar. In the article, “Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction”, discussed various relationships that can bring a movie’s revenue up. One of the relationships was looking at ratings these movies held. They looked into different types of users and how a user may change their ratings over time.

I want to go into something a bit simpler which can be used in the future. Which organization that gives back public ratings can be used to indicate which one is more accurate with the revenue received. Which brings me to a research question: Which organization has a more accurate representation of how well a movie is and did.

I will be comparing with three different popular organizations: IMdB data, Rotten Tomatoes Critics data, and Rotten Tomatoes Audience data.

IMdB users can be consisted of anyone who creates an account on the website. These users can give any rating to any movie they desire.

Rotten Tomatoes Critics have more requirements to have their review accepted. For example, one of the requirements to even be eligible for an individual are “Consistent output for a minimum of two years. Demonstrated film/TV coverage at a publication outside of a self-published website. However, we will consider critics who solely self-publish if their site and work reflect our key values.” There are a lot more requirements, but as you can see these users have high knowledge of how to rate a movie.

Rotten Tomatoes Audience consist of ratings of users who have been verified to have bought a ticket for the movie.

## Theory

After investigating each of the organization, I believe that IMdB’s users will have a better representation of reflecting a movie’s revenue.

## Data

Now to begin, I needed to find some data to work with. I found a dataset that contained ratings from the three I am focusing on and more. It also contained the movie titles from the years 2016-2017. This data set provided me with plenty of data for me to do some calculations with. But I was missing the revenue for these movies. After some more digging, I found another data set that contained these. This dataset had a wider range of movies over different years.

library("readxl");  
  
movieRevenueTable <-read\_excel("AllMoviesDetailsCleaned.xlsx");  
movieRatingsTable <-read\_excel("movie\_ratings\_16\_17.xlsx");  
  
head(movieRevenueTable);

## # A tibble: 6 x 22  
## id budget genres imdb\_id original\_langua~ original\_title overview  
## <dbl> <dbl> <chr> <chr> <chr> <chr> <chr>   
## 1 2 0. Drama~ tt0094~ fi Ariel "Taisto~  
## 2 3 0. Drama~ tt0092~ fi Varjoja parat~ "An epi~  
## 3 5 4.00e6 Crime~ tt0113~ en Four Rooms "It's T~  
## 4 6 0. Actio~ tt0107~ en Judgment Night "While ~  
## 5 8 4.20e4 Docum~ tt0825~ en Life in Loops~ "Timo N~  
## 6 9 0. Drama tt0425~ de Sonntag im Au~ <NA>   
## # ... with 15 more variables: popularity <dbl>, production\_companies <chr>,  
## # production\_countries <chr>, release\_date <chr>, revenue <dbl>,  
## # runtime <dbl>, spoken\_languages <chr>, status <chr>, tagline <chr>,  
## # title <chr>, vote\_average <dbl>, vote\_count <dbl>,  
## # production\_companies\_number <dbl>, production\_countries\_number <dbl>,  
## # spoken\_languages\_number <dbl>

head(movieRatingsTable);

## # A tibble: 6 x 15  
## movie year metascore imdb tmeter audience fandango n\_metascore n\_imdb  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 10 C~ 2016 76 7.2 90 79 3.5 3.8 3.6   
## 2 13 H~ 2016 48 7.3 50 83 4.5 2.4 3.65  
## 3 A Cu~ 2016 47 6.6 40 47 3 2.35 3.3   
## 4 A Do~ 2017 43 5.2 33 76 4.5 2.15 2.6   
## 5 A Ho~ 2016 58 6.1 70 57 3 2.9 3.05  
## 6 A Mo~ 2016 76 7.5 87 84 4 3.8 3.75  
## # ... with 6 more variables: n\_tmeter <dbl>, n\_audience <dbl>,  
## # nr\_metascore <dbl>, nr\_imdb <dbl>, nr\_tmeter <dbl>, nr\_audience <dbl>

These two datasets contained quite a bit of information. I needed to clean the data to only show what I wanted.

library("dplyr");  
  
movieRatingsTable <-select(movieRatingsTable,movie,imdb,tmeter,audience);  
movieRevenueTable <- select(movieRevenueTable,revenue,title);

I decided to combine these two datasets into one based on movie titles. After the merge, I noticed that there were duplications over many rows of the same movie. I cleared out these duplications which made the dataset even smaller.

merged <- merge(movieRatingsTable,movieRevenueTable,by.x = "movie",by.y = "title");  
  
noDuplicates <- merged[!duplicated(merged$movie), ];  
head(noDuplicates);

## movie imdb tmeter audience revenue  
## 1 10 Cloverfield Lane 7.2 90 79 108286421  
## 2 A Cure for Wellness 6.6 40 47 24209725  
## 3 A Dog's Purpose 5.2 33 76 0  
## 4 A Hologram for the King 6.1 70 57 0  
## 5 A Monster Calls 7.5 87 84 40919166  
## 6 A Street Cat Named Bob 7.4 77 79 35070000

## Methodology

After finishing cleaning the data, it was now time to do some calculations on the ratings. After looking at each rating column, I made the assumption was calculated differently. I figured the best way to make this easier was to normalize the data into a range. I decided to scale each rating to the range of 0 to 5.

library("BBmisc");  
  
justRatings <- select(noDuplicates,imdb,tmeter,audience);  
normalizedValues<- normalize(justRatings, method = "range", range = c(0, 5));  
moviesInfo <- mutate(noDuplicates,normalizedValues$tmeter, normalizedValues$imdb,normalizedValues$audience);  
  
head(moviesInfo);

## movie imdb tmeter audience revenue  
## 1 10 Cloverfield Lane 7.2 90 79 108286421  
## 2 A Cure for Wellness 6.6 40 47 24209725  
## 3 A Dog's Purpose 5.2 33 76 0  
## 4 A Hologram for the King 6.1 70 57 0  
## 5 A Monster Calls 7.5 87 84 40919166  
## 6 A Street Cat Named Bob 7.4 77 79 35070000  
## normalizedValues$tmeter normalizedValues$imdb normalizedValues$audience  
## 1 4.545455 3.7 4.146341  
## 2 2.020202 3.1 2.195122  
## 3 1.666667 1.7 3.963415  
## 4 3.535354 2.6 2.804878  
## 5 4.393939 4.0 4.451220  
## 6 3.888889 3.9 4.146341

After this, I got rid of the old ratings and kept the normalized ratings. I also adjusted the ratings to only show one decimal point.

moviesInfo <- select(moviesInfo,movie,imdbN=`normalizedValues$imdb`,rTMCN=`normalizedValues$tmeter`,rTMUN=`normalizedValues$audience`,revenue);  
moviesInfo$rTMCN <- round(moviesInfo$rTMCN,digits=1);  
moviesInfo$rTMUN<- round(moviesInfo$rTMUN,digits=1);  
  
head(moviesInfo);

## movie imdbN rTMCN rTMUN revenue  
## 1 10 Cloverfield Lane 3.7 4.5 4.1 108286421  
## 2 A Cure for Wellness 3.1 2.0 2.2 24209725  
## 3 A Dog's Purpose 1.7 1.7 4.0 0  
## 4 A Hologram for the King 2.6 3.5 2.8 0  
## 5 A Monster Calls 4.0 4.4 4.5 40919166  
## 6 A Street Cat Named Bob 3.9 3.9 4.1 35070000

Then it was time to do some linear regression over each of these ratings against the revenues to study the relationship it provides.

library("jtools");  
movieReg <- lm(moviesInfo$revenue ~ moviesInfo$imdbN+ moviesInfo$rTMCN+ moviesInfo$rTMUN,data = moviesInfo);  
summ(movieReg);

## MODEL INFO:  
## Observations: 207  
## Dependent Variable: moviesInfo$revenue  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(3,203) = 7.68, p = 0.00  
## R² = 0.10  
## Adj. R² = 0.09   
##   
## Standard errors: OLS  
## -------------------------------------------------------------------  
## Est. S.E. t val. p  
## ---------------------- -------------- ------------- -------- ------  
## (Intercept) -35482790.33 45034476.72 -0.79 0.43  
## moviesInfo$imdbN 1357237.45 34950089.65 0.04 0.97  
## moviesInfo$rTMCN -6804534.17 14942663.26 -0.46 0.65  
## moviesInfo$rTMUN 58896199.29 28367710.51 2.08 0.04  
## -------------------------------------------------------------------

## Results

Looking at the summary from the linear regression above, this can help give a clearer image of the relationships.

For each increase of a positive rating for IMdB, there was an increase of revenue of 1,357,237.45.

For each increase of a positive rating for RT Critics, there was a decrease of revenue of 6,804,534.17.

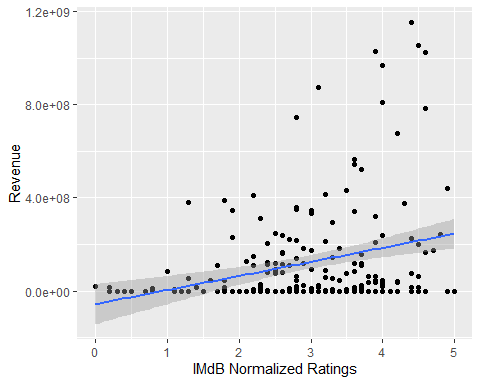
For each increase of a positive rating for RT Audiences, there was an increase of revenue of 58,896,199.29.

These numbers surprised me quite a bit. Another shocking piece of information is looking at the p values that were returned. The difference between IMdB and the RT Audience are vastly different. IMdB sees to give back a confidence level of 3% while RT Audience is giving back 96%.

Below you will find graphing of each individual regression

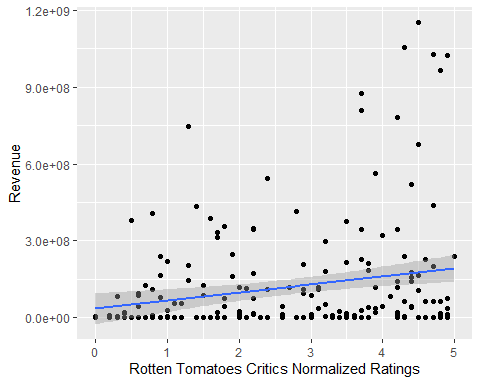
library("ggplot2");  
  
ggplot(movieReg, aes(x=moviesInfo$imdbN, y=moviesInfo$revenue)) + geom\_point() +  
 geom\_smooth(method="lm") + labs(x="IMdB Normalized Ratings", y="Revenue")

## `geom\_smooth()` using formula 'y ~ x'



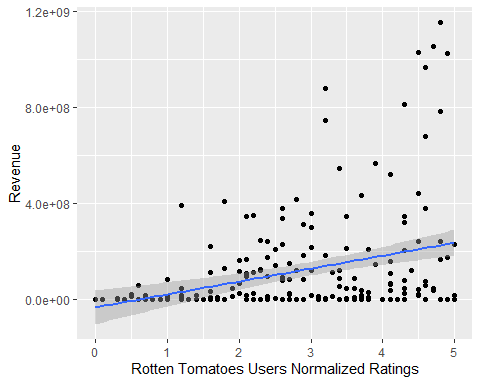
ggplot(movieReg, aes(x=moviesInfo$rTMCN, y=moviesInfo$revenue)) + geom\_point() +  
 geom\_smooth(method="lm") + labs(x="Rotten Tomatoes Critics Normalized Ratings", y="Revenue")

## `geom\_smooth()` using formula 'y ~ x'



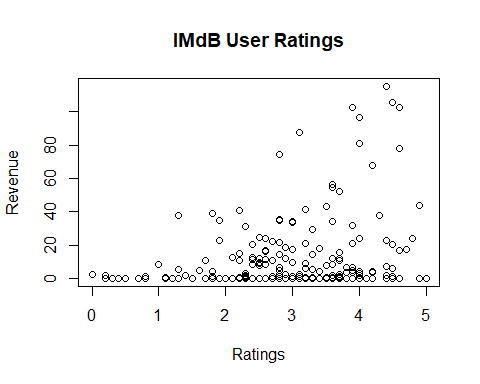
ggplot(movieReg, aes(x=moviesInfo$rTMUN, y=moviesInfo$revenue)) + geom\_point() +  
 geom\_smooth(method="lm") + labs(x="Rotten Tomatoes Users Normalized Ratings", y="Revenue")

## `geom\_smooth()` using formula 'y ~ x'

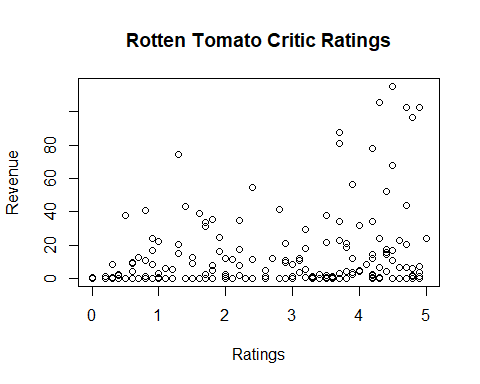


I also decided to graph the non-linear gression graphs to show what is currently in the dataset. The revenue is reduced to help make the numbers more clear.

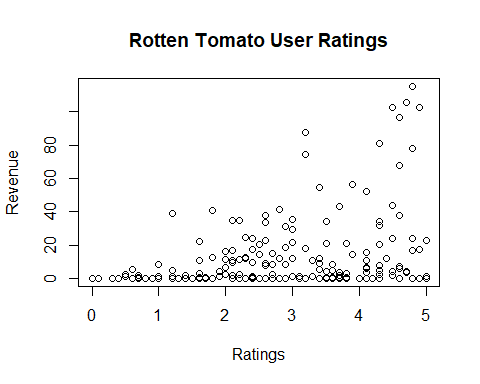
moviesInfo$revenue <- moviesInfo$revenue/10000000;  
  
plot(moviesInfo$imdbN,moviesInfo$revenue, main="IMdB User Ratings", xlab="Ratings", ylab="Revenue");



plot(moviesInfo$rTMCN,moviesInfo$revenue, main="Rotten Tomato Critic Ratings", xlab="Ratings", ylab="Revenue");



plot(moviesInfo$rTMUN,moviesInfo$revenue, main="Rotten Tomato User Ratings", xlab="Ratings", ylab="Revenue");



## Implications

The results of this project have shown that the Rotten Tomato Audience ratings show a better relationship between the rating of a movie and its revenue. I would want to try this project again with a wider range of movies. There would need to be some recalculation for movies from past decades for inflation.

## Conclusion

This project has shown me that my original theory was incorrect. I believe there could also be some factors as to why this is. For RT Audience members, they were required to have purchased a ticket for a movie. This directly affects the revenue for the movie. For IMdB, a user could have given a movie a rating years after its release by watching it on their television. This factor does not tie to affecting the revenue.

## References

AllMoviesDetailsCleaned: <https://www.kaggle.com/stephanerappeneau/350-000-movies-from-themoviedborg/data>

movie\_ratings\_16\_17: <https://github.com/mircealex/Movie_ratings_2016_17/blob/master/movie_ratings_16_17.csv>

“Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction”: <https://www.researchgate.net/publication/259557588_Dynamic_Effects_Among_Movie_Ratings_Movie_Revenues_and_Viewer_Satisfaction>

Rotten Tomato Links: <https://www.rottentomatoes.com/about> (Audience Definition) <https://www.rottentomatoes.com/help_desk/critics> (Critics Definition)