April 20 - Turning in Final Report Document

due April 20, 2022 by 11:59 PM

Matthew Distell, Bo Aldridge, Haris Adnan, Lucan Franzblau, Devin Obee

4-19-2022

Load Packages

Load Data

Introduction: Motivation and Context for the Research

Movies and film have served not only as a form of entertainment in the modern era, but a reflection of who we are as a society, and what our culture values. Analyzing not only films themselves, but the broader types of films that succeed can grant us many insights into our collective psyche.

One interesting idea that we sought to research was the difference between the preferences of fans and critics for a variety of metrics. The two main research questions, that we will dive into later, center around the differences between fans and critics with regards to genre preference and box office success.

This dichotomy is fascinating because it represents a cultural difference between factions of our society. While the average American might view certain types of movies as entertaining or meaningful, those who call themselves "critics" might very well come from a far different background than the average American, and thus carry a different persepetive. Analyzing these differences can illuminate broader social and cultural differences in our country.

Dataset

This dataset is from a csv called "IMDB-Movie-Data.csv". It was loaded into this project as "myMovieData." This dataset was found from data.world, created by user "promptcloud." This dataset looks uses scrapes publicly available data from IMDb, or the International Movie Database, and contains movie data from 2006 to 2016.

```
## Rows: 1,000
## Columns: 12
## $ Rank                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ Title               <chr> "Guardians of the Galaxy", "Prometheus", "Split",~
## $ Genre               <chr> "Action,Adventure,Sci-Fi", "Adventure,Mystery,Sci~
## $ Description         <chr> "A group of intergalactic criminals are forced to~
## $ Director            <chr> "James Gunn", "Ridley Scott", "M. Night Shyamalan~
## $ Actors              <chr> "Chris Pratt, Vin Diesel, Bradley Cooper, Zoe Sal~
## $ Year                <dbl> 2014, 2012, 2016, 2016, 2016, 2016, 2016, 2016, 2~
## $ `Runtime (Minutes)` <dbl> 121, 124, 117, 108, 123, 103, 128, 89, 141, 116, ~
## $ Rating              <dbl> 8.1, 7.0, 7.3, 7.2, 6.2, 6.1, 8.3, 6.4, 7.1, 7.0,~
## $ Votes               <dbl> 757074, 485820, 157606, 60545, 393727, 56036, 258~
## $ `Revenue (Millions)` <dbl> 333.13, 126.46, 138.12, 270.32, 325.02, 45.13, 15~
## $ Metascore           <dbl> 76, 65, 62, 59, 40, 42, 93, 71, 78, 41, 66, 74, 6~
```

This myMovieData set contains the title, genre, description, director, actors, year of release, runtime, fan rating from IMDb (International Movie Database) (Rating), number of fan votes, box office revenue, and critic rating (Metascore) for 1000 movies.

Research Questions

From these broader ideas regarding movies and the relationships between certain types of them, we narrowed our research into two specific areas:

1. What is the relationship between fan ratings and critic ratings with regards to box office success? Do higher fan or critic ratings lead to more box office success? And if so, which factor contributes more?
2. What is the relationship between fan ratings and critic ratings with regards to certain genres of movies? Specifically, is there any difference between how fans and critics view, for example, action movies, or dramas?

Methodology Part 1

Before getting started, we wanted to only look at "popular" movies. We arbitrarily defined this as having 8000 or more reviews on the IMDb database. Those movies with lower than that total are filtered out. This excluded 133 of the initial 1000 movies from myMovieData.

Next, we created a new variable called newRating, which adjusts the "out-of-10" scale value of the IMDb rating, and puts it on an "out-of-100" scale, similar to that of the Metascore rating.

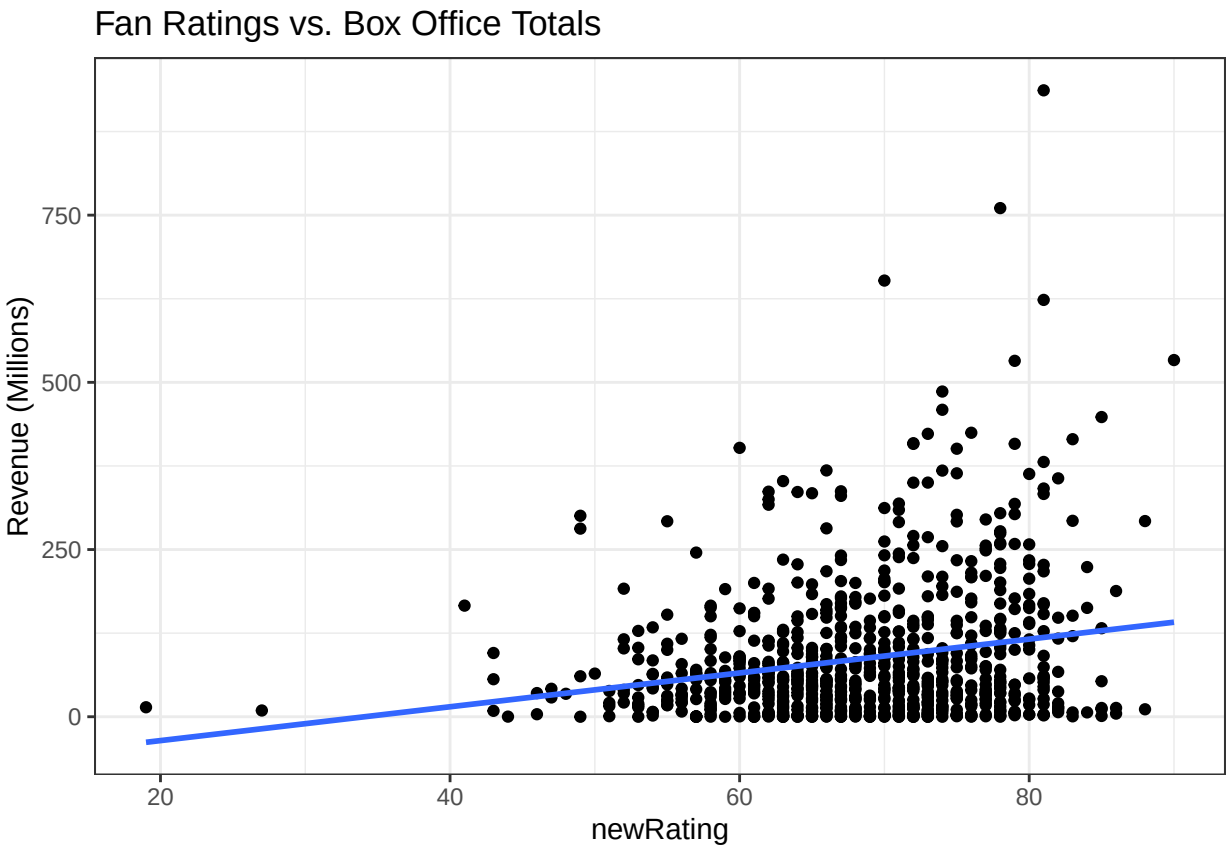
Part 1: The Relationship Between Fans vs. Critic Scoring and Box Office Revenue

```
##
##  Spearman's rank correlation rho
##
## data:  popularMovies$newRating and popularMovies$`Revenue (Millions)`
## S = 81357124, p-value = 7.772e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##     rho
## 0.13696
```

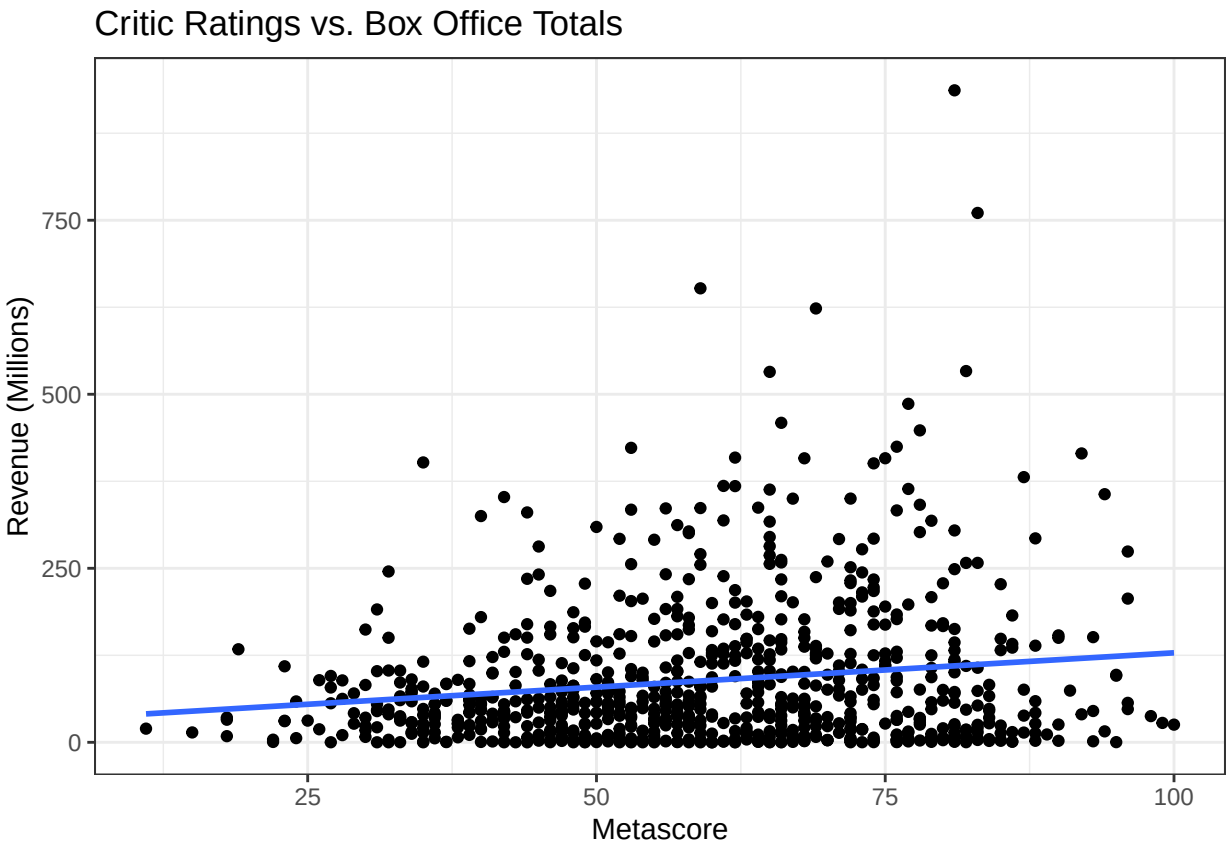
```
##
##  Spearman's rank correlation rho
##
## data:  popularMovies$Metascore and popularMovies$`Revenue (Millions)`
## S = 77160187, p-value = 0.02056
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.08207753
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  -83.8      31.3      -2.68  0.00749
## 2 newRating      2.39      0.594     4.03  0.0000622
## 3 Metascore      0.146     0.300     0.486 0.627
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
```



Part 1 Results and Discussion

In our correlation model, we found that there was a stronger relationship between fan ratings and box office success (.137) as compared to critic ratings and box office success (.082).

Our linear model confirms some of these findings. The null hypothesis for each parameter in this model is that each specific parameter does not have a relationship with the box office revenue of each specific movie.

The alternative hypothesis, on the other hand, is that each specific parameter does have a relationship with the box office revenue of each specific movie.

The formula for our linear model is as follows:

$$\widehat{Revenue} = -83.8 + 2.39\ fanRating + 0.15\ criticRating$$

The only predictor that was significant at the alpha level of .05 was newRating, with a p-value of .00006 and a high t-value.

The newRating slope indicates that with all else held constant, for each 1 unit this movie increases in rating, the movie will earn, on average, $2.39 million more at the box office. The criticRating slope indicates that with all else held constant, for each 1 unit this movie increases in rating, the movie will earn, on average, $0.15 million more at the box office.

This does not necessarily mean that the other predictor, Metascore, was uncorrelated with the outcome, but it does mean that once you account for the one factor that was statistically significant (newRating), the Metascore predictor in it of itself was not making as large of a difference there. Metascore predictor failed to reject the null hypothesis, while newRating rejected the null hypothesis.

Graphically, these same findings can be seen as well. There is a slightly stronger/similar positive correlation between newRating and box office as compared to Metascore and box office graphically. This, however, does not take into account the fact that the p-value was not found to be significant, meaning that while there still is a positive trend between Metascore and box office, we cannot comment on its specific strength of relation to box office with nearly as much certainty as we can with user ratings in newRating.

These findings, logically, make sense. The fact that there is not only a stronger correlation, but also a more statistically significant one between fan ratings and box office success makes sense given that fans using these rating services should, as a whole, represent a larger portion of the population, which would, in turn, represent a larger portion of the ticket-purchasing constituency as compared to critics, who might come from a less diverse background (many critics might have gone to film/art school, or have formal education in film history, etc).

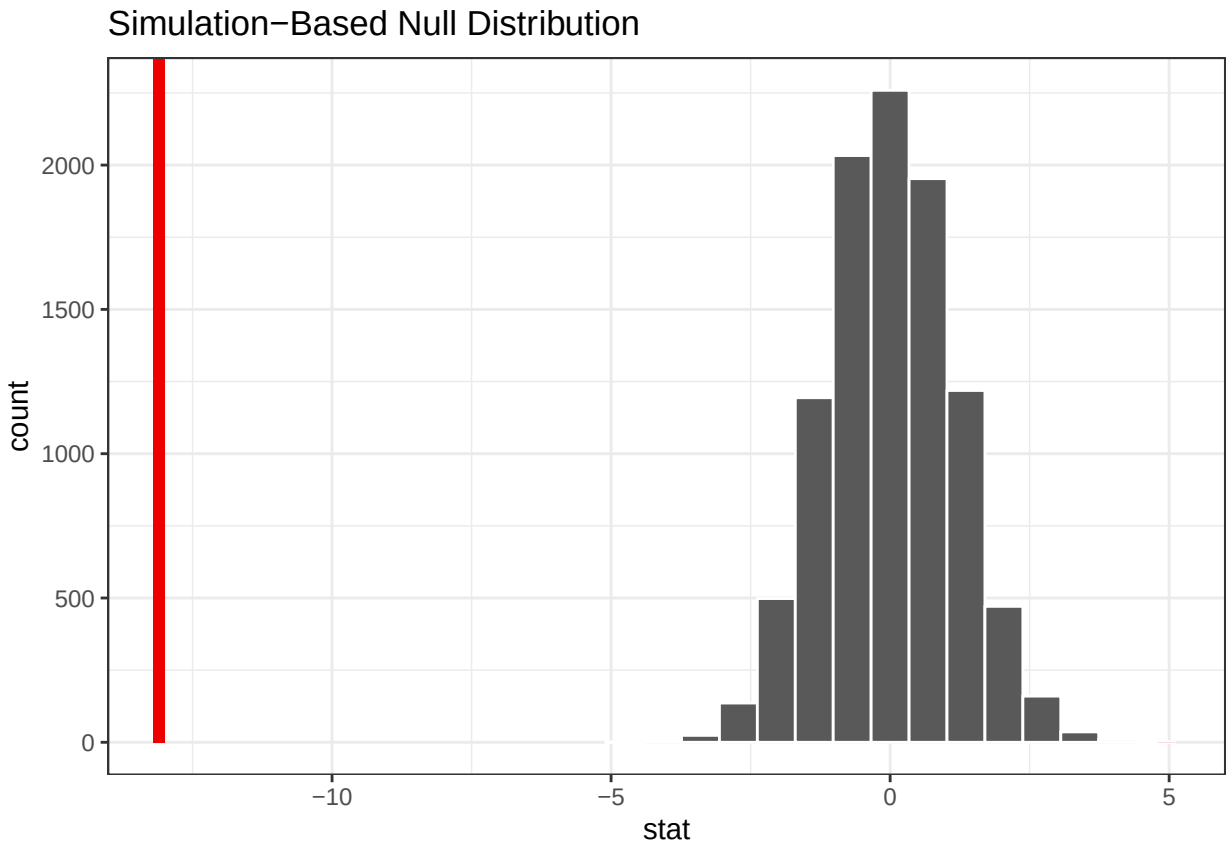
Something interesting to consider would be the differences/correlation if we tested everybody who just left the theater about a certain movie. Since critics can be assumed to have a somewhat more uniform "film background," to a certain extent, it could also be assumed that users who took the time to personally rate movies on IMDb also do not represent the average viewer's level of interest or engagement with film in general. By sampling more people directly at the source, we could have a more in-depth discussion regarding personal views and film.

Methodology Part 2

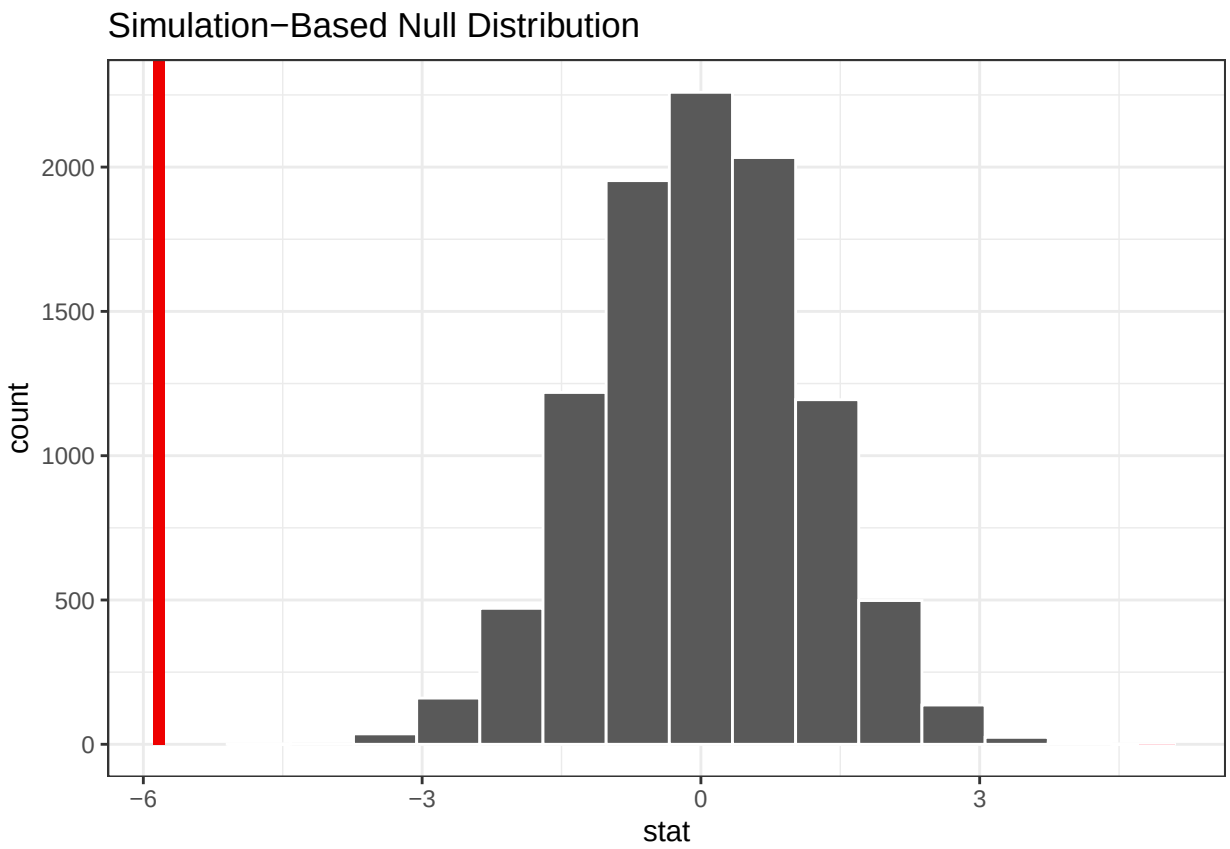
Before getting started with part 2, we wanted to split up action movies from drama movies. The code below splits these two categories up, and then filters out any movie that might be labeled as both an action movie and a drama movie. This decreased the amount of movies in our dataset from 877 to 596. These 596 are all categorized as either dramas or actions, but not both. We then created a new variable, critic_dif_fan, that showed the difference between the critics view of a movie and the fans view of a movie numerically. This will be used in our hypothesis testing to generate an outlook into the relationship between these variables.

Part 2: Hypothesis Testing for the Difference in Genre Preferences Between Fans and Critics

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```



Part 2 Results and Discussion

There will be two separate null hypotheses for this part of the report.

The first null hypothesis is that there is no difference between the mean fan and critic views of action movies.

- $H_0 : \mu_{criticaction} - \mu_{fanaction} = 0$

The first alternative hypothesis is that there is a difference between the fan and critic views of action movies.

- $H_1 : \mu_{criticaction} - \mu_{fanaction} \neq 0$

The second null hypothesis is that there is no difference between the fan and critic views of drama movies.

- $H_0 : \mu_{criticdrama} - \mu_{fandrama} = 0$

The second alternative hypothesis is that there is a difference between the fan and critic views of drama movies.

- $H_1 : \mu_{criticdrama} - \mu_{fandrama} \neq 0$

This simulation was conducted using 10000 repetitions for both in which the difference in means was tested. The p-value for both of these tests was 0, indicating that we can reject both null hypotheses. This means that there is a difference between the fan and critic views for both action and drama movies. What we found is that critics are harsher, on a "to-100" scale, for both dramas and action movies relative to fans. However, critics much preferred dramas over action movies as compared to fans, seeing as the observed statistic for the mean difference in action between critics and fans was -13.1 points, while the mean difference for dramas was -4.8 points. This can be seen graphically, as in both representations of the distributions, our red line representing our observed statistic is located to the left of our actual distribution. However, that observed statistic was far closer to the distribution in the drama graph as compared to the action graph.

This logically seems to make sense. Building off of our previous explanations of the "film background" of fans vs. critics, it would make sense that fans might prefer a traditional blockbuster movie, as compared to a critic who might prefer a slower-paced drama. Because critics might care more than fans about more subtle aspects of film making such as cinematography, movies, such as dramas, that are slower, might focus on these aspects more and cater to a critic's interest.

Conclusion

Overall, fan ratings correlated at both a higher and more significant rate than critic ratings with regard to box office success. Fan ratings and critic ratings were also statistically different with regards to mean scoring between types of genres. Fans prefer both dramas and action movies as compared to critics, but critics preferred dramas over action movies relatively speaking.

In the future, it would be interesting to change the pool of data as a whole. As mentioned in a previous paragraph, the type of people that seek to review ratings on IMDb and the type of person who becomes a professional film critic probably do not fully represent the audience as a whole. Interviewing people as they left the theater might present interesting insights into this analysis.

This dataset also only considered films made between 2006 and 2016, which, considering the whole history of film, is a very short timespan. By adding more movies to this dataset as a whole, which could be done with this publicly accessible data, new insights into broader cultural shifts over time, rather than cultural differences between subsets of American society, could be analyzed and offer key insights.