BST 270 Final Project

Randy Williams

1/18/2023

Introduction

For my final project, I chose to reproduce the visualizations for the article, "Be Suspicious Of Online Movie Ratings, Especially Fandango's", which was published by FiveThirtyEight. I also critiqued the reproducibility of the article.

Data for every film pulled from Fandango on Aug. 24, 2015 can be downloaded from this FiveThirtyEight GitHub repository (use `fandango_scrape.csv`). Data pulled on Aug. 24 2015 for every film that has a Rotten Tomatoes rating, a RT User rating, a Metacritic score, a Metacritic User score, and IMDb score, and at least 30 fan reviews on Fandango can be downloaded from the FiveThirtyEight GitHub repository (use `fandango_score_comparison.csv`). I uploaded the Fandago Scrape and Fandago Comparison datasets directly from FiveThirtyEight's github repository.

Data Dictionary

The data dictionary for each file used for the article can be found in the github repository (https://github.com/fivethirtyeight/data/tree/master/fandango).

Figure 1

Data Wrangling the Fandango Scrape Data

Use the Fandango Scrape data (`fandango_scrape.csv`) to create Figure 1. First, I reduced the list of films to ones that came out in 2015. There was no variable for movie year. Movie year was recording in the variable for film name. So, I used the grep function find films with the 2015 in the film name variable. I also had reduce the observations to films with 30 or more reviews. The author of the article stated they found 209 films. However, I found only 192 with the method that I used. I created a frequency table the number of Fandango reviewed film with a certain star count.

Plotting Figure 1

Despite having fewer observations, the graph I created was comparable to the original graph in the article.

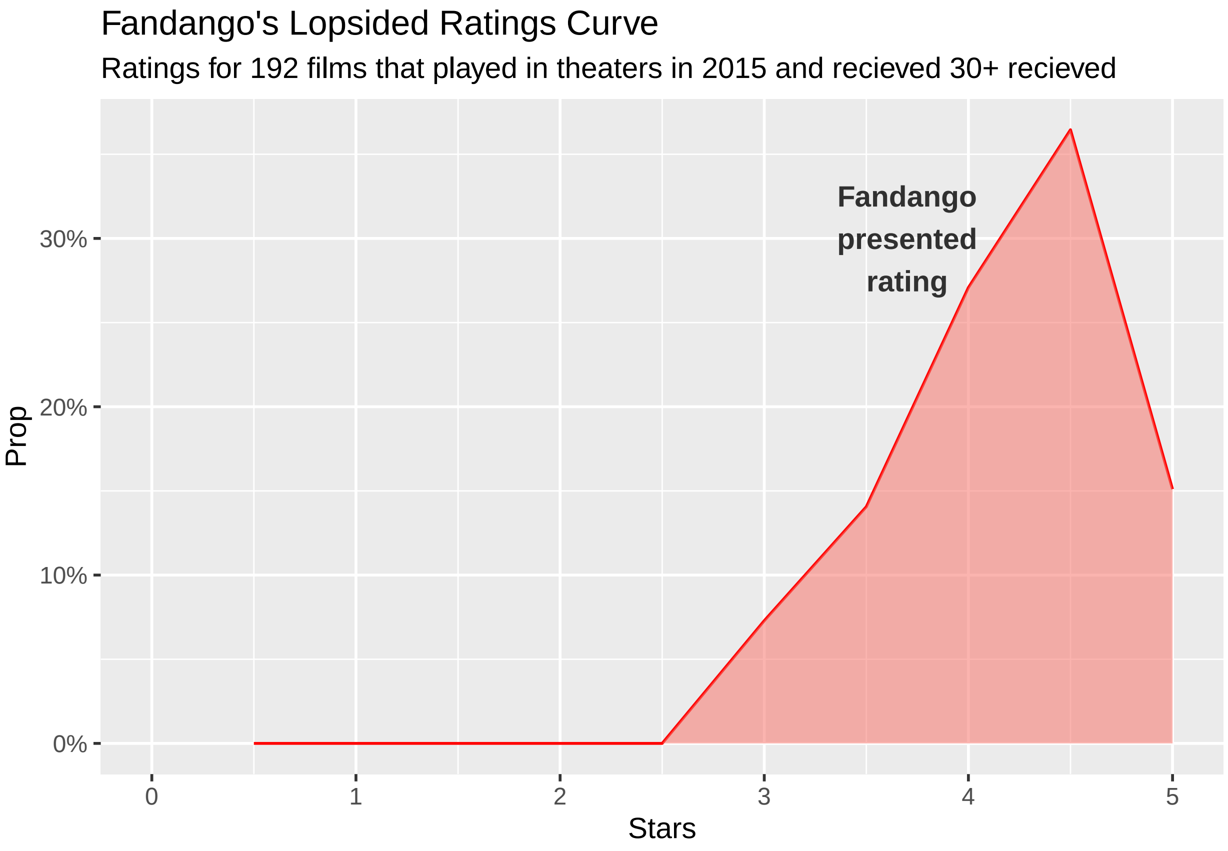


Figure 2

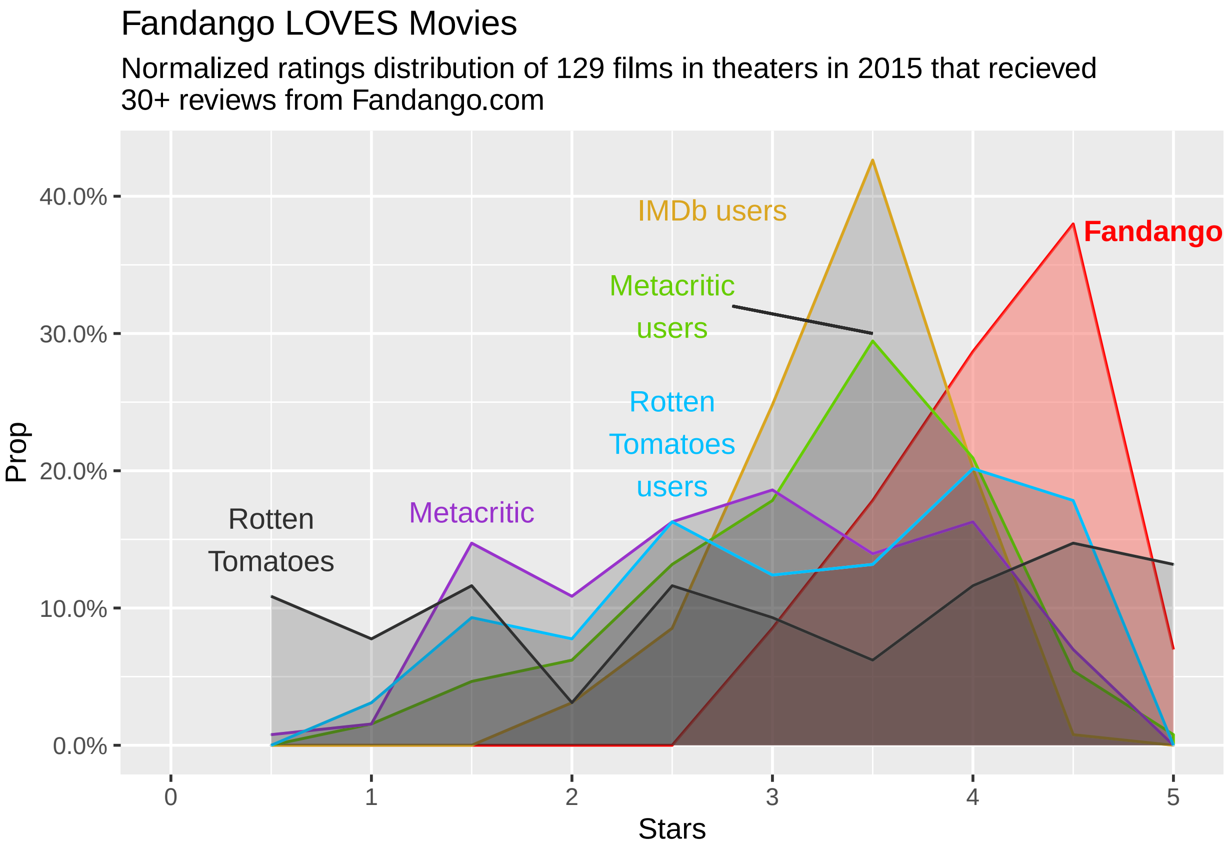
Data Wrangling the Fandango Comparision Data

Figure 2 was created using the Fandango Comparison dataset (`fandango_score_comparison.csv`). To create Figure 2, I also needed reduced the list of films to ones that came out in 2015. Similar to the Fandango scrape dataset, the Fandango comparison dataset had no variable for movie year. Therefore, I had to use a similar approach to filter the data to movies released in 2015. According to the article, not all the movies sold by Fandango had IMDb and Metacritics. Therefore, the number of observations was smaller for the Fandango comparison dataset compared to the Fandago scrape dataset. The author of the article claimed of have found 146 films in 2015 from the subtitle of Figure 2. However, I found only 129 films with the method that I used. I would assume that there was an error made by the author of the article because there are only 146 observations in total for the Fandago Comparison Datasets. The author would have used the whole dataset to produce Figure 2 and not all of the movies are from 2015. The dataset consists of films like "Two Days, One Night (2014)", "Into the Woods (2014)", "A Most Violent Year (2014)", "The Hobbit: The Battle of the Five Armies (2014)" and more which were released earlier than 2015.

I also created a frequency table of the number of films with a certain star count for each film review site using the Fandango comparison dataset. These frequency tables were used to construct the curves.

Plotting Figure 2

There are some differences between the graph I generated compared to the original graph. For example, the peak of the curve for Metacritic users reviews was slightly smaller than 30% in my graph, while the original graph had the peak of the curve for Metacritic users reviews was higher than 30%. This was probably because there was a difference between number of films in 2015 I found and the number films they reported in the article. Nevertheless, the shape of the curves for each category was very similar to my graph compared to the original.



Fandango LOVES Movies

Normalized ratings distribution of 129 films in theaters in 2015 that recieved 30+ reviews from Fandango.com

Final Remarks on Reproducibility

I was able to reproduce data visualizations of curves with a similar shape. However, I wasn't able reproduce the same number of films after filtering by year due to poor reproducible data science practices when designing the data. There wasn't a separate variable created for year which made the process of looking for films in 2015 not straightforward and error prone. Furthermore, there was a strong possibility that the author made an error in the Fandango Comparison dataset because of this issue. Also, the author mentioned that he normalized the rating from other review sites to the Fandango 5 star scale and provided both the normalized ratings rounded and unrounded as variables in the dataset. However, the author failed to disclose the details on the normalization that was done or what normalization method was used.

In conclusion, despite being able to reproduce similar looking graphs with the data provided, there was clear evidence of poor reproducible data science practices that were used throughout the investigation by the FiveThirtyEight article.