Mini-Project 4 – Large Data Set Analysis

Cameron T. Crites

November 22, 2024

Abstract

Data analytics is a powerful tool that can be used to draw conclusions based on not-so-obvious trends that may be taking place. By using tools to create visuals for large data sets, conclusions can be drawn and hypotheses can be scrutinized. Looking at movies, data analytics can be used to determine what makes a good movie. Of course there is a subjective nature involved in deciding how good a movie is, but by analyzing the characteristics of well rated movies, rigid rules can be created for the subjective nature of this realm.

Introduction

Everybody loves going to see movies. Whether you see them in the theater, or you wait until you can buy the DVD or stream it online, they are a large part of our culture. We glorify the people in them, and we regard their creators as the master play-writes of our time.

Problem Statement

The movie industry has a tremendous amount of influence over the general public. They themselves have critics which rate movies, which in turn either promotes or demotes the movie t the general public. While these 'critics' may be the experts, I believe that the better gauge of a movie's success is the rating that the public gives to the movie. In order to investigate this difference between critics and general audiences, I will be analyzing the ratings reported by both *iMDb* and *Rotten Tomatoes*. Furthermore, I'd like to discover the characteristics of a well rated movie.

Methods

In order to complete this data analysis, the main tool will be the *pandas* Python package. This package has many tools to upload, edit, retrieve, and save data from csv files. The most important pieces of information from the files regarding to Rotten Tomatoes and iMDb are as shown in the table below:

Title	Year	Genre	Director	Cast	Description	iMDb Audience Score	iMDb Metascore	RT Audience Score	RT Critic Score
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 1: Example of Important headers.

To get a single table like the one in figure 1, the data from both sources must be 'cleaned' and then merged together. To do this, many of the not so necessary or redundant columns were deleted from each file, such as *Runtime* and *Number of Reviews*. While they do carry relevant information, it was not important for the scope I was planning on taking for this investigation. From each data set, these 'unnecessary' columns were removed and one singular data set was created by mergin the two resulting data sets. The primary modes of visualization for the data are 2D Histograms (heatmaps), Bar graphs, Box plots, and Pie charts. These are only some of the tools available through the use of *pandas* and *matplotlib* packages, but they provide a comprehensive view of the data.

Results

In this section, I will discuss the trends that were noted from data sets.

Rotten Tomatoes v. iMDB

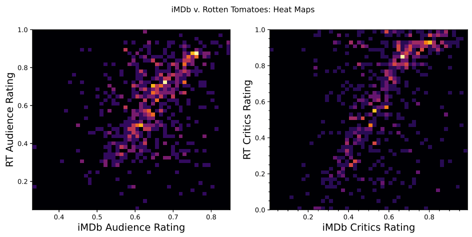

Figure 2: Heatmap of Rotten Tomatoes and iMDb ratings

When comparing Rotten Tomatoes to iMDb ratings, there is clearly a linear relationship. For the audience ratings, this relationship has a slope closer to 1 than the critics ratings. This can be seen from the pattern of a diagonal line (0,0) to (1,1) in the first plot of Figure 2. When it comes to the critics scores, it seems that the Rotten Tomatoes' critics are harsher in their scoring. They do not rank movies as high as iMDb critics. This leads to the steeper slope of the second plot.

Audience v. Critics

Still comparing the Rotten Tomatoes ratings to the iMDb ratings, we can see how closely the critic ratings and the audience rating match each other.

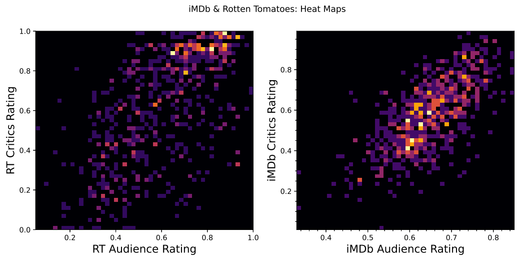

Figure 3: Heatmap of Audience v. Critic Scores

In this figure, we can see that there is much more correlation for iMDb than for Rotten Tomatoes. This is not to say there is no correlation for Rotten Tomatoes, but the hot-spot is much wider meaning there is less uniformity in the ratings for audiences and critics. Similar to Figure 2, the hot-spot is higher on the plot, indicating that the Rotten Tomatoes critics are again rating movies more harshly than the comparably data. I believe it is safe to say that Rotten Tomatoes is a tougher grade, but probably a more accurate one. I believe this is because Rotten Tomatoes has individual critics, whereas iMDb compiles their critic ratings from an amalgamation of other critic sites. It is for this reason that Rotten Tomatoes will be used when considering ratings for the majority of the rest of this paper.

Influence of Movie Descriptions

I was curious as to whether the inclusion of an actor's name in the movie's description would lead to higher movie ratings. There were not many movies that actually did this, but of the ones that did, I was able to produce another heatmap.

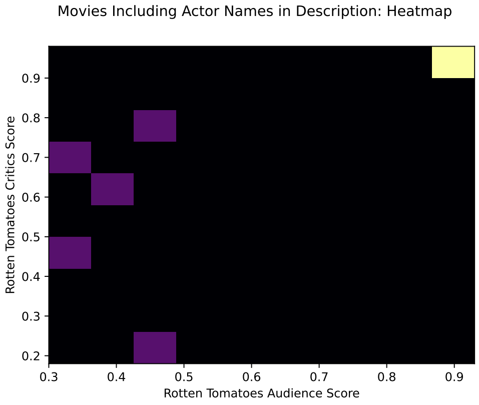

Figure 4: Heatmap of Audience v. Critics Ratings

From Figure 4, it can be seen that there is no clear correlation between audience rating and critics rating anymore. I believe the use of actor names in the description had the opposite affect of what I was expecting. Instead of the names leading to higher ratings, it is possible that they just set high expectations for the audience that were unable to be met by the film. This shouldn't affect the critics, as they are supposed to unbiased reviewers of films. The plots shows multiple movies that were rated highly by the critics, but rated poorly by the audiences. Given that the audience rating seems to be more relatable to how well a movie does, it would be my conclusion that including actor names in descriptions is not a good idea.

Genre Trends

The files of movie information included many genres that the movie could be categorized as. Such as ”Action, Comedy, Drama... etc.” for practicality, only the first listed genre was considered.

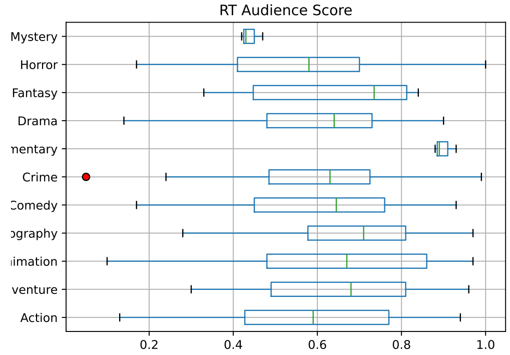

Figure 5: Box plot of Audience Scores by Genre

Most average ratings fall between 0.6 and 0.8, with a few being less and one being more. From Figure 5, It can be seen that the most widely spread genres of movie are *horror* and *animation*, with *horror* having a high end whisker reaching a perfect score of 1. There are also two genres which have very narrow spreads. *Mystery* and *Documentary* films both have spans which are less than 0.1 in width, indicating a consistent view of that genre from the public. *Documentary* films are rated around 0.9, hinting that they may be a safe bet for a movie, as there is very little variation in audience perception of the movie. I believe that this is because there is very little room for interpretation when making a documentary, so audience members know exactly what to expect when they sit down in front of the screen.

Certificate Trends

Movies are given a certificate in order to tell audiences what age group of people the movie is suitable for. Some of the movies included in the data sets were not American films, and thus they did not have American certificates. Part of the data cleaning process was to convert these international certificates to American equivalences.

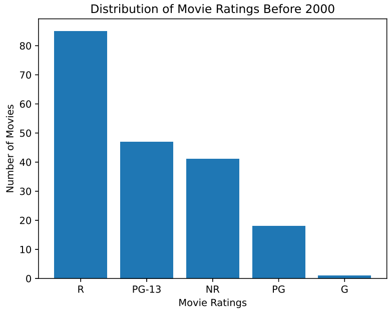

Figure 6: Distribution of certificates before 2000

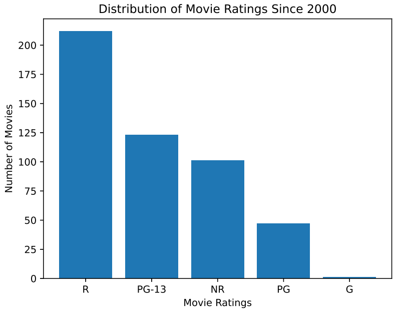

Figure 7: Distribution of certificates after 2000

Looking at the distribution of movie certificates both pre and post 2000 show the same distribution of movie certificates. It is commonly said by people of older generations that movies today are much raunchier than they were in the past. This is clearly not the case as the distribution is the same. There is a caveat that as time progresses things that were once taboo become mainstream, so it is possible that the certificates have gotten less strenuous, but the rate of movies being created of certain certificates has not changed.

Pie Charts

One very easy way to visualize data is to create a pie chart. This is because the abstract concept of a quantity is materialized into a quantity that can be seen in relation to other sizes.

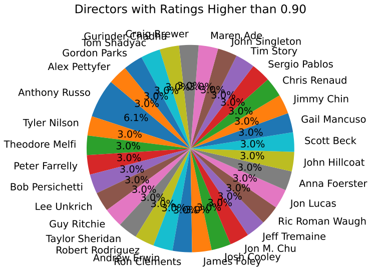

Figure 8: Pie chart of top directors

Figure 8 shows the names of the directors who were involved in movies that received a rating greater than 0.90. Of these names, all are equal at 3.0% corresponding to a single movie, except for Anthony Russo who has two movies in that rating range. Anthony Russo is the director of multiple MARVEL movies as well as romantic comedies such as ”You Me & Dupree”. Being the only director with more than one movie in this upper echelon speaks to his directing abilities.

It would also be useful to see which genres have the largest number of movies that are rated greater than 0.90.

Figure 9 shows that action movies are clearly the most popular. Action movies in the top 0.10 rating number almost twice that of the second category: animation. From this chart we can also see that the lowest genres of top rated movies include: documentary, horror, and genre. It is interesting to see that documentary is numbered the least when Figure 5 showed that they are the most consistently highly rated. It is important to note that Figure 9 is just the number of movies, so it does not account for the percent of movies in the top 0.10 ratings.

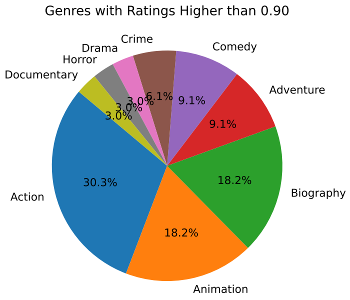

Figure 9: Pie chart of top genres

Discussion/Conclusion

As stated previously, movies are a subjective experience. What may be enjoyable for one individual may be boring to another. And what may be inappropriate to one, may be hilarious to another. The above data can still be used to draw some useful conclusions. Most notably, a ’formula’ can be created that could lead to a successful movie. In terms of Director, having Anthony Russo direct the movie lends a greater chance that it will end up being highly rated. Most highly rated movies are action movies, so making the movie an action film also lends towards it being higher rated. The items below are a rough blueprint for a well rated movie:

- Director: Anthony Russo
- Genre: Action
- Certificate: R
- Cast list: Absent from description