Wikidata Movie Analysis

CMPT 353
Tyler Pham
301137122

Introduction

Using data from Wikidata and Rotten Tomatoes, I want to see how well each cast member, director or genre does with the metrics of a movie during a certain period. The metrics I am interested in are Rotten Tomatoes “Average audience rating”, “Percentage of the audience who ‘liked it’”, “Average critic rating” and “Percentage of critics who gave a positive review” and the movies return on investment high return on investment (reported box office divided by the reported budget). Initially the different categories were compared together, however the results were difficult to compare and will be compared individually. I will focus

Data

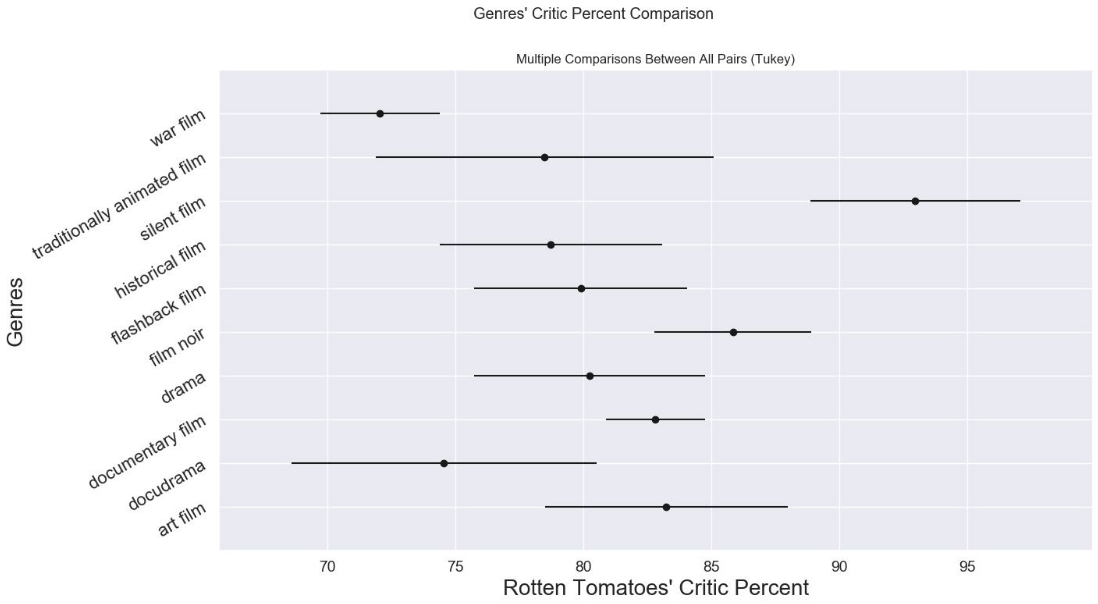
The data was obtained from Wikidata’s JSON data dump cleaned to output movie data. Cast members, directors and genres also obtained from the data dump are given as ‘wikidata ids’ (eg. Q2853003), which need to be mapped later to represent something their name. This data was then combined with the Rotten Tomatoes data to obtain the movie’s ratings.

To make comparisons across the different categories, I needed to narrow down the total number of movies to inspect. I focused on the genres and critic percentage to start. I started by taking average scores of all the genres that appeared in the data. I removed any genre that appeared less than 40 times. This is done so that each genre could be assumed to be normally distributed. Among the remaining genres, I took the 10 best rated by critic percentage.

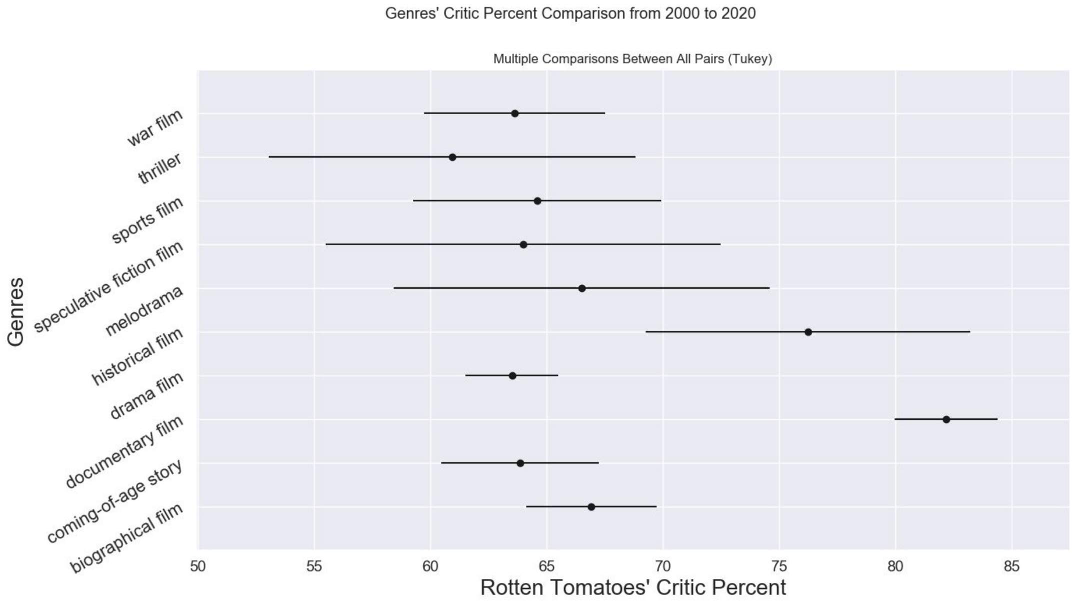
Movies were then selected only if they have that genre. Movies that have multiple genres were counted individually. For example, the score for “Avengers: Infinity War” would count towards both a “superhero film” and a “adventure film”. This process is the same for any category and metric

Analysis

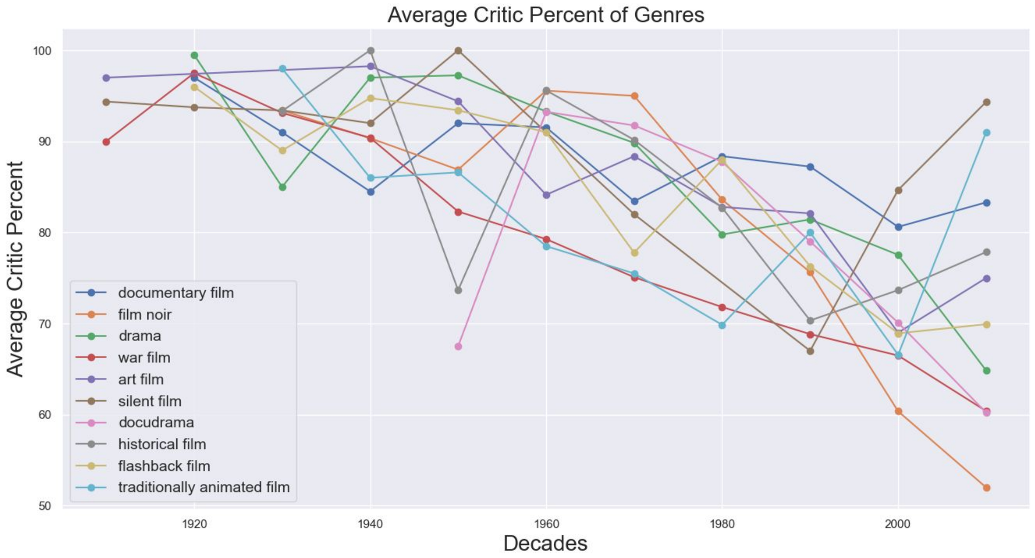
After the data is organized, the genres are tested to determine if their means differ. If the p-value is less than alpha (in this case is which 0.05), we can proceed with Tukey's Honest Significant Difference.



From this plot, we can see that silent films have the highest average critic percentage of the group. This is a little interesting since there can't be many silent films in recent history. We can see how that compares movies that come after 2000

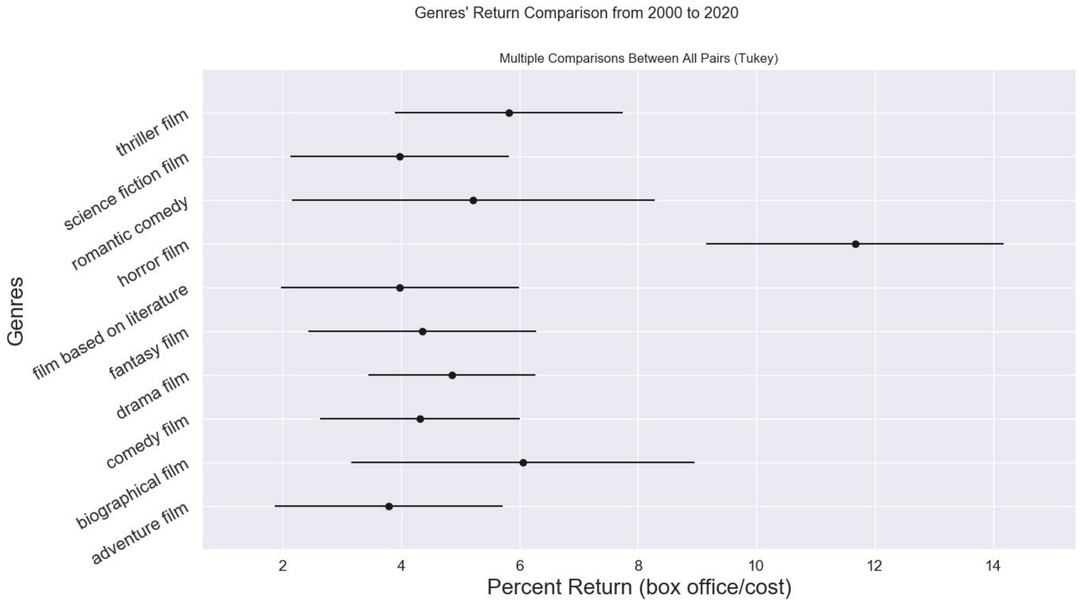


Looking only at movies published from 2000-2020, we can see that it changes quite a bit. Older types of films have disappeared, such as art, noir and silent films.


Average Critic Percent of Genres

However if we look at the average critic percent every decade, we can see that certain genres were much more critically acclaimed in the 20s. Many of them, except silent films, have now fallen during the 2000s, possibly due to the rise of other genres.

We can do the same to look at the genres' return on investment.


Genres' Return Comparison from 2000 to 2020

Multiple Comparisons Between All Pairs (Tukey)

Horror films have the highest return of all the genres. Typically horror films are known to be lower budget but are still very successful in the box office (for example, Saw 1 had a budget of $1.2 million with a box office of $103.09 million)

Conclusions

Using Tukey's Honest Significant Difference, we quickly compare the different genres with the highest average with a certain score across a time period. The inclusion of a time period can change the scoring significantly as there can be trends in what is popular in that decade.

Unfortunately, this approach fails for cast members and directors as it is unlikely that a person is involved in more than 40 movies to assume normality. Even if we try to lower the number of movies required, we can not confirm that the means are different.

Limitations

- I would've liked to have gone further in my initial objective of being able to use cast members, directors and genres to predict any of the scores. I was worried about the time constraints and initial bad results that I had to pivot to a somewhat more simpler idea.
- I spent too much of the project trying to write functions that were more flexible and capable of producing different results depending on the parameters. In retrospect this made the problem less defined which made analysing and drawing conclusions much more difficult. It feels more like I made a tool rather than doing research
- In my opinion, the most important metric for a movie would be the return on investment. However, the budget and box office values are not accurate, these values are only reported. Of the 40430 movies, only 791 have data for both the budget and box office. I was not satisfied with the amount of data to play with. If I had more time I would have looked for more resources to find these values.
- Not being able to confirm the differences in means and normality with the cast members and directors seems obvious afterwards. A person involved with over 40 movies is very unlikely

Project Experience Summary

Movie Data Analysis - CMPT 353 Computational Data Science

- Wrote pandas/python scripts to clean and organize data from wikidata.org
- Analyzed different properties of a movie, such as genres or cast member, to see what would yield the best results in terms of critic ratings or return on investment
- Used stats tests (Tukey's HSD) to understand and confirm the different results