CMPT 353 Movies Project

Leo Chen || lca113@sfu.ca || 301276015

My original goal of this project was to measure how ratings affect the success of a movie. I combined all of the movie json files into one big Dataframe and manipulated my data based on it. The first thing was cleaning out all of the unnecessary columns out. Many datasets had NaN values or serial numbers in place of a name. For example in wikidata_movies.json the country of origin column used Q145 to represent the United Kingdom. I could not find the resource to translate the many codes to their respective country without googling each one.

The main datasets I used to squeeze out answers include the imdb id, audience average, critic average, publication date, omdb awards and omdb genres.

ML Movies

The Machine learning part of the course was one of the most interesting things I learned. I wanted to apply the algorithms I learned on the movies dataset and see if I can predict anything interesting.

I cleaned out many datasets and decided to use these columns: audience average, audience total ratings, critic average, critic percent and total awards/nominations

I used Naïve Bayes, K Nearest Neighbors (KNN), and SVC but SVC was too slow so I eventually removed it from my list of models. The result was unfortunately less than satisfying.

I tried out different y values and used the remainder columns as X trainers. Predicting critic averages and awards did not go well. Accuracy score would be < 0.2. Fortunately, KNN actually worked very well with y = audience ratings. accuracy scored was around ~0.70 but bayes accuracy was still around ~0.2. In a way I was p-hacking this accuracy score because the audience rating scale is only between 0 and 5 so there are a smaller range of values to predict. I noticed that KNN worked better than Naïve Bayes overall.

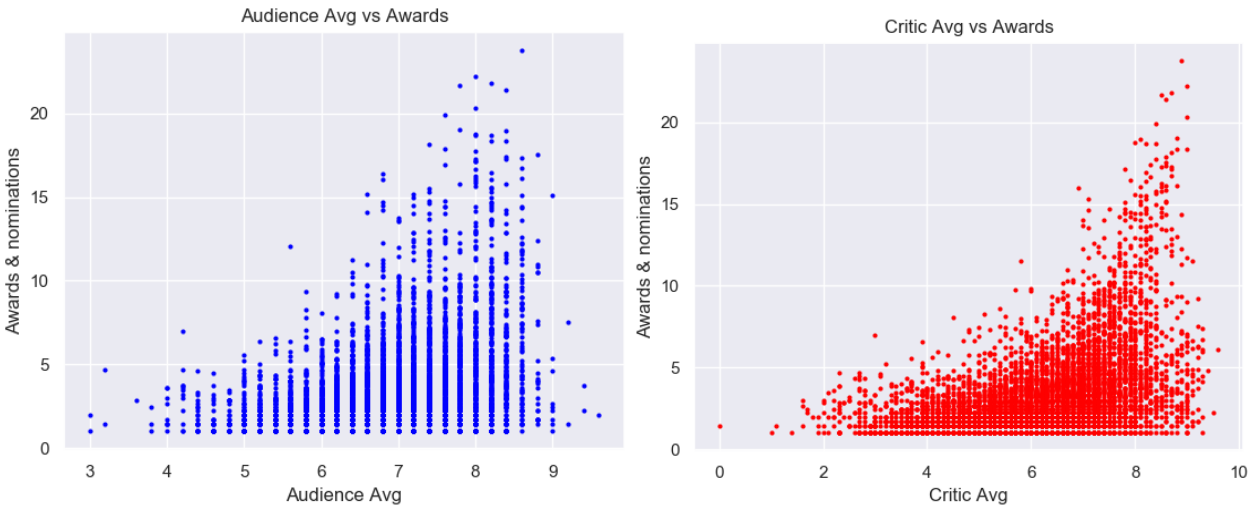
I concluded that it is possible to predict audience scores at a 70% accuracy rate when using KNN.

Ratings vs Awards

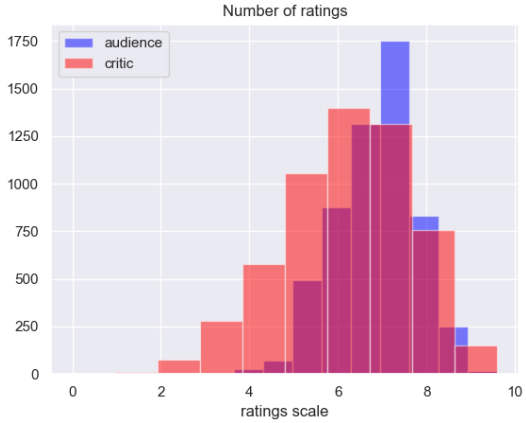
Do higher ratings relate to higher awards and nominations?

Many data in the omdb awards column were labelled N/A so I interpreted that as either the movie received no awards or that information is not available. In order to remove ambiguity, I decided to only keep movie awards and nominations greater than zero. I added all of the awards and nominations up line by line using a regex expression to make my data easier to work with.

I had trouble deciding how to compare the data to the ratings. There were many datasets such as audience ratings, critic ratings, audience percent and critics percent. A histogram showed a skewed distribution with the percentage scores while the average ratings had a pleasant normal distribution. I decided to use the normally distributed average ratings for both audience and critics. One thing to note is that I doubled the audience scores so they match with the 10-number scoring scale as critics.



NOTE: Awards & nominations values are square rooted for better visual representation
The average audience rating is 6.87 while the average critic average is 6.20. The surprising thing here is even though audience scored over 0.5 points higher, the graphs shows that critic ratings have more of a skew to the left. This means that higher critic scores correlate to more awards and nominations compared to audience ratings.

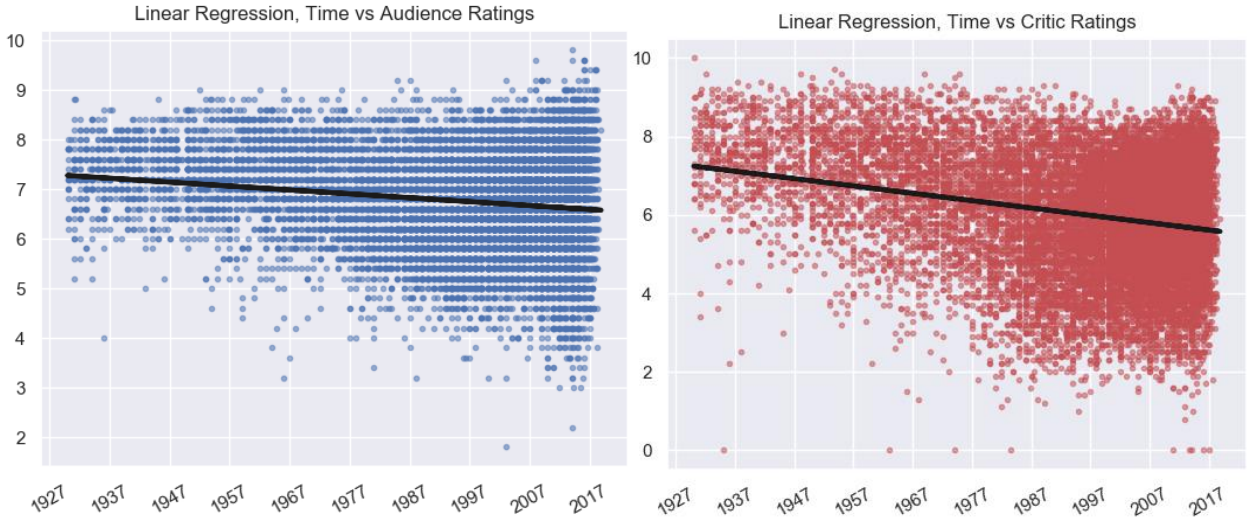


Here is another graph to show distribution of average ratings. You can See that audience bars are slightly to the right of the critic bars. Audience also have a smaller variance but this might be due to me doubling the scores to match with critics.

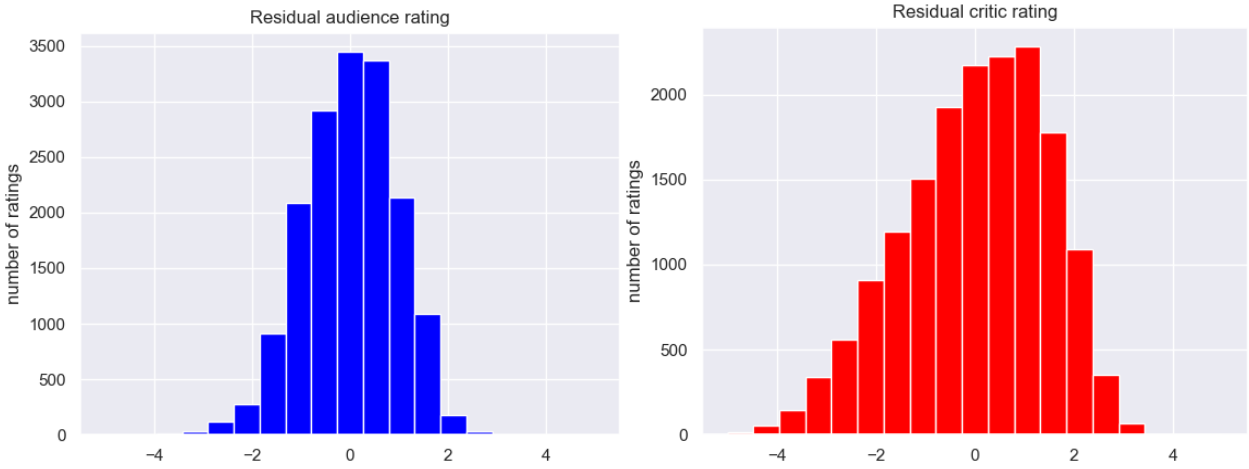
Time vs Ratings

Out of my curiosity for history, I wanted to see how ratings differ over time.

Most of this analysis is based on the dog rates assignment. I converted the publication date column to datetime values, then to a timestamp so it can be graphed. There was little data before 1930 so I filtered that out too. Therefore, the movies that remain span from 1930 to 2018. Furthermore, I used linear regression to make out a trend line over time.



Our graphs show that ratings have decreased overtime from 1930 to now. Audience ratings dropped from 7.3 to 6.5 while critic ratings dropped from 7.1 to 5.6. This lower critic rating shows a steeper negative linear regression slope compared to audience ratings.



The residual graph shows that audience ratings have a smaller distribution which means that values are closer to the average.

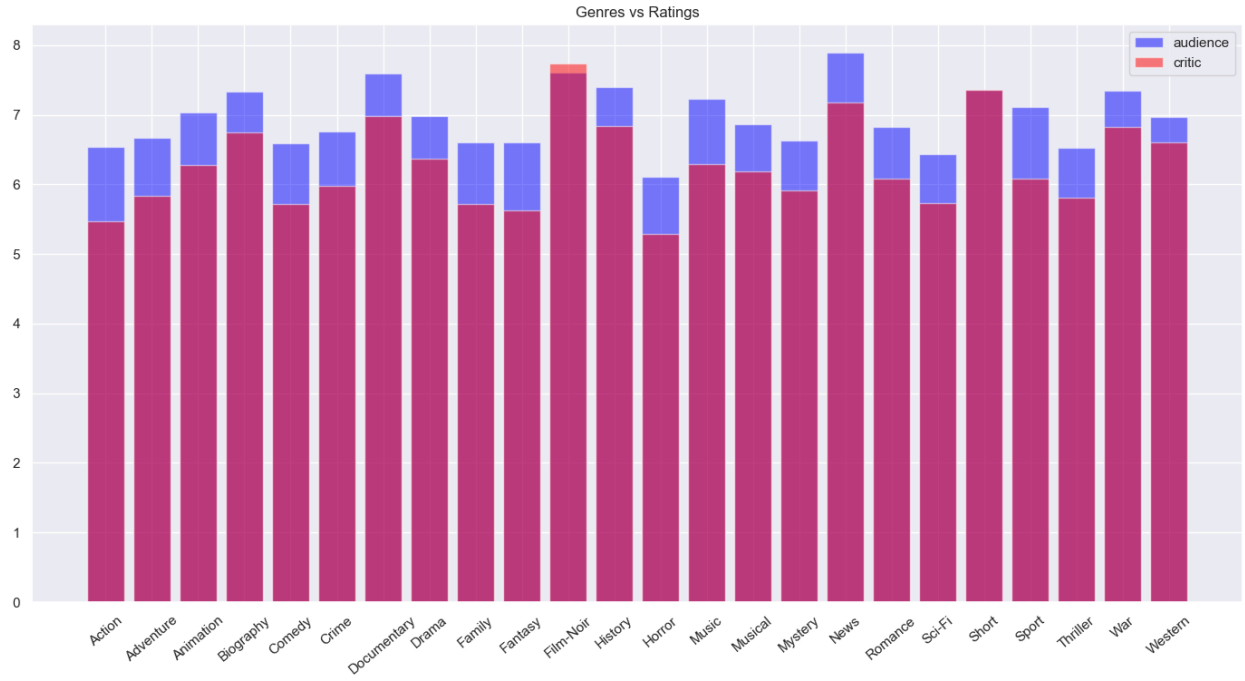
There are many reasons for the negative ratings overtime. Perhaps our data is flawed and we kept track of only the good movies from the past while we keep many movies presently. Another explanation is that critics and audience are harder raters and do not give out good ratings as easily as before. Another plausible reason is that movies are getting worse over time like those horrible 50 Shades of Grey or Superman movies.

Genre vs Ratings

I wanted to see if audience and critics have a similar or difference preference to movie genres. I compared the ratings that each genre received.

The three genres columns from all the combined movies data were genre_label, omdb_genres and genre. genre_label from genres.json had zero matches with the rotten tomatoes ratings data. The genre column from wikidata was labeled as a serial id numbers so that was unusable. Thankfully there were about 6000 datasets for the omdb_genres that matched with audience and critic ratings.

The hardest part about this assignment was reading the lists from the genre rows and extracting all the genres and putting them into a separate dataset while preserving the associated ratings. I used np.repeat and np.concatenate to do this. Then I used groupBy() and mean() to get the final dataset.



Interestingly enough, there are no major discrepancies between the movie preferences of the audience and critics. They both gave high scores for short films and Film-Noir while equally disliking horror movies. The higher averages for audience ratings are very visible across the board. I did not sort or put a trend line on this graph because the X values are nominal.

I concluded that certain genres like short and Film-Noir generally scored higher averages than films like horror. Also, critics and audience share the same taste in movies.

Project Experience Summary

- Used Python libraries such as numpy, pandas, Sklearn, and matplotlib to manipulate and graph movies data
- Predicted average movie audience ratings to an accuracy of 70% using the K Nearest Neighbors machine learning algorithm
- Used visual interpretation and mathematical processes to predict the positively correlated behaviour between movie ratings and movie awards
- Applied Linear Regression techniques to find out the decreasing rating averages over time
- Proved that audience and critics enjoy and dislike the same genre of movies