Wikidata, Movies, and Success

In this report we will be addressing two main questions, the affect of movie genres on ratings, and the relationship between different factors of movie success (profit, reviews, popularity). We used various statistical testing techniques and machine learning models to make sense of the data.

The Effect of Movie Genres on Ratings

Part 1

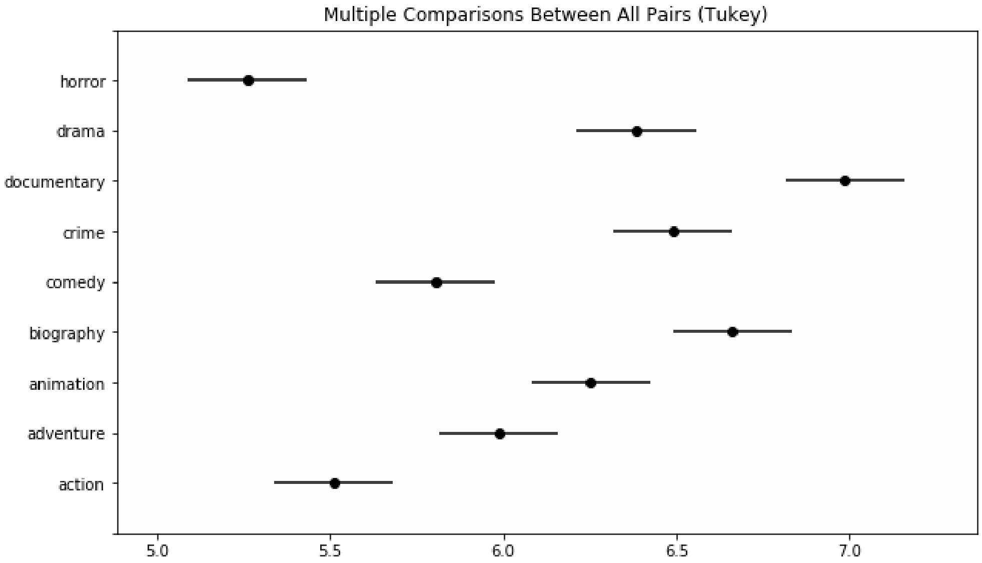
The question that we are addressing here is whether there is a difference between the average ratings that movies receive based on the genre. For example, do action movies tend to be more highly rated than comedy movies? We will be using the critic ratings to investigate this.

We used the data sets provided in the data.zip. To answer this question, we first needed to clean and gather the data we needed. In the provided data set, the omdb data and wikidata data both had genres columns, so we had to choose one of them. We chose to go with the genres listed in the omdb data as they were more general, whereas as the wikidata data would sometimes be very specific in identifying genre (for example, wikidata lists the movie “Krampus: The Devil Returns” as being a “Christmas movie”, while the omdb data will list it as a “action” movie, which will be better for our analysis). The omdb genres would give a list of genres for a particular movie, however, to clean this up a bit and be more specific, we took the first genre listed in that list as being the genre for the movie. This is fine as omdb lists the genre that most relates to the film first (it is not listed alphabetically), thus for example a “Comedy and War” movie will be listed as a “Comedy” movie instead. I believe that this will give us better results, for example, it would be better to list a movie such as “Tropic Thunder”, as a comedy movie rather than a war movie. Lastly, to further refine our data set, we only took genres that were listed more than 40 times in the dataset, reducing the number of genres from 22, down to the 9 most popular (an example of genre being excluded is “film-noir”, which only had two occurrences in the data set).

Once we had gathered and cleaned our data, the next step is to analyze it. Our data sets of the individual genres were not of the same length, for example we had more data for action movies than for biographies, thus we had to take make the data sets all of equal length so that we can use an ANOVA test. Our hypothesis for the test is:

- Null hypothesis: The mean of the critic ratings between the genres is the same
- Alternate Hypothesis: The means of the critic ratings between the genres is different

We used 0.05 as our alpha value. After ensuring that our data met all the requirements of ANOVA, we did the ANOVA test and got a p-value significantly lower than our alpha, thus there is a difference in means between the genres. We then used the Tukey HSD test to see what genres had different means. Our results are shown in the figure below.

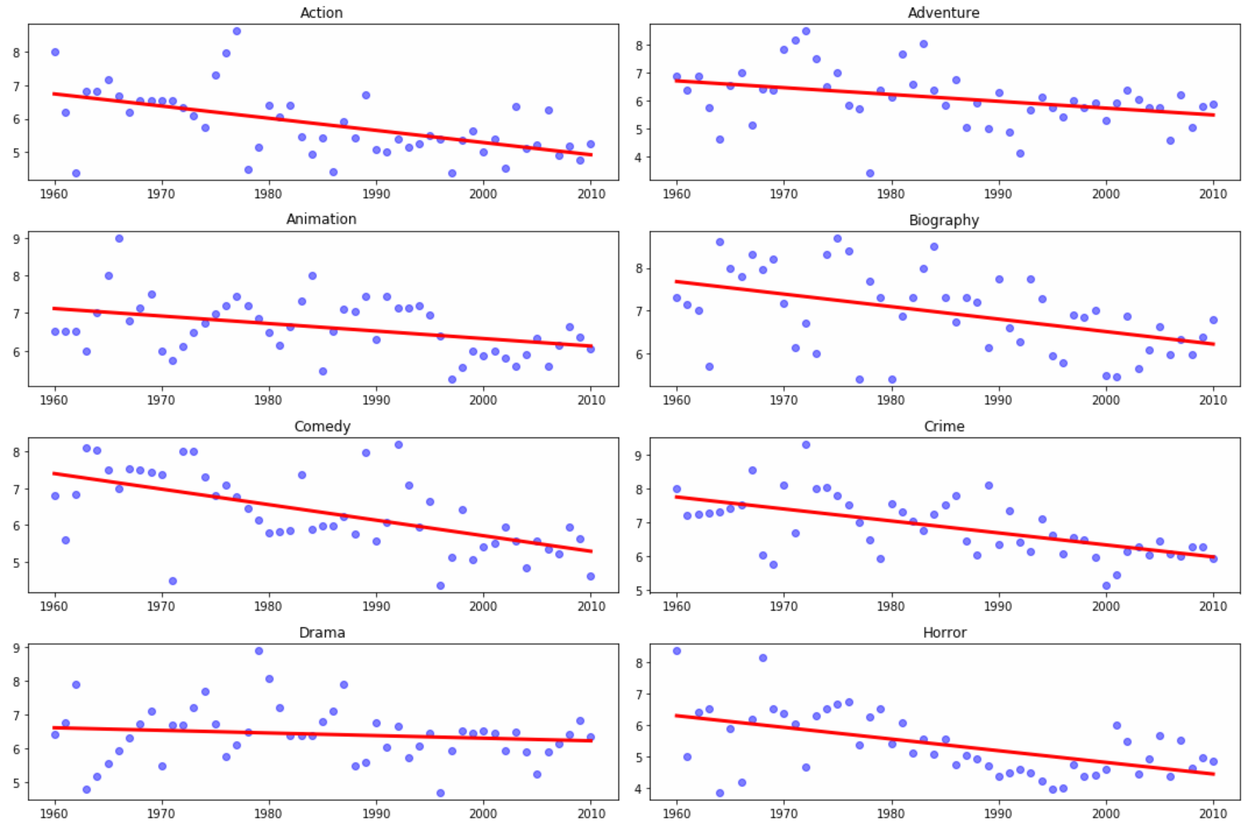


After doing all the required cleaning and testing, we are given the above result. Thus, we can conclude that there is a difference in ratings that a movie receives depending on the genre. From the above plot, we can clearly see that horror and action movies are on average, rated lower in comparison to documentaries and crime movies. We can also see that adventure and comedy movies receive about the same ratings.

Part 2

After answering the question about the effect of genre on ratings, we had a follow up question; Has the ratings of certain genres changed over the years? We used the same data as in the first part; however, some additional data cleaning was required.

In addition to our data from part 1, we added a date column to our dataset. We will look at the average rating of a genre per year, thus we extracted this information from the publication_date column in the wikidata data, grouped by the year, and calculated the average rating for a genre in a given year. In order to have an equal comparison of ratings for genres over the years, we need to have our data all within the same date range. This was difficult, as we had data for some genres staring in 1914, and other genres will have their data start in 1940. There were also gaps in between points, for example, we may have average ratings for horror movies in 1950, but not for 1951. We thus limited our data set to ratings between 1960 – 2010, as most of the genres had data during this period. To fill in these gaps in our data, we interpolated the points, and filled the rest of the missing values with the mean of the data set. We had to exclude the “documentary” genre from our analysis, as we had only 25 of the 50 data points (1960 – 2010 is 50 points) we were looking for, thus interpolating and filling nulls for half the values will be of no use. The plots of the ratings vs year are show below.

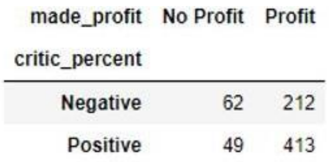


We can see visually see some definite trends from our regression line, such as the decrease in the ratings of horror and comedy movies over the years, and how adventure and drama movies have remained fairly consistent in their ratings over the years. However, there are also cases where the regression line may not be giving a clear indication on the ratings, such as in the Biography and Animation genres, where the ratings look to be increasing at some points, and decreasing at other points.

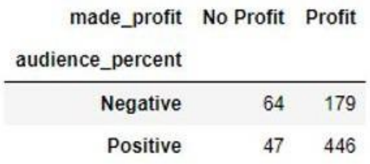
Some limitations of this analysis are our dataset. We had to interpolate and fill in missing values, and had to lose some data so that we can have the same date range for all of the genres. Had we had a larger data set and more values to base the average yearly rating of a genre of off, we may have had better results. Due to these results, I had also used a machine learning model see if we can predict genres based on ratings (all ratings, rather than just critic ratings) and year, to see if there is some type of correlation that can be found that helps us predict. Using the GradientBoostingClassifier machine learning model, I was able to get a score of %33, thus there does not seem to be much of a relationship between ratings, year, and the genre, as I had initially anticipated.

The Relationships between Movie Reviews, Profits, and Popularity

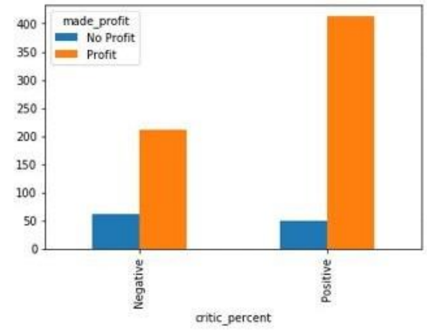
Using the data we previously cleaned up, we were able to easily merge the “made_profit” column from the wikidata data and the data from Rotten Tomatoes. Due to the need for critical and audience reviews, we are using most of the data from the Rotten Tomatoes dataset. Our goal was to analyze the relationship and correlation between critic reviews, audience reviews, whether the movie made profits, and the popularity of the movie. Since our data only tells us if the data made a profit or not, we decided to use a chi-square test to analyze the data.

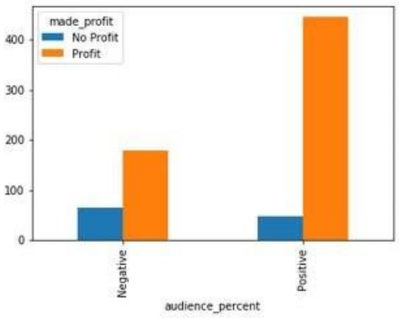


made_profit critic_percent	No Profit	Profit
Negative	62	212
Positive	49	413



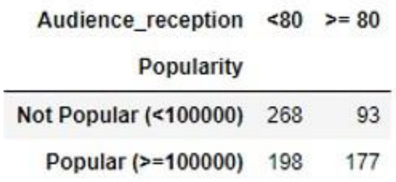
made_profit audience_percent	No Profit	Profit
Negative	64	179
Positive	47	446



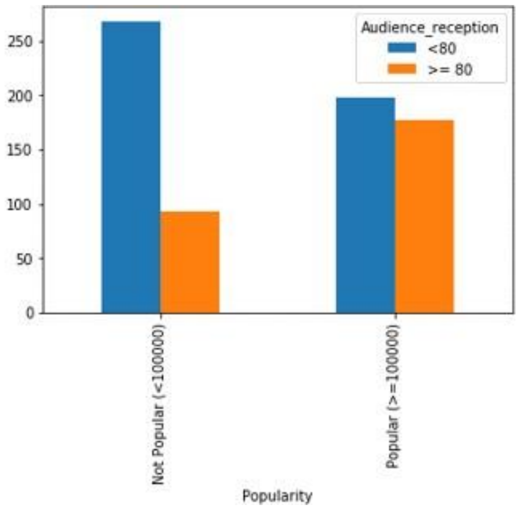


Using the Rotten Tomatoes’ system of dividing between positive and negative reviews, we labelled critic and audience scores that were 60% or above as positive, and the rest as negative. Our null hypothesis would be that critic or audience reviews and whether a movie makes profit are independent. The p-values we get from the chi-square test are 1.715 x 10$_{-5}$ and 4.07 x 10$_{-9}$ for critics and audiences respectively, well below our alpha of 0.05. As a result, we can reject the null hypothesis and conclude that reviews, from either critics or audiences, do influence the chances of a movie making profit. However, when looking at the crosstab tables, we notice that it appears that most movies make profit regardless of positive or negative reviews. This could indicate that our data is skewed or unreliable, and our results are not conclusive. Alternatively, it could mean that the movie industry is a profitable one where most movies succeed, making it lackluster marker for statistical analysis. To truly analyze the impact of reviews on profits, we would need to compare the amount of profit gained, through numbers or percentages, rather than a binary yes profits/no profit. Another limitation was that there were a lot of entries in the database with no data about profits, reducing the sample size greatly.

Next, we investigated the relationship between how popularity (number of audience ratings) a movie is and whether audiences liked a lot (>=80%). There have been many popular movies in history that were initially panned by critics, but due to sheer popularity, became beloved fan favourites and popular classics. The cut-off point we used for a “popular” movie was 100,000 ratings, which is approximately average. Then, we did a crosstab and chi-square test to see if there’s a connection between the popularity of a movie and a more positive audience score than critic score.

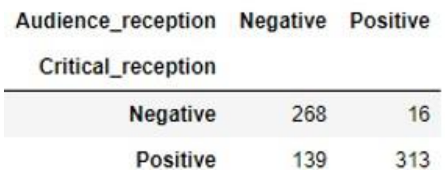


Audience_reception	<80	>= 80
Popularity		
Not Popular (<100000)	268	93
Popular (>=100000)	198	177

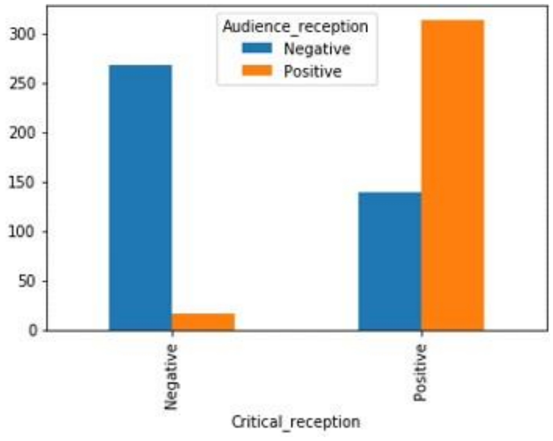


The null hypothesis would be popularity of a movie is independent from it being beloved by the audience. The p value we obtain from the chi-square test is 2.5 x 10$_{-9}$, which is below our alpha of 0.05 and allows us to conclude that the popularity of a movie is related to overwhelmingly positive audience reception. This is a logical conclusion because most popular movie franchises become popular in the first places due to positive reception from moviegoers. Likewise, movies that are unpopular generally were not warmly received by audiences.

It is a popular belief that critics and audiences do not share the same opinions on movies. Critics are seen as snobbish and elitist, whereas audiences are seen as easily entertained and shallow. Our next analysis is to find if there is a relationship between critic reception and audience reception. For critic reception, we followed Rotten Tomatoes' definition of a "fresh" movie, which is 60% positive reviews. For audience reception, we were more stringent, categorizing a movie as positive for audiences when it has 75% positive reviews from audiences.



Audience_reception	Negative	Positive
Critical_reception		
Negative	268	16
Positive	139	313



The null hypothesis is that critic reception and audience reception are independent from each other. The p-value obtained from the chi-square test is 1.70 x 10$_{-63}$, which is below our alpha of 0.05 and solidly concludes that there is a relationship between critic and audience reception. Looking at the crosstab table and the bar graph, there is a very strong relationship between critic and audience reception. In fact, the Pearson correlation between critic averages and audience averages is 0.7377 and the Pearson correlation between critic percentage and audience percentages is 0.8064. From these results, it appears that critics and audiences do agree on which movies deserve positive reception.

As a result, we cleaned up the data to get the specific factors we needed and performed chi-square tests on them to determine the relationship between these factors, illuminating the correlation that exists within the data. Since these success factors could determine if a movie is a hit or a flop, data analysis of this sort is essential for the movie-making industry.