Homework 1.1

Movie Dataset

Statistical Data Analysis - 20th of October, 2019

Group 5

Antoñanzas Martínez, Guillermo
Pueblas Núñez, Rodrigo
Martínez Capella, Ignacio
Burrell, David

Brief data description, preparation and manipulation

We are using the movies dataset[1]. The data set is comprised of 651 randomly sampled movies produced and released before 2016, including information about release date, rating according to multiple websites and cast. Upon investigation it was discovered there were several null values found in columns: **"Runtime"**, **"Director"**, **"Studio"**, **"DVD release date"**, **"Actors"**. Using online sources (see attached code for references) these values were added. Additionally, a film (Hurt Locker) was found to be awarded with a Best Picture Oscar, but not nominated - this error was corrected.

During the analysis of the data two new columns were added to the data set: **thtr_rel_date** (computed date using thtr_rel_day, thtr_rel_month and thtr_rel_year) and **thtr_rel_decade** (a factor variable that holds the decade the film was released). Also, during the dataset up stage, it was decided to remove the columns: **top200_box**, **dvd_rel_day**, **dvd_rel_month**, **dvd_rel_year**, **imdb_url** and **rt_url** as they were not required for our analysis.

Research questions, plots and findings

1. Is there a relationship between the type, genre and score?

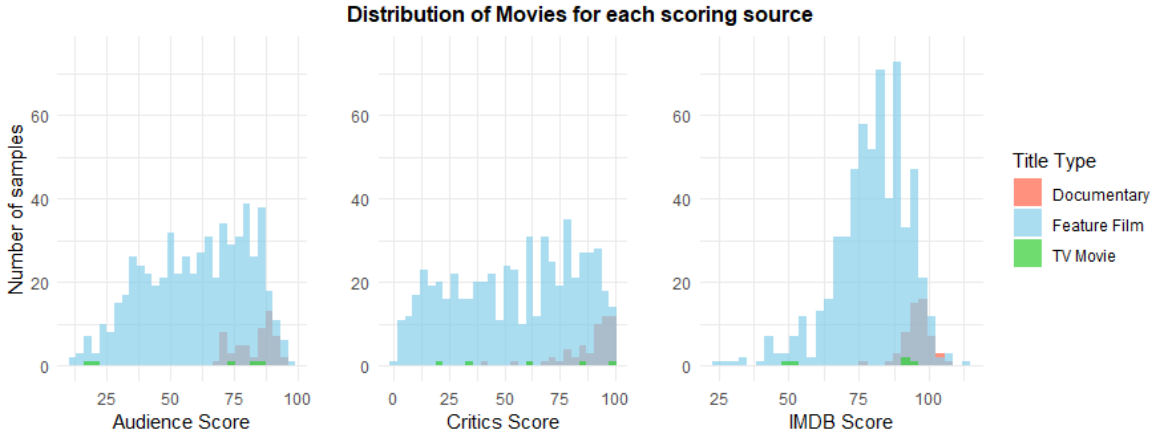

Figure 1.1 Distribution of the score of the films by type, grouped by source.

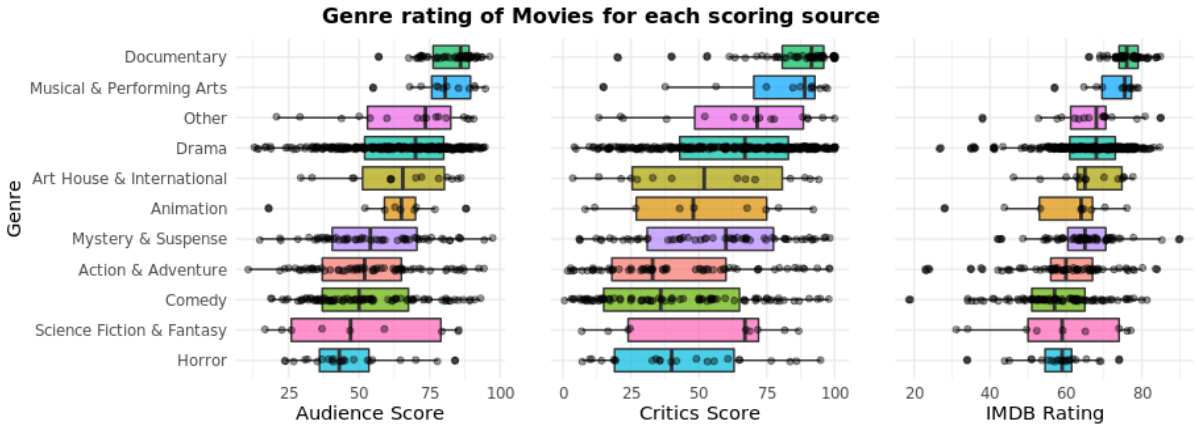

Figure 1.2 Distribution of scores sorted by genre, grouped by source.

[1] https://moodle.upm.es/titulaciones/oficiales/pluginfile.php/1947512/mod_resource/content/4/_site/movies.html

Figure 1.1 shows that a relationship exists between the movie type and score. With it clearly shown Documentary having the greatest score. Figure 1.2 also shows us that genre has an influence on score, with certain ones having better results than others. This may be due to random distribution, but the available data show Documentary and Musicals & Performing Arts being the highest scoring genres. While others, like Horror, tend to have lower ratings.

2. How do the score sources differ for genre and MPAA rating?

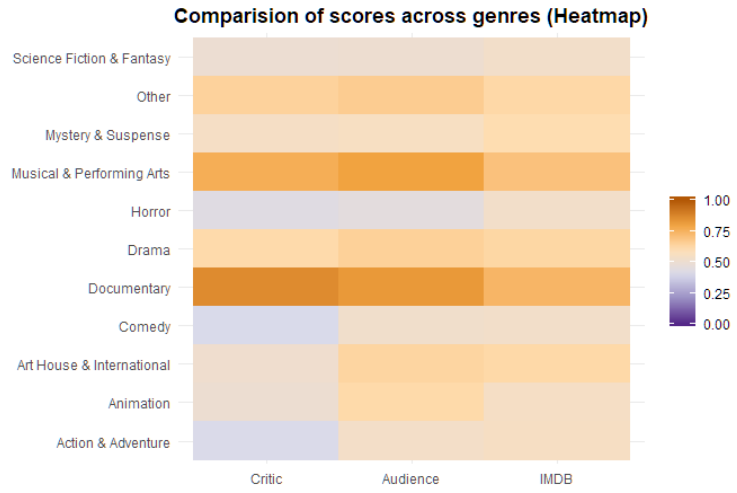

Figure 2.1 Heatmap comparing the differences between sources by genre.

Genre

Figure 2.1 shows the average score value for each genre by source. From the three sources Critics seem to be the most negative, while Audience and IMDB are aligned giving higher results. This is clearly seen in Action & Adventure genre where Critics have a far lower score than the other two groups. This may be due to Critics having a different criterion on what constitutes a "good" movie but without details on the required metrics this is only speculation. Interestingly though IMDB scores are most unaligned with the other groups for the two highest scoring genres, Documentary and Musical & Performing Arts.

MPAA rating

If we make a similar comparison for the MPAA ratings (Figure 2.2) we also find that overall IMDB and Audience tend to give more positive reviews than Critics. However, the scores given by all sources are not that different one from another in the majority of the MPAA ratings (exceptions being the extreme Critics score for NC-17 and PG-13). Therefore, there is not enough evidence to say that there is a significant difference in the score according to the MPAA rating.

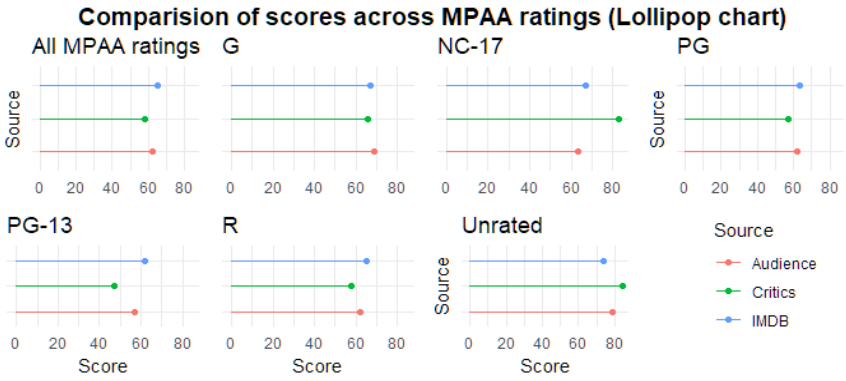

Figure 2.2 Lollipop plot comparing the differences between sources by MPAA ratings.

3. Are Oscar nominated films preferred? Which award improves the score?

We can see (Figure 3.1) that movies nominated for Best Movie receive higher scores, which makes sense. Since all the Best Movies in the dataset had an actor/actress or director who had received an award, we were interested in seeing if there was a relationship between these awards and the Best Movie Oscar. It turns out that only the Best Director award seems to be improving the score. This deduction makes sense, since actors are only responsible for playing their roles whereas the director's decisions involve all the elements of a film.

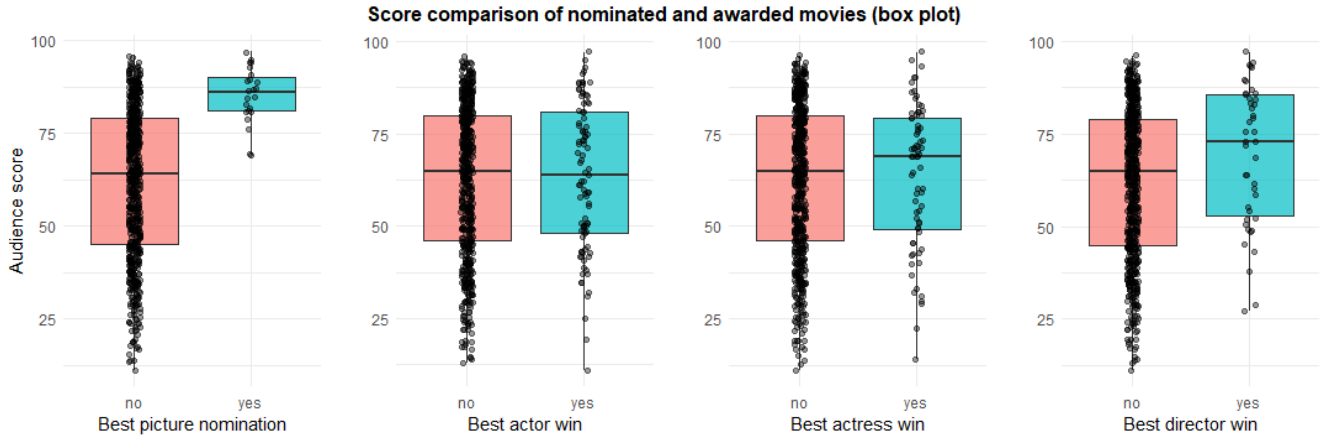

Figure 3.1 Mean audience score between films that have won or not different awards.

4. What are the trends over the years?

Which is the preferred month for releases? How does this change over the years?

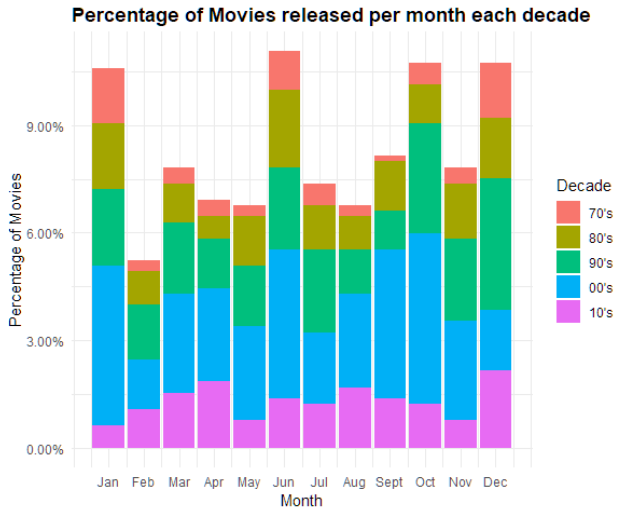

Figure 4.1 Percentage of movies released by month each decade.

Figure 4.1 shows the percentage of movies from the data set that were released each month (coloured by decade) to see if there is a trend. We can see a pattern emerge with the majority of releases occurring in the months of January, June, October and December (holiday seasons). To investigate if this trend exists across decades, we broke down the results into multiple plots (Figure 4.2). By comparing these graphs, it is clearly seen that some decades vary in the percentage of releases for each month, but that overall, almost all decades share the previously stated months as the most popular for releases.

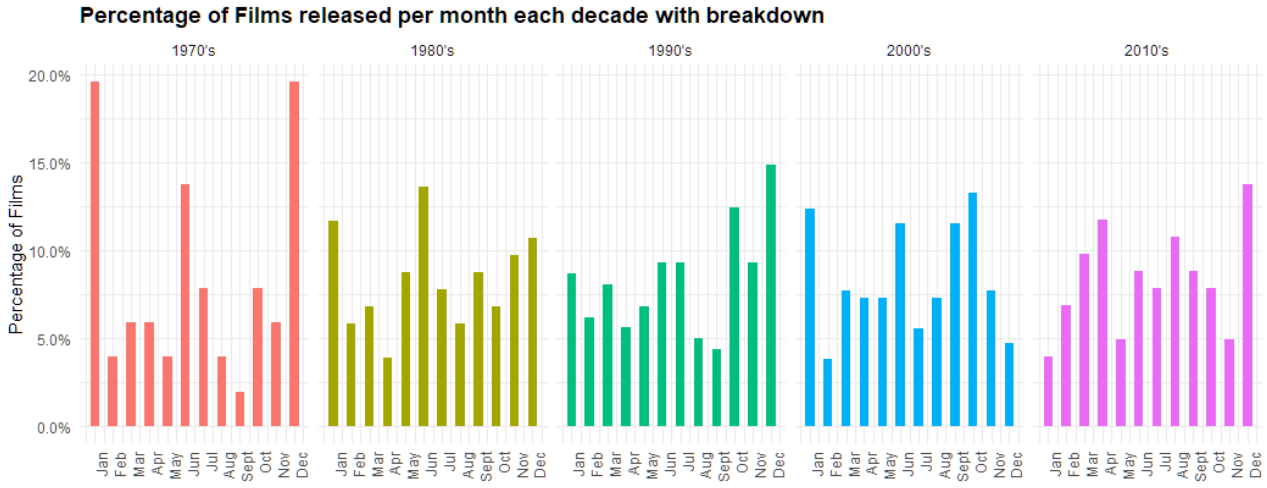

Figure 4.2 Percentage of movies released by month separated by decade.

Do older films tend to have a higher score/number of IMDB votes?

We have plotted the number of IMDB votes available in the data set against the date of each film. Additionally, we have plotted the ratings over the years to see if there was any kind of drastic change.

Looking at Figure 4.3.a we can see a higher variance in the number of votes for movies created before the introduction of IMDB in 1990. We believe the variance is caused due to certain older movies being regarded as "classics" meaning people from future generations have greater exposure to them, thus more people are able to review the film. While movies from the same era that are not as well regarded will not have the same level of exposure, thus have a lower number of votes. As the graph moves through the years toward more recent times, when IMDB was in existence at the same time of the movie releases, the average number of votes increases which is expected with our reasoning. The decrease in number at the end of the graph could be explained that the movie had not been released long before the dataset was created so had not yet had great exposure.

It is shown in Figure 4.3.b that the larger number of votes in later years has not had a great effect on movie rating. There has been little fluctuation over the years, and any variances we perceive is due to random variation in the ratings, as the score difference over the years has been minimal (approximately 0.6).

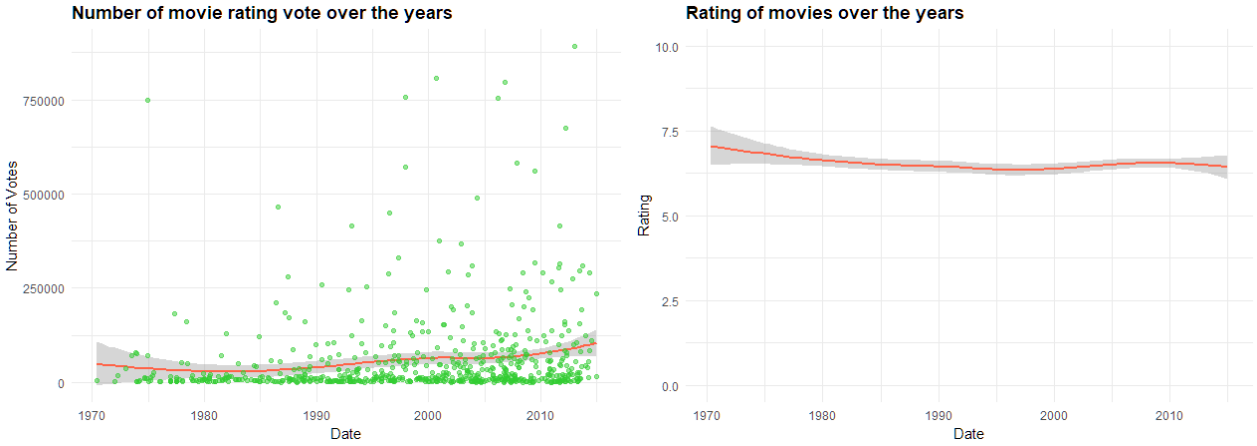

Figure 4.3 Left: number of votes by film over the years by release date. Right: rating by film over the years.

5. Do scores of movies directed by Oscar winning directors have any trend?

To answer this question, we visualized if the likelihood of winning a director Oscar could be predicted by the tread of a director review scores. By plotting each director's movie releases against the three different score categories we can see that the comparisons between the groups vary but they approximately show the same positive trajectory for the directors over the years. To see if these trends indicated likelihood of winning an Oscar, the years each director has won the best direct award were added to the plot using vertical lines. Unfortunately, due to the lack of data available in the dataset it is not possible to confidently determine any information, thus we are not able to answer the question we proposed. Due to space restrictions the graph has not been included in this report but can be seen within the attached code.

Data analysis plan

Question 1: Is there a relationship between the type/genre and score?

Regarding type, to formally address the question we would use a **hypothesis test**. First, we should define our null hypothesis, Documentary type movies have the same score as other Films, and the alternative, that the overall score of the Documentaries type is higher than the overall score of Films: $H_0: \boldsymbol{\mu_D} = \boldsymbol{\mu_F}, \boldsymbol{H_1}: \boldsymbol{\mu_D} > \boldsymbol{\mu_F}$. Next, we

would choose the **Difference Between Means** as sample statistic because the size of the population is much larger than the sample size and samples are independent: $\bar{x}_D - \bar{x}_F$.

After applying the statistic of choice to the set of randomized samples, we would compute the ***p*-value**, which would tell us the proportion of samples on the distribution with a statistic as extreme or more than the observed sample. If the *p*-value is lower than a 5% **significance level** (α=0.05 is used as a default level), the conclusion would be that the results are statistically relevant, thus rejecting H0. If the *p*-value is not lower than the significance level, it would mean that we cannot reject H0, and the results would be not statistically relevant.

For genre, we face the same situation, but with many different values for the qualitative explanatory variable. We would repeat the analysis but using a one-versus-all approach, computing the *p*-value of each of the different genres and comparing all of them individually against the significance level. Also, a **Kruskal-Wallis one-way analysis**[2] could be used, indicating if at least one sample dominates the rest of the population stochastically.

Question 2: How do the score sources differ for genre and MPAA rating?

To know which is the variable that has the most effect we could perform a **Principal Component Analysis** to measure the percentage of variance explained of each of the three variables (audience_score, critics_score and imdb_rating) to see which of them is the highest. We should compare the three vectors of the n-dimensional data space, after rescaling, which we expect to be highly correlated.

As the graphs suggest that critics tend to evaluate the lowest, we could perform a hypothesis test on each of the genres under the null and alternative hypothesis of the critics score being respectively equal or lower than the rest of the scores: $H_0: \mu_C = \mu_O$, $H_1: \mu_C > \mu_O$

Question 3: Are oscar-awarded films more liked?

To answer this question we can apply the same process used before (Question 1), with a null hypothesis, $H_0: \mu_O = \mu_N$, stating that being nominated for the Best Picture award does not affect the overall score, and an alternative hypothesis, $H_1: \mu_O > \mu_D$, stating that winning an award increases the score.

Question 4: What are the trends over the years?

To find out if older films tend to have a higher score and/or a higher number of votes, we would first perform a **linear regression model** to predict a film's score based on relevant variables, after **normalizing**. Once we have done this, we would obtain the *p*-value of each of the variables based on the coefficients of the regression line. If the *p*-value of the date variable is lower than 0.05, we would consider this result statistically relevant and therefore conclude that there is a relationship (positive or negative, based on the coefficient) between the two variables.

Question 5: Do scores for movies directed by Oscar winning directors have any trend?

This is a somewhat more complex question, as we are not well versed on the **Time Series**[3] topic. However, we will try: our approach would be to divide the data for each occurrence (directors and actors who have won an Oscar) before and after winning the award. First, we should do some research to find the year they won the Oscar. Once we have our data separated, we would perform a **Time Series Regression** for each of the occurrences before and after. By comparing the regression slope of the variable score on the regression model, we could perform a **hypothesis test** to see if the change is relevant.

[2] https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance
[3] https://www.mathworks.com/discovery/time-series-regression.html