Using Data to Find the Best Certificate for Movies

By Jason Woo

Overview

Hi, I am currently a senior majoring in BAIT and minoring in Statistics. I have been currently curious about the rankings of movies on IMDB and what determines its high rankings. Would genre play a huge role or would the certificate play a huge role? Maybe for upcoming directors, could this information be of use to determine the higher chance of getting yourself to the top 250 movies?

Data Set

The dataset used for this assignment was taken from Kaggle and the data comes from IMDB, which is one of the largest online databases for movies and television shows providing ratings, reviews, or any information you want to learn about the film. The ratings are widely used as a benchmark for the movie’s popularity and success. The data was taken from 2021 of the most popular and highest rated movies of recent times. We can gain insight and see certain trends in movie ratings and genres.

Basics

To develop a basic understanding of our dataset, we pull some basic values:

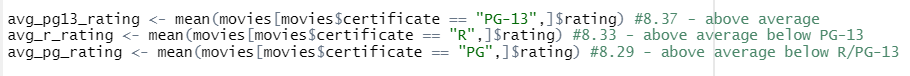


We run this code in R, we see that the average rating of the 250 movies is 8.31. We used the three certificates (PG-13, PG, and R) as a baseline for ratings of movies because they show the highest volume of all the other certificates within the 250. We also use these because they are the most commonly known within the movie industry. A boxplot is created to show the average ratings based on their certificate (Some are shown as a black bar indicating the certificate of the movie appeared once in the list).

Chart, box and whisker chart

Description automatically generated

Here, there is a wider distribution for R rated movies, however the average rating for PG-13 was slightly higher than PG and R. Here is the code to verify my observation:



Hypothesis

With these ratings in mind, I came up with the hypothesis:

**Null Hypothesis:** PG-13 rated movies have the same average rating as PG rated movies.

**Alternative Hypothesis:** PG-13 rated movies have a higher average rating than PG rated movies.

To test this, we ran the permutation test:



Chart, histogram

Description automatically generated

The result I obtained was a p-value of 0.0377. This is below the standard level of significance which is usually 0.05, so we can **reject the null hypothesis.** We can also run a hypothesis on comparing rated R movies to PG rated movies.

**Null Hypothesis:** R rated movies have the same average rating as PG rated movies.

**Alternative Hypothesis:** R rated movies have a higher average rating than PG rated movies.

We can run another permutation test:



Chart, histogram

Description automatically generated

The result we obtained was a p-value of 0.1251. This time, since it is below the standard level of significance, we **do not reject the null hypothesis.** We have shown there is a slight chance that the observed difference is indeed not random.

Conclusions

If you are planning on becoming a director and want your movie in the top 250 on IMDB, you should want to create PG-13 rated movies, however that might not be easy to do as it sounds. Some of our data could have been biased because although it is the top 250, the ratings given to the movies are all bias. The data source is trustworthy since it is data from IMDB, the most reputable data source for movies and rankings.