Analysis of Hollywood Movie

Anjali Patil, Kunal Mehrotra, Foram Shah, Nidhi Bodar

Summary: Every year thousands of movies are released and huge revenue is generated contributing to the global economy. We can say that the film industry plays a significant role in the worldwide economy. However, online platforms like Netflix, Prime, etc. might have impacted this revenue. We want to perform an analysis on various factors like the film genre, the release date, the rating of movies, main actors, IMDB rating, etc.

Proposed Plan: We want to perform exploratory data analysis on the movie dataset. We would like to see if there is any correlation between different attributes. The main objective is to:

1) Find the most profited month of the year

2) Analyze the month with more number of movie releases

3) Analyze the relationship between the number of movies released in a month and the most profited month in a year.

4) Analyze which features are affecting our target variable "success" the most.

5) Based on those features we will train a model to accurately predict the success of a movie.

Preliminary Results:

After analysis, we have decided to drop the following columns released, director, writer, star, and company as they cannot be encoded and therefore used for predicting. The columns which we will encode with categorical values are rating, genre, and country.

Columns like Rating, budget, and gross have missing values and need to be imputed. For missing values for Rating, we can fill them by Not Rated which is already present as a value, for missing values of Budget we will take the Production company average of the budget and assign accordingly, for Gross we first find a threshold number of votes for which the movies are profitable, not profitable and loss-making, so if the movie is profit-making we set a value 25% higher than the budget, if not profitable then same as budget and if loss-making then we set 25% below the budget.

Total there are 7669 rows after all the preprocessing steps.

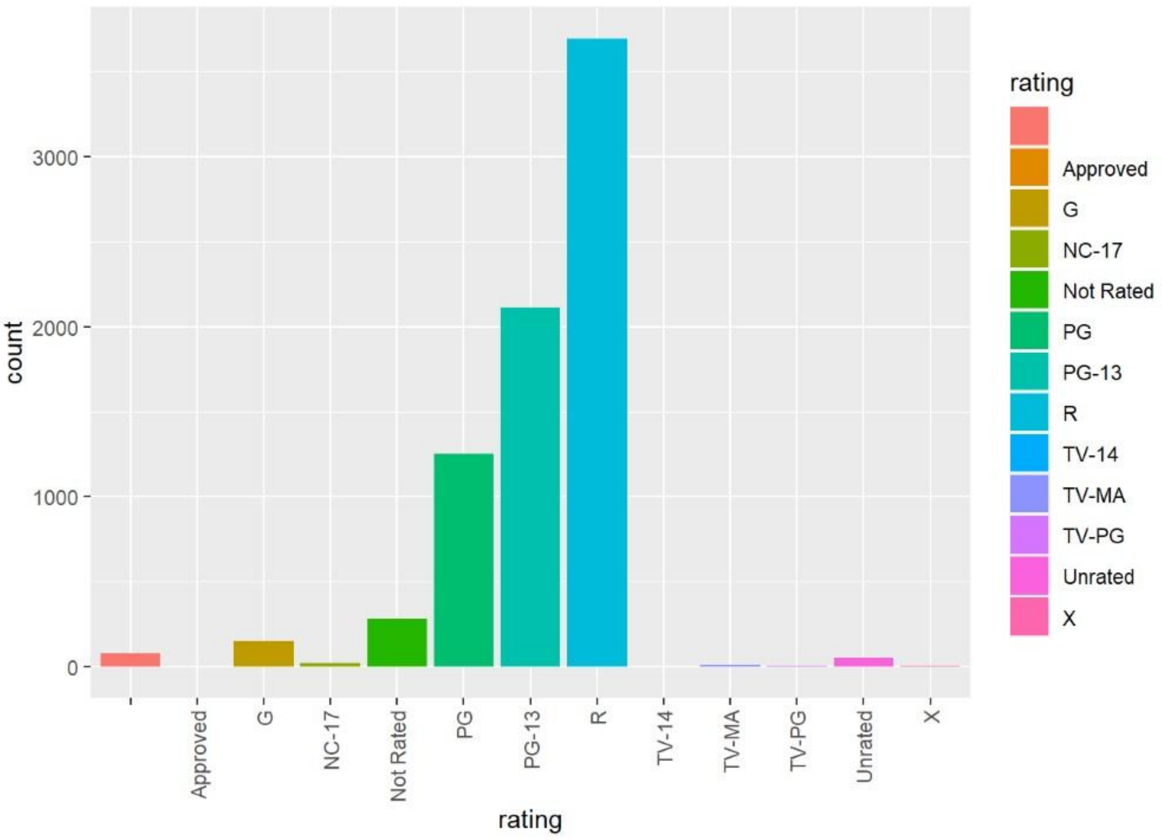
After imputing missing values, our objective is to determine what is the success rate of a movie which will be considered as a target variable. For computing that, we will add one column for computing success rate by using the formula gross/budget. For setting, if the ratio of the computed equation is greater than 1 then the movie can be considered a hit or if the ratio is less than one then it can be considered a flop movie.

Also, we need to determine which predictors are strongly correlated to the new column success of the movie.

References :

Grijalva, Daniel. "Movie Industry." *Kaggle*, 23 July 2021,https://www.kaggle.com/datasets/danielgrijalvas/movies.

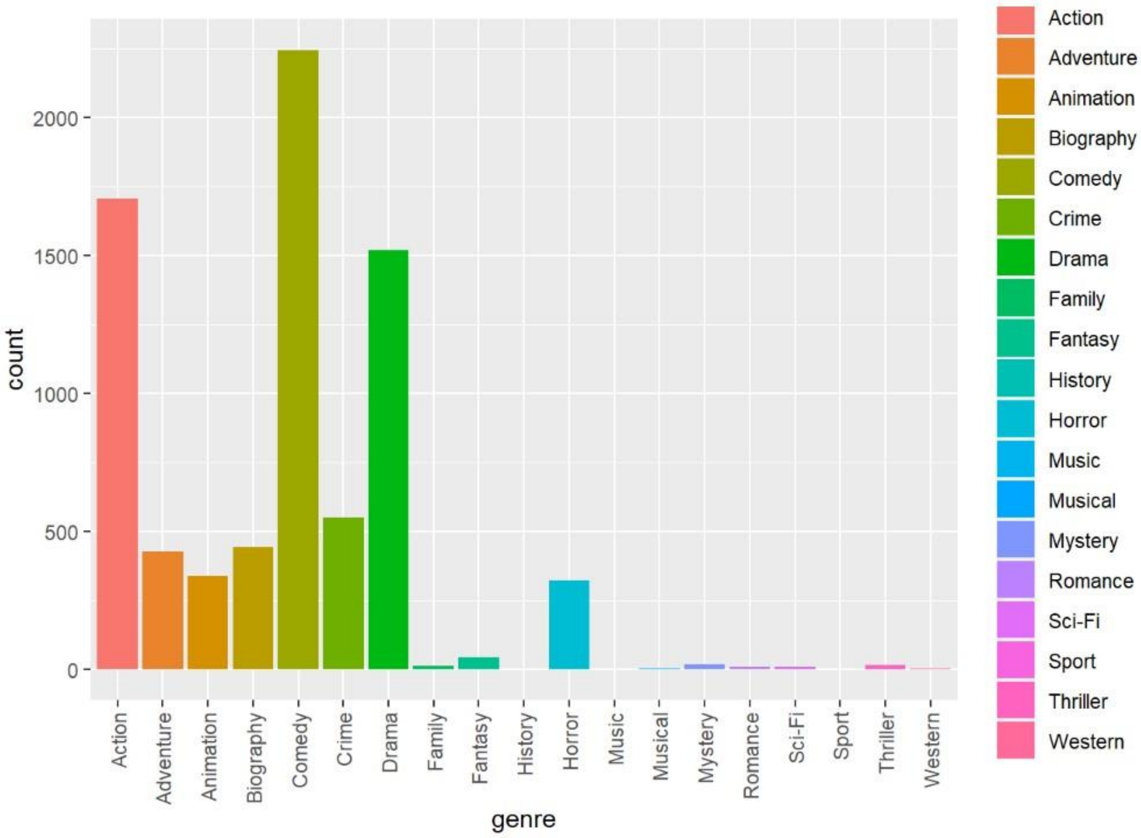
https://www.canr.msu.edu/news/what_do_movie_ratings_mean



From the above graph, we can observe that R(Restricted, Children Under 17 Require Accompanying Parent or Adult Guardian.) movie has the highest rating.

Analysis of Hollywood Movie

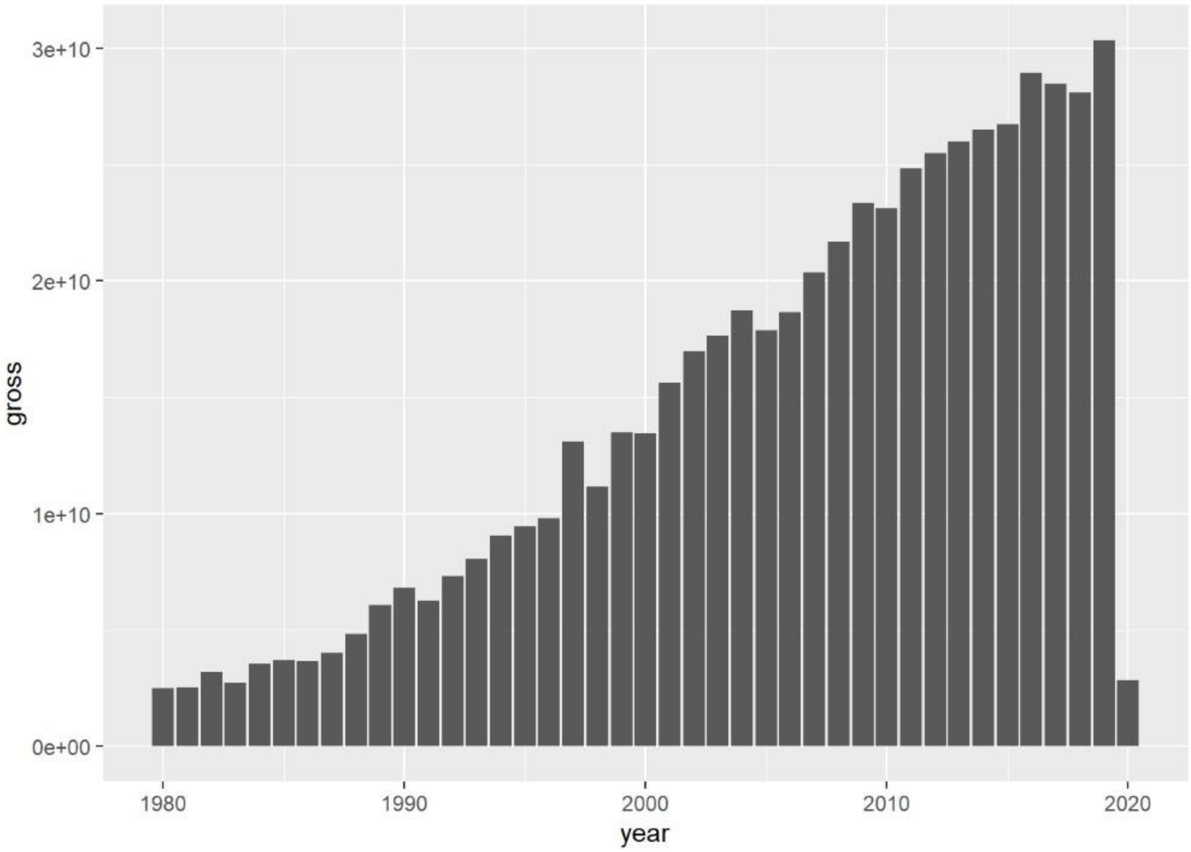
Anjali Patil, Kunal Mehrotra, Foram Shah, Nidhi Bodar



From the graph we can observe that people tend to like Comedy, Action and Drama genre most. Also, after performing appropriate exploratory analysis, we can represent which genre has the highest rating in which category.

Analysis of Hollywood Movie

Anjali Patil, Kunal Mehrotra, Foram Shah, Nidhi Bodar



This graph represents that Gross tends to increase over years but suddenly a drop can be observed during 2020.