Report - TMDB Analysis: Comparative Study of Comedy vs Horror Genres

Introduction

This study is aimed at making a comparative study of Comedy vs Horror genre films, on parameters of popularity, revenue, budget and count of films. The study is particularly focused on the trends of the parameters over period of time.

Some of the questions of interest are:

- Which genre was more popular in each time frame (A time interval of 5 years is considered)?
- Which genre of films were more produced during those time frames?
- Budget vs popularity relation of all the movies in the dataset.

Data Investigation steps and Data Cleaning Steps

The minimum and maximum release year was found out from the dataset. Based on those values, each movie was categorized into release year group. The year groups, are grouped into interval of 5 years (1960-1964, 1965-1970 and so on). The dataset was checked to see if any row was having 0 or null as value in release year column, and such rows were eliminated. The budget and revenue was formatted to suppress scientific notations.

Two Subsets of genre comedy and horror was extracted from the dataset. Point to be noted here, the selecting criteria was that the genre list must be having at least one value in comedy or horror and may contain other genres combined with either of these two.

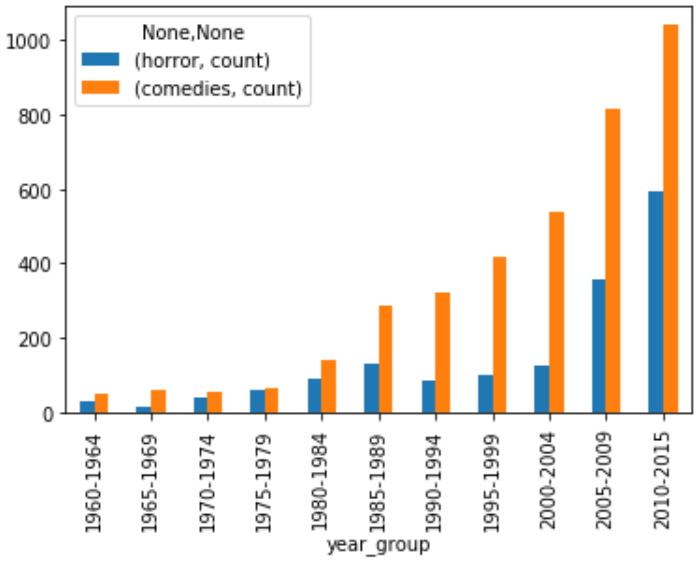
Aggregate functions count was applied on the two subsets and corresponding bar graphs were drawn to do a visual analysis of number of movies of each genre produced in each group of release year. Then the mean of popularity of each year group was taken of both the genres, and bar graph were drawn. A comparative dual bar graph was created on the parameters of count and mean of popularity of each year group. From the graphs conclusions were drawn.

To study the relation between popularity and budget of a movie, first all the rows where budget and revenue was 0 or NA were removed. On subsequent subset formed a scatter plot was drawn and r squared coefficient was calculated to determine how the two are related.

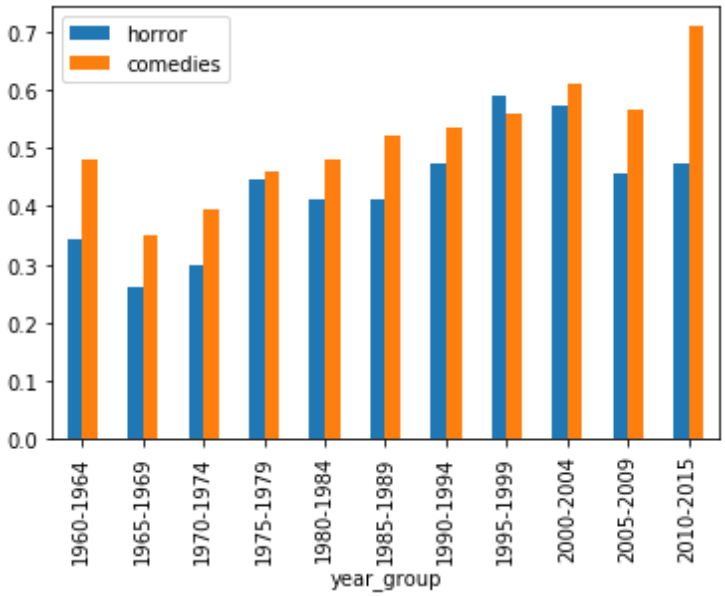
Plots and Graphs

The graphs of various comparisons are below.

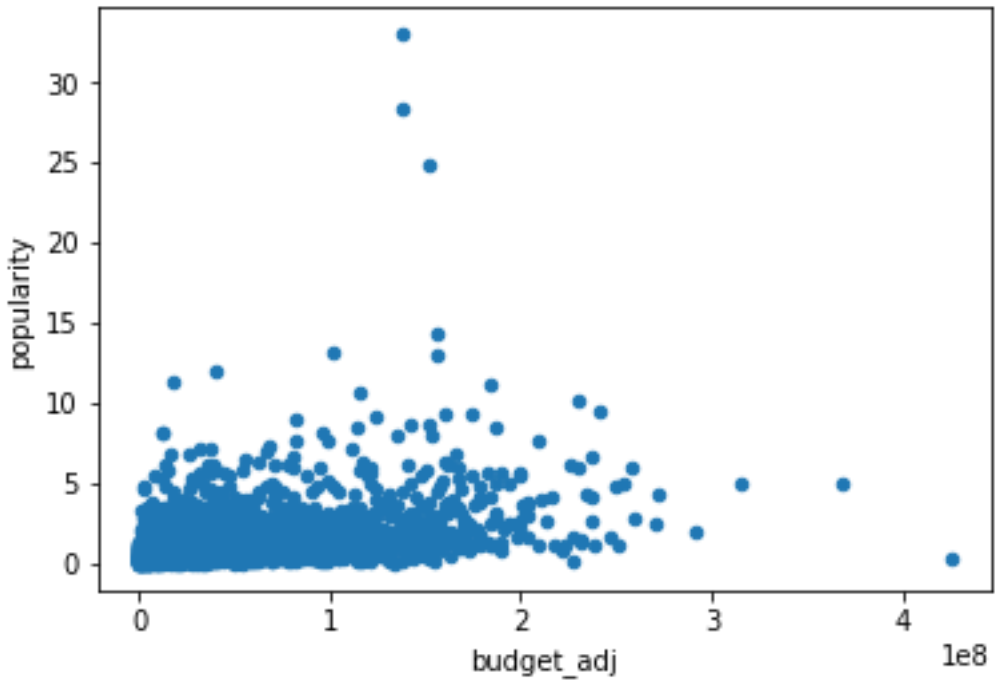
Count Graphs: Number of movies produced of each genre



Popularity Graphs: Mean Popularity of movies of each genre year wise



Scatter Plot: Between budget adjusted to inflation and popularity



R-Squared coefficient of above plot: 0.1595

Assumptions and Limitations:

- The analysis done is not exclusively for movies of types comedy or horror. There could be a mix of other genres, for eg: Science Fiction and Comedy or Horror and Adventure.
  Any movie with at least one genre category in comedy is classified as comedy, and one genre as horror is classified as horror movie.
- The dataset is considered to be accurate and contains all the movies released between period of 1960 to 2015. This is important because of the year wise grouping and measuring their parameters. Some movies missing will cause variation in the actual results

Conclusions

From the above analysis the following conclusions are drawn about the above questions.

- Comedy genre movies are produced more in number in all the year groups compared to horror movies. It is evident from the graph.

- Comedy movies are more popular when compared to horror movies in general, over period of time. The mean of the popularity is considered of each year group.

- The r squared value between budget adjusted and popularity is very small. This implies there is little variation in popularity with respect to budget of the movies (Ignoring the extreme outliers).

References:

1. Stack overflow
2. Data camp intro to python for data analyses course