

Data Science Research Methods Assignment-3

Sahil Singh

University of Sussex

2021

Introduction

The objective of the report was to analyze a movie dataset from imdb to make recommendations on the type of movie the studio should consider making. Our budget for production was 1.5 million.

Data Analyses

- First step was to control for budget and only consider movies with a budget of less than 1.5 million.
- Next, a new feature, 'profit_percentage' was calculated from the data such that,

$$\text{profit_percentage} = \left(\frac{\text{gross}}{\text{budget}} - 1 \right) \times 100$$

where 'gross' is just the total earning of the movie
and 'budget' is the budget of the movie

- We try to analyze the gross and profit of different genres through a boxplot.

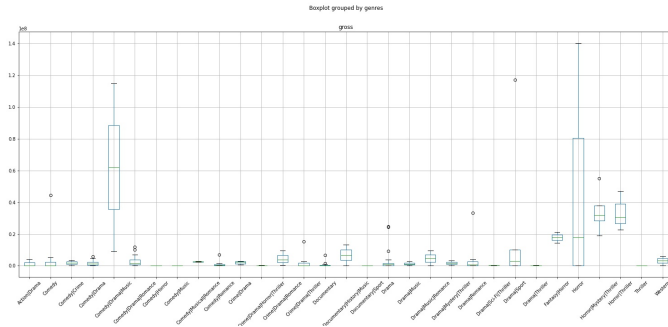


Figure: Boxplot of Gross grouped by Genres

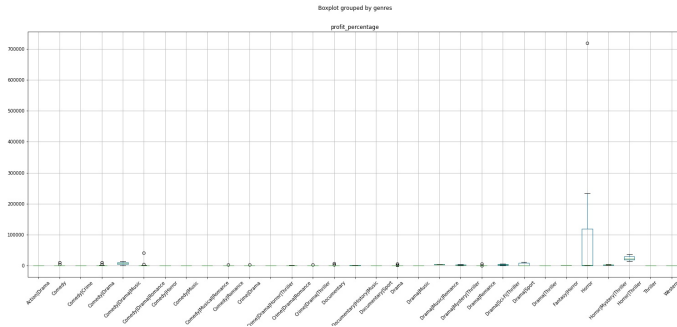


Figure: Boxplot of profit_percentage grouped by Genres

- Tried to analyze most prevalent subplots in these movies.

plot_keywords	frequency of occurrence
friend	16
love	9
independent film	8
drugs	7
friendship	7

Hypothesis Generated from Data

- On analysing the subplots, one sees a clear pattern that certain movie genres, for instance, Comedy/Drama/Music have higher earnings.
- With respect to profit_percentage however, one sees that genres like Horror and Horror/Thriller have higher return on investment.
- We try to test if the above mentioned genres have higher returns on average compared to other genres.

Example of a Hypothesis

μ_1 = Mean of profit_percentage of genre 'Horror'

μ_2 = Mean of profit_percentage of all genres **except Horror**

$H_0 = \mu_1 \leq \mu_2$, Null Hypothesis

$H_1 = \mu_1 > \mu_2$, Alternate Hypothesis

The above hypothesis is tested using a right-sided t-test as the sample size might be too small for a z-test.

If the p-value from the above test is less than 0.05(our confidence value), we conclude that Horror movies indeed earn higher than average compared to other genres. Aspect Ratios are tested in a similar way.

Bootstrapped Hypothesis Testing

Another type of hypothesis test was performed using bootstrapping to confirm the results obtained from t-test. In this version, bootstrapped sample of both the sets that are to be compared were taken and then the mean was calculated for each sample. These samples of means were then compared.

Testing Feature's Impact on Profit

- To test whether features like 'Director Facebook Likes', 'actor_1_facebook_likes' impact movie profits, a different method is adopted. Features are divided into percentiles like 10,20,30 up until 100.
- Then its tested if the profit of movies in higher percentile for that feature(say higher than 60 percentile) is higher than that for movies in the lower percentile(less than 60 percent).
- In this way we are testing if higher values of a particular feature means higher earnings.

