report

xiaofan xia

12/5/2019

Overview

Introduction

This project focuses on digging into the dataset with the information of more than 5000 movies. My main objective of this project is to find out how much the movie's reputation is influenced by other related variables such as the popularity of the casts, and movie's budget.

Motivation

I am a very big fan of many kinds of movies. During my spare time, I always choose to watch movies to relax. Also, many friends of mine are working on the film industry right now. Therefore, I generate some questions about movies and its rating system, and that is why I choose this dataset and this project.

Dataset: IMDB 5000 Movie Dataset from Kaggle.

The IMDB 5000 Movie Dataset contains 28 variables such as movie names, director names, IMDB scores, and movie Facebook likes, etc. This is an open database from Kaggle.

Research Questions

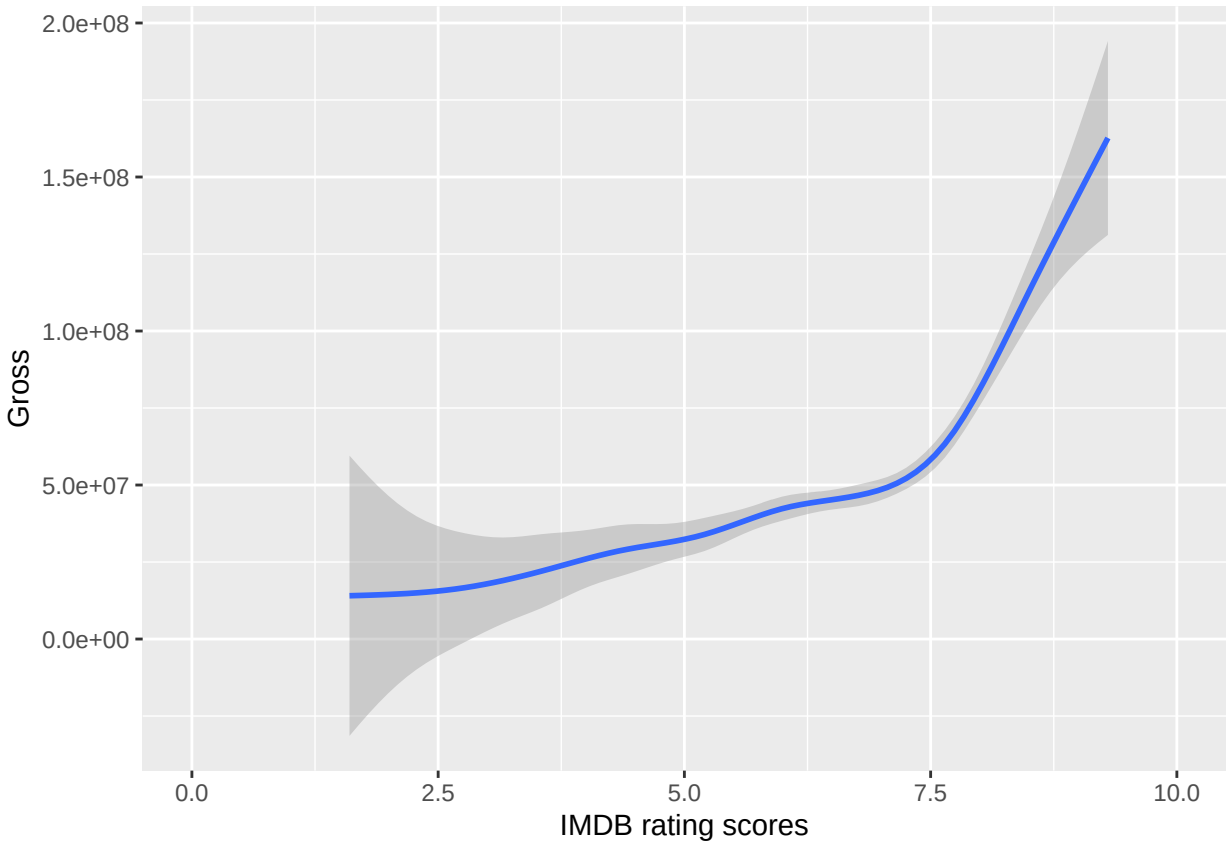
1. Is there any relationship between the popularity of casts and the income of the movie?
2. Is there any relationship between the popularity of casts and the reputation of the movie?
3. Is there any difference in the ratio of the budget (cost) to the gross regarding to different movie genres?
4. What are the patterns of budget and cast in movies with high reputation?
5. What are the patterns of budget and cast in movies with high income?
6. What's the relationship between the number of face in the poster and the movie's reputation?
7. Is there any difference in above questions regarding to different language or country of the movies?

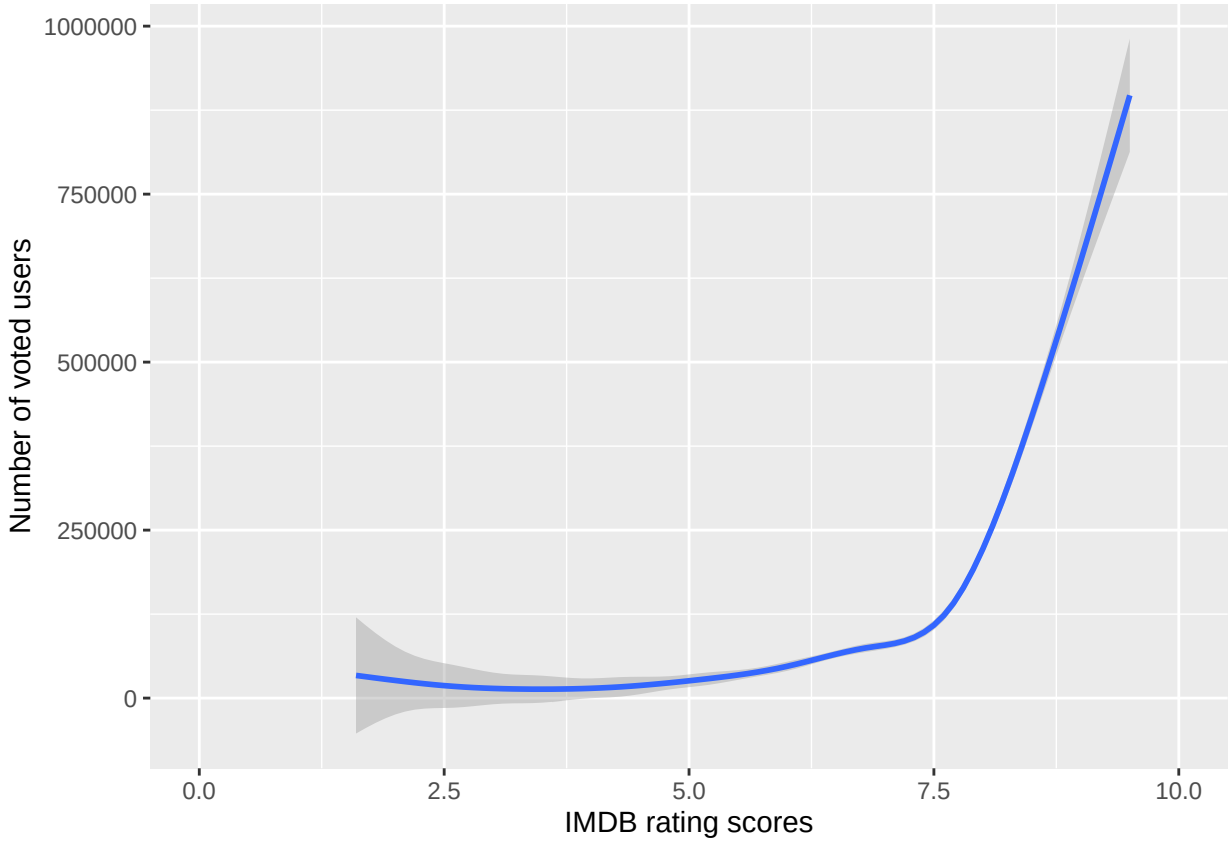
Concerns

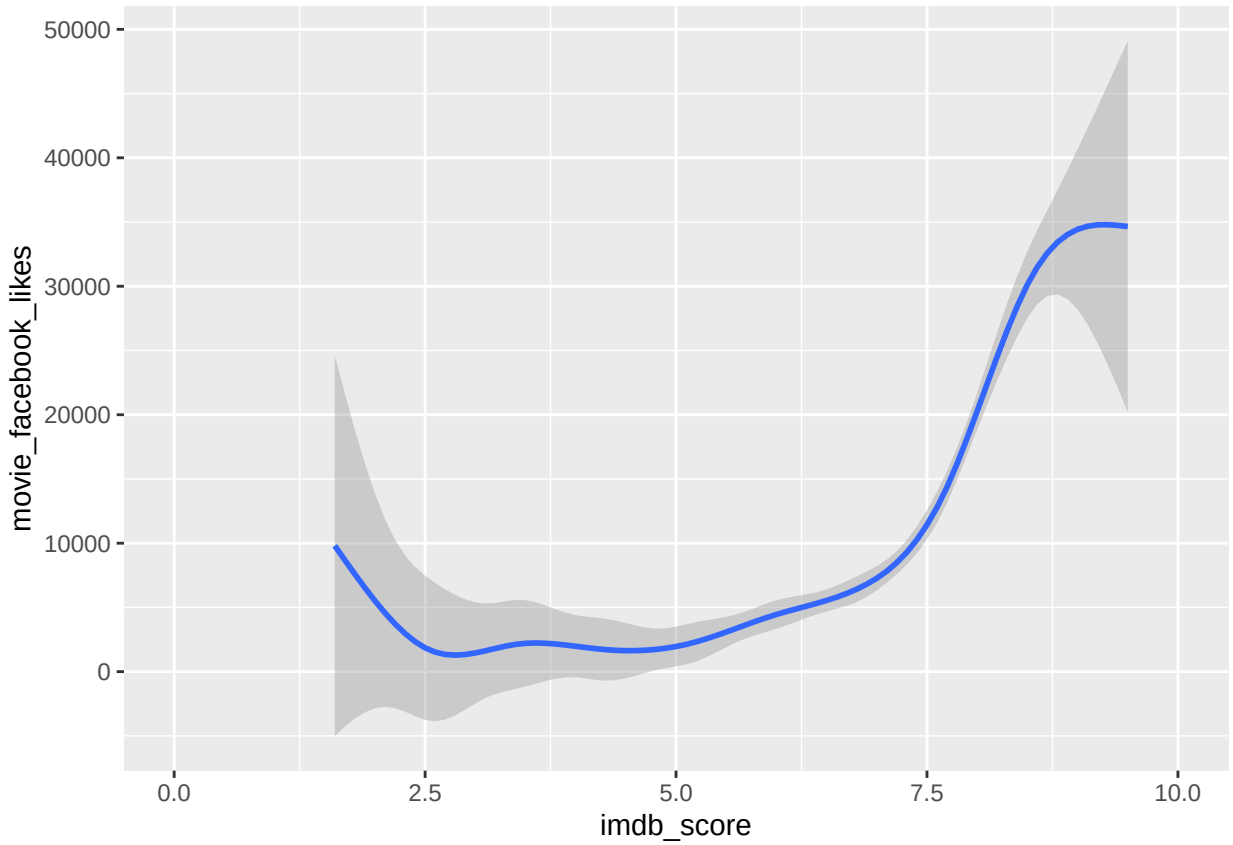
1. The dataset contains 28 variables and more than 5000 entries. It takes time to figure out how to clean the data and possibly categorize the data.
2. The dataset does not contain the year of the movie and one of the variables is Facebook likes of the movie, so the likes of some old movies might not be very timely.

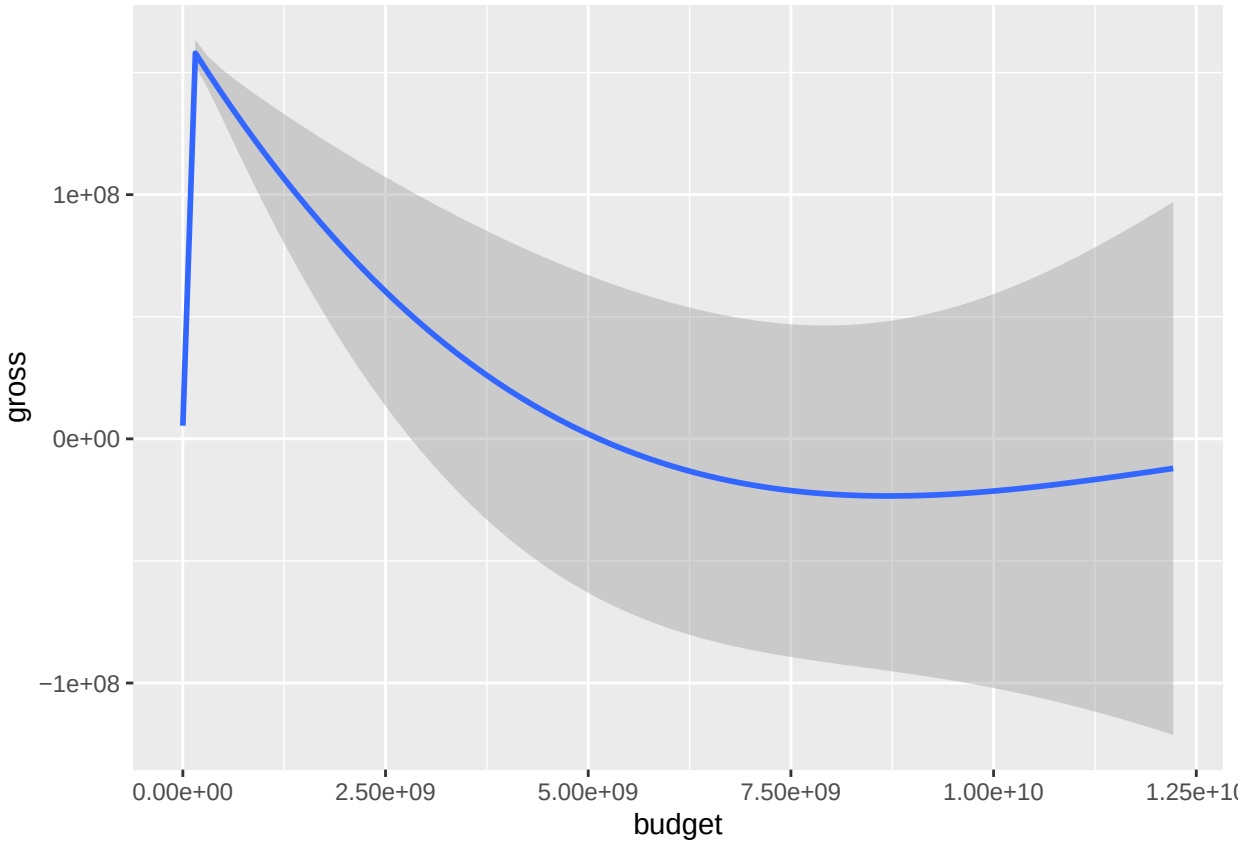
EDA

After


2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
Gross
0.0
2.5
5.0
7.5
10.0
IMDB rating scores


1000000
750000
500000
250000
0
Number of voted users
0.0
2.5
5.0
7.5
10.0
IMDB rating scores


movie_facebook_likes
50000
40000
30000
20000
10000
0
0.0
2.5
5.0
7.5
10.0
imdb_score


gross
1e+08
0e+00
−1e+08
0.00e+00
2.50e+09
5.00e+09
7.50e+09
1.00e+10
1.25e+1(
budget

Medthod

Analysis

Conclusion and Discussion

Appendix