IMDB Movie Project Report

Xiaofan Xia

12/7/2019

1. Overview

1.1 Introduction

This project focuses on digging into the dataset with the information of more than 5000 movies including IMDB scores, number of movies' Facebook likes and basic information about movies' casts. 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. The reputation of the movie would be quantified as IMDB rating scores, and the popularity of the movie would be quantified as number of Facebook likes and number of voted users. In this project, I firsly cleaned the dataset and did a Exploratory Data Analysis before digging into the modeling analysis. The major model I would use to conduct the analysis and prediction in this project is multi-level linear regression model.

1.2 Motivation

The initial idea of choosing the IMDB movie topic comes from my personal interest. 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. I hope that I could solve my own questions about movies during this project as well as help my friends to learn their own industries from a different perspective with data analysis.

1.3 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.

Source: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

1.4 Research Questions

a. Is there any relationship between the popularity of casts (Facebook likes) and the income of the movie (gross)?

b. Is there any relationship between the popularity of casts (Facebook likes) and the reputation of the movie(IMDB rating scores)?

c. Is there any difference in the ratio of the budget to the gross regarding to different movie genres?

d. What are the patterns of budget and cast in movies with high reputation?

e. What are the patterns of budget and cast in movies with high income?

f. What's the relationship between the number of face in the poster and the movie's reputation?

g. Is there any difference in above questions regarding to different language or country of the movies?

h. How can we predict the IMDB rating score using other information of the movie?

2. Exploratory Data Analysis

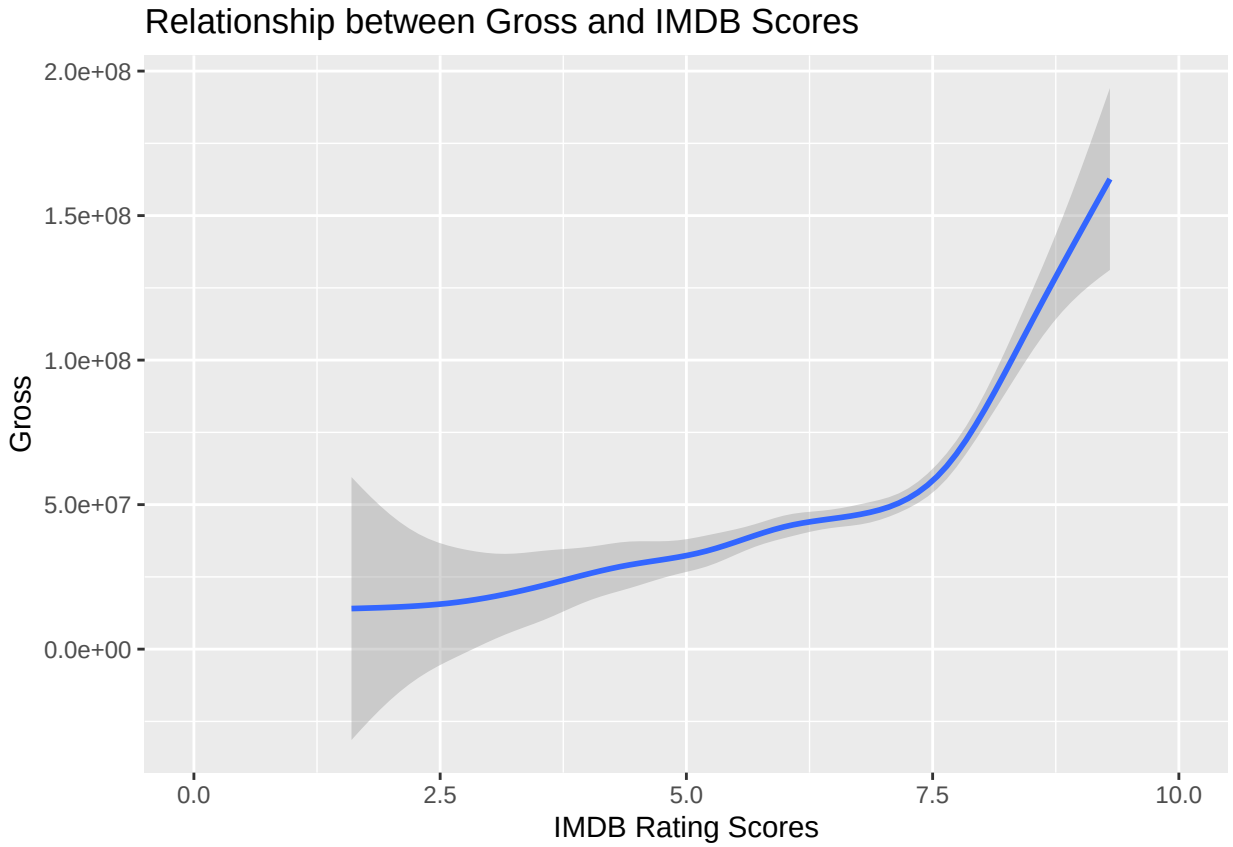
2.1 Basic Statistics of Major Variables

After cleaning the dataset, I firstly explore the basic statistics of major variables I am interested in. In this part, I am assuming the gross and budget are in the same currency with same unit. There are 66 different countries and 48 different languages in this dataset, which shows diverse background of movies. Here's a table of statistics (values of mean, maximum, minimum and median) for varibales including number of Facebook likes, IMDB rating scores and gross.

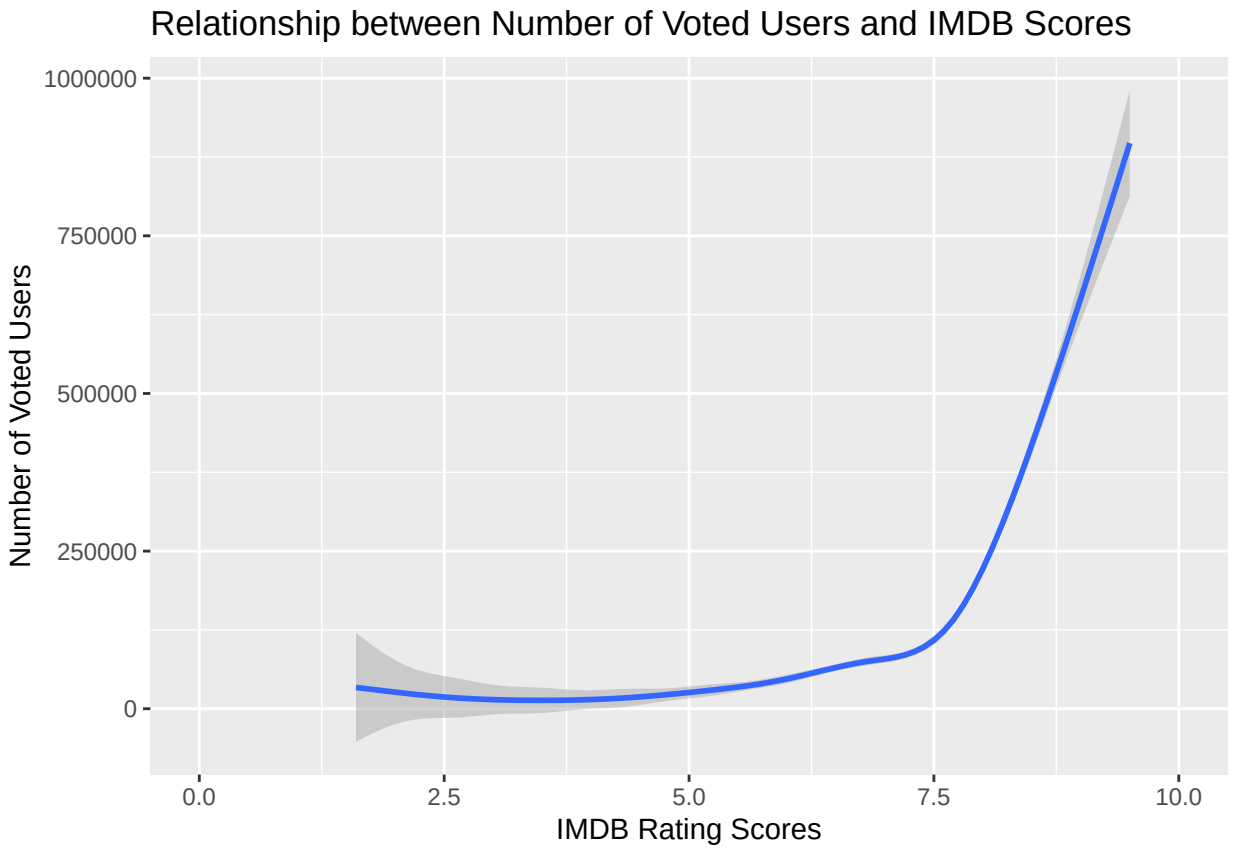
	Mean	Maxium	Minium	Median
Number of Facebook likes	7.525965e+03	3.490000e+05	0.0	166.0
Number of Cast Total Facebook likes	9.699064e+03	6.567300e+05	0.0	3090.0
Number of voted users	8.366816e+04	1.689764e+06	5.0	34359.0
IMDB rating scores	6.442138e+00	9.500000e+00	1.6	6.6
Gross	4.846841e+07	7.605058e+08	162.0	25517500.0
Budget	3.975262e+07	1.221550e+10	218.0	20000000.0
Duration Time in minutes	1.072011e+02	5.110000e+02	7.0	103.0
Number of Critic Reviews	1.401943e+02	8.130000e+02	1.0	110.0

2.2 Graph Presentation

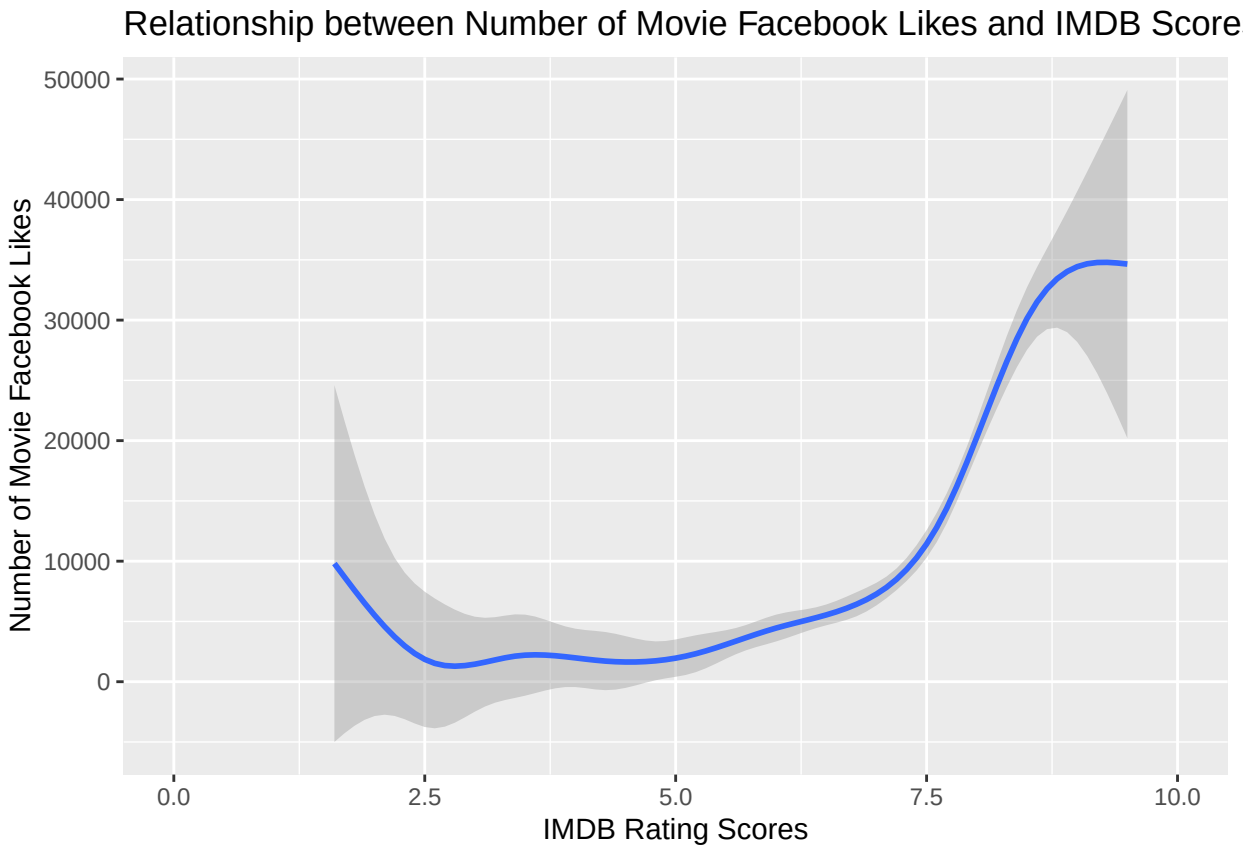
The first graph shows the relationship between movie gross and movie IMDB rating scores. Generally speaking, with higher IMDB rating scores, movies tend to have higher gross. Around IMDB score 7.5, the range of gross is relatively small, which indicates that movies with IMDB score 7.5 tend to have similar profit. We can see that people would spend money on movies with good reputation and movies with good reputation could gain more and more profit. Keeping a great reputation is kind of significant for a movie to survive in the movie market today. Indeed, more and more filming production teams put great effort in advertising their work to the public than before.



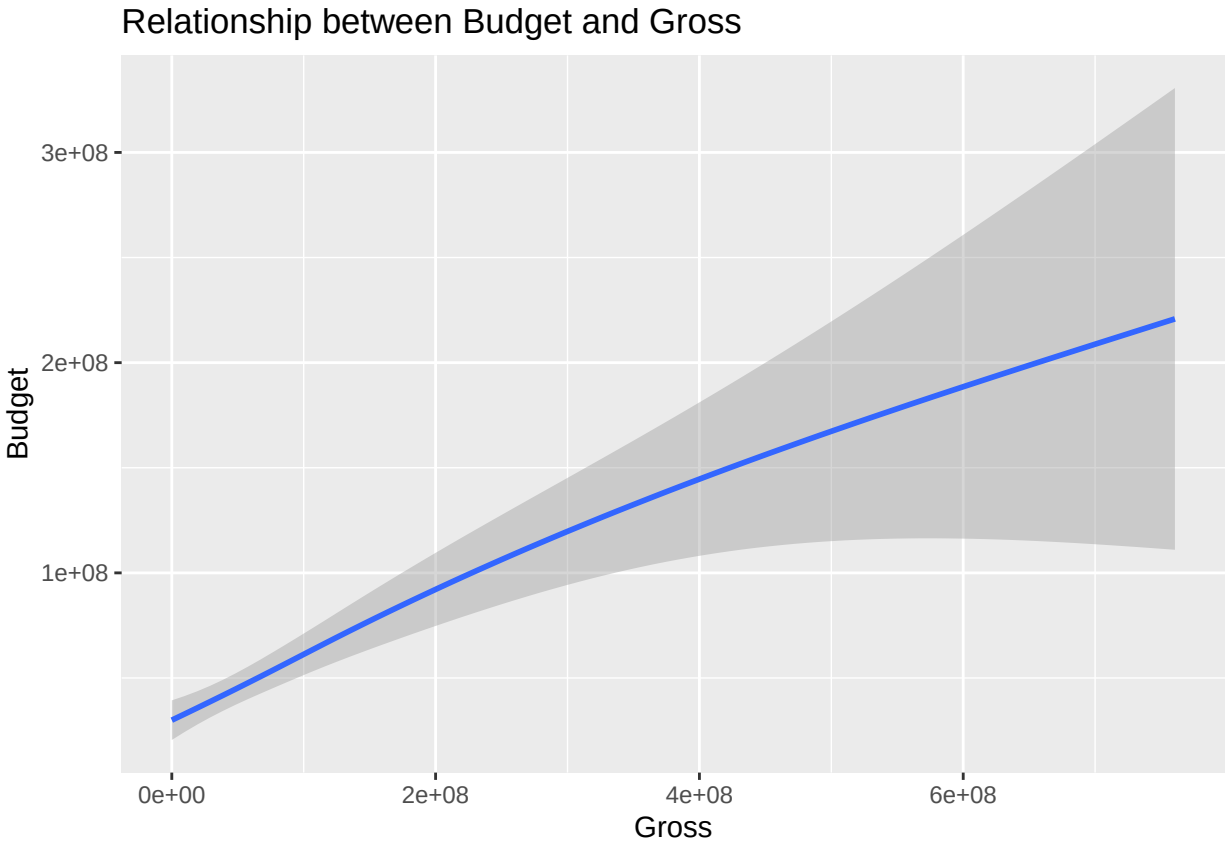
The second graph shows the relationship between IMDB rating scores and number of voted users. It is pretty obvious that movies with higer IMDB scores tend to have more voted users, which indicates that most users have similar tastes in movies and higher-rated movies have more audiences willing to vote for them.



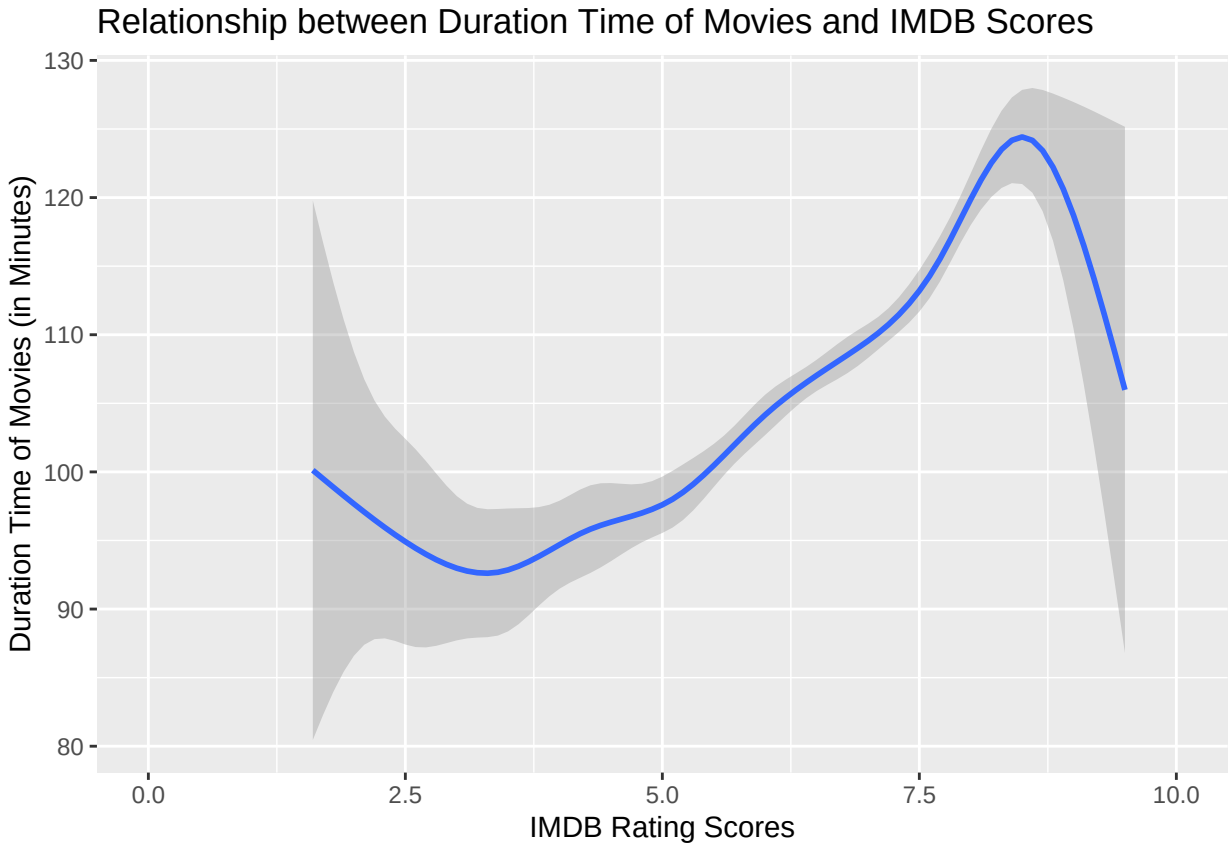
The third graph gives a visual interpretation of the relationship between number of movie Facebook likes and IMDB rating scores, which also could indicate the relationship between movies' popularity and their reputation. Although some movies with lower IMDB scores recieve relatively more Facebook likes, we can see that movies with higher IMDB scores generally have more Facebook likes.



From the fourth graph, we can see that generally movies with higher gross tend to have higher budget in the begining phase of filming production. Furthermore,movies with higher gross tend to have wider range of budget than movies with lower gross. Generally speaking, better script of movies could earn more budget at the begining phase of filming production and better script have larger chance to gain more gross. These days, movies with low budget could be famous and win many big rewards with novel ideas and cutting-edge shooting techniques, which should explain the wide range of budget in higher gross movies.



The fifth graph here interpret the relationship between the movie duration time and the IMDB rating scores visually. We can see that movies with most high scores and most low scores tend to have wider range of duration time, but movies with scores from around 5.0 to 8.0 have relatively stable pattern in duration time. Behind the graph, duration time is not a main factor affecting IMDB rating scores since the duration time of a movie is generally between 90 minutes to 120 minutes.



3. Medthod

In this project, I focus on the relationship between the popularity of movies and the reputation of movies. Therefore, I use the multi-level linear regression model to analyze that if there is any relationship between the numbers of movie Facebook likes and the IMDB rating scores.

The

4. Modeling and Analysis

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## imdb_score ~ movie_facebook_likes + (movie_facebook_likes | primarygenre)
##    Data: df
##
## REML criterion at convergence: 14737.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.0130 -0.5313  0.1093  0.6455  3.2574
##
## Random effects:
##  Groups       Name                Variance  Std.Dev.  Corr
##  primarygenre (Intercept)         1.708e+00 1.307e+00
##               movie_facebook_likes 3.414e-11 5.843e-06 -1.00
##  Residual                         1.065e+00 1.032e+00
```

```
## Number of obs: 5043, groups:  primarygenre, 21
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          6.266e+00  3.003e-01  20.865
## movie_facebook_likes 1.441e-05  1.538e-06   9.369
##
## Correlation of Fixed Effects:
##             (Intr)
## mv_fcbk_lks -0.879
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
##             (Intercept) movie_facebook_likes
## Action         6.087209         1.520971e-05
## Adventure      6.361377         1.398396e-05
## Animation      6.560027         1.309584e-05
## Biography      7.048167         1.091347e-05
## Comedy         6.138920         1.497852e-05
## Crime          6.803946         1.200533e-05
## Documentary    7.158380         1.042073e-05
## Drama          6.667752         1.261422e-05
## Family         5.664661         1.709883e-05
## Fantasy        6.325101         1.414614e-05
## Film-Noir      7.087716         1.073666e-05
## Game-Show      4.190923         2.368760e-05
## History        7.023388         1.102425e-05
## Horror         5.575405         1.749788e-05
## Music          6.679546         1.256150e-05
## Musical        5.975747         1.570803e-05
## Mystery        6.399792         1.381222e-05
## Romance        5.811943         1.644036e-05
## Sci-Fi         5.909130         1.600586e-05
## Thriller       5.602739         1.737567e-05
## Western        6.512019         1.331048e-05
```

5. Conclusion and Discussion

5.1 Results

5.2 Limitations

6. Appendix