**Project Proposal: IMDB 5000 Movie Dataset**

Yueming Zhang

## Background

A commercial success movie not only entertains audience, but also enables film companies to gain tremendous profit. A lot of factors such as good directors, experienced actors are considerable for creating good movies. However, famous directors and actors always can bring an expected box-office income but cannot guarantee an acceptable profit, and this is because film companies cost a lot for hiring them at the beginning.

Other factors that turn out to be essential are whether the film has an appropriate release data or has a classic tagline. Based on these factors, we will predict what kind of films will generate high box-office income and will be highly rated.

## Data

The dataset is from Kaggle website. It contains **28** variables for **5043** movies, spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. “imdb\_score” is the response variable while the other 27 variables are possible predictors.

The original dataset has been replaced in Kaggle, here’s the link for the original dataset from Dataworld: <https://data.world/data-society/imdb-5000-movie-dataset>

Data description:

|  |  |  |
| --- | --- | --- |
| 1 | color | Specifies if it was color/black & white movie |
| 2 | director\_name | Name of movie director |
| 3 | num\_critic\_for\_reviews | Number of critics who reviewed |
| 4 | duration | Duration of the movie (minutes) |
| 5 | director\_facebook\_likes | Number of likes on director’s FB page |
| 6 | actor\_3\_facebook\_likes | Number of likes on 3rd actor’s FB page |
| 7 | actor\_2\_name | Name of second actor |
| 8 | actor\_1\_facebook\_likes | Number of likes on 1st actor’s FB page |
| 9 | gross | Gross earning by the movie ($) |
| 10 | genres | Genres of the movie |
| 11 | actor\_1\_name | Name of the first actor |
| 12 | movie\_title | Title of the movie |
| 13 | num\_voted\_users | Number of users voted on IMDB |
| 14 | cast\_total\_facebook\_likes | Total facebook likes for all cast members |
| 15 | actor\_3\_name | Name of the third actor |
| 16 | facenumber\_in\_poster | Number of the actor who featured in the movie poster |
| 17 | plot\_keywords | Keywords describing the movie plot |
| 18 | movie\_imdb\_link | IMDB link of the movie |
| 19 | num\_user\_for\_reviews | Number of users who gave a review |
| 20 | language | Language of the movie |
| 21 | country | Country the movie was produced in |
| 22 | content\_rating | Content rating of the movie |
| 23 | budget | Budget of the movie ($) |
| 24 | title\_year | Year the movie released in |
| 25 | actor\_2\_facebook\_likes | Number of facebook likes for actor 2 |
| 26 | imdb\_score | IMDB score for the movie (out of 10) |
| 27 | aspect\_ratio | Aspect ratio the movie was made in |
| 28 | movie\_facebook\_likes | Number of facebook likes |

## Problem Statement

Based on the massive movie information, it would be interesting to understand what are the important factors that make a movie more successful than others. So, we would like to analyze what kind of movies are more successful, in other words, get higher imdb score. We also want to show the results of this analysis in an intuitive way by visualizing outcome using ggplot2 in R.